Chapter 5

Searching for Primary Studies

The focus of this chapter is on the identification of relevant primary studies. This process forms the first step of the conduct phase of the systematic review process, as highlighted in Figure 5.1.

An important element of any systematic review or mapping study is to devise a search strategy that will find as many primary studies as possible that are relevant to the research questions. The likelihood is that the strategy will involve a combination of search methods. One widely used method is automated searching of the literature using resources such as digital libraries and indexing systems. Other methods include manual searching of selected journals and conference proceedings, checking papers that *are cited* in the papers included in a review (backwards snowballing) and checking papers that *cite* the papers included in a review (forwards snowballing). The search strategy will aim to achieve an acceptable level of *completeness* (see Section 5.1) within the review's constraints of time and human resources. The level of completeness that might be targeted will depend on the type of review being undertaken. Generally, for a quantitative systematic review, that is, one which compares software engineering technologies, a high level of completeness

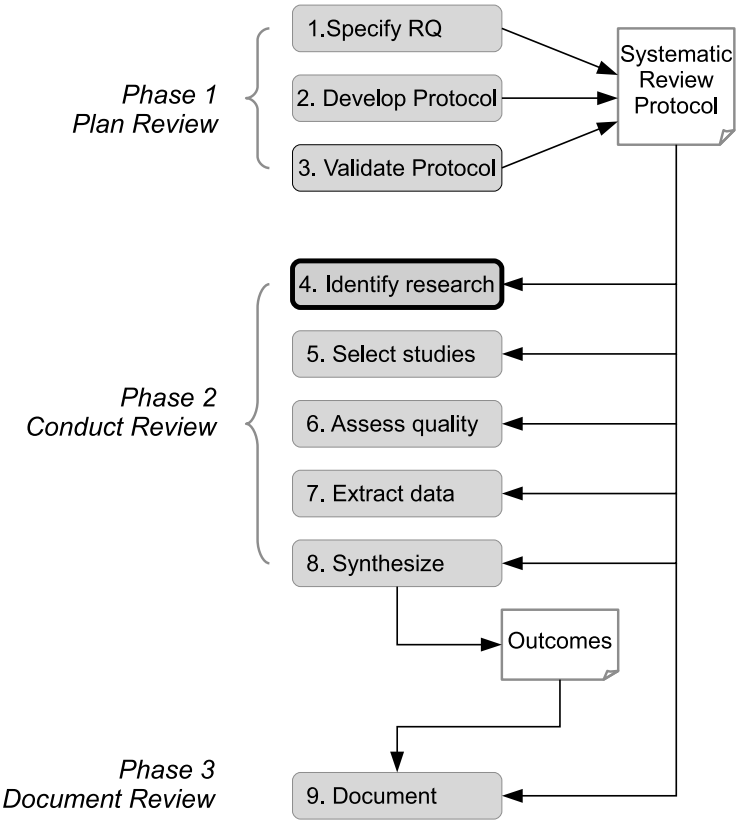

FIGURE 5.1: Searching stage of the systematic review process.

is essential. For other types of review, such as qualitative systematic reviews which assess risks, benefits, motivating factors, etc. or which review research processes, and for mapping studies, a lower level of completeness may be acceptable. This point is discussed in more detail in the following section.

Once a set of candidate papers have been found, and duplicate copies of the same paper have been removed, references can be managed in, for example, a spreadsheet or a database. Tools to support the systematic review process are discussed in Chapter 13.

Within the following sections we look at assessing the completeness of the set of papers found by the search process and discuss a range of searching methods that can be used as part of a strategy. Also, examples of strategies are presented for different types of systematic reviews and mapping studies.

5.1 Completeness

We consider two aspects of the completeness of the set of papers found by following a search strategy. The first relates to *completeness target*: how 'complete' should the set of papers be? The second relates to *completeness*

assessment: once we have a target, how will we know whether we have achieved it?

How complete?

A big question facing many reviewers is when to stop searching; and of course the answer is 'it depends'.

For *quantitative systematic reviews*, completeness is crucial. If we look at the examples described in Chapter 3, which relate to methods of estimating software development effort (Jørgensen 2004), pair programming versus solo programming (Hannay et al. 2009) and perspective-based reading (PBR) compared to other approaches to reading (Basili et al. 1996) we see that 15 studies are included in the cost estimation review, 18 in the pair programming review and 21 in the PBR review (with many of the primary studies in the last of these considered to be replications rather than independent studies). Given the highly focused nature of these reviews and the small numbers of included studies, missing only a few of these could substantially affect the outcomes of the reviews.

