# Chapter 4

## Planning a Systematic Review

We and other researchers have found that undertaking a systematic review or mapping study is an extremely time-consuming activity requiring a great deal of attention to detail (Babar & Zhang 2009). As with any complex project, planning is a key factor in achieving a successful outcome.

In this chapter we look at the tasks that need to be performed before and during Phase 1 of a systematic review or mapping study (see Figure 4.1 which highlights the planning phase of a review). The focus here is particularly on the development of a *review protocol*. The protocol plays a key role in planning a review, providing a framework within which to make and document the necessary study design decisions. The aim is to minimise bias by defining in advance the steps that will be followed and the criteria against which decisions will be made during the conduct of a review. We note though, that although it is important to agree and document a review design in advance, it is sometimes necessary to modify that design, and hence the protocol, during the conduct phase.

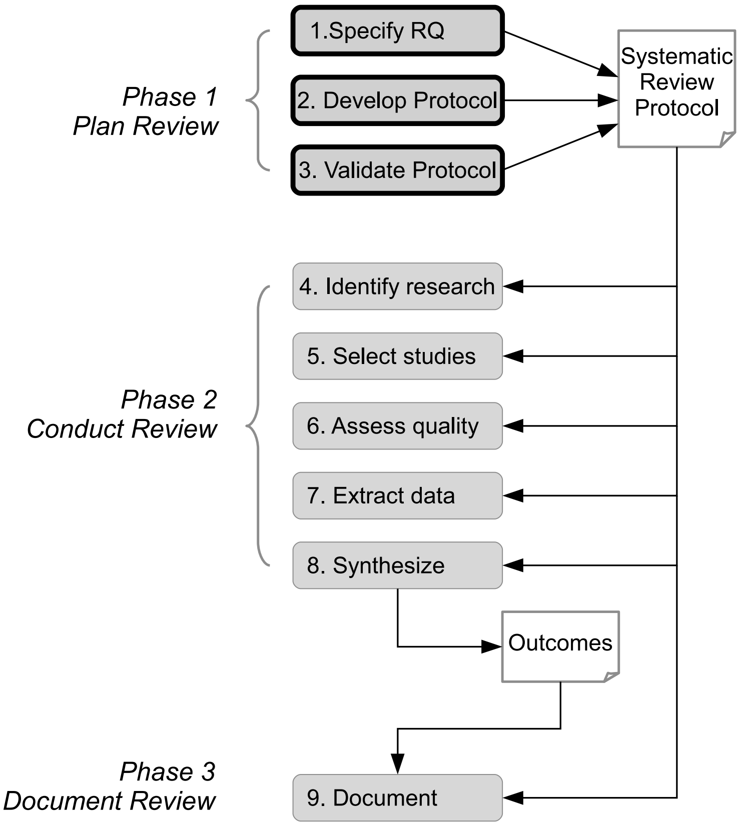

**FIGURE 4.1**: Planning phase of the systematic review process.

Even before developing and validating a protocol, reviewers should ensure that a review is both needed and feasible. We briefly consider these issues as well as aspects of managing the review process, before addressing the three main planning tasks:

1. Specifying the research questions,
2. Developing the protocol,
3. Validating the protocol.

## 4.1 Establishing the need for a review

To date, systematic reviews and mapping studies in software engineering have been largely motivated by the requirements of researchers (that is, to achieve academic goals) rather than by real problems from practice. Researchers undertake reviews to summarise evidence about some particular phenomenon in a thorough and unbiased manner. A recent survey by Santos & da Silva (2013) found that the four main factors that have motivated systematic reviewers in software engineering are:

- To gather knowledge about a particular field of study,
- To identify recommendations for further research,
- To establish the context of a research topic or problem,
- To identify the main methodologies and research techniques used in a particular research topic or field.

