HAL
open science

Automating Systematic Mappings: adding Quality to Quantity

Regina Motz, Genoveva Vargas-Solar, Umberto Souza, Javier A. Espinosa-Oviedo, Martin A. Musicante, José-Luis Zechinelli-Martini, Alberto Pardo

To cite this version:

Regina Motz, Genoveva Vargas-Solar, Umberto Souza, Javier A. Espinosa-Oviedo, Martin A. Musicante, et al.. Automating Systematic Mappings: adding Quality to Quantity. 39th International Conference on Software Engineering. (ICSE'17), May 2017, Buenos Aires, Argentina. hal-01517187

HAL Id: hal-01517187

https://hal.science/hal-01517187v1

Submitted on 9 May 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Poster: Automating Systematic Mappings, adding Quality to Quantity

Regina Motz[1], Genoveva Vargas-Solar[2], Umberto Souza da Costa[3], Javier Alfonso Espinosa-Oviedo[4], Martin A. Musicante[3], José Luis Zechinelli-Martini[5], Alberto Pardo[1]

I. INTRODUCTION

This paper introduces an approach for semi-automating some of the steps of the Systematic Mapping Studies (SMS) [4], [3] and enhancing them with quality criteria[1]. The methodology for producing SMS defines the following workflow: (i) defining research questions; (ii) querying bibliography data sources using a key-word complex query[2]; (iii) selecting relevant documents from data sources; (iv) keywording of documents and defining a classification scheme; (v) classifying the documents and (vi) producing the mapping and answers to the research questions.

Frequently these steps are performed manually and somehow empirically. Thus, they can be time consuming and prone to errors. Existing work has been done for developing tools that automate some of the steps of the methodology [1], [2] and enabling the analyst to include some qualitative criteria for selecting and analyzing sources. These criteria can be, for example, the H-index of authors, sources classification, publication freshness. The majority do not address the conditions in which the steps are done regarding the criteria and guidelines used to perform them. The objective of our work is to enhance the systematic mapping (SM) methodology: (a) Adding domain knowledge and quality criteria for guiding key-word selection, query expression, sources selection and expanding and refining bibliographic collections; (b) Automating systematic mapping using data analytics and information retrieval techniques.

This poster gives an overview of our approach that models the knowledge domain and sources classification according to different quality measures. Our approach also uses data processing algorithms to automate the steps of SMS methodology guiding it with user objectives (e.g., the type of study she wants to perform) and quality preferences.

[1] INCO, Universidad de la República, Montevideo, Uruguay. {rmotz,pardo}@fing.edu.uy
[2] CNRS, LIG-LAFMIA, Saint Martin d'Hères, France. genoveva.vargas@imag.fr
[3] Universidade Federal do Rio Grande do Norte, Natal, Brazil. {umberto,mam}@dimap.ufrn.br
[4] Barcelona Supercomputing Center, Barcelona, Spain. javiera.espinosa@gmail.com
[5] Universidad de las Américas-Puebla, LAFMIA, Cholula, Mexico. joseluis.zechinelli@udlap.mx

[1]SMS provide an overview of a research area to pinpoint trends, strong points and research opportunities

[2](This key-word complex query is named search string in the SMS methodology jargon.

II. AN AUTOMATIC QUALITY ORIENTED SMS

Figure 1 presents the general approach that we propose for semi-automating the systematic mapping workflow and adding quality concerns to it. Our work first considers the problem of identifying resources relevant to a research question that involves one or several knowledge domains. It consists in identifying those resources exhibiting their multidisciplinary dimension and also other quality measures such as reputation of the authors, of the publication in which they are published and their provenance.

For example, let us say the topic is *King Arthur* and the research question is *Why are there many kings Arthur?* The challenge here is to select the resources (i.e., first documents providers and then documents themselves) that can contribute to answer the question. In this case, we can be interested in exploring the search space consisting in the set of articles stored in scientific databases, such as ACM, DBLP and Springer databases to identify those that are related to a specific topic in digital humanities and data science. Once a search space has been retrieved, according to SM methodology, it will be filtered, aggregated and classified in order to produce and analytic view that can answer target research questions (*Why are there many kings Arthur?* in our example).

