Improvements in the StArt Tool to Better Support the Systematic Review Process

Sandra Fabbri, Cleiton Silva, Elis Hernandes
Federal University of São Carlos
Rod. Washington Luís, km 235, CEP 13565-905, São Carlos, SP, Brazil
+55 16 3351 9488
{sfabbri, cleiton.silva, hernandes}@dc.ufscar.br

Fábio Octaviano, André Di Thommazo, Anderson Belgamo
Federal Institute of São Paulo
Rod. Washington Luís, km 235, CEP 13565-905, São Carlos/Piracicaba, SP, Brazil
+55 16 3351 9458
{foctaviano, andredt, anderson}@ifsp.edu.br

ABSTRACT
Context: Systematic Review (SR) is a methodology used to find and aggregate relevant existing evidence about a specific research topic of interest. It can be very time-consuming depending on the number of gathered studies that need to be analyzed by researchers. One of the relevant tools found in the literature and preliminarily evaluated by researchers of SRs is StArt, which supports the whole SR process. It has been downloaded by users from more than twenty countries. *Objective:* To present new features available in StArt to support SR activities. *Method:* Based on users' feedback and the literature, new features were implemented and are available in the tool, like the SCAS strategy, snowballing techniques, the frequency of keywords and a word cloud for search string refining, collaboration among reviewers, and the StArt online community. *Results:* The new features, according to users' positive feedback, make the tool more robust to support the conduct of SRs. *Conclusion:* StArt is a tool that has been continuously developed such that new features are often available to improve the support for the SR process. The StArt online community can improve the interaction among users, facilitating the identification of improvements and new useful features.

Categories and Subject Descriptors
D.2.m [**Software Engineering**]: Miscellaneous – *Evidence-Based Software Engineering (EBSE)*

General Terms
Design, Standardization.

Keywords
Systematic review, systematic literature review, evidence-based software engineering, StArt tool, tool support.

EASE '16, June 01-03, 2016, Limerick, Ireland

DOI: http://dx.doi.org/10.1145/2915970.2916013

1. INTRODUCTION
Evidence-Based Software Engineering (EBSE) refers to research methods used to build a body of knowledge in the discipline of software engineering (SE) [1]. In this context, systematic reviews (SR) (a.k.a. systematic literature reviews (SLR)) have been providing mechanisms to identify relevant research evidence [2].

Since its introduction to the SE field in 2004, reviewers from different areas have been reporting challenges based on their experiences when conducting SRs, including the time spent to perform them. This challenge is aggravated by the fact that SRs are usually conducted without a tool developed to support SRs that could help speed up the process. Marshal and Brereton [3] report on the availability of several tools. StArt (State of the Art through Systematic Review) [4] is one of them, and it has been developed and improved by this research group. StArt currently has more than 3000 downloads from distinct users worldwide, including countries like Brazil, the United States, Italy, Spain, the United Kingdom and others.

The objectives of this paper are: i) to describe how StArt currently supports the whole SR process and ii) to present the new features (available from version 3.0) implemented in the tool to help SR practitioners.

The remainder of this paper is organized as follows. Section 2 presents the background and related work. Section 3 presents a description of how StArt provides support to the SR activities. Section 4 presents the new features available in the tool, and, finally, Section 5 presents the conclusions.

2. BACKGROUND AND RELATED WORK
Kitchenham and Charters [2] outline three phases to conduct an SR: (i) planning; (ii) execution and (iii) summarization/reporting. In the first phase, reviewers must identify the need for a review and create a review protocol containing the important information of the SR. In the second phase, they should identify and select relevant primary studies, perform the data extraction and synthesize the extracted data. Finally, in the third phase, they should summarize and report the results of the SR to relevant communities.

An SR can be very time consuming because of the required rigor needed for performing its activities and can be error prone because of the subjectivity of some activities. Thus, reviewers could use some tool developed to support part of or the whole SR process, aiming at minimizing the required effort. In this context, Marshall and Brereton [3] report a mapping study about existing tools to support the SR process. They identify 11 tools of which three of them (SLR-Tool, SLuRp and StArt) support the whole process. The remaining tools support specific parts of an SR. Marshall, Brereton and Kitchenham [5] identify a fourth tool called SLRTOOL that supports the whole process. They compare and evaluate the four tools through a feature analysis based on a calculated score for each tool. StArt was classified as the second highest score among the tools. This study helped our research group to improve StArt in order to fix some reported issues like the lack of security, quality criteria and collaboration.