The results of the survey largely support the outcomes of a study by Zhang & Babar (2013) which found the most important motivators for performing systematic reviews and mapping studies to be (1) obtaining new research findings and (2) describing and organising the state-of-the-art in a particular area.

Whatever the motivation, before investing the substantial time and effort needed to carry out a thorough systematic review or mapping study, it is important to consider:

- whether it is likely to contribute to our knowledge about the topic,
- whether it is feasible, given the resources available within a review team.

Whether a review is needed and is feasible depends on a range of factors. For example, it may not be *needed* if a good quality review addressing the same or a similar topic already exists. The problem of multiple systematic reviews addressing the same topic is handled in other disciplines by researchers registering their intention to undertake a review with a central authority. For example, the Cochrane Collaboration provides a facility for such registration[1]. However, at present there is no such central authority for software engineering reviews. In fact, there are at least two examples of (pairs of) reviews addressing the same software engineering topic (Kitchenham, Mendes & Travassos 2007, MacDonell, Shepperd, Kitchenham & Mendes 2010, Verner, Brereton, Kitchenham, Turner & Niazi 2012, Marques, Rodrigues & Conte 2012). It may not be *feasible* to undertake a review if, for example, there are too many primary studies to analyse with the available resources or if there are too few good quality studies to make the synthesis or aggregation of their results meaningful.

In the examples below, we summarise the motivations for some published reviews and note that in each case the authors had previously undertaken research in the topic area and had first hand knowledge of the research issues. Further discussion about establishing the need for a systematic review or a mapping study can be found in Part III.

[1] http://www.cochrane.org/cochrane-reviews/proposing-new-reviews

## Examples of justifications for systematic reviews

Hall, Beecham, Bowes, Gray & Counsell (2012) state that fault prediction modelling is an important area of research which has been the subject of many studies. They note that published fault prediction models are both complex and disparate and that before their review there was no up-to-date comprehensive picture of the state of fault prediction. They indicate that their results will enable researchers to develop models based on best knowledge and will enable practitioners to make effective decisions about which models are best suited to their context.

Kitchenham et al. (2007) argue that accurate cost estimation is important for the software industry, that accurate cost estimation models rely on past project data and that many companies cannot collect enough data to construct their own models. Thus, it is important to know whether models developed from data repositories can be used to predict costs in a specific company. A number of studies had addressed this issue but had come to different conclusions. They indicate that it is necessary to determine whether, or under what conditions, models derived from data repositories can support estimation in a specific company.

## Examples of justifications for mapping studies

Zhang & Budgen (2012) recognised that the concept of design patterns for developing object oriented systems is valued by experienced developers. However, during preliminary investigations they found that much of the literature on patterns was in the form of advocacy or experience reports rather than empirical studies about effectiveness. They carried out the mapping study to try to identify studies that evaluate aspects of design patterns.

The mapping study by Penzenstadler, Raturi, Richardson, Calero, Femmer & Franch (2014), which focuses on software engineering for sustainability, updates an earlier mapping study on the same topic. The authors indicate that the updated study was motivated by:

- The wide range of journals, conferences and workshops which address this topic,
- A high level of research activity in recent years,
- A desire to broaden the scope of the review.

In a tertiary study, Cruzes & Dybå (2011*b*) review the methods used in systematic reviews to synthesise the outcomes of the primary studies that they include. The authors point out that “comparing and contrasting evidence is necessary to build knowledge and reach conclusions about the empirical support for a phenomenon”. The motivation for the study therefore stems from the needs of systematic reviewers to address the challenges associated with integrating evidence from multiple sources, especially where there is a high degree of heterogeneity in the research methods used for the contributing studies.

## 4.2 Managing the review project

At the start of a review, it is important to consider how the review project as a whole will be managed. This is distinct from planning and specifying the technical aspects of the review process. During the planning phase, management activities include:

- Organising the development, validation and signing off of the review protocol,
- Specifying the time scales for the review,
- Assigning the tasks specified in the protocol to team members,
- Deciding what tools to use for managing data and for supporting collaboration (see Chapter 13).