For other types of review, a lower level of completeness may be acceptable. For example, the *qualitative systematic review* by Beecham et al. (2008) aimed to 'plot the landscape of current reported knowledge in terms of what motivates developers'. The review includes 92 papers which report motivators, many of which are common across many of the papers (most of which report some form of survey). Failing to include some of these 92 papers would not have substantially affected the 'landscape of knowledge'.

Another situation where completeness may not be critical is where *mapping studies* are performed during the early stages of a research project (such as a PhD project). The value of the mapping study may come from acquiring a broad understanding of the topic and from identifying clusters of studies rather than from achieving completeness (Kitchenham, Brereton & Budgen 2012). However, a point to note here is that if a mapping study provides the basis for a more detailed and focused analysis (for example, where the presence of a cluster indicates that quantitative analysis may be feasible and valuable) it should not be assumed that the set of papers identified is complete. In this case a more focused search should be performed unless it can be demonstrated that the mapping study is of high quality in terms of completeness and rigour (Kitchenham, Budgen & Brereton 2011).

It can be argued that in some cases the level of completeness of *tertiary studies* should be high. Where a tertiary study aims to provide a catalogue and detailed analysis of systematic reviews across the software engineering domain (or across a broad sub-domain such as global software development), it can provide a key reference document for the community and as such should be as complete as possible. The argument for a high level of completeness may not be quite so compelling where a tertiary review is performed as a preliminary study, for example, to identify related reviews in advance of a

more focused mapping study or systematic review. In the end, knowing when to stop searching depends on what level of completeness is needed in order to provide satisfactory answers to the research questions being addressed by a review.

Completeness assessment

There are two fundamental ways of assessing the completeness of the set of studies found by searching the literature. One is to use personal judgement. This may be the judgement of members of the review team, especially if they are experienced researchers on the topic being reviewed or it may involve external researchers whose views are sought by a review team at some point in the process. Whatever the source of personal knowledge, it is difficult to quantify the level of completeness achieved using this subjective approach. The alternative is to use some objective measure of the level of completeness.

Two key criteria for assessing the completeness of an automated search are *recall* (also termed sensitivity) and *precision* (Dieste, Grimán & Juristo 2009, Zhang, Babar & Tell 2011).

The *recall* of a search (using particular search strings and digital libraries/indexing systems) is the proportion (or percentage) of all the relevant studies that are found by the search.

The *precision* of a search is the proportion (or percentage) of the studies found that are relevant to the research questions being addressed by a review. These can be calculated as follows:

$$Recall = \frac{R_{found}}{R_{total}} \tag{5.1}$$

$$Precision = \frac{R_{found}}{N_{total}} \tag{5.2}$$

where:
R_{total} is the total number of relevant studies
N_{total} is the total number of studies found
R_{found} is the number of relevant studies found

Of course the practical problem in calculating recall is that the denominator, that is, the total number of relevant studies (R_{total}), is not known. Ideally, a search should have high recall, that is, it should find most (if not all) of the relevant studies. Precision is also important and high precision is desirable. High precision means that the burden on reviewers to check papers that turn out not to be relevant is low. If precision is reduced, for example as a consequence of efforts to improve recall, the reading load on reviewers will increase.

In the following section we look at how these measures can be used to validate a search strategy by assessing the completeness of the set of studies found.

5.2 Validating the search strategy

Developing a search strategy is an iterative process, involving refinement based on some determination of the level of completeness achieved. An essential basis for the subjective assessment of completeness is having a set of papers which are known to report relevant studies. This *known set* can be obtained in a number of ways:

- Through an informal automated search using a small set of digital libraries or indexing systems, or a manual search of a small set of relevant conferences and journals,
- Using the personal knowledge of researchers who have experience in the topic of the review,
- Using a previous systematic or traditional literature review which addresses a similar or overlapping topic,
- Through the construction of a quasi-gold standard. The use of a quasi-gold standard to assess completeness is discussed later in this section.

If the number of studies in the known set is considered to be large (although of course it is not easy to decide what constitutes a large known set) then a search process that finds all of these may be judged adequate. The argument here is that if these are found then it is likely that most of the other relevant studies have also been found. Personal judgement, based on knowledge of the topic of a review, has to be used to decide whether the number of known papers can be considered large enough. To give an idea of the numbers of studies that might be included in a review, we note that the numbers included in the reviews catalogued by the third broad tertiary study (da Silva et al. 2011) range from 4 to 299, although well over half fall in the range 15–80.

