Analyzing the Impact of the Search Phase in a Systematic Mapping Study

João Carbonell[1], Luciano Marchezan[1], Aníbal Neto[1], Elder Rodrigues[1], Maicon Bernardino[1], Yury Alencar Lima[1]

[1]Federal University of Pampa (UNIPAMPA) – Alegrete, RS – Brazil.

{joaocarbonellpc, lucianomarchp, netoiiung, yuryalencar19}@gmail.com
elderrodrigues@unipampa.edu.br,bernardino@acm.org

Abstract. *Collect data from a research field is a crucial task in academic works to understand a scientific area. Systematic Mapping Studies (SMS) provide a formal and well designed protocol to select, classify and present relevant data. To select a digital library database or database set and to define the search strings used in database engines on digital libraries are search phase activities aimed to delimit the retrieved works quantity. This work analyzes the impact of definition of the digital libraries that may be used in the search phase on systematic mapping studies conduction, as well as the definition of the search strings. The results related with exclusive studies from each base were compared and presented. In addition, the difference in the results of the retrieved studies according to changes in the search string are presented. Based on these results, five activities that may give support to inexperienced researches when conducting SMS.*

Resumo. *Coletar dados de um campo de pesquisa é uma tarefa crucial em trabalhos acadêmicos para compreender uma área científica. Mapeamentos Sistemáticos da Literatura fornecem uma maneira formal e bem definida de se selecionar, classificar e apresentar dados relevantes. Selecionar bases de dados e definir as strings de buscas para estas são atividades importantes da fase de busca em um mapeamento sistemático. Estas atividades são utilizadas para se delimitar a quantidade de trabalhos pesquisados. Neste trabalho, o impacto da escolha das bases das bibliotecas digitais nas fases de busca de um Mapeamento Sistemático é analisado, bem como a definição das strings. Foram comparadas e demonstrados os resultados relacionados aos estudos exclusivos de cada base, além das diferenças nos resultados com a troca de termos e caracteres especiais em strings de buscas. Também foram mapeadas cinco atividades que podem ajudar pesquisadores com pouca experiência conduzindo mapeamentos sistemáticos.*

1. Introduction

Systematic Mapping Studies (SMS) are important in the scientific research field to collect overall information about a topic of interest. In addition, SMS may be used to identify contributions for an area and give support to other researchers to find primary studies of the specific area.

To perform a SMS, there is a well-defined protocol proposed by [Petersen et al. 2008] that may be followed, increasing the chance of obtaining

relevant results. In the search phase of this protocol, databases must be selected, and search strings have to be formulated. These decisions have impact in the search phase, because they are directly related to the possible set of retrieved studies. However some researchers often uses scientific databases which are known to index minor databases works.

The protocol executed when performing a SMS is extremely important to achieve satisfactory results. Thus, several decisions made during the preparation of the protocol may impact in the final set of selected studies. The objective of this work is to investigate, analyze and provide evidence about the impact of some decisions related with the search phase of the SMS protocol. The impact of these decisions may be summarized as two Research Questions (RQ):

RQ1. What is the impact of the database selection?
RQ2. What is the impact of a poor formulated search string?

To answer these RQs we observed and documented the information directly related with these decisions impact: (i) how researches made these decisions; (ii) which information and reasons were used to make the decisions, and; (iii) what was the impact of these decisions in the final result of the SMS. The result of this analysis indicate the problems that may appear if some aspects of the SMS are not carefully planed.

The rest of this work is structured as follows. Section 2 gives a description of systematic studies and their protocols. Section 3 details our motivation and the research methodology used to achieve our results. Section 4 presents and discuss the results of our research. Finally, Section 5 concludes this work.

2. Background

2.1. Systematic Studies

Systematic studies are a research methodology commonly used in scientific field [Connolly et al. 2012, Stuck et al. 1999, Šmite et al. 2010]. These studies follow guidelines and well-defined protocols [Kitchenham 2004], executing process activities to improve the analysis of studies and obtained results, thus reducing the bias and mitigating the loss of data and threats to the research. They aggregate the experience, date and information from different studies in order to answer a specific research question defined by the researchers [Budgen and Brereton 2006].