Generally, reviews are performed by two or more reviewers who constitute the review team. One of the reviewers acts as the team leader, taking responsibility for ensuring the management activities are planned, monitored and refined when necessary. If a review forms part of PhD, ideally, the student will take the lead role.

## 4.3 Specifying the research questions

Specifying research questions is a critical part of planning a systematic review or mapping study and the factors that motivate the questions should be fully explained. The questions drive the entire review process providing the basis for:

- Deciding which primary studies to include in a review, and hence driving the search strategy,
- Deciding what data must be extracted and how the data is synthesised or aggregated in order to answer the questions.

The nature of the research questions depends very much on the type of review being carried out.

For *systematic reviews*, questions are about evaluating a particular software engineering technology or research process. The term 'technology' is used in a broad sense here to encompass software engineering methods or processes, or particular management-related characteristics such as the attributes of software engineers or of software engineering teams. The research questions are formulated in one of two ways:

- As a quantitative comparison of two (or more) technologies to determine which one is more effective or efficient (or is in some other way 'better') than the others within some specific context.
- As a qualitative evaluation of a specific software engineering technology (including management-related characteristics) or an approach or procedure used in software engineering research, with respect to benefits, risks, value, impact or some other aspect of adoption.

In both cases, the questions will be driven by some underlying model of the topic, involving, for example, a comparison of a new model (or technology) with a traditional (control) model or the identification of consequences of adopting a new model.

For *mapping studies*, research questions are broader and concerned with classifying the literature in some way. The research questions for mapping studies are the most likely to change as a review progresses and new categories emerge (that is, the underlying model evolves).

We note also that mapping studies and qualitative systematic reviews are usually less focused than quantitative systematic reviews and hence tend to have a greater number of research questions.

It is important in any review to ask the right question(s). For systematic reviews, ideally, this should be one that:

- Is meaningful and important to practitioners as well as researchers. For example, researchers might be interested in whether a specific analysis technique leads to a significantly more accurate estimate of remaining defects after design inspections. However, a practitioner might want to know whether adopting a specific analysis technique to predict remaining defects is more effective than expert opinion at identifying design documents that require re-inspection.
- Will lead either to changes in current software engineering practice or to increased confidence in the value of current practice. For example, researchers and practitioners would like to know under what conditions a project can safely adopt agile technologies and under what conditions it should not do so.
- Will identify discrepancies between commonly held beliefs and reality.

Nonetheless, as indicated earlier, many systematic reviewers ask questions that are primarily of interest to researchers. This is particularly the case for mapping studies which often ask questions that lead to the identification of opportunities for future research activities. For mapping studies, research questions should be ones that:

- Enable the literature on a particular software engineering topic to be classified in ways that are interesting and useful to researchers. For example, a mapping study undertaken as part of a PhD can provide the basis for the research student's work by enabling the student to show how the proposed research fits into the current body of knowledge.
- Are likely to identify clusters of research as well as gaps in the literature. Clusters can provide researchers with some indication of where there is a sufficient body of work to warrant a more focused systematic review. Gaps in the literature can indicate that further primary studies may be usefully performed in order to fill the gaps.

As described in Section 3.2, a tertiary study is a special form of mapping study that classifies or maps reviews relating to some aspect of software engineering. Research questions for tertiary studies are aimed at identifying trends in systematic reviews focusing, for example, on:

- identifying the topics addressed by the reviews,
- the specific review procedures or approaches used by researchers.

Further details and advice about specifying research questions can be found in Part III.

## Examples of research questions from quantitative systematic reviews

The review by Mitchell & Seaman (2009) covers studies that compare the cost, duration and product quality for two approaches to software development. These are (1) the 'waterfall' approach and (2) iterative and incremental development (IID). The research questions posed in this review are:

> "What is the development cost of software produced using waterfall or its variations versus using IID?"
>
> "What is the development duration for software produced using waterfall or its variations versus using IID?"
>
> "What is the quality of software produced using waterfall or its variations versus using IID?"