In addition to the improved features for fixing the issues, StArt has some particularities, like the ranking of primary studies according to their relevance based on a calculated score and the use of visualization resources to support some SR activities [4]. They are mentioned in the next sections.

3. HOW StArt SUPPPORTS AN SR PROCESS

3.1 The Protocol

The StArt tool allows researchers to fill in an SR protocol whose fields are based on the ones described in [2] and in [6]. It has information about all the SR phases, including the goal, research questions, search and selection strategies, inclusion and exclusion criteria, a data extraction form, quality criteria and a strategy for synthesizing the extracted data.

The new features related to the protocol are as follows. i) It is not mandatory anymore to structure the research questions based on the PICOC criteria [7]. This was done to meet what was recently mentioned in [8]. ii) Data extraction form fields may be created by choosing among three kinds of fields: a text field, pick-one list or pick-many list. iii) Definition of quality criteria and a numeric scale to assign values to them (if desired) in order to rank studies based on a quality level.

3.2 The Initial Selection Activity

Once the protocol is defined, the next activity is to load the bibliographic information from the studies into the tool. StArt supports the main online search databases, including Scopus, IEEE, ACM and Web of Science. Besides, there are other online databases that users from distinct areas asked to be added, like PubMed. The file formats that StArt can understand are BibTex, RIS, Medline and Cochrane.

After loading the information from studies, the initial selection activity should be performed. Each study has a score assigned to it. The score is calculated based on the occurrences of the keywords (defined in the protocol) found in the title, the abstract and the keywords of the studies. This score is used in different features of the tool and provides the ranking of studies according to their supposed relevance to the context of the SR. Based on the reading of the titles and abstracts, the researcher must set the studies as accepted or rejected, choosing the corresponding inclusion and exclusion criteria that are met for each study. It is also possible to define a reading priority for each study, classifying it as very high, high, low or very low. This can guide the researcher in reading the full texts in the next activity. Besides, StArt has text mining resources that help in the automatic detection of duplicated studies and in the detection of similar studies of a selected study. This can be helpful in case of classifying a study as relevant and searching for studies with similar contents.

3.3 The Data Extraction Activity

The data extraction activity follows the initial selection. The researcher should read the full text of the studies accepted in the initial selection activity. StArt shows only the accepted studies in the extraction activity, and the data extraction form (defined in the protocol) is displayed for each study. The researcher has to fill the extraction form out for each study because these data will be used in the data synthesis activity.

It is possible to assign quality criteria to the studies too. The quality criteria (defined in the protocol) are displayed in the extraction activity, allowing the researchers to assign them to the studies and choose a value from a numeric scale in order to have a quality value calculated for each study (if desired). Thus, the researchers may reject a study if it does not meet a minimum quality value or just use the quality value to rank the studies.

3.4 The Data Synthesis Activity

The last activity that must be done is the data synthesis. StArt lists all the studies considered relevant (accepted by reviewers) after the full text reading. It is possible to check the information of each study (data extraction form, inclusion and exclusion criteria, quality criteria, comments) and make the data synthesis. StArt has a built-in text editor that allows a researcher to synthesize the data. Besides, some visualization resources help in this activity also, as presented in [4]. Finally, StArt allows exporting all the protocol and study information in an Excel format for additional analysis if desired.

Figure 1 shows a screen montage containing some important screens of StArt. The left menu shows all SR phases supported by the tool (planning, execution and summarization), highlighted by a red color, and the main screen shows information about the papers, like their scores (highlighted by a red circle), the status of selection and extraction, and reading priority (see Figure 1-A). Details of studies used for the initial selection are displayed in Figure 1-B. Part of a data extraction form is shown in Figure 1-C.

4. StArt - NEW FEATURES AVAILABLE

4.1 The SCAS Strategy

The SCAS (Score Citation Automatic Selection) strategy [9] is implemented in StArt. It is a strategy to semi-automate the initial selection of studies in SRs based on two features calculated for each study: the score and number of citations. When combined, the studies are classified into four quadrants: i) Quadrant 1 – studies with high scores and at least one citation; ii) Quadrant 2 – studies with high scores and no citation; iii) Quadrant 3 – studies with low scores and at least one citation; and Quadrant 4 – studies with low scores and no citation. We would like to mention that the study age is under investigation to complement the number of citation feature.