2.2. Systematic Literature Review and Systematic Mapping Study

The Systematic Literature Reviews (SLR) are inspired on medical research field activities. Their main goal is to perform a deep analysis of study methodologies and results to collect and evaluate evidence pertaining to a focused topic [Biolchini et al. 2005]. According to [Budgen and Brereton 2006], a systematic review is a means of identifying, evaluating and interpreting research relevant data from a topic area, or phenomenon of interest.

Similar to a SLR, a Systematic Mapping Study (SMS) uses guidelines and protocols to find relevant studies. In the SMS, the results are categorized, summarized and presented in a map [Petersen et al. 2008] which shows that data in form of graphs and tables.

2.3. Protocols of Systematic Studies

The protocols of systematic studies are processes in which every step aims to ensure the relevance of the data collected and to avoid rework. An example of mapping protocol is presented by [Petersen et al. 2008] (see Figure 1). Some protocol activities, such as definition of exclusion and inclusion criteria, improve the depth and relevance of primary studies. Others activities, such as definition of scope, research questions and search strings, helps to delimit the boundaries of the review.

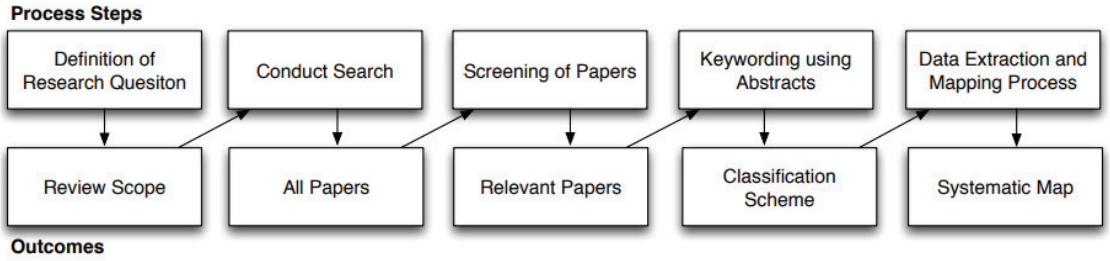

Figure 1. Systematic Mapping Protocol [Petersen et al. 2008]

The definition of the search strings must be in accordance with the search engine limitations of each database. In the structure of a search string, terms and logical operators are inserted to be read according to the syntax of each search engine. According to [Petersen et al. 2008] a good way to construct the search string is to structure them with PICO (population, intervention, comparison and outcome) [Kitchenham 2004].

3. Motivation and Methodology

In the Software Engineering context, both SLR and SMS have great importance in a proposal research methodology. Although there are several protocols for both, more commonly used for SLR is presented in [Kitchenham 2004] and for SMS was propose in [Petersen et al. 2008]. These protocols detailed the whole processes to perform these systematic reviews/mappings, however, there is still some decisions that have to be made by who is executing those protocols. These decisions are related with the search phase, which is similar for both SLRs and SMSs. We could summarize these decisions with two questions:

1. Which databases should be used for the search of studies?
2. Which search string should be used in each database?

Answer these questions is not a trivial task. To collect evidence that may help us obtain those answers we observed and registered the reasons and results related with those questions during the search phase of a real SMS conducted between July and September 2017. The objective of this SMS was to present and overview of DSL supporting tools. During the search phase of this SMS we registered: (i) the number of studies return from each database, unique studies of each database and the number of studies from each database included in the final selection; (ii) generic string, strings for each database and impact of minor changes on them; The results of the collected data are presented in the following sections.

4. Results and Discussion

By analyzing the data collected during the execution of the SMS we could answer the following RQ.

4.1. RQ1. What is the impact of the database selection?

To decide which databases should be used to perform the search some characteristics related with those databases have to be considered:

1. The capacity of the search engine related with the number of terms, synonyms and operators (AND, OR, NOR) that can be included in the search string;
2. If the databases indexes exclusive studies;

Although some researchers may argue that some databases such as SCOPUS and Compendex index works from other databases, that are still works which are exclusive to a specific database. During the conduction of the SMS used to answer the RQ of this work, the databases used were IEEEXplore, ACM Digital Library, Compendex (Engineering Village), Science Direct, SCOPUS and SpringerLink. We documented the total number of studies from each database, the number of studies exclusive to a database and also how many of the exclusive studies were part of the final selection.

