TECHNICAL REPORT

RT – ES 679 / 05

Systematic Review in Software Engineering

Jorge Biolchini
(jorgebio@cos.ufrj.br)

Paula Gomes Mian
(pgmian@cos.ugfrj.br)

Ana Candida Cruz Natali
(anatali@cos.ufrj.br)

Guilherme Horta Travassos
(ght@cos.ufrj.br)

Systems Engineering and Computer Science Department

COPPE / UFRJ

Rio de Janeiro, May 2005

Systematic Review in Software Engineering

1. What is a Systematic Review

The term Systematic Review (SR) is used to refer to a specific methodology of research, developed in order to gather and evaluate the available evidence pertaining to a focused topic.

In contrast to the usual process of literature review, unsystematically conducted whenever one starts a particular investigation, a SR is developed, as the term denotes, in a formal and systematic way. This means that the research conduction process of a systematic type of review follows a very well defined and strict sequence of methodological steps, according to an aprioristically developed protocol. This instrument is constructed around a central issue, which represents the core of the investigation, and which is expressed by using specific concepts and terms, that must be addressed towards information related to a specific, pre-defined, focused, and structured question. The methodological steps, the strategies to retrieve the evidence, the focus of the question are explicitly defined, so that other professionals can reproduce the same protocol and also be able to judge about the adequacy of the chosen standards for the case.

Synonyms of this methodology that are to be found in the literature include the following terms: overview, research review, research synthesis, research integration, systematic overview, systematic research synthesis, integrative research review, and integrative review.

The type of acceptable evidence to be gathered in a systematic review is stated beforehand. The retrieved evidence is thoroughly reviewed, comparable to other types of evidence previously and elsewhere retrieved.

The evidence data are normalized in such a way as to make results from different studies comparable, in terms of their magnitude of effect, even when they are presented in diverse ways but related to compatible concepts. It is then possible, e.g., to compare studies which evidence is expressed by absolute risk reduction with others where it is expressed by relative risk.

Besides comparing results of individual studies, different kinds of syntheses can be done. The election mode allows the researcher to look for each study separately and counting them as “votes” about the question focus. For instance, in a specific SR conducted in the field of medicine, the researcher could find that, among 35 valid studies, 29 showed a positive result, while 5 showed no result, and one study showed a negative result. Internal comparison of studies, based on their specific parameters, can show contrasts and other kinds of differences that may elucidate distinct aspects of the question. In the same example, one could find that the negative effect must be due to a different dosage scheme, while the five studies that showed no result were conducted in subjects that had a different age distribution in comparison to the 29 positive ones.

Another type of research synthesis is known as meta-analysis, where the original individual studies are treated as if they were parts of one larger study, by having their data pooled together in one single and final result that summarizes the whole evidence. By selecting studies that are compatible in their quality level, and by taking strict care with their specific details, this methodological procedure can produce evidence as well as reveal aspects that the original studies are not individually able to elucidate. For instance, meta-analysis may prove that the results are statistically significant when small studies give inconclusive results with large confidence intervals. Besides that, when

conflicting results arise from different individual studies, meta-analysis may reconcile the data in a synthetic result, while each individual study can then be weighted and compared with it, so that other kinds of conclusions might be derived from these discrepancies.

1.1. Systematic and Unsystematic Reviews: Differences, Advantages, and Disadvantages

A literature review is usually an initial step in any research and development enterprise. From the viewpoint of scientific methodology it is in fact a recommended and necessary step for the professional to endeavor whenever starting a research project. Since science is a cooperative social activity and the scientific knowledge is the result of a cumulative process of this cooperation, the literature review is the means by which the researcher can perform a mapping of the existing and previously developed knowledge and initiatives in the field. The review can provide material to be used by the researcher in the work that is being designed, and locate it in relation to the different regions of the field and approaches to the issue in focus. It also permits both an analysis of the previous findings, techniques, ideas and ways to explore the topics in question, as well as their relevance in relation to the issues of interest, and a synthesis and summarization of this information. It can help planning the new research, avoiding unnecessary duplication of effort and error, and orient the investigation process. Due to the growth of scientific production, the role of literature reviews has been proportionally growing larger, and "their importance grows as a direct function of the number of documents on a topic" [Cooper and Hedges, 1994]. Due to its important role in the scientific enterprise, general rules for performing literature overviews have been developed, in order to warrant the investigator good quality of information from the covered material.

The systematic review consists in a specific scientific methodology that goes one step further than the simple overview. It aims to integrate empirical research in order to create generalizations. This integrative enterprise involves specific objectives, which allows the researcher to critically analyze the collected data, to resolve conflicts detected in the literature material, and to identify issues for planning future investigation. Due to these particular aims, the systematic review is not considered to be a phase of a research enterprise, a role that is performed by the usual literature review. As a matter of fact, the integrative review is a different methodological procedure of research in its own, comprising distinct investigation aims as well as specific methodological features, requirements, and procedures. From the epistemological perspective, it represents a different approach to the relevant issues in a research area that opens up a new field of possibilities for generating new types of knowledge in a scientific domain.

In practice, the distinction between ordinary review articles and systematic review ones can be done by comparing their underlying semantic structures, as evidenced by the types of contents in their respective abstracts as well as in the titles of the respective article sections. In the medical field, for instance, a simple overview article refers in its abstract to key points about the subject, without discussing or emphasizing the methodology of the review itself. The article sections include titles that refer to topics that are very similar to the sections that are usually found in a textbook chapter, such as the natural history of the disease referred to its different phases of evolution and expression, the characteristics of the symptoms and signs and the differential diagnosis with other diseases, causal mechanisms or hypotheses of the disease, the goals of the treatment, the types of drugs that might be used or recommended, other types of

intervention, and so on [Gross, 2001]. In contrast, a systematic review article abstract contains a specific pattern of sections, such as background, purposes, data sources, study selection, data extraction, data synthesis, discussion and conclusion. The article sections expand this same abstract section structure, including in its titles terms such as 'methods', 'data synthesis', 'efficacy', 'discussion', as well as describing and discussing the methodology of the research review itself. It also presents considerations about the specific requirements that were aprioristically defined and explicated in order to include or exclude the primary studies in the review material. It also includes tables containing quantitative information, such as the data extracted from the individual studies, results of each study weighted to account for the relative size of the study, a row entitled 'total', and sometimes individual study numbers reassessed as a new, aggregated pool of patients.

Despite the importance of the literature reviews, even when they are conducted according to their corresponding 'good practice' rules, they suffer from lack of scientific rigor in performing its different steps. The unsystematic conduction of this type of review might introduce, as it usually does, some research biases in different stages of the review process, ranging from question formulation, through data collection, data evaluation, analysis, interpretation, summarization, and presentation. The development of a systematic approach of research review aims to establish a more formal and controlled process of conducting this type of investigation, avoiding the introduction of the biases of the unsystematic review. Besides this central aspect, the systematic review does not consist on a simple rearrangement of the already known or published data. It is at the same time a new type of methodological approach for doing research, with an integrative purpose. Therefore it emphasizes the discovery of general principles, in a higher level of conceptual abstraction in the research field, the diagnosis and analysis of the relative external inconsistencies when comparing individual studies with contrasting results between themselves, as well as it helps to illuminate new aspects and issues in the field and guide future research lines and possibilities. For the classical approach of literature review, variation among studies tends to represent a source of noise, a disturbing factor for interpretation and judgment. For the systematic review methodology, on the contrary, variety is a stimulating factor for understanding the whole scenario of the particular issue that is under investigation, allowing the researcher to moderate the relative influences of the different individual studies, by viewing them as probabilistically distinct possibilities of result.

Like any other scientific methodology, the integrative and systematic review presents its potentials and also its limitations [Feinstein, 1995], [Liberati, 1995]. When compared to primary research, the unique contributions of research synthesis include the improvements in precision of the data and the reliability of the information, as well as the three aforementioned ones: testing hypotheses that possibly have never been, or could never be tested in primary studies; using consistent and explicit rules for an evidence-based process of moderating influences of primary studies; and, in a recursive way in relation to cumulative scientific knowledge, addressing questions about the research enterprise itself, such as trends of issues, concepts, methods, or results over time, as well as questions about the field research contexts in a broader sphere.

The main limitations of research integration are related to the nature of review-generated evidence and of post hoc hypothesis test. The first one can happen when the researcher compares the results of primary studies that used different procedures to test the same hypothesis. Because the antecedent variable is not aprioristically controlled, such as randomly assigning the issue of interest to the different types of study procedure, confounding variables are liable to exist and consequently interfere in the

integrated results. The consequence is that, in this situation, causal inferences are not possible to be done with the same degree of confidence. At the same time, it can provide the researcher with some good suggestions related to the future orientation of new primary studies.