It is important to highlight that a score is considered high or low based on a calculated cut-off value. There are two techniques to calculate the cut-off value: i) the 50% technique, which ranks the studies by score and considers the score of the study ranked in the middle of the list as a cut-off value candidate, and ii) the J48 technique that considers the value of the score that is 17.14% of the value of the highest score as a second cut-off value candidate. The rate of 17.14% was obtained through the J48 decision tree available in the Weka tool after inputting data from some published SRs. StArt compares the values of the candidates and determines that the cut-off value is the lowest score between the

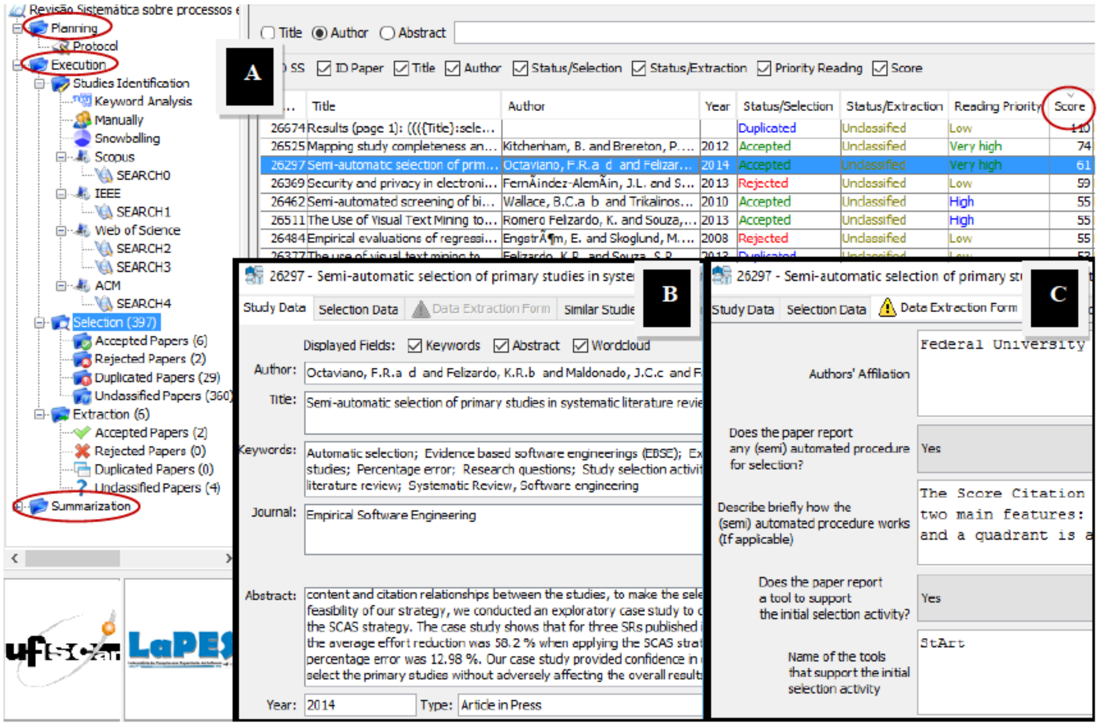

Figure 1. An Overview of StArt.

two chosen candidates. Consequently, a larger number of studies are classified with a high score (probably relevant studies).

After classifying the studies into the quadrants, SCAS suggests that the studies belonging to quadrant 1 should be accepted, studies belonging to quadrant 4 should be rejected, and studies belonging to quadrants 2 and 3 should be manually reviewed. StArt has the option to apply the SCAS recommendations and automatically set the studies belonging to quadrants 1 and 4 as accepted and rejected respectively. The tool automatically sets the criteria "Included by SCAS recommendation" or "Excluded by SCAS recommendation" to the studies belonging to those mentioned quadrants. Figure 2 shows an example of SCAS usage in the tool. The red arrow points out the button that must be pushed to generate the quadrants, and the red circle shows the column where the generated quadrant for each study is displayed.

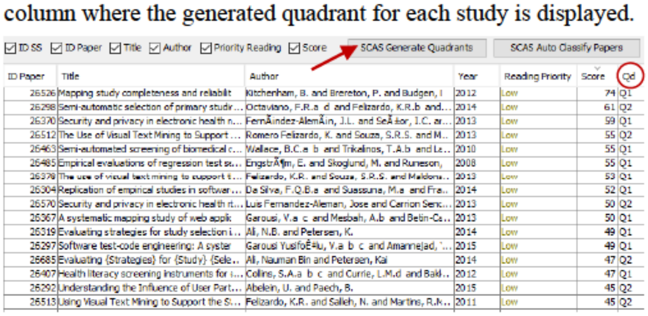

Figure 2. Generation of SCAS Quadrants in StArt.