Table 1. Number of relevant studies retrieved from each database

Database	Retrieved Studies		
	Total	**Exclusive**	**Final Selection**
ACM Digital Library	401	89	3
Compendex	424	29	1
IEEE Xplore	355	33	3
Science Direct	101	23	1
SCOPUS	960	352	4
SpringerLink	154	107	0

Table 1 presents these numbers, where SCOPUS was the database with higher number of retrieved studies, 960. Science Direct and SpringerLink had the lower number of retrieved works, 101 and 154 respectively. Considering the exclusive studies, SCOPUS was also the database with the higher number, 352. Compendex and Science direct on other hand had the lower number, 29 and 23. Even with the higher number of exclusive retrieved studies, SCOPUS, should not be used and the only database during a systematic study. The reason is the exclusion of a possible high number of exclusive studies from other database. In this case, the total of 281 would not be considered because they were not indexed in SCOPUS. Considering the works for the final selection, ACM and SCOPUS were the databases with more exclusive studies selected, 3 and 4 respectively. These results indicated that even if only one database, such as IEEE, would not be included in the SMS, some studies would not be mapped in the SMS.

In addition, some of the retrieved studies contained insufficient data related to our topic of interest, which was tools to develop DSLs. In this case, we have to further search in the grey literature about additional information of the tools extracted from the studies.

4.2. RQ2. What is the impact of a poor formulated search string?

The first step was to define a generic search string that would be used to find studies about domain specific languages development tools. Figure 2 presents the generic search string.

The string have terms and synonyms used on related researches of the Domain Specific Languages area. The term "DSML" is related to Domain Specific Modeling Languages, which are languages that have graphical notations. The terms "Little Language" and "Small Language" are both less used synonyms, but still would be used because in SMSs we need to cover the research area as widely as possible.

```
(DSL OR DSML OR Domain Specific Language OR Domain-Specific
Language OR Domain-Specific-Language OR Domain-Specific
Modeling Language OR Domain Specific Modeling Language OR
Domain-Specific-Modeling-Language OR Small Language OR
Small-Language OR Little Language OR Little-Language AND
Tool OR Language Workbench)
```

Figure 2. Search String

However, six databases were used during this research, then it was necessary to derive each search string for each specific database as shown in Table 2. The IEEEXplore database have a specific differences about the search engine, this database limits the search to 15 terms in each string. Therefore the composite search strategy was to use two search strings, then we split the generic search string into two compound terms, which are: DSL and DSML. Keeping the other common terms in each compound search.

For others databases the string were derived accord to each search engine. It is important to know that for some databases the use of plural it makes a difference about the retrieved works.

These derived strings, however, are not free from problems, as some databases have specific rules that may have great impact in the results. For instance, if we include quotation marks in the terms *language workbench* in the SpringerLink string, the results would change from 154 to 1,257 returned studies. This small change in the string, could cost a lot of meaningless effort during the SMS because of these additional 1,103 studies, none would be included in the final selection.

A similar problem occurred with ACM Digital Library, where by removing some of the DSL synonyms, *Little Language* and *Small Language*, in this case, we obtained a total of 273 studies, 128 less than the original 401.

4.3. Preliminary Activities Before Conducting a SMS

As final result of our analysis during the conduction of a SMS, we documented five activities to be performed before conducting the SMS. This preliminary activities may help inexperienced researches when preparing to conduct both SMS or SLR. As illustrated in Figure 3, the first activity is to select the databases that will be used during the SMS or SLR. This activity is really important because as showed in previous sections, most databases index exclusive works from conferences and journals. Thus, the not addition of a database will exclude any of these exclusive works.