When reviewers are not confident that the number of known studies can be considered large, a quasi-gold standard can be constructed and used to assess completeness (Zhang et al. 2011). The quasi-gold standard is determined by performing a manual search across a limited set of topic-specific journals and conference proceedings over a restricted time period. The set of relevant papers found is then used to assess the completeness of an automated search. The approach has been evaluated through two participant-observer case studies with promising results (Zhang et al. 2011). The approach, shown in Figure 5.2, has the following steps:

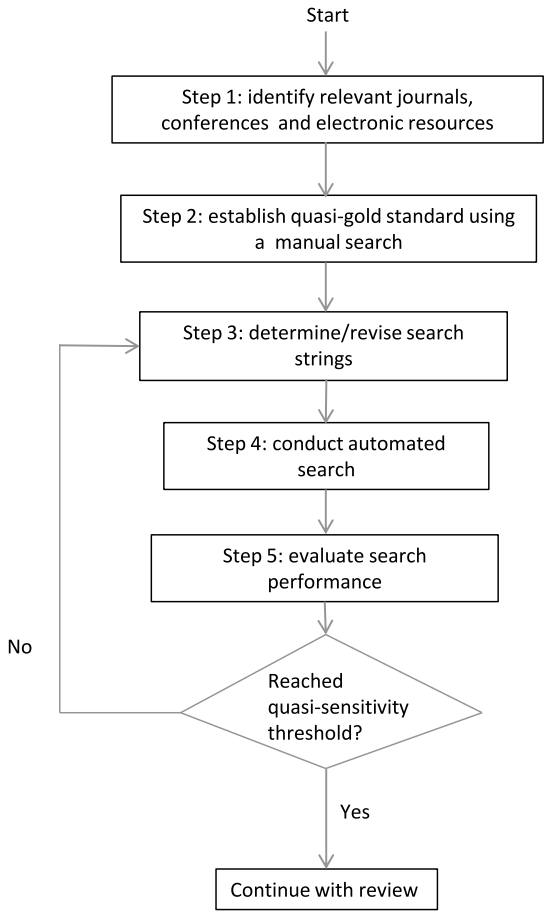

FIGURE 5.2: A process for assessing search completeness using a quasi-gold standard.

Step 1: Identify relevant journals, conferences and electronic resources

In this step, reviewers decide which journals and conference proceedings will be searched manually (in Step 2) and which digital libraries and indexing services to use for the automated search (Step 4). Manual searching is quite time consuming so the aim is to chose those outlets that are most likely to publish relevant papers. Selecting electronic sources for the automated search is discussed later in this chapter (see Section 5.3).

Step 2: Establish quasi-gold standard using a manual search

This step involves performing a manual search of the selected journals and conference proceedings over the chosen (and limited) time period. Essentially, the review team screen all of the papers in the selected sources and apply the inclusion and exclusion criteria, which should be defined in advance. Screening

can be applied initially to the title and abstract of a paper (keywords could be considered too) and then, if a decision cannot be made, other parts of a paper, possibly the whole paper, can be read. The development and use of inclusion and exclusion criteria are discussed in more detail in Chapter 6.

Step 3: Determine/revise search strings

Zhang et al. suggest two ways of defining the strings to used to search the selected electronic resources. These are:

1. Subjective search string definition based on domain knowledge and past experience,
2. Objective elicitation of terms from the quasi-gold standard using a text analysis tool.

Search strings can also be derived from the research questions being addressed by a review (see Part III, Section 22.5.2.2 for practical advice about constructing search strings).

Step 4: Conduct automated search

Here, the selected electronic resources (digital libraries or indexing systems) are searched using the strings determined in Step 3 and for the chosen time period. Automated search is discussed in more detail in Section 5.3 and in part III, Section 22.5.2.

Step 5: Evaluate search performance

In this step, the results of the automated search are compared to the results of the manual search (the quasi-gold standard) and quasi-sensitivity is calculated. For the calculation, using equation 5.1, R_{found}, is the number of relevant studies found by the automated search (step 4) that are published in the venues used in Step 2 (the manual search) during the time period covered by the manual search. R_{total}, the total number of relevant studies for the selected venues and time period, is the number of papers found by the manual search (Step 2). Similarly quasi-precision can be calculated using equation 5.2 where N_{total} is the total number of papers found by the automated search.

Zhang et al. suggest that a sensitivity (recall) threshold (i.e. a completeness target) of between 70% and 80% might be used to decide whether to go back to Step 3 (and to refine the search terms) or whether to proceed to the next stage of the review. These percentages are based on the scales developed by Dieste & Padua (2007) who in turn based their scales on research in the medical domain. Clearly this is a judgement that must be made on a case by case basis and will depend on a number of factors such as the completeness target and the available human resources.