4.2 Applying Snowballing Techniques

Snowballing is a technique for identifying primary studies in SRs and systematic mappings (SM). Wohlin [10] describes how to perform snowballing as the main strategy to search for primary studies with a heterogeneous initial set of studies. Snowballing can also be used to complement the search strategy in online databases when performing SRs or SMs as suggested in [8, 11].

StArt allows the execution of the two mentioned approaches. For running the snowballing backward technique [10], StArt gets references using the full text of the primary study. This is possible by using the full text in a PDF format or the reference file in the BibTex format as exported by most online databases, such as Scopus. Resources for running the snowballing forward technique [10] are still under development in StArt. Currently, the researcher must manually search for studies in the online databases and import them into the tool.

Due to the exponential growth in the number of studies found by snowballing, the technique is very time-consuming [10]. To minimize this problem, StArt shows a ranking of references, suggesting what references should be investigated. The prioritization is done using text processing algorithms, and the ranking is created based on the frequency of relevant keywords, the frequency of recurrent authors in the SR, the frequency of recurring conferences in the SR and the number of times a single reference is cited by other studies in the SR. The researcher can choose to follow the ranking and read only the best-ranked studies, which is less time-consuming, or perform a reading of all new references found, which is more complete but very time-consuming. Figure 3 shows the ranking of references used to prioritize the reading of studies in StArt. In the example, the studies were ranked by authors (see red ellipse), but it allows the

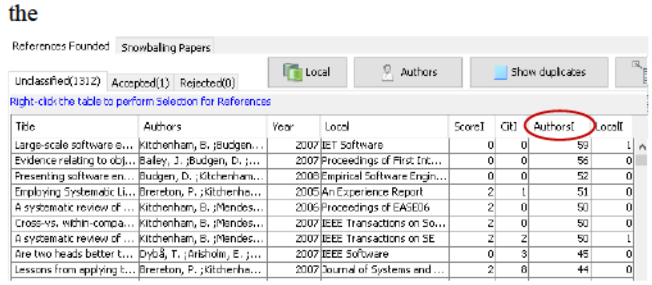

Figure 3. The Ranking of References in StArt.

ranking of studies by score, number of citation or scientific event, just clicking on the respective column.

4.3 Frequency of Keywords and Word Cloud

The frequency of keywords is a new feature available in StArt that allows researchers to search for terms that could be potentially missing in the search strings they ran in the online databases. The tool detects and shows the number of times the authors' keywords existing in the studies is found in the SR, ordering them by frequency. The tool allows three options for verifying the frequency of keywords considering: (i) all studies; (ii) only the studies accepted in the initial selection; and (iii) only the studies rejected in the initial selection.

When option (i) is chosen, the tool ignores the status of the studies, even if a sample was already classified, and displays the frequency of keywords, considering all the studies. The researcher can have an overview of the most frequently used keywords in the SR context. Option (ii) is recommended when the researcher has a sample of studies already classified. Then, he/she can analyze the most frequently used keywords in the relevant (accepted) studies, aiming to identify new relevant terms that could be used to refine his/her previous search string. Option (iii) is recommended when the researcher wants to identify "bad keywords", that is, keywords that appear many times in the rejected studies and can help in identifying possible irrelevant studies for the context of the SR, which should be rejected. This may help in the detection of study rejection trends based on the identified bad keywords, like "every time the keywords x and y are found in a study, there is a high possibility of rejecting this study".

StArt highlights the keywords with green, yellow or orange circles. When a keyword has a green circle, it means that the term was already used in the search string. When a keyword has an orange circle, it means that the term was not used in the search string. When a keyword has a yellow circle, it means that the term can be potentially covered by a similar term. Figure 4 shows the screen in StArt regarding the frequency of keywords. In the example, only the studies accepted in the initial selection were considered. The term "systematic review" was used in the search string, the term "data mining" was not used and the term "systematic review (topic)" was possibly covered by another term ("systematic review" in this case). In addition to the numeric value corresponding to the frequency of keywords, StArt can show the information through a word cloud. This is a visualization mode to detect relevant keywords for a set of studies.