Jørgensen (2007) reports a review of evidence about the use of expert judgement, formal models and a combination of these to estimate software development effort. The research questions for the review are:

> "Should we expect more accurate effort estimates when applying expert judgement or models?"

> "When should software development effort estimates be based on expert judgement, on models, or on a combination of expert judgement and models?"

MacDonell & Shepperd (2007) review studies that compare the use of cross-company and within-company data within effort estimation models. Their research question is:

> "What evidence is there that cross-company estimation models are at least as good as within-company estimation models for predicting effort for software projects?"

## Examples of research questions from qualitative systematic reviews

The technology focused study by Beecham, Baddoo, Hall, Robinson & Sharp (2008) reviews studies on motivation in software engineering. Research questions are:

> "What are the characteristics of Software Engineers?"

> "What (de)motivates Software Engineers to be more (less) productive?"

> "What are the external signs or outcomes of (de)motivated Software Engineers?"

> "What aspect of Software Engineering (de)motivate Software Engineers?"

> "What models of motivation exist in Software Engineering?"

The research-oriented review by Kitchenham & Brereton (2013) focuses on primary studies that address aspects of the systematic review process in software engineering. the research questions addressed are:

> "What papers report experiences of using the systematic review methodology and/or investigate the systematic review process in software engineering between the years 2005 and 2012 (to June)"?

> "To what extent has research confirmed the claims of the systematic review methodology?"

> "What problems have been observed by software engineering researchers when undertaking systematic reviews?"

> "What advice and/or techniques related to performing systematic review tasks have been proposed and what is the strength of evidence supporting them?"

### Examples of research questions from mapping studies

Walia & Carver (2009) report a technology focused mapping study about the sources of requirements faults. The high level research question addressed by this review is:

> "What types of requirements errors can be identified from the literature and how can they be classified?"

This is decomposed into four more specific questions (some with sub-questions). The four specific questions are:

> "Is there any evidence that using error information can improve software quality?"
>
> "What types of requirement errors have been identified in the software engineering literature?"
>
> "Is there any research from human cognition or psychology that can propose requirement errors?"
>
> "How can the information gathered in response to the above questions be organized into an error taxonomy?"

Another technology focused mapping study by Marshall & Brereton (2013) identifies and classifies tools developed to support the systematic review process in software engineering. The research questions for this mapping study are:

> "What tools to support the systematic review process in software engineering have been reported?"
>
> "Which stages of the systematic review process do the tools address?"
>
> "To what extent have the tools been evaluated?"

The study by Ampatzoglou & Stamelos (2010) maps research relating to software engineering for games development. It addresses the following research questions:

> "What is the intensity of the research activity on software engineering methods for game development?"
>
> "What software engineering research topics are being addressed in the domain of computer games?"
>
> "What research approaches do software engineering researchers use in the domain of computer games?"
>
> "What empirical research methods do software engineering researchers use in the domain of computer games?"

### Examples of research questions from tertiary studies

The tertiary study by  Marques et al. (2012) maps reviews about distributed software development (DSD). The research questions are:

> "How many systematic literature reviews have been published in the DSD context?"
>
> "What research topics are being addressed?"
>
> "What research questions are being investigated?"
>
> "Which individuals and organizations are involved in systematic literature review-based DSD research?"
>
> "What are the limitations of systematic literature reviews in DSD?"

The study by Cruzes & Dybå (2011*b*) focuses on the synthesis stage of the systematic review process and addresses three questions:

> "In terms of primary study types and evidence that is included, what is the basis of software engineering systematic reviews?"
>
> "How, and according to which methods, are the findings of systematic reviews in software engineering synthesized?"
>
> "How are the syntheses of the findings presented?"