The second limitation can derive from the fact that, in many cases, the researcher has in advance a reasonable knowledge about the empirical evidence related to the issue of interest. If the hypothesis stated for a specific systematic review is derived from the same data that will be integrated through this methodology, the researcher cannot use this same evidence to test the hypothesis so generated. In order to avoid a vicious circularity of the evidence base, the data to generate the review hypothesis and to test it must be independent.

These limitations reinforce the idea that primary research and systematic review are complementary processes of knowledge production. The second methodological approach cannot be considered to be a substitute for primary evidence production, in a competing way. On the contrary, the enhancement of precision and reliability provided by the systematic review process helps to improve and to better direct future primary research, through a positive feedback relationship between them.

1.2. The Origins of Systematic Review

Early works to integrate results can be traced back from the beginning of the 20th century [Cooper and Hedges, 1994]. Pearson, in 1904, calculates the average of results of the correlation between inoculation for typhoid fever and mortality in order to better estimate this type of effect and to compare it with that of the inoculation for other kinds of disease.

In the 1930s, methods for combining estimates are developed in other fields of research, such as the physical sciences [Birge, 1932] and in the statistical sciences [Tippett, 1931], [Fisher, 1932], [Cochran, 1937], [Yates and Cochran, 1938], and later applied to other fields, such as agriculture, with few methodological unfolding in the following decades. The use of synthesis techniques gains momentum in the 1970s, when new methodological proposals for integrating research results are developed as well as several applications are mainly developed in the social sciences.

In the methodology sphere, Feldman [1971] describes steps in the literature reviewing process; Light and Smith [1971] develops a methodological treatment of the variations in the outcomes of studies; Taveggia [1974] describes common problems in literature reviews and proposes to treat “contradictory findings” of individual researches in a particular topic as “positive and negative details of a probabilistic “distribution of findings” rather than “inconsistencies”; Glass [1976] defines the term meta-analysis as “the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings”.

In the application field, the main studies include the fields of clinical psychology [Smith and Glass, 1977], industrial/organizational psychology [Schmidt and Hunter, 1977], social psychology [Rosenthal and Rubin, 1978], education [Glass and Smith, 1979], and cardiology [Chalmers et al., 1977].

The field of research synthesis spreads from the social sciences to medicine in the 1980s, while books devoted to this methodology are published [Glass et al., 1981], [Hunter et al., 1982], [Rosenthal, 1984], [Hedges and Olkin, 1985], which are followed by others in the same decade. At this time, the research synthesis methodological approach becomes a new and independent specialty and achieves legitimacy as a field of research. The research review and the meta-analytic approaches are integrated, new

methods and techniques are developed, and a more rigorous level of methodology is achieved. Research synthesis also spreads to the social policy domain, to help the decision-making process, where Light and Pillemer [1984] emphasize the relevance of uniting numeric data and narrative information for the effectiveness of result interpretation and communication.

Since the late 1980s, systematic research synthesis and meta-analysis reach an especially distinctive methodological status in the health sciences domain [Piantadosi, 1997]. From then on, health policy agencies and legislation have fostered and relied on this methodology as a fundamental requirement to develop, publish, and recommend guidelines on clinical practice in the various medical specialties and application areas.

1.3. Examples of Systematic Reviews in Medicine

In this section, we present a sample of some titles of systematic review studies, developed in the medical area, that illustrate the specific nature of the results that are obtained by using this type of methodology.

The type of evidence that can be derived from primary studies differs according to the research design that is used for conducting them. The degree of confidence that is possible to obtain from an investigation experiment directly depends on the degree of experimental control that the researcher can exert over the object that is under study [Warren and Mosteller, 1993].

In the medical field, with the purpose of ascribing different values for the quality and scientific reliability of studies, a scale of distinct types of study design has been developed, in order to serve as a reference of the different confidence degrees of evidence that can be produced. Control measures of the experiment, such as blindness of either or both the patients and the clinicians that directly assess them, as well as comparison of subgroups inside the study, and also random assignment of the patients to the different subgroups, contribute to increase the reliability degree in the evidence that can be produced.

Therefore, the following evidence-grading reference system has been developed in the health field in order to help professionals to judge the quality of articles reporting scientific studies (Table 1).

Table 1. Level and Source of Evidence [Sackett et al., 2000]

Level of Evidence	Source of Evidence
1 a	Systematic Review of Randomized Controlled Trials
1 b	Individual Randomized Controlled Trial
1 c	“All or None” Case Series
2 a	Systematic Review of Retrospective Cohort Studies
2 b	Individual Retrospective Cohort Study; or Low Quality Individual Randomized Controlled Trial
2 c	“Outcomes” Research
3 a	Systematic Review of Case-Control Studies
3 b	Individual Case-Control Study
4	Case Series; or Low Quality Cohort Studies; or Low Quality Case-Control Studies
5	Expert Opinion without explicit critical appraisal, or based on physiology, bench research or “first principles”

The analysis of the syntactic structure of these titles shows that they contain two basic information units, one referring to the technology that is being studied, and the second one standing for the target problem and/or population.

This type of knowledge corresponds to what is known as foreground knowledge, specifically addressed for providing support to decision making processes, in contrast to the background knowledge type, which is acquired by the professional during one's training and educational process.

The italicized words were done by the authors, in order to evidence and make it explicit the difference between the linguistic pieces of the titles' text that refer to the aforementioned parts.

- Carotid endarterectomy *for* symptomatic carotid stenosis [Cina et al., 2000];
- Medium-dose aspirin or other antiplatelet drugs *for* patients at high risk *of suffering* some occlusive vascular disease *over the next few months or years* [ISIS-2, 1998];
- Meta-analysis of exercise testing *to detect* coronary artery disease *in* women [Kwok et al., 1999];
- Streptokinase or other "clot-busting" drugs *as* emergency treatment *for* patients *who are suffering* an acute heart attack [Early Breast Cancer Trialists' Collaborative Group, 1994];
- Chronic hepatitis B virus infection: treatment strategies *for the next millennium* [Malik and Lee, 2000];
- Lack of significant benefit of magnesium infusion *in* suspected acute myocardial infarction [ISIS-4, 1991];
- A systematic review of randomized controlled trials of pharmacological therapy *on* osteoarthritis of the knee, *with an emphasis on* trial methodology [Towheed and Hochberg, 1997] ;
- Hormonal adjuvant treatments *for* early breast cancer [Early Breast Cancer Trialists' Collaborative Group, 1992] ;
- A systematic review of newer pharmacotherapies *for* depression *in* adults: evidence report summary [Williams, et al., 2000].

2. Systematic Reviews in Software Engineering

Several primary studies have been conducted in the field of software engineering in the last years, accompanied by an increasing improvement in methodology. However, in most cases software is built with technologies for which developers have insufficient evidence to confirm their suitability, limits, qualities, costs, and inherent risks. It is difficult to be sure that changing software practices will necessarily be a change for the better. It is possible that research syntheses can provide the mechanisms needed to assist practitioners to adopt appropriate technologies and to avoid inappropriate technologies. Thus, the development of research syntheses in this field is still an area of investigation that remains to be explored and that could well bring many benefits.

In this context, there are few initiatives that question how Software Engineering would benefit from adopting the evidence approach. Kitchenham et al [2004] discuss the possibility of evidence-based Software Engineering by using an analogy with medical practice. Nevertheless in order to obtain evidence, it is necessary to perform systematic reviews. So, Kitchenham [2004] evolutes the idea of Evidence-Based Software Engineering and proposes a guideline for systematic reviews that is appropriate for software engineering researchers. The guideline has been adapted to reflect the specific problems of software engineering research and covers three phases of a systematic review: planning the review, conducting the review and reporting the

review. However, it is described at a relatively high level. It does not consider the impact of question type on the review procedures, nor does it specify in detail the mechanisms that are needed to undertake meta-analysis.

Like all knowledge areas that have previously developed this research methodology, developing this investigation approach in the software engineering field implies in adapting the conceptual and methodological dimensions of research synthesis to the domain, taking into account its specificities as a scientific knowledge area [Pressman, 2002].

Differently to the medical area, Software Engineering has some specificity that would make it difficult for the research synthesis to obtain evidence.

One major difference between medicine and software engineering is that most software engineering methods and techniques must be performed by skilled software practitioners that are aware of the methods and techniques that are being applied. In contrast, although medical practitioners are skilled individuals, the treatments they prescribe (e.g. medicines and other therapeutic remedies) do not necessarily require awareness of their effective presence in order to be skillfully administered by the professional or received by the patient. The reason why skill presents a problem in the software engineering field is due to the fact that it prevents adequate blinding of practitioners during the study. In medical experiments (particularly drug-based experiments), the gold standard experiment is a double-blind randomized controlled trial (RCT). In a double-blind experimental trial neither the doctor nor the patient knows which treatment the patient is being administered. The reason why double-blinded trials are required is to prevent patient and doctors expectations biasing the results. Such experimental protocols are impossible to conduct in software engineering experiments, which rely on a subject performing a human-intensive task.