A refinement of the quasi-gold standard approach has been proposed by Kitchenham, Li & Burn (2011) who suggest that the set of known papers is divided into two sets with one being used to construct the search strings and the other to evaluate the effectiveness of the search process.

5.3 Methods of searching

As we have indicated in the introduction to this chapter, there are a number of ways of searching for relevant primary studies. In practice, methods are often combined in some way to achieve good coverage. In this section we describe the most commonly used methods and in the following section illustrate their use across a range of systematic reviews and mapping studies. In addition to the methods described here, reviewers can consider contacting researchers directly where they are known to be actively engaged in research in the specific topic area being addressed by the systematic review or mapping study.

We also note that it can be hard to find papers when the topic of a review is secondary to that of many of the relevant primary studies. This might arise, for example, where a review is about tool usage or about research methodology. In these circumstances the best method to choose for searching might be a manual search (looking at particular sections of a paper) or alternatively an automated search where the searching process accesses the complete text of a candidate paper (as opposed to just the title and abstract).

Automated search

This approach has been widely adopted by software engineering reviewers and involves the use of electronic resources such as digital libraries and indexing systems to search for relevant papers. In order to perform an automated search reviewers have to address two elements of the process. They have to decide which electronic resources to use and they have to specify the search strings that will drive the search.

Two key publisher-specific resources are the IEEE Digital Library (IEEEXplore) and the ACM Digital Library which together cover the most important software engineering (and more general computing) journals and conference proceedings. A tertiary review focusing on the period mid-2007 to end of 2008 found that IEEEXplore was used in 92% of the 38 reviews that were included and 81% used the ACM Digital Library (Zhang et al. 2011). ScienceDirect (Elsevier) was also quite extensively used for the systematic reviews included in the tertiary study.

General indexing services such as Scopus and Web of Science will find papers published by IEEE, ACM and Elsevier (although not necessarily the

most recent conference proceedings). They also index papers published by Wiley and Springer and hence such services reduce the need for searching some publishers' sites.

Although some publishers provide open access to some papers, many require a payment, or a subscription, to obtain copies of full papers. Many universities now subscribe to publishers' packages of journals and conference proceedings, and there is also a growth in open access journals. Also, at some academic institutions, authors are required to put pre-publication versions of their papers into the University's open access catalogue. Additionally, pre-publication versions can sometimes be found by looking at an author's website. The publishing landscape for academic journals and conference proceedings is changing quite rapidly at the moment so we suggest that reviewers check with their library services and with publishers' websites to get an up-to-date-picture of their best route to acquiring access to full papers.

Generally, digital libraries and indexing systems provide mechanisms for exporting the bibliographic details of papers in a range of formats such as BibTeX, EndNote and Refworks.

Defining and refining search strings is an iterative process as illustrated in the quasi-gold standard approach described in the previous section. An initial set of keywords can be determined in a number of ways, such as:

- Extracting software engineering concepts and terms from the research questions,
- Reviewing terms used in the known papers,
- Identifying synonyms of the key terms.

As indicated earlier, it is a tricky balance between a search which finds most of the relevant papers (that is, having a high recall/sensitivity) and one which achieves a good level of precision (that is, not generating a large number of irrelevant papers). Even if the quasi-gold standard approach is not used, some iteration will be needed to ensure that all known papers that can be found by an automated search (that is, those that are indexed by the electronic systems being used) are included in the list of papers generated by the search.

Manual search

Manual searching of software engineering journals and conference proceedings can be very time consuming and onerous especially if the topic of a review is broad (so that the papers are not limited to a few specialist's outlets) or where the topic is quite mature (so that a large time span needs to be covered). The key decisions here are identifying the most appropriate journals and conferences and determining the date from which to start the search. Manual search can be particularly valuable for multidisciplinary reviews (see for example the mapping study by Jorgensen & Shepperd, summarised in Section 5.4, which addresses the topic of cost estimation). In general it is useful to

have team members from the different domains covered by a multi-disciplinary review.

If the search validation mechanism is strong, for example where an independent search is performed by two or more reviewers and the agreement between them is high, a manual search can provide what is effectively a gold standard set of relevant papers. Achieving a gold standard set of papers in this way may not be practical except perhaps where the topic is highly focused, reviewers are experts in the subject area, and the time span for the search is not large.