Kitchenham, Pretorius, Budgen, Brereton, Turner, Niazi & Linkman (2010) report a broad research-focused tertiary study of systematic reviews and mapping studies in software engineering. Research questions are:

> "How many systematic reviews were published between 1st January 2004 and 30th June 2008?"
>
> "What research topics are being addressed by systematic reviews in software engineering?"
>
> "Which individuals and organisations are most active in research on systematic reviews?"

## 4.4 Developing the protocol

A systematic review or mapping study protocol is a documented plan describing, as far as possible, all of the details about how a review will be conducted. A protocol is particularly valuable because: (1) it can help to reduce the probability of researcher bias by limiting the influence of researcher expectations on, for example, the selection of individual (primary) studies or the synthesis of results; (2) it can be evaluated by other researchers who can provide feedback about the design of a review in advance of its conduct; and (3) it can form the basis of the introduction and method sections of a report of a review.

It is important that a protocol is structured in such a way that it can be easily used as a reference document by a review team and can be updated as necessary during the conduct of a review. We stress that a review protocol is a living document that is likely to be updated during the conduct of a review. An example template for systematic review and mapping study protocols is shown in Figure 22.5.

As well as covering all of the technical elements of a review, a protocol can provide information about the management of a review project. This can include the allocation of roles, mechanisms for resolving disagreements and the project schedule.

The following subsections summarise the main components of a protocol.

### 4.4.1 Background

The background section of a protocol provides a summary of related reviews and the justification for a review. Establishing the need for a review is discussed in Section 4.1.

### 4.4.2 Research questions(s)

This is a critical component of a protocol because the research questions drive the later stages of the review process. Specifying the research questions is discussed in Section 4.3.

### 4.4.3 Search strategy

The strategy for finding appropriate studies will describe and justify the way in which specific searching methods, such as automated searching, manual searching, snowballing and contacting key researchers, are combined. If an automated search is planned, this component will include a description of the search strings and resources, such as digital libraries or indexing services, that will be used. For a manual search, suitable journals and conference proceedings should be specified and their selection justified. This part of a protocol will also include a description of the mechanism for validating the search process. Chapter 5 and Part III provide further details and advice about search strategies and approaches to validation. Management decisions that are specific to the search process, such as the allocation of members of the review team to the searching tasks and the approach to resolving disagreements can also be recorded here.

### 4.4.4 Study selection

In this component, reviewers specify (1) the study selection criteria for determining whether a primary study is included in or excluded from a review and (2) procedures that will be followed to apply the criteria. The inclusion and exclusion criteria relate closely to the research questions and hence will be formulated to ensure the inclusion of those studies that can contribute to answering these questions.

The likelihood is that criteria are applied in a number of stages. For example, initial decisions can be based on the title or the title, abstract and keywords of a paper in order to exclude those that are clearly irrelevant. In later stages, reviewers will read candidate papers in full. Marginal papers, or those for which inclusion/exclusion is uncertain, can be kept in the inclusion set with the final decision being made during data extraction. This situation is most likely to arise for qualitative systematic reviews. For quantitative systematic reviews the criteria are usually easier to apply and for mapping studies leaving out a few papers, or including a few extra papers is not usually critical. Plans might also address the allocation of team members to the stages of study selection and the resolution of disagreements. There is further discussion of study selection in Chapter 6 and in Part III.

### 4.4.5 Assessing the quality of the primary studies

This is a particularly challenging task relying on two key decisions. One is to decide on the criteria against which quality will be assessed and the other is to establish the procedures for applying the criteria. The *criteria* will usually be expressed as one or more checklists depending, at least in part, on the range of evaluation methods used in the primary studies. Evaluation methods may include experiments, surveys, case studies and experience reports. One approach is to use separate checklists for each study type. The alternative is to use a generic checklist across all study types. Each of these approaches has limitations. For mapping studies, where the goal is to map out a domain of interest, assessing the quality of the individual studies may not be needed.