Another difference between software engineering and medicine is that most software engineering techniques impact a part of the lifecycle, in such a way that it makes the individual effect of a technique difficult to be isolated. The target techniques interact with many other development techniques and procedures. In general, it is difficult to determine a linear causal link between a particular technique and a desired project outcome, when the application of the technique and the final outcome are temporally removed from one another, while at the same time there are many other tasks and activities involved in the study that could also affect the final outcome.

And also, differently in software engineering, medical researchers and practitioners look for already published systematic reviews, i.e., papers that have already assembled all relevant reports of a particular topic. Medical researchers have a large amount of technological and scientific infrastructure to support them. There are several organizations (in particular, the international Cochrane Collaboration - www.cochrane.com) that assemble systematic reviews of studies of drug and medical procedures. To provide a central information source for evidence, the Cochrane Collaboration publishes systematic reviews in successive issues of The Cochrane Database of Systematic Reviews. These reviews are continually revised, both as new experimental results become available and as a result of valid criticisms of the reports.

There is no equivalent to the Cochrane Collaboration in the Software Engineering area. Instead of it, there are many abstracting services that provide access to software engineering articles. Currently, available evidence related to software engineering technologies is [Kitchenham et al., 2004]:

- *Fragmented and limited*. Many individual research groups undertake valuable empirical studies. However, because the goals of such works are either to produce individual publications and/or to generate post-graduate theses, sometimes there is

little sense of an overall purpose to such studies. Without having a research culture that strongly advocates systematic reviews and replication, it is easier for researchers to undertake research in their own areas of interest rather than contribute to a wider research agenda.

- *Not properly integrated*. Currently, there are no agreed standards for systematic reviews. Thus, although most software engineering researchers undertake reviews of the "State of the Art" in their topic of interest, the quality of such reviews is variable, and they do not as a rule lead to published papers. Furthermore, if we consider "meta-analysis", which is a more statistically rigorous form of systematic review, there have been few attempts to apply meta-analytic techniques to software engineering not least because of the limited number of replications. In general there are few incentives to undertake replication studies in spite of their importance in terms of the scientific development of the area.
- *Without agreed standards*. There are no generally accepted guidelines or standard protocols for conducting individual experiments. Kitchenham et al. [2004] proposed some preliminary guidelines for formal experiments and surveys. However, they do not address observational, as well as investigative studies. Furthermore, because they attempt to address several different types of empirical study, the guidelines are not as specific, nor as detailed as they are found in the medical area.

3. How Systematic Reviews Can Be Conducted

From a broader conceptual perspective, related to the general categories of the units of study, the systematic review process conduction can be understood as a three-step approach (Figure 1). The first phase of the research starts from concepts, which explicitly and formally represent the issue in question, and goes to studies, which material potentially contains the information that can provide evidence about the specific topic of the investigation. The second phase starts from these studies, which are dissected in their contents, compared among themselves, and sometimes reassembled in their constituent parts, leading to results, which represent the emergence of a new type of evidence. The third phase goes from these results, through a process of analysis and synthesis of the new arrangements of data that are made possible through this methodology, towards the conclusions, which implicate in acquiring new knowledge about the issue in question as well as supporting some decision making related to it.

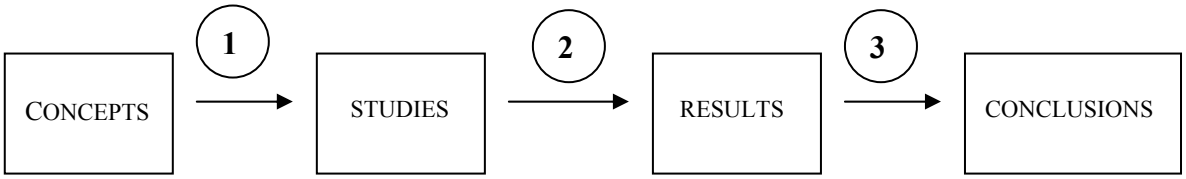

Figure 1. The Systematic Review Three-Step Approach.

From a more specific and operational point of view, the systematic review process can be defined as a five-step approach. The first stage is related to the Problem Formulation, in which the central issue refers to what kind of evidence should be included in the review. It then consists in constructing definitions that allow the researcher to establish a distinction between the relevant and the irrelevant studies for the specific purpose of the investigation. At this stage, narrow concepts might represent

a source of potential invalidity in the integrative review conclusions, since these can be made less robust and definitive. In a similar way, superficial operational detail in constructing definitions can also lead to conclusion invalidity, because it can help to hide interacting variables.

The second stage of the overview conduction is related to the Data Collection, when the main focus is addressed to define what procedures should the researcher use in order to find the relevant evidence that was defined in the preceding stage. As a major point, it includes determining the sources that may provide the potentially relevant studies to be included in the research, preferentially considering multiple channels for accessing primary studies as well as the how these channels might complement each other in their corresponding material. At this stage, potential invalidity in the systematic review can arise if the studies that are accessed correspond to a qualitatively different nature when compared to the target population of studies as defined in the protocol. The same type of problem can result if the specific study units and elements contained in the accessed studies differ from those related to the target population of the research. In medicine, for instance, it can derive from accessed studies that include people that happen to be different from the target population of people to be investigated.

The third stage of the systematic overview is related to the Data Evaluation, in which the nuclear issue refers to what retrieved evidence should be included in the review. It then consists in applying quality criteria to separate studies that can be considered valid from those that are to be considered invalid. It also consists in determining guidelines for the kind of information that should be extracted from primary research reports. The sources of potential invalidity in the systematic review conclusions, at this stage, include non-quality factors that can result in inadequate weighting of the study information, as well as omissions in study reports that might lead to unreliable conclusions.

The fourth stage of the integrative review is related to the Analysis and Interpretation process, when the main focus is addressed to define what procedures should the researcher use in order to make inferences about the collected data as a whole. As a relevant point, it includes synthesizing the valid retrieved studies so that generalizations about the issue in question might be possible to be done. At this stage, rules that are inappropriate for distinguishing patterns from noise can represent a source of potential invalidity in the systematic review conclusions. The same type of problem can derive from initiatives to use review-generated evidence in order to infer causality relationships.

The fifth stage of the systematic overview is related to the Conclusion and Presentation process, in which the central issue refers to what information should be included in the systematic review report. It then consists in applying editorial criteria in order to determine a clear separation between the important and the unimportant information. At this stage, omitting the overview procedures can represent a source of potential invalidity in the systematic review conclusions, since it leads to irreproducibility of the research itself as well as of its conclusions.

As we see, systematic reviews conduction is a tree-step approach. The main steps composing the systematic review process, as shown in Figure 2 are review planning, review execution, and result analysis.

During the *planning* phase, research objectives are listed and a review protocol is defined. Such protocol specifies the central research question and the methods that will be used to execute the review. The execution stage involves primary studies identification, selection and evaluation in accordance with the inclusion and exclusion criteria established in the review protocol. Once studies were selected, data from the

articles are extracted and synthesized during the *result analysis* phase. Meanwhile which one of these phases is executed, their results must be stored. Therefore, systematic review *packaging* is performed through the whole process. There are two checkpoints in the proposed systematic review process. Before executing the systematic review, it is necessary to guarantee that the planning is suitable. The protocol must be evaluated and if problems are found, the researcher must return to the planning stage to review the protocol. Similarly, if problems regarding web search engines are found during the execution phase, the systematic review may be re-executed.

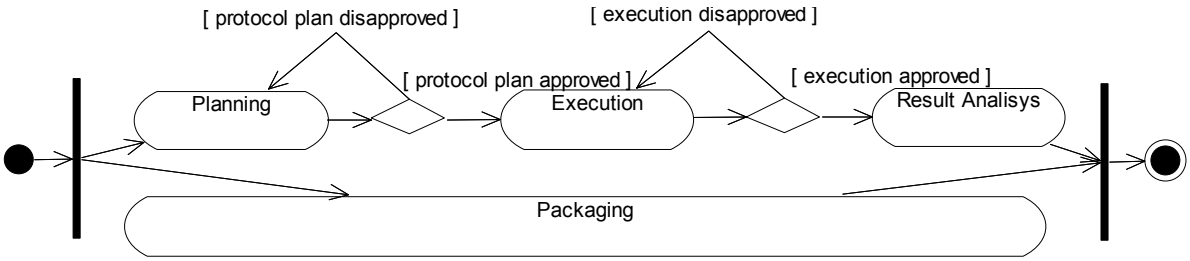

Figure 2. Systematic Review Process.