Table 2. Search Strings Derived for Each Database

Database	Search String
ACM Digital Library	((acmdlTitle:("DSL" "DSML" "Domain Specific Language" "Domain-Specific Language" "Domain-Specific-Language" "Domain Specific-Language" "Domain-Specific Modeling Language" "Domain Specific Modeling Language" "Little-Language" "Little Language" "Small-Language" "Small Language"))) OR (recordAbstract:("DSL" "DSML" "Domain Specific Language" "Domain-Specific Language" "Domain-Specific-Language" "Domain Specific-Language" "Domain-Specific Modeling Language" "Domain Specific Modeling Language" "Little-Language" "Little Language" "Small-Language" "Small Language")) AND ((acmdlTitle:("Tool" "Tools" "Language Workbench")) OR (recordAbstract:("Tool" "Language Workbench")))
Compendex (Engineering Village)	(((DSL) OR (DSML) OR ("Domain Specific Language") OR ("Domain-Specific Language") OR ("Domain-Specific-Language") OR ("Small Language") OR ("Small-Language") OR ("Little-Language") OR ("Little Language") OR ("Domain Specific Modeling Language") OR ("Domain-Specific Modeling Language") OR ("Domain-Specific-Modeling-Language") OR ("Domain-Specific Modeling-Language" OR "Domain Specific Modeling-Language") WN KY) AND (tool) OR ("Language Workbench") WN KY) AND ((English) WN LA)
IEEE Xplore - DSL	(((((DSL OR "Domain Specific Language" OR "Domain-Specific Language" OR "Domain-Specific-Language" OR "Domain Specific-Language" OR "Small Language" OR "Small-Language" OR "Little-Language" OR "Little Language") AND (Tool OR "Language Workbench")))))
IEEE Xplore - DSML	(((DSML OR "Domain-Specific Modeling Language" OR "Domain Specific Modeling Language" OR "Domain-Specific Modeling-Language" OR "Domain Specific Modeling-Language") AND (Tool OR "Language Workbench")))
Science Direct	TITLE-ABSTR-KEY(DSL or DSML or "Domain Specific Language" or "Domain-Specific Language" or "Domain-Specific-Language" or "Domain Specific-Language" or "Little-Language" or "Little Language" or "Small-Language" or "Small Language" or "Domain Specific Modeling Language" or "Domain-Specific Modeling Language" or "Domain-Specific-Modeling-Language" or "Domain-Specific Modeling-Language") and TITLE-ABSTR-KEY(tool or "Language Workbench")
SCOPUS	(TITLE-ABS-KEY (dsl OR dsml OR "domain specific language" OR "domain-specific language" OR "domain-specific-language" OR "domain specific language" OR "domain specific modeling language" OR "domain-specific modeling language" OR "domain specific modeling-language" OR "domain-specific-modeling-language") AND TITLE-ABS-KEY (tool OR "language workbench")) AND (LIMIT-TO (LANGUAGE , "English"))
Springer Link	((DSL OR DSML "domain specific language" OR "domain-specific language" OR "domain-specific-language" OR ("Small Language") OR ("Small-Language") OR ("Little-Language") OR ("Little Language") OR ("Domain Specific Modeling Language") OR ("Domain-Specific Modeling Language") OR ("Domain-Specific-Modeling-Language") OR ("Domain-Specific Modeling-Language" OR "Domain Specific Modeling-Language") AND (tool or language workbench))

The second activity is the definition of the search string based on the terms and synonyms related with the research area. A crucial point here is the difference between SMS and SLR search strings. In SMS generic strings should be less restrictive to achieve more wider results while SLR aims to specific areas or topics.

During the third activity, the generic string is derived into specific strings for each database search engine. Then, a preliminary search with the variations of the search string is performed in the search engines. If problems are found with the strings, there is a possibility that the generic strings have problems as well, so it should be re-defined. It

is difficult to say if a search string have problems or not. A possible hint to identify this problem is a huge or tiny number of retrieved studies.

During the last activity the returned studies are briefly analyzed. This analysis may be performed only in the titles and abstracts of the studies. If there is a great number of studies outside the scope of the review, there is a possibility that there are still problems with the search strings. If the number of studies outside scope are small or zero, then researchers may continue to the other phases of their review.