*Procedures* for applying the quality criteria are specified in a way that aims, as far as is possible, to ensure the reliability of the assessment. Mechanisms that can be used for this include having all, or a sample of, assessments checked by another person or having two reviewers perform the assessment independently. As well as describing who will undertake the quality assessment and the mechanism for resolving disagreements, a protocol can record decisions about the use of forms or tools to manage both individual scores and the outcomes of the resolution process. Whatever the type of review, it is important to consider the purpose of assessing the quality of the primary studies and to justify the approaches taken. There is further discussion of quality checklists, their limitations and assessment procedures in Chapter 7 and Part III.

### 4.4.6 Data extraction

This part of a protocol defines the data that will be extracted and the procedures for performing the extraction and for validating the data. The data will include publication details for each paper plus the information that is needed to answer the research questions. Extracting qualitative information presents a particular challenge since specific pieces of text need to be extracted and linked to specific research questions. Where qualitative synthesis is planned, data extraction and synthesis can be combined within a single process. For mapping studies in particular, data extraction may be iterative since important trends and ways of categorising papers may only become evident as individual papers are read. These challenges have led to an interest in the use of textual analysis tools to support data extraction and other aspects of the systematic review process (see Chapter 13).

A protocol should also define how data will be recorded (for example, using a review support tool or spreadsheet), who will perform the data extraction and how disagreements will be resolved. One approach is for a review leader to extract standard publication data and for two reviewers to extract data that is specific to a review. Strategies for resolving disagreements include discussion and using a third reviewer. The data extraction strategy (i.e. the selected data items and the procedures) should be justified. There is further discussion and advice about data extraction in Chapter 8 and in Part III.

### 4.4.7 Data synthesis and aggregation strategy

This section of a protocol defines the strategy for summarising, integrating, combining and comparing the findings from the primary studies included in a review. For quantitative data there is usually little opportunity to undertake a formal meta-analysis for software engineering studies. However, where meta-analysis is planned, details of the techniques to be used should be included. More commonly, for systematic reviews in software engineering, primary studies are too heterogeneous for statistical analysis and a qualitative approach (such as vote counting) has to be used.

The studies included in a review are often qualitative in nature and use a wide range of empirical methods. For textual data, synthesis is generally an iterative process because authors use different terminology to describe the same concepts (and sometimes use the same terminology to describe different concepts). Also, if the text is to be coded, the codes will be derived after reading the papers and need to be agreed on by all of the members of the review team who are performing the coding. Combining findings across multiple methods is especially challenging ((Cruzes & Dybå 2011*b*), (Kitchenham & Brereton 2013)). Common approaches to synthesis include narrative and thematic synthesis where data is tabulated in a way that is consistent with the research questions. For mapping studies, the goal should be to classify the findings in interesting ways and to present summaries using a variety of tabular and graphical forms. Further information and advice about a range of approaches to data synthesis and aggregation is provided in Chapters 9, 10 and 11, and in Part III.

### 4.4.8 Limitations

This section can be used to document the limitations of a review that are inherent to its context. Essentially these are limitations that have not or cannot be addressed by the review design. One example of this type of validity problem is where the data is extracted from papers that were written by the reviewers. The data could be based on the reviewers' understanding of their own research rather than the information actually reported in the papers.

### 4.4.9 Reporting

It is useful to consider, in advance, the approach that will be taken to disseminating the findings of a review. Usually a review is reported as a detailed technical report, as a conference paper (or papers) and/or as a journal paper. A technical report (or a chapter in a PhD thesis) and a journal paper can include all, or at least most, of the information that is needed to provide traceability from individual primary studies to the results and conclusions of a review and to demonstrate rigour in applying the review process. Conference papers, however, are usually limited in size and hence may need to provide links to additional information. A protocol should record agreements about the list of authors for each publication and about the target audience. Further details about reporting can be found in Chapter 12 and Part III.