The stages listed above may appear to be sequential, but it is important to recognize that many of the stages involve iteration. In particular, many activities are initiated during the protocol development stage, and refined when the review proper takes place. For example [Kitchenham et al., 2002]:

- The selection of primary studies is governed by inclusion and exclusion criteria. These criteria are initially specified when the protocol is defined but may be refined after quality criteria are defined.
- Data extraction forms initially prepared during construction of the protocol will be amended when quality criteria are agreed.

Data synthesis methods defined in the protocol may be amended once data has been collected.

4. The Developed Template for Systematic Reviews in Software Engineering

Despite it importance, conducting systematic reviews is not a simple task. A systematic review uses specific concepts and terms that may be unknown to researches used to conduct informal reviews. Besides, systematic reviews require an additional conduction effort. The review must be planned before execution and the whole process must be documented, including, intermediary results.

To facilitate systematic reviews planning and the execution, a review protocol template has been developed. This template was based on systematic review protocols developed in the medical area, on the guidelines proposed by [Kitchenham, 2004] and the protocol example found in [Mendes and Kitchenham, 2004]. This template can be seen in Appendix 1.

The objective of this template is to serve as a guideline to Software Engineering researchers when conducting the systematic review. Therefore, the template lead researchers through each step of the systematic review process presented previously, defining clearly the content of each protocol section.

To illustrate this guidance process, we describe each phase of the systematic review process in terms of template sections.

4.1. Review Planning

In this phase, it must be defined the research objectives and the way the review will be executed, which includes to formulate research questions and to plan how the sources and studies selection will be carry out. The sections of the protocol template that guide the planning phase are shown bellow.

1. **Question Formularization:** in this section, the research objectives must be clearly defined. It is composed by the following items: Question Focus and Question Quality and Amplitude.
   1.1. **Question Focus:** defines the systematic review focus of interest, i.e., the review research objectives. Here, the researcher must decide what he/she expects to be answered in the end of the systematic review.
   1.2. **Question Quality and Amplitude:** this section aims at defining the syntax of the research question (the context in which the review is applied and the question the study must answer) and its semantics specificity (or question range) described by the remaining items of this section - intervention, control, effect, outcome measure, population and application. Each one of them described bellow:
      - **Problem:** defines the systematic review target, describing briefly the research context.
      - **Question:** research question to be answered by the systematic review. It is important to highlight that, if the systematic review context is too wide, it may be necessary to decompose the research question in secondary questions to narrow the research target.
      - **Keywords and Synonyms:** list of the main terms that compose the research question. These terms will be used during the review execution (in case the search by keywords is chosen as study selection methodology).
      - **Intervention:** what is going to be observed in the context of the planned systematic review.
      - **Control:** baseline or initial data set that the researcher already posses.
      - **Effect:** types of results expected in the end of the systematic review.
      - **Outcome Measure:** metrics used to measure the effect.
      - **Population:** population group that will be observed by the intervention.
      - **Application:** roles, professional types or application areas that will benefit from the systematic review results.
      - **Experimental Design:** describes how meta-analysis will be conducted, defining which statistical analysis methods will be applied on the collected data to interpret the results. Examples of statistical calculations application for result analysis can be found in [Juristo and Moreno 2001].
2. **Sources Selection:** the objective of this section is to select the sources where searches for primary studies will be executed.
   2.1. **Sources Selection Criteria Definition:** defines which criteria are going to be used to evaluate studies sources, i.e., which characteristics make these sources candidate to be used in the review execution.
   2.2. **Studies Languages:** it defines the languages in which obtained primary studies must be written. This item belongs to this section, and not to "Studies Selection", because the chosen language may restrain the sources identification.

2.3. **Sources Identification:** this item aims at selecting sources for the review execution.
   - **Sources Search Methods:** describes how to execute the search for primary studies (for instance, manual search, search through web search engines).
   - **Search String:** case one of the selected search methods includes using keywords in search engines it is necessary to create search strings to be run at such engines. This item presents a set of logical expressions that combine keyword and its synonymous arranged in a way that highest amount of relevant studies is obtained from search engines.
   - **Sources List:** initial source list in which the systematic review execution will be run.

2.4. **Sources Selection after Evaluation:** which element of the initial sources list, must be evaluated according to the source selection criteria. If the source fits all criteria, it must be included in the final sources list, presented in this session of the protocol.

2.5. **References Checking:** one or more expert must evaluate the sources list obtained from the previous item. Case the experts find the need to add new sources or to remove some of them, the result of such evaluation must be described in this item.

3. **Studies Selection:** once the sources are defined, it is necessary to describe the process and the criteria for studies selection and evaluation.

3.1. **Studies Definition:** this item defines the way studies will be selected.
   - **Studies Inclusion and Exclusion Criteria Definition:** presents the criteria by which studies will be evaluated to decide if they must be selected or not in the context of the systematic review. It is necessary to define these criteria because a search executed in web engines may find a great number of articles that do not answer to the research question. The main reason for this to happen is that a keyword may have different meanings or be used in studies that do not deal with the systematic review research topic. Therefore, it is necessary to define what makes an article a potential candidate to be selected or to be excluded from the review. Criteria can be found in literature, as in [Kitchenham et al., 2002], or be defined by the researchers.
   - **Studies Types Definition:** it defines the type of primary studies that are going to be selected during the systematic review execution. For instance: *in-vivo*, *in-vitro*, *in-virtuo* or *in-silico* studies [Travassos and Barros, 2003]; qualitative or quantitative studies; observation, feasibility or characterization studies.
   - **Procedures for Studies Selection:** it describes the procedure by which the studies will be obtained and evaluated according to exclusion and inclusion criteria. If the selection process has more then one stage, all of them must be described. Examples of studies selection procedures are reading the article abstract and reading the full study.

4.2. Planning Evaluation

Before executing the systematic review, it is necessary to evaluate the planned review. A way to perform such evaluation is to ask experts to review the protocol. Another way to evaluate the planning is to test the protocol execution. The review is executed in a reduced set of selected sources. If the obtained results are not suitable, the protocol must be reviewed and a new version must be created.

4.3. Review Execution

After evaluating the planning, the systematic review execution can be initiated. During this phase, the search in the defined sources must be executed and the studies obtained must be evaluated according to the established criteria. Finally, the relevant information to the research question must be extracted from the selected studies.

3.2. **Selection Execution:** this section aims to register the primary studies selection process, reporting the obtained studies and the results of their evaluation.
- **Initial Studies Selection:** the search in itself is executed and all the obtained studies must be listed for further evaluation.
- **Studies Quality Evaluation:** the procedures for studies selection are applied to all obtained articles in order to verify if the studies fit the inclusion and exclusion criteria. Moreover, it must be checked if the studies belong to the types selected during the planning phase. The objective of this section is to register the results of this evaluation.
- **Selection Review:** studies selection must be reviewed to guarantee that the studies quality evaluation does not eliminate relevant articles. Here, independent reviewers may be useful. The results of the review must be recorded in this item.

4. **Information Extraction:** once primary studies are selected, the extraction of relevant information begins. In this protocol section, extraction criteria and results are described.

4.1. **Information Inclusion and Exclusion Criteria Definition:** criteria by which the information obtained from studies must be evaluated.

4.2. **Data Extraction Forms:** to standardize the way information will be represented, the researcher must create forms to collect data from the selected studies. These forms may vary depending on the systematic review's objective and context.

4.3. **Extraction Execution**: two kinds of results can be extracted from the selected studies: objective and subjective results.
- **Objective Results Extraction:** objective results are those that can be extracted directly from the selected studies. Such results must be organized as follows:
  i) **Study Identification:** studies identification includes the publication title, its authors and the source from which it was obtained.
  ii) **Study Methodology:** methods used to conduct the study.
  iii) **Study Results:** effect obtained through the study execution.
  iv) **Study Problems:** study limitations found by the article's authors.
- **Subjective Results Extraction:** subjective results are those that cannot be extracted directly from the selected studies. There are two ways to obtain such results:
  i) **Information through Authors:** reviewers contact the study's authors to solve doubts or to ask for to more details about it.
  ii) **General Impressions and Abstractions:** reviewers raise their own conclusions after the reading the study.

4.4. **Resolution of divergences among reviewers:** if reviewers don't agree on the information extracted from the studies, the divergences must be recorded. The reviewers must reach a consensus on this matter and register it in this section.

4.4. Execution Evaluation

Our experience on conducting systematic reviews showed that, during the execution phase, several problems may occur to due web search engines limitations. Each one of them deals with logical operators differently or presents restrictions on terms combination. It's not possible to identify those issues until we execute the search in theses engines.