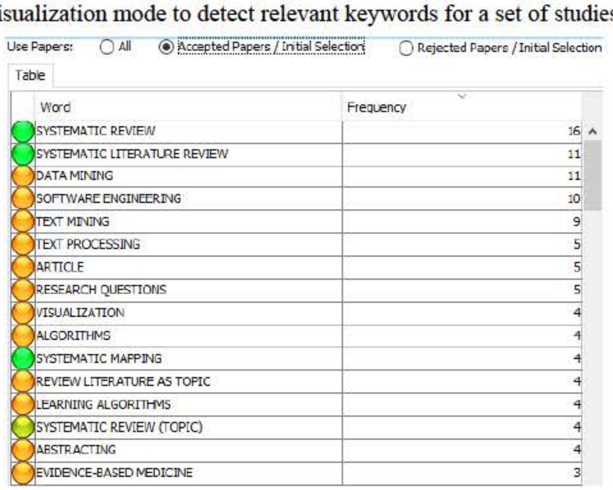

Figure 4. Frequency of Keywords Available in StArt.

4.4 Performing a Collaborative SR

Another new feature available in the tool is the possibility of performing collaborative SRs. In this context, the reviewers are registered in the tool, and a group coordinator who is responsible for some specific activities is chosen from among them. The studies should be split among the registered reviewers for the initial selection activity, and a number of studies in common should be shared among them in order to measure the agreement level. An advantage of StArt is that the score feature can be used to achieve a more balanced and controlled distribution of studies among reviewers based on the possible relevance of the studies. This intends to promote a balanced distribution of studies instead of a randomized one, trying to avoid one reviewer receiving only irrelevant studies and being prevented from performing his/her task quickly and another receiving only relevant studies and taking a long time to complete his/her tasks. Besides, it is also possible to choose how the shared number of studies is sent to reviewers based on the score feature, sending only possibly relevant studies (by choosing the top n studies descending ordered by score), only possibly irrelevant studies (by choosing the bottom n studies from the ordered list) or sending both possibly relevant and irrelevant studies (by choosing studies from the top, the bottom and the middle of the ordered list). This aims to compare the decisions made by each reviewer with respect to the distinct relevance of studies.

After performing the initial selection, the individual reviewers' files, previously sent to the coordinator, are merged, and the existing conflicts related to the studies shared in common need to be resolved. The tool allows a manual or automatic merge action, indicating default actions to be taken for the possible conflicts. If the manual option is chosen, every time a conflict arises, the coordinator will be asked to make a decision, showing the chosen option by default but allowing changes. If the automatic option is chosen, StArt applies the default chosen options for conflicts wherever possible. For the initial selection, three kinds of conflicts may arise in StArt: status (accepted or rejected), inclusion and exclusion criteria and reading priority (very high, high, low or very low). Figure 5 shows the merge screen for the initial selection activity in StArt.

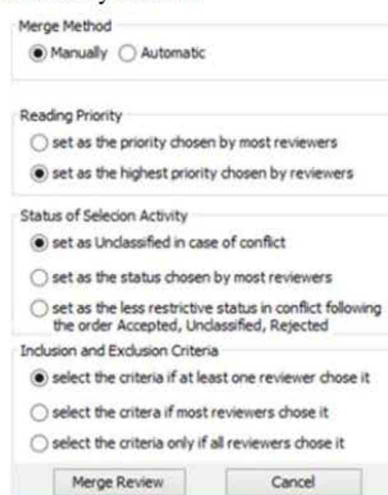

Figure 5. Merge Options for the Initial Selection in StArt.

The same split and merge actions are required for the extraction activity. Studies are split by score and sent to the reviewers, using a number of studies in common to assess the agreement level. The reviewers work individually and send the results back for merging, which can be done manually or automatically by the tool. For the data extraction activity, three kinds of conflicts may arise in StArt: status after full text reading (accepted or rejected); extraction form fields, in particular for the pick-one and pick-many lists (see Section 3.1) that have default values, and quality criteria. Once the final merge is done, the next task is to perform the data synthesis. The coordinator is responsible for leading and performing the split and merge tasks and coordinating the data synthesis.