Advice about selecting appropriate sources to use for a manual search can be found in Part III, Section 22.5.3.

Snowballing

Snowballing, also referred to as citation analysis, can take one of two forms. *Backwards snowballing* is where a search is based on the reference lists of papers that are known to be relevant (the included set). It is usually used as a secondary method to support automated search. *Forwards snowballing* is the process of finding all papers that cite a known paper or set of known papers. This approach is particularly useful where there are a small number of seminal papers that are likely to be cited by most of the subsequent papers on the topic. Skoglund & Runeson (2009) compare the recall (sensitivity) and precision of two snowballing approaches based on citation analysis with those of three historic reviews, where two had used automated searching and the other had used a manual search. The outcomes were quite varied across the three example reviews and no general conclusions were reached. A study by Jalali & Wohlin (2012) compared automated search and backwards snowballing for a review on Agile practices in global software engineering. They found that:

- Precision was better when using the snowballing approach,
- Although different papers were found by the two approaches there was a substantial degree of overlap,
- Conclusions drawn using each of the approaches were very similar.

5.4 Examples of search strategies

We summarise a range of strategies reported by researchers who have performed systematic reviews and mapping studies.

Examples of search strategies for systematic reviews

Kitchenham et al. (2007) report a quantitative systematic review of studies that compare cross-company and within-company cost estimation. The authors carried out their search in two stages. Initially, an automated search of six electronic databases and seven individual journals and conference proceedings, chosen because they had published known relevant papers, was performed. The set of known papers was also used to validate the automated search. For the second stage, the authors:

1. carried out backwards snowballing for the papers included after the initial search,

2. contacted researchers who have either authored relevant papers found by the initial search or who they believed to be working on the topic.

The qualitative systematic review by Beecham et al. (2008) focused on the motivation of software engineers. Following a piloting exercise, eight electronic resources were used in an automated search and a manual search was undertaken "directly on key conference proceedings, journals and authors". Additionally, for included papers, backwards snowballing was performed and the corresponding authors of the papers were asked whether they had any relevant material 'in press'.

Examples of search strategies for mapping studies

Jorgensen & Shepperd (2007) describe a mapping study which addresses a set of eight research questions about research on software cost estimation. The authors report a manual search of all volumes (up to their search date) of more than 100 peer-reviewed journals. Journals were identified by reading through the reference lists of known papers (important because it is a multidisciplinary topic), by searching for relevant journals and using their own experience. The reviewers constructed independent lists of potential journals and merged their lists.

da Silva et al. (2011) performed a research-focused broad tertiary study of systematic reviews and mapping studies in software engineering published between 1st July 2008 and 31st December 2009. The authors used a search strategy that combined automated search, manual search and backwards snowballing. The automated search was performed by two of the authors using six search engines and indexing systems (ACM Digital Library, IEEEXplore Digital Library, ScienceDirect, CiteSeerX, ISI Web of Science and Scopus). All of the searches except for the ISI Web of Science were based on the full texts of the published papers. The search process was validated using a set of known papers found by two earlier broad tertiary studies (Kitchenham, Brereton, Budgen, Turner, Bailey & Linkman 2009, Kitchenham, Pretorius,

Budgen, Brereton, Turner, Niazi & Linkman 2010). In parallel with the automated search, three of the authors performed a manual search of 13 journals and conference proceedings, selected because they had been used by the earlier tertiary studies (except where two of the conferences had been merged). The reviewers checked titles and abstracts. The two sets of candidate papers were merged and duplicates removed. Backwards snowballing was applied to the papers remaining after the study selection stage.

Zhang et al. (2011) replicated a published tertiary study (Kitchenham, Brereton, Budgen, Turner, Bailey & Linkman 2009) to evaluate a search strategy based on a quasi-gold standard (as described in Section 5.2). In Step 1 (identify relevant journals, conferences and electronic resources), the authors used personal experience and published journal and conference rankings to inform their selection of nine outlets for the manual search and four digital libraries for the automate search. In Step 2 (establish quasi-gold standard using a manual search), two of the authors performed independent manual searches of the selected outlets to establish a quasi-gold standard (after resolving disagreements). In Step 3 (determine/revise search strings) and Step 4 (conduct automated search), the search string was based on the authors' knowledge and on the papers in the quasi-gold standard and was coded to fit the syntax requirements of each of the search engines. In Step 5 (evaluate search performance), the quasi-sensitivity was calculated to be 65% which was considered to be below the required threshold (70–80%). The search string was reviewed and revised to include additional terms. This increased the quasi-sensitivity to 85% which was deemed acceptable.