Therefore, the systematic review process presented in this technical report suggests evaluating web search engines at the execution phase to verify if they are capable of executing the search strings previously defined during the planning phase. If there are approved, the process may go on. Otherwise, it may be necessary to exclude a digital source selected or to reform the search strings.

4.5. Result Analysis

After the systematic review execution, the results must be summarized and be analyzed using the statistical methods defined during the planning phase.

5. **Results Summarization:** this systematic review protocol section aims to present the data resulting from the selected studies.
   - 5.1. **Results Statistical Calculus:** statistical methods chosen in the "Experimental Design" section are applied to analyze data and to understand the complexity relations between obtained results.
   - 5.2. **Results Presentation in Tables:** the results obtained from the systematic review must be displayed in tables to facilitate analysis. Tables allow to classify studies according to different criteria and to organize them under different perspectives.
   - 5.3. **Sensitivity Analysis:** result robustness must be verified, investigating if there were uncertainties about including or excluding certain studies. Sensitivity analysis is more important when a complete meta-analysis is performed.
   - 5.4. **Plotting:** a data plotting strategy may be chosen to present the results. Likewise sensitivity analysis, plotting is indicated when meta-analysis is performed.
   - 5.5. **Final Comments:** this item presents reviewers final comments about the systematic review results.
     - **Number of Studies:** quantity of obtained and selected studies.
     - **Search, Selection and Extraction Bias:** if any search, selection or information extraction biases that can invalidate the systematic review results are identified by the reviewers, they must be described here.
     - **Publication Bias:** it refers to the problem that positive results are more likely to be published than negative results since the concept of positive or negative results sometimes depends on the viewpoint of the researcher.
     - **Inter-Reviewers Variation:** conflict resolution between reviewers regarding the systematic review results.
     - **Results Application:** defines how the obtained systematic review results can be applied.
     - **Recommendations:** reviewers' suggestions on how the systematic review results must be applied.

5. The Template Evaluation

To evaluate the developed protocol template, a pilot study was conducted. This study aimed to verify if the proposed template could be efficiently used to drive the researchers through the systematic review process and observe how researches conduct such process. It also intended to capture improvement opportunities in the template structure.

The study was conducted with four master students who had no prior experience on executing systematic reviews but have already read scientific articles about it. To characterize the subjects` level of experience in conducting systematic reviews, a *Researcher Characterization Form* (Appendix 2) was developed.

The template was used by these students to perform systematic reviews about different Software Engineering research topics. After concluding the systematic review execution, the subjects answered the *Follow Up Form* (Appendix 3) that was created to capture data about their experience in using the protocol template. This form intended to capture the effort in each phase of the systematic review process and the tools and methodologies adopted to perform the process` activities. In addition, the difficulties found during the review conduction and its benefits should also be reported. It's important to highlight that the collected data didn't considers the result analysis phase. A summary of the information collected from the four systematic reviews analyzed is presented below:

- *Software Architectures* [Barcelos and Travassos 2005]: the objective of this review was to identify evaluation approaches for software architectures. The search was executed in web search engines and libraries. First, the abstract of the obtained studies was read to filter the ones that were no relevant. Then, the "Introduction" section of the remaining articles was read. 80 studies were obtained and 54 were selected. Time spent for planning, evaluation, creating new protocol versions and execution was 12, 2, 4 and 23 hours.
- *Process Models Evaluation* [Cabral and Travassos 2004]: this systematic review was conduct to identify to existing initiatives for review and verification of such models. The search was executed in digital libraries. First, the abstract of the obtained studies was read to filter the ones that were no relevant. Then, the full text of the remaining articles was read. 125 studies were obtained and just 1 was selected. Time spent for planning, evaluation, creating new protocol versions and execution was 8, 1, 12 and 100 hours.
- *Software Test Planning* [Dias Neto and Travassos, 2004]: The objective of this systematic review was to find initiatives of test planning. The search was executed in web search engines and libraries. First, the abstract of the obtained studies was read to filter the ones that were no relevant. Then, the full text of the remaining articles was read. 56 studies were obtained and 8 were selected. Time spent for planning, evaluation, creating new protocol versions and execution was 12, 2, 4 and 23 hours.
- *Reading Techniques for Requirement Documents* [Mafra and Travassos, 2005]: the objective of this review was to identify, analyze and evaluate experimental studies regarding reading techniques for requirement documents. The search was executed in web search engines. To select the studies, the abstract of the obtained article was read. 278 studies were obtained and 38 were selected. Time spent for planning, evaluation, creating new protocol versions and execution was 20, 1, 1 and 44 hours.

We observed that all subjects were able to conduct their systematic reviews using the protocol template as guiding material. They reported that the template was helpful in

guiding the review execution since it provides formalization for the "ad hoc" literature review process.

Figure 3 presents the average effort time the subjects spent in executing some systematic review process´ activities. As shown in Figure 3, systematic reviews require addition application effort when considering informal reviews execution. This extra effort is due to the necessity time to plan the review, to evaluate the protocol and to review it, creating new versions according to modifications suggested in the evaluation phase.

Despite the addiction time spent planning and documenting the results of the systematic review process, the pilot study results show that most part of the efforts are concentrated at the execution phase of the process. As shown in Figure 3, searching and evaluating studies still represent a bottleneck in the literature review process.

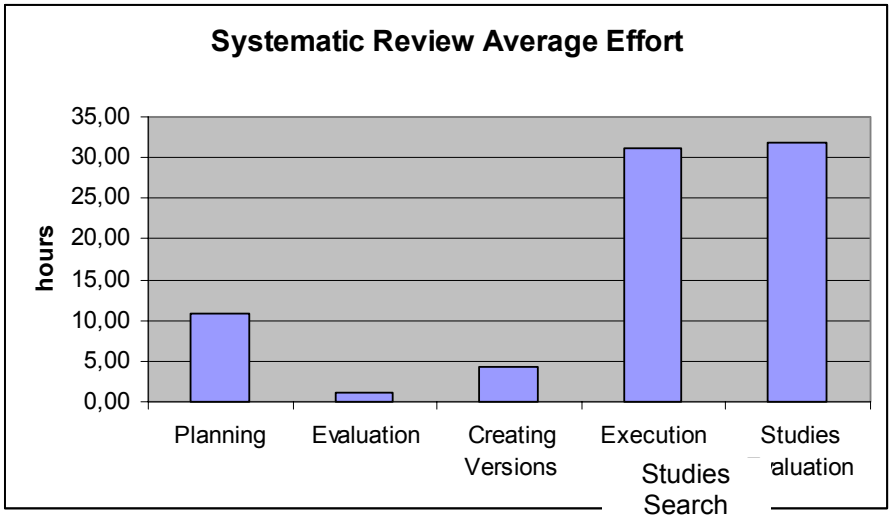

Figure 3. Average effort time for conducting Systematic Reviews.

Another result of this pilot study was regarding the tools and methodologies used by the subjects to perform some activity of the systematic review process. These results may serve as suggestions to researchers in future review executions.

In the Review Planning stage, all subjects used word processors as supporting tool to this phase, as shown in Figure 4. In addition, half of the used web search engines (of the sources selected in section 2.4) to help building the search strings (section 2.3).

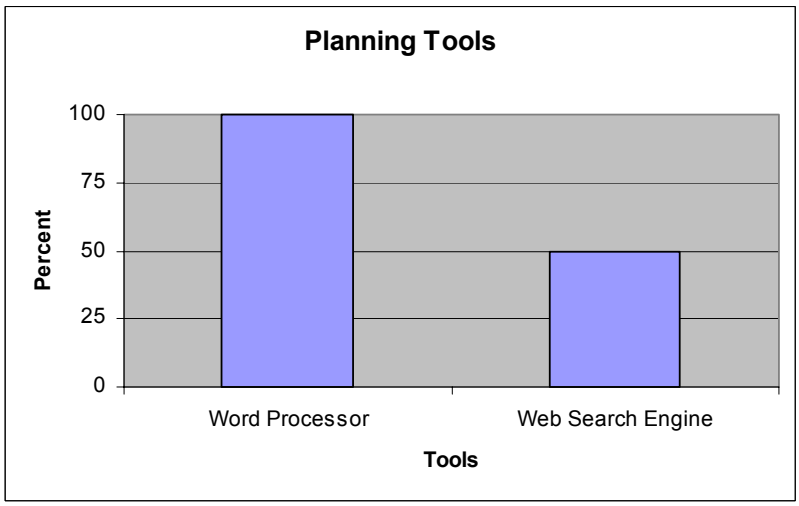

Figure 4. Tools used during the Review Planning Phase.