4.5 The StArt Online Community

The StArt Online Community was created to improve the communication and interaction among StArt users around the world and the communication between StArt users and the StArt development team. In the StArt Community, users are able to report bugs, provide suggestions and comments, see news and updated information about the tool and be informed about new version releases and tutorials. Besides, it will be possible for users to upload their SRs supported by StArt and share the SR information, if desired. The idea is to build an SR repository where StArt users will be able to upload their SRs, keep the files as backup (like a cloud computing service) and look for SRs already performed and shared by other users, thus avoiding performing an SR already done and facilitating the update of an existing SR. This is possible because all information regarding an SR performed in StArt is packed only in a .start file. Thus, it is only necessary to upload the .start file, and the SR will be stored in the repository. It is possible to send PDF files (full texts) as well by compressing them and uploading only compressed files.

It is important to mention that even if a user uploads an SR into the StArt Community repository, it will only be shared and visible to other users if that user chooses the option to share his/her SR. Otherwise, it will only be available for himself/herself as a backup file. The StArt Community link is available at the official StArt webpage: http://lapes.dc.ufscar.br/tools/start_tool.

5. CONCLUSION

An SR can be a very time-consuming process depending on the number of primary studies that need to be reviewed. A way to minimize this problem is by using some tool specially developed to support SRs. StArt is one of these tools. It has been being developed by this research group since 2006 and has thousands of users worldwide.

New features are available in StArt to fix some issues pointed out in [5] and to improve the support provided by the tool for SR practitioners. Among these new features, we highlight the SCAS strategy, the snowballing techniques, the frequency of keywords for improving search strings, the support for collaborative SRs and the creation of the StArt online community, which is supposed to improve the interaction among StArt users, including the possibility of uploading and sharing (if desired) their SRs performed in StArt. Thus, this paper aimed to provide a feature update of StArt to the EBSE community, reporting what is new and available for use in the tool. The StArt 3.0 version is available for download at http://lapes.dc.ufscar.br/tools/start_tool.

For future work, we intend to perform additional tests on the features available in StArt and to keep developing the tool and the StArt community by organizing and maintaining an SR repository. In addition, we intend to post new tutorial videos regarding the new available features on the official LaPES Youtube channel at https://www.youtube.com/user/lapesufscar, which will also be available at the StArt Online Community.

6. REFERENCES

[1] Kitchenham, B. A., Dybå, T., and Jørgensen, M. 2004. Evidence-based software engineering. In *Proceedings of the 26th International Conference on Software Engineering* (Edinburgh, Scotland, May 23 - 28, 2004). ICSE'04. IEEE Computer Society, Washington DC, USA, 273–281.

[2] Kitchenham, B. and Charters, S. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report. Keele University and University of Durham, version 2.3.

[3] Marshall, C. and Brereton, P. 2013. Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study. In *Proceedings of the 7th International Symposium on Empirical Software Engineering and Measurement* (Baltimore, USA, October 10 – 11, 2013). ESEM'13. IEEE, Washington DC, USA, 296-299.

[4] Fabbri, S., Hernandes, E., Di Thommazo, A., Belgamo, A., Zamboni, A., and Silva, C. 2012. Using Information Visualization and Text Mining to Facilitate the Conduction of Systematic Literature Reviews. In *Proceedings of the 14th International Conference on Enterprise Information Systems* (Wroclaw, Poland, June 28 – July 01, 2012). ICEIS'12. Springer, 243–256.

[5] Marshall, C., Brereton, P., and Kitchenham, B. 2014. Tools to Support Systematic Reviews in Software Engineering: A Feature Analysis. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (London, UK, May 13 – 14, 2014). EASE'14. ACM, New York, NY, 139-148.

[6] Biolchini, J., Gomes, M., Cruz, N., and Horta, T. 2005. *Systematic Review in Software Engineering*. Technical Report. Federal University of Rio de Janeiro.

[7] Petticrew, M. and Roberts, H. 2006. How to appraise the studies: An introduction to assessing study quality. In *Systematic Reviews in the Social Sciences: A Practical Guide.* Malden: Blackwell Publishing, 125–163.

[8] Kitchenham, B. and Brereton, P. 2013. A systematic review of systematic review process research in software engineering. *Information and Software Technology* 55 (Aug. 2013), 2049–2075.

[9] Octaviano, F., Felizardo, K., Maldonado, J., and Fabbri, S. 2015. Semi-automatic selection of primary studies in systematic literature reviews: Is it reasonable? *J. Empirical Software Engineering* 20, 6. (Dec. 2015), 1898–1917.

[10] Wohlin, C. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (London, UK, May 13 – 14, 2014). EASE'14. ACM, New York, NY, 1–10.

[11] Petersen, K., Vakkalanka, S., and Kuzniarz, L. 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology* 64 (Aug. 2015), 1-18.