### 4.4.10 Review management

This section covers management decisions, for example relating to scheduling and to tool support, that have not been recorded in other parts of the protocol. Further details about tool support can be found in Chapter 13.

## 4.5 Validating the protocol

In this component, reviewers specify the steps that will be taken, both internally and externally, to validate the protocol. Internal validation will include trialling specific aspects of the review plan such as the search strings and the data extraction forms to be used as well as the processes to be followed for data synthesis and/or aggregation. Also, we have emphasised the key role played by a protocol so it is important that it is evaluated by researchers who are external to a review team. PhD students should at least have their protocol evaluated by members of their supervisory team and might also call upon independent researchers, particularly if their supervisors have limited experience of the process. Evaluators can check a protocol against review guidelines, looking to confirm that the main elements are covered, that the decisions made are justified, that validation is adequately addressed and that a protocol is internally consistent. Authors of a protocol should provide evaluators with a checklist or set of questions addressing each of the elements of a protocol. Table 4.1 lists some examples of questions about each of the elements.

**TABLE 4.1**: Example Questions for Validating a Protocol

| Components | Example questions |
|---|---|
| Background | Is the motivation for the review clearly stated and reasonable?<br>Are related reviews summarised? |
| Research questions | Do these address a topic of interest to practitioners and/or researchers?<br>Are they clearly stated? |
| Search strategy | Is the strategy justified and is it likely to find the right primary studies without the reviewers having to check or read a large number of irrelevant papers?<br>For automated searches, is there likely to be a substantial level of duplication of papers found across the set of electronic resources used?<br>Has the strategy been validated?<br>Is it clear which members of the review team will perform the searching? |
| Study selection | Are inclusion/exclusion criteria clearly defined and related to research questions?<br>Is a staged process defined?<br>Is a validation process specified?<br>Are the roles of the team members defined for each stage of the process and is the mechanism for resolving disagreements specified?<br>Is there a process for handling marginal and uncertain papers (especially for qualitative reviews), and for managing multiple reports of individual studies? |
| Quality of primary studies | If quality is to be assessed, is it clear that the outcomes will be used in the later stages of the review?<br>Are criteria for assessing quality provided and justified (given the range of primary study types anticipated in the review)?<br>Is a validation process specified?<br>Are the roles of team members and the process for resolving disagreements specified? |

**TABLE 4.1**: Example Questions for Validating a Protocol

| | |
|---|---|
| Data extraction | Does the data to be extracted properly address the research questions?<br>Are the methods of recording the data appropriate for the types of data to be extracted (e.g. using forms, tables, spreadsheets or more advanced tools)?<br>Have these been adequately piloted?<br>Are there mechanisms for iteration where data is qualitative and categories are not (or cannot be) fully defined in advance of the extraction?<br>Is a validation process specified?<br>Are roles and strategies for resolving disagreements specified?<br>If textual analysis tools are to be used, is their use justified?<br>Will the data extracted by each reviewer, and any agreed values where reviewers differ, be appropriately stored for later analysis? |
| Data aggregation and synthesis | Will the process enable the research questions to be answered?<br>Are the methods proposed for qualitative and quantitative data appropriate?<br>Have they been piloted?<br>Has consideration been given to combining results across multiple study types?<br>Is the approach to aggregation and synthesis justified with reference to appropriate literature? |
| Reporting | Has this been considered?<br>If the aim is to publish the review (or even if it is not!), has sufficient attention been paid to completeness, general interest, validation, traceability and the limitations of the review?<br>Has the authorship of reports been considered? |
| Review management | Is the proposed schedule realistic?<br>Have roles and responsibilities been defined for the stages in review?<br>Are the tools that will be used for managing papers, studies and data specified and appropriate (and available)?<br>Is the management of the many-to-many relationship between papers and studies addressed? |