To evaluate the protocol defined in the Planning phase, the methodologies applied by the subjects differ (Figure 5). Most of them submitted the protocol to experts' evaluation. More experience researchers were consulted and reviewed the protocol. One of them, however, reviewed the protocol by himself. Finally, one subject performed a "pilot execution" to evaluate to protocol. The search was executed in one of the sources selected and, based on the results found, the planning was considered satisfactory or not.

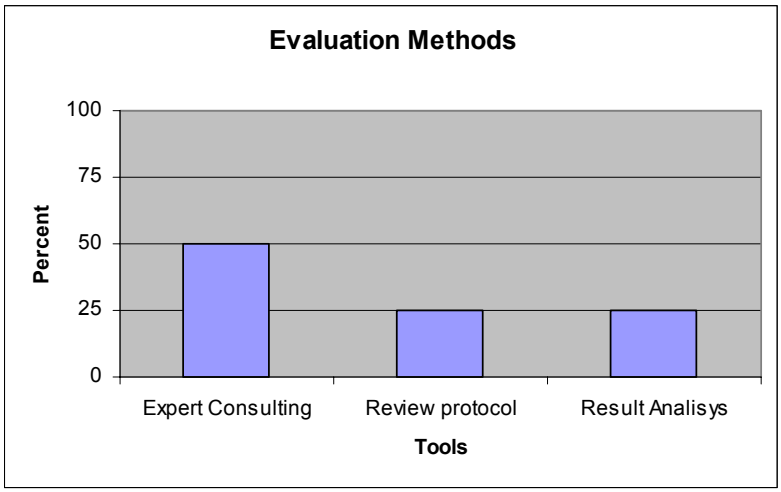

Figure 5. Methods for Planning Evaluation.

All subjects chose digital libraries as studies source. Therefore all of them used web search engines as execution tool, as shown in Figure 6. Only one of them executed the search in libraries. But it is interesting to highlight that one of them used an organization reference tool to organize the obtained and selected studies. This tool facilitated the retrieving of the selected articles when the execution was concluded.

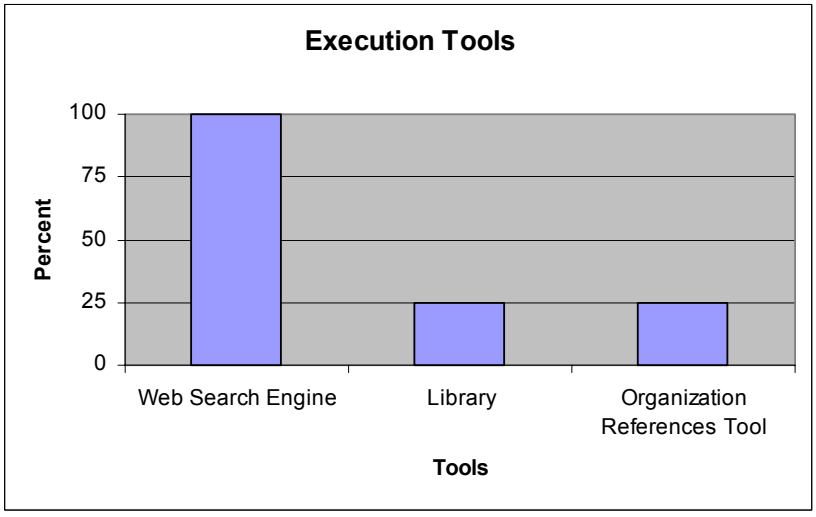

Figure 6. Review Execution Tools.

Finally, Figure 7 shows the main procedures to evaluate if a study was relevant to the research topic. All of them read the abstract of the article. However, if the abstract wasn't clear enough to detect the study relevance, most of then adopted a second procedure to evaluate the articles (reading the study introduction section or reading the full article).

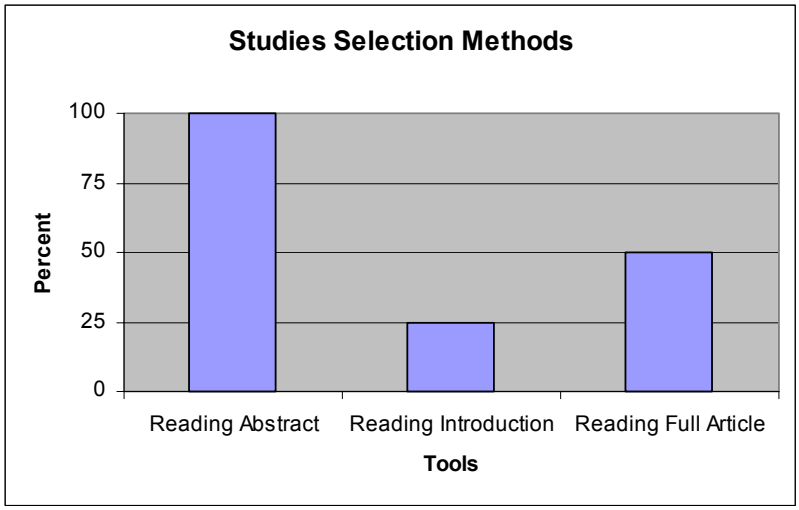

Figure 7. Methods for Studies Selection.

After using the material, the subjects suggested a few modifications of the template. Among the improvements suggested are: to include examples and the sections "keyword" and "search strings' in the template and to define metrics to measure review effort. In the template version presented here, some of the suggestions made (including new sections) were performed.

5.1. Evaluation Biases

When planning the template evaluation, we were able to identify some biases of this pilot study, which are listed below. We believe, however, that theses biases don't compromise the obtained results, since we were interested in observing how researches conduct systematic reviews using the template, not in comparing the proposed systematic review process with other approaches:

- Subjects took an Experimental Software Engineering course before the reviews were planned. In this course, the systematic review subject was discussed.
- Subjects had prior access to other material regarding systematic reviews in Software Engineering [Kitchenham, 2004]. We were not capable of calculating the impact of this fact on the subjects' learning curve.
- Subjects could share information about how to fill in the template among each other.
- Subjects had access to the researchers who created the template and could ask to them questions about the systematic review process.
- Subjects used different methodologies to evaluate studies quality. Such methodologies have influence over the systematic review execution time. For instance, if one chose "reading the full article" to decide if it is relevant to the research topic, the time spent will be higher than if "reading the abstract" had been chosen. Therefore, we cannot compare their execution time.

6. An Example of a Systematic Review in Software Engineering

Next, an example of a systematic review conduction using the developed template is presented. The following systematic review was executed in the context of a master degree work regarding methodologies to software process models evaluation. This work's premise is that inspections can be used to support process models verification [Cabral and Travassos, 2004]. Thus, a systematic review was conduct to identify to

existing initiatives for review and verification of such models. The protocol for this systematic review is presented below.

1. Question Formularization

1.1. Question Focus: To identify initiatives and experience reports in Software Engineering related to review and/or verification execution regarding software processes description.

1.2. Question Quality and Amplitude

- **Problem:** Software processes description guides the processes execution. It is necessary to verify, prior to the process execution, if its description is suitable. The objective of such verification is to prevent that distortions in the process activities affect the quality of the final product.
- **Question:** $\mu 0$: Which initiatives have been carried out to evaluate processes description in the context of the Software Engineering?
- **Keywords and Synonyms:**
  Evaluation: review, verification, analysis;
  Process description: process definition, process model, process programming;
  Software engineering.
- **Intervention:** Evaluation of software processes description.
- **Control:** None.
- **Effect:** Identification of initiatives related to review and verification.
- **Outcome Measure:** Number of identified initiatives.
- **Population:** Publications regarding software processes review and definition.
- **Application:** Software processes managers.
- **Experimental Design:** none statistical method is going to be applied.

2. Sources Selection

2.1. Sources Selection Criteria Definition: Availability to consult articles in the web; presence of search mechanisms using keyword and publishing companies suggested by experts.

2.2. Studies Languages: English.

2.3. Sources Identification

- **Sources Search Methods:** Research through web search engines.
- **Search String:** (Evaluation OR review OR verification OR analysis) AND ("process description" OR "process definition" OR "process model" OR "process programming") AND "software engineering"
- **Sources List:**
  ACM SIGSOFT Software Engineering Notes
  Empirical Software Engineering
  Journal of Systems and Software
  Software Engineering Journal
  The complete sources list can be found in [Cabral and Travassos, 2004].

2.4. Sources Selection after Evaluation: *A priori*¸ all the listed sources had satisfied the quality criteria.

2.5. References Checking: All sources were approved.

3. Studies Selection

3.1. Studies Definition

- **Studies Inclusion and Exclusion Criteria Definition:** The studies must present initiatives to evaluate the description (model) of software processes prior to its execution. This research will not select studies describing evaluations carried out during the process.