Our approach supports the iterative and interactive research process promoted by the SM methodology. In order to do so, it introduces quality concerns involving domain knowledge, quality criteria, data analytics and information retrieval techniques. Our approach organizes the systematic mapping workflow steps into two groups (see Figure 1):

1. The one that involves steps i - iii devoted to build the search space (i.e., collection of resources). We propose to add guidelines related to the way research questions are translated into key-word queries (step (i)). Indeed, frequently key-words are chosen empirically or based on previous knowledge of a scientist. We propose to use ontologies to validate/enrich the vocabulary used to define keywords for one or several domains and keep track of this choice. Depending on the domains, we assume that either ontologies have been validated by experts (the Stanford Encyclopedia of Philosophy in our example https://plato.stanford.edu/entries/fiction/) or built out of crowdsourcing processes performed on the communities working on such domains. For example vocabulary flashcards of *King Arthur* (https://quizlet.com/58126560/king-arthur-vocabulary-list-3-flash-cards/) in our example. Ontologies and vocabularies can help to define key-word queries that

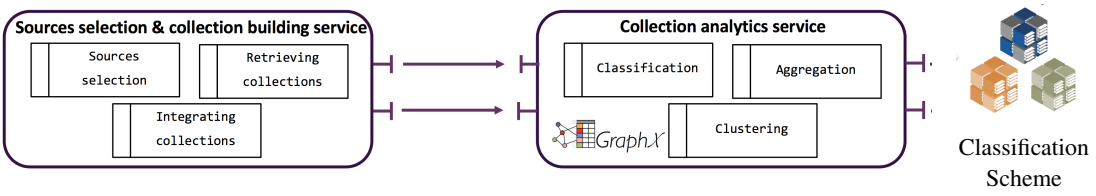

Fig. 1. SMS automation proposal.

can help to retrieve resources from sources (e.g., specialized and general purpose search engines).

Similarly, criteria used for choosing sources are not explicitly stated and they remain implicit in the SM process (step (ii)). We propose to use catalogs where sources reputation and provenance are reported or can be deduced (publication tool, author or research centre). In our example the catalog of Arthurian Legends can give references to prominent journals, anthologies, books and other resources (http://d.lib.rochester.edu/camelot/text/sources-for-the-study-of-the-arthurian-legends). Of course catalogs are directly associated to the discipline of the research.

Regarding step (iii), since resources and in particular papers belong to knowledge communities and they have related publications stemming from similar journals, conferences, books, and authors (e.g. H index), and scientific groups. The search space can be extended and completed considering other resources (i.e., papers) produced according to these criteria to increase the probability of having a broad view of resources that can potentially answer the research question.

2. The one that involves step iv - vi devoted to aggregate, classify and analyze the search space. Regarding resources relevance, the SM methodology relies on the intelligence and flare of the person who filters documents and decides whether their content is related to her research or not. This can be a subjective method that depends on the expertise of the analyst. For step (iv) the SM methodology does not state guidelines for choosing facets and dimensions and how to combine them for answering the research query. This is responsibility of the person applying the methodology for her research. The choice of the dimensions and facets and the best multi-dimensional classification can be understood and supported by quantitative and qualitative arguments. We propose to apply text analysis methods for extracting frequent terms in the search space and then clustering and classifying them using reference ontologies. The facets can be for example the most frequent terms or more general or more specific terms in the ontology. Going back to our example, assume that *King Arthur* is a frequent term in a collection of documents. An ontology can propose *Arthurian legends* as a more generic concept, which could be in fact a facet, and *King Arthur* a dimension in the facet. Of course, this is fundamental knowledge for an expert but a novice scientist might not know. The process can be programmed and provide a wide Besides other possibilities might be possible for defining such facets and an expert could have different classifications automatically generated thanks to this strategy.

Thus, we propose to

- Use data analytics and information retrieval techniques to estimate the topic of the resource, its pertinence with respect to the query, clustering it with similar resources, classifying it with respect to the concepts of different related domains.
- Derive facets and dimensions and populating the papers database. Using ontologies and data analytics to help the user build a classification scheme.

III. Concluding Remarks

We believe that systematic mapping, requires a qualitative and "less empirical" perspective. We feel that choosing key words in the second phase of the methodology can be empirical. Thus, using vocabularies of the knowledge domain, can help to have a more representative choice. Moreover, quality guidelines can be introduced by explicitly adding filtering and clustering criteria related to the provenance of resources, the impact factor of the conference/journal where they appear, authors reputation (given for example by their H factor) their institution and country. Without discarding the quantitative analysis, adding these criteria can increase the quality and value of the analysis. We are currently working in providing tools that can help to add quality to the systematic mapping method.

References

[1] Manolo J Cobo, Antonio Gabriel López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. *Journal of Informetrics*, 5(1):146–166, 2011.

[2] Manolo J Cobo, Antonio Gabriel López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, 62(7):1382–1402, 2011.

[3] David Gough, Sandy Oliver, and James Thomas. *An introduction to systematic reviews*. Sage, 2012.

[4] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic mapping studies in software engineering. In *12th international conference on evaluation and assessment in software engineering*, volume 17. sn, 2008.