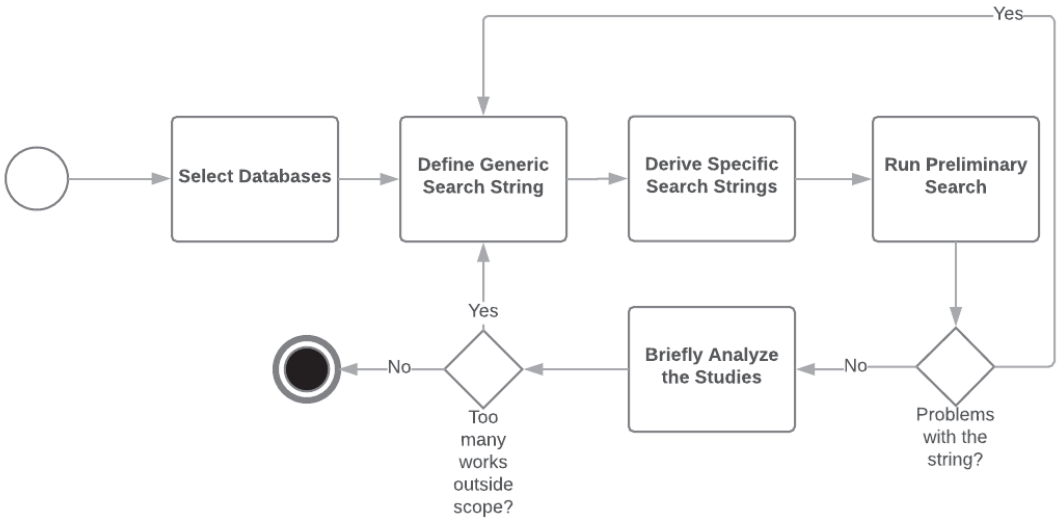

Figure 3. Activities conducted before the DSLs SMS

5. Conclusion

This work presented evidences related to the impact of the search phase in a SMS conducted in the DSL area. This evidence could be used to analyze issues that may have negative impact in a SMS such as: researchers decided to use only one digital library (*e.g.* SCOPUS), thus involuntary excluding exclusive works from other databases; and researchers used a poor formulated search string which did not cover the area as expected.

The evidence was collected by observing and documenting data during the conduction of the search phase of the SMS in the DSL area. We registered the different numbers of studies for each database, as well as exclusive works of a database that were included in the final SMS selection. Based on this numbers we have made some decisions, such as reformulating a search string.

The results of the data collected during our SMS gave us evidence to achieve two conclusions:

1. **The researches should include as many databases as possible**: this selection may have small but important impact in the final selection. For instance, if during our SMS we did not search works in SCOPUS, 4 studies would not be included in the final selection. Although this is a small number, the not inclusion of these works would have a huge impact in the SMS analysis, such as, not including emerging DSL tools;

2. **Minor changes in the search string may change the retrieved studies numbers to more than 1000**: in addition, the IEEE Xplore string had to be split because of the limitation of the database terms. When deriving the generic string for each database we also had to be careful, as some databases had returned a complete different number of studies with minor changes in the strings.

In addition, we documented five activities that may give support to researches with low experience conducting a SMS or SLR. This activities should be performed before the conduction of the systematic review and can be considered preliminary activities.

For the future, we are currently working to expand this analysis into the additional phases of the SMS protocol. We plan to construct a set of guidelines based on this research and develop a tool to give support to SMS and SLR conduction with the implementation of such guidelines.

References

[Biolchini et al. 2005] Biolchini, J., Mian, P. G., Natali, A. C. C., and Travassos, G. H. (2005). Systematic review in software engineering. *System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES*, 679(05):45.

[Budgen and Brereton 2006] Budgen, D. and Brereton, P. (2006). Performing systematic literature reviews in software engineering. In *Proceedings of the 28th international conference on Software engineering*, pages 1051–1052. ACM.

[Connolly et al. 2012] Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., and Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, 59(2):661–686.

[Kitchenham 2004] Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

[Petersen et al. 2008] Petersen, K., Feldt, R., Mujtaba, S., and Mattsson, M. (2008). Systematic mapping studies in software engineering. In *EASE*, volume 8, pages 68–77.

[Šmite et al. 2010] Šmite, D., Wohlin, C., Gorschek, T., and Feldt, R. (2010). Empirical evidence in global software engineering: a systematic review. *Empirical software engineering*, 15(1):91–118.

[Stuck et al. 1999] Stuck, A. E., Walthert, J. M., Nikolaus, T., Büla, C. J., Hohmann, C., and Beck, J. C. (1999). Risk factors for functional status decline in community-living elderly people: a systematic literature review. *Social science & medicine*, 48(4):445–469.