- **Studies Types Definition:** All kinds of studies related to the research topic will be selected.
- **Procedures for Studies Selection:** The search strings must be run at the selected sources. To select an initial set of studies, the abstract of all obtained studies from web search engines is read and evaluated according to inclusion and exclusion criteria. To refine this initial set of studies, their full text is read.

3.2. Selection Execution

- **Initial Studies Selection:** The complete studies list can be found in [Cabral and Travassos, 2004].
- **Studies Quality Evaluation:** Just one of the obtained studies completely fit all inclusion and exclusion criteria defined previously.
- **Selection Review:** The study selection was approved.

4. Information Extraction

4.1. Information Inclusion and Exclusion Criteria Definition: The extracted information from studies must contain techniques, methods, strategies or any kind of initiative to evaluate the software processes description.

4.2. Data Extraction Forms: The information forms defined for this systematic review can be found in [Cabral and Travassos, 2004].

4.3. Extraction Execution

- **Objective Results Extraction**
  - **i) Study Identification:** Hao, J.-K.; Trousset, F.; Chabrier, J.J.; Prototyping an inconsistency checking tool for software process models, Proceedings of 4th International Conference on Software Engineering and Knowledge Engineering, 15-20 June 1992, Pages: 227 – 234.
  - **ii) Study Methodology:** The authors built a prototype tool (MASP Definition Language) to carry out consistency verification of software process models using a formal approach for processes description.
  - **iii) Study Results:** The prototype tool identified inconsistencies on the process models evaluated.
  - **iv) Study Problems:** Although the results of inconsistencies identification were positive, language limitations were briefly discussed. The study population also was restricted to process models used in projects the authors worked at.
- **Subjective Results Extraction**
  - **i) Information through Authors:** It was not requested.
  - **ii) General Impressions and Abstractions:** The authors see model processes as process programming, according to Lehman's point of view, cited in the references.

4.4. Resolution of divergences among reviewers: There were no divergences.

5. Results Summarization

5.1. Results Statistical Calculus: Statistical calculi were not used.

5.2. Results Presentation in Tables

Quantity of Studies by Initiative

Type of Initiative	*Tool*	*Technique*	*Method*	*Strategy*	*Others*	***Total***
# of Studies	*1*	*0*	*0*	*0*	*0*	***1***

5.3. Sensitivity Analysis: It wasn't applied.

5.4. Plotting: It wasn't applied.

5.5. Final Comments

- **Number of Studies:** Studies found: 125; Studies selected: 1.
- **Search, Selection and Extraction Bias:** None was defined.
- **Publication Bias:** None was defined.
- **Inter-Reviewers Variation:** There was no variation.
- **Results Application:** The research results suggest that new initiatives to review and verify software processes models before their execution must be developed.
- **Recommendations:** None.

7. Final Comments

Systematic review is a scientific methodology that can be used to integrate empirical research on SE. Though its importance, conducting systematic reviews is not a simple task. It evolves performing complex actives and understanding specific concepts and terms that may be unknown to researches.

The difficulties found during the systematic reviews execution pointed the need to investing research efforts in developing systematic reviews planning and execution methodologies. Therefore, we defined a systematic review conduction process. This process aims at guiding researches when performing systematic reviews.

To facilitate the execution of this process, we developed a review protocol template. This template was based on systematic review protocols developed in the medical area, on the guidelines proposed by [Kitchenham, 2004] and the protocol example found in [Mendes and Kitchenham, 2004]. Its objective is to serve as a guideline to SE researchers when conducting the systematic reviews. The template leads researchers through each step of the systematic review process, defining clearly the content of each protocol section.

Another issue is that systematic reviews require an additional conduction effort. The review must be planned before execution and the whole process must be documented, including, intermediary results. However, our experience on conducting systematic reviews shows that most part of the efforts is concentrated at the execution phase. Searching and evaluating studies still represent a bottleneck in the literature review process.

Acknowledgements

We would like to thank Prof. Barbara Kitchenham (Keele University/UK) and Prof. Emilia Mendes (University of Auckland/NZ) for sending to us all the initial materials regarding systematic reviews. We also recognize the effort of the ESE team members (Reinaldo Cabral, Somulo Mafra, Rafael Barcelos and Arilo Claudio Dias Neto) in accomplishing the systematic reviews from where the data and examples have been acquired.

This work has been develop under the scope of CNPq Universal eSEE Project. Dr. Travassos has been granted by the CNPq – The Brazilian Research Council.

References

Barcelos, R.F. and Travassos, G.H. *Arquitetura de Software: Identificando as metodologias que avaliam a sua qualidade.* Final Course Monograph COS 722, PESC - COPPE/UFRJ, 2005.

Birge, R.T. The calculation of errors by the method of least squares. *Physical Review*, 40:207-227, 1932.

Cabral, R. and Travassos, G. H. *Inspeção em Modelos de Processo de Software*. Final Course Monograph COS 722, PESC - COPPE/UFRJ, 2004.

Chalmers, I., Enkin, M. and Keirse, M.J.N.C. *Effective care in pregnancy and childbirth*. Oxford: Oxford University Press, 1989.

Cina, C.S., et al. *Carotid endarterectomy for symptomatic carotid stenosis*. Cochrane Database of Systematic Reviews 2: CD001081, 2000.

Cochran, W.G. Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*. 4:102-118,1937.

Cooper, H. and Hedges, L.V. eds. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 1994.

Dias Neto, A.R. and Travassos, G. H. *Planejamento de Testes de Software*. Final Course Monograph COS 722, PESC - COPPE/UFRJ, 2004.

Early Breast Cancer Trialists' Collaborative Group. *Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomized trials involving 31,000 reccurrences and 24,000 deaths among 75,000 women*. Lancet 339: 1-15 (Part I) and 71-85 (Part II), 1992.

Early Breast Cancer Trialists' Collaborative Group. *Indications for fibrinolytic therapy in suspected acute myocardial infarction: Collaborative overview of early mortality and major morbidity results from all randomized trials of more than 1000 patients*. Lancet 343: 311-322, 1994.

Feinstein, A.R. Meta-Analysis: Ststistical Alchemy for the 21st Century. *Journal of Clinical Epidemiology*, 48(1): 71-79, 1995.

Feldman, K.A. Using the work of others: Some observations on reviewing and integrating. *Sociology of Education*, 4:86-102, 1971.

Fisher, R.A. *Statistical methods for research workers*. 4th ed.,London: Oliver & Boyd, 1932.

Glass, G.V. Primary, secondary, and meta-analysis. *Educational Researcher*, 5:3-8,1976.

Glass, G.V. and Smith, M.L. Meta-analysis of research on the relationship of class size and achievement. *Educational Evaluation and Policy Analysis*,1:2-16, 1978.

Glass, G.V., McGaw, B. & Smith, M.L. *Meta-analysis in social research*. Beverly Hills: Sage, 1981.

Gross, R. *Decisions and Evidence im Medical Practice*. St. Louis: Mosby, Inc., 2001

Hedges, L.V. and Olkin, I. *Statistical methods for meta-analysis*. Orlando: Academic Press, 1985.

Hunter, J.E., Schmidt, F.L. and Jackson, G.B. *Meta-analysis: Cumulating research findings across studies*. Beverly Hills: Sage, 1982.

ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. *Randomized trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2*. Lancet ii: 349-360, 1988.

ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. Protocol for a large, simple study of the effects of oral mononitrate, of oral captopril, and of intravenous magnesium. *American Journal of Cardiology* 68: 87D-100D, 1991.

Juristo, N. and Moreno, A., *Basics of Software Engineering Experimentation*, Kluwer Academic Press, 1st edition, 2001.

Kitchenham, B.A., Pfleeger, S.L., Pickard, L.M., Jones, P.W., Hoaglin, D.C., El Emam, K. and Rosenberg, J., Preliminary Guidelines for Empirical Research in Software Engineering, *IEEE Transactions on Software Engineering*, vol. 28, n° 8, 2002.

Kitchenham, B. A., Dyba, T. and Jorgensen, M., Evidence-based Software Engineering, *26th International Conference on Software Engineering (ICSE 2004)*, Scotland, 2004.

Kitchenham, B., *Procedures for Performing Systematic Reviews*. Joint Technical Report Software Engineering Group, Department of Computer Science Keele University, United King and Empirical Software Engineering, National ICT Australia Ltd, Australia, 2004.

Kwok, Y., et al. Meta-analysis of exercise testing to detect coronary artery disease in women. *American Journal of Cardiology,* 83(5):660-666, 1999

Liberati, A. Meta-Analysis: Ststistical Alchemy for the 21st Century: Discussion. A Plea for a More Balanced View of Meta-Analysis and systematic Overviews of the Effect of Health Care Interventions. *Journal of Clinical Epidemiology,* 48(1): 81-86, 1995.

Light, R.J. and Smith, P.V. Accumulating evidence: Procedures for resolving contradictions among research studies. *Harvard Educational Review*, 41:429-471, 1971.

Light, R.J. and Pillemer, D.B. *Summing up: The science of reviewing research*. Cambridge: Harvard University Press, 1984.

Mafra, S.N. and Travassos, G. H. *Revisão Sistemática sobre Técnicas de Leitura de Documentos de Requisitos*. Final Course Monograph COS 722, PESC - COPPE/UFRJ, 2005.

Malik, A.H. and Lee, W.M. Chronic hepatitis B virus infection: treatment strategies for the next millennium. *Annals of Internal Medicine,* 132:723-731, 2000.

Mendes, E. and Kitchenham, B. *Protocol for Systematic Review*. Available at http://www.cs.auckland.ac.nz/emilia/srspp.pdf. Last accessed by 05/10/2005, 2004.

Pearson, K. Report on certain enteric fever inoculation statistics. *British Medical Journal*, 3:1243-1246, 1904.

Piantadosi, S. Clinical Trials – A Methodologic Perspective. New York: John Wiley & Sons, Inc., 1997

Pressman, R. S. *Software Engineering: A Practtioner´s approach*, 5th edition. Mc Graw Hill, 2002.

Rosenthal, R. and Rubin, D.B. Interpersonal expentancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 3:377-386, 1978.

Rosenthal, R. *Meta-analytic procedures for social research*. Beverly Hills: Sage, 1984.

Sackett, D.L., Straus, S.E., Richardson, W.S., Rosenberg, W., Haynes and R.B. *Evidence-Based Medicine*. Edinburgh: Churchill Livingstone, 2000.

Schmidt, F.L. and Hunter, J.E. Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 62:529-540, 1977.

Smith, M.L. and Glass, G.V. Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32:752-760, 1977.

Taveggia, T.C. Resolving research controversy through empirical cumulation: Toward reliable sociological knowledge. *Sociological Methods and Research*, 2: 395-407.

Tippett, L.H.C. *The methods of statistics*. London: Williams & Norgate, 1931.

Towheed, T.E. and Hochberg, M.C. A systematic review of randomized controlled trials of pharmacological therapy on osteoarthritis of the knee, with an emphasis on trial methodology. *Semin Arthritis Rheum,* 26(5):755-770, 1997.

Travassos, G.H. and Barros, M.O. Contributions of in virtuo and in silico experiments for the future of empirical studies in software engineering. *Proc of 2nd Workshop on Empirical Software Engineering the Future of Empirical Studies in Software Engineering*, Rome, 2003.

Warren, K.S. and Mosteller, F. eds. Doing More Good Than Harm – The Evaluation of Health Care Interventions. New York: *Annals of the New York Academy of Sciences*, v.703, 1993.

Williams, J.W., et al. A systematic review of newer pharmacotherapies for depression in adults: evidence report summary. *Annals of Internal Medicine*, 132:743-756, 2000.

Yates, F. and Cochran, W.G. The analysis of groups of experiments. *Journal of Agricultural Science*. 28:556-580, 1938.

Appendix 1 – Systematic Review Protocol Template

1. Question Formularization
   1.1. Question Focus
   1.2. Question Quality and Amplitude
      - Problem
      - Question
      - Keywords and Synonyms
      - Intervention
      - Control
      - Effect
      - Outcome Measure
      - Population
      - Application
      - Experimental Design
2. Sources Selection
   2.1. Sources Selection Criteria Definition
   2.2. Studies Languages
   2.3. Sources Identification
      - Sources Search Methods
      - Search String
      - Sources List
   2.4. Sources Selection after Evaluation
   2.5. References Checking
3. Studies Selection
   3.1. Studies Definition
      - Studies Inclusion and Exclusion Criteria Definition
      - Studies Types Definition
      - Procedures for Studies Selection
   3.2. Selection Execution
      - Initial Studies Selection
      - Studies Quality Evaluation
      - Selection Review
4. Information Extraction
   4.1. Information Inclusion and Exclusion Criteria Definition
   4.2. Data Extraction Forms
   4.3. Extraction Execution
      - Objective Results Extraction
         i) Study Identification
         ii) Study Methodology
         iii) Study Results
         iv) Study Problems
      - Subjective Results Extraction
         i) Information through authors
         ii) General Impressions and Abstractions
   4.4. Resolution of divergences among reviewers
5. Results Summarization
   5.1. Results Statistical Calculus
   5.2. Results Presentation in Tables
   5.3. Sensitivity Analysis
   5.4. Plotting
   5.5. Final Comments
      - Number of Studies
      - Search, Selection and Extraction Bias
      - Publication Bias
      - Inter-Reviewers Variation
      - Results Application
      - Recommendations

Appendix 2 – Researcher Characterization Form

Name:_____ Degree (MS.c/D.Sc.): ____

I –Language Competence

Please, inform your competence to use work material in English.
I consider English as a language in which:

My ability to read and comprehend texts in English is **(choose the most suitable option)**:
( ) none
( ) low
( ) medium
( ) high
( ) I'm a native speaker

My capability to work with and/or follow instructions written in English is **(choose the most suitable option):**
( ) none
( ) low
( ) medium
( ) high
( ) I'm a native speaker

II – Bibliographic Research

What is your former experience in accomplishing bibliographic research? **(choose the most suitable options):**
( ) I've never accomplished bibliographic research before **(if you choose this item, go to section III).**
( ) I've accomplished bibliographic research as part of a thesis, dissertation or monograph.
( ) I've been accomplishing bibliographic research as a researcher in the academy.

Please, explain your answer by describing your experience in accomplishing bibliographic research. Include the number of semesters or years of relevant experience (For instance, "I've accomplished bibliographic research during one semester as part of my master thesis")

_____
_____
_____

What were the difficulties in accomplishing your last bibliographic research? **(choose the most suitable options)**
( ) selection of reliable reference sources.
( ) great amount found references.
( ) selection of relevant reference sources.
( ) establishing the quality of the found references.
( ) accessing relevant references sources.

If you had additional difficulties in accomplishing your last bibliographic research, please list them bellow.

_____
_____
_____

III – Systematic Reviews

Please, answer the next question using the following scale:
1 = Never heard about it
2 = I've heard about it, but I've never read about it
3 = I've read a scientific article about it
4 = I've made an exercise in classroom
5 = I've used it once as a research mechanism
6 = I've used it several times

Knowledge about Systematic Review

- Degree of Knowledge about Systematic Review 1 2 3 4 5 6

Please, answer the next questions using the following scale:
1 = none
2 = one
3 = between 2 and 5
4 = more than 5

Experience in Building Systematic Reviews

- Number of systematic reviews that you have built 1 2 3 4
- Experts helped you to build how many systematic reviews 1 2 3 4

Experience in Executing Systematic Reviews

- Number of systematic reviews that you have executed 1 2 3 4
- Number of systematic reviews built by others that you executed 1 2 3 4

Appendix 3 – Follow up Questionnaire

Name: _____ Degree (MS.c/D.Sc.): ____

Material used to guide the construction of a systematic review:
( ) Technical Report (Kitchenham, 2004) – Appendix 1
( ) ESE Template– Appendix 2
( ) None of these **(if you choose this item, please answer only questions 1 to 6)**

1. How long did you take to plan this systematic review (in **hours**)? What tools did you used to plan it? Consider "Planning" the phase in which the first systematic review version was made.

_____
_____
_____
_____
_____
_____
_____

2. How many versions of this review were created before the final execution? What approach did you used to evaluate the need for creating new versions? How long did you take to evaluate the versions (in **hours**)? And how long did you take to create the new versions (in **hours**)?

_____
_____
_____
_____
_____
_____
_____

3. If you already executed this systematic review, how long did you take to do it (in **hours**)? Which were the kind of sources did you used to find studies (search machines, libraries, etc)? Consider "Execution" executing the search in the selected sources and obtaining the articles.

_____
_____
_____
_____
_____
_____
_____
_____

4. How long did you take to refine the search results (in **hours**)? What mechanism did you used to do it? Consider "Refining" the selection of relevant studies among those obtained.

_____
_____
_____
_____
_____
_____
_____
_____

5. Did you find difficulties in building the systematic review? Which ones?

6. Did the systematic review make the search process easier? In which way?

7. What are the strengths of the material used to guide the construction of a systematic review?

8. What are the weaknesses of the material used to guide the construction of a systematic review?

9. In your opinion, should something be added to the material?

10. Which parts do you see as unnecessary or duplicated?

11. Are there terms unclear or open to misinterpretation? Which ones?

12. Do you think that the usage of this material could be improved somehow? In which way?