Protocol for the systematic literature search for:
A Literature Review on Methods for the Extraction of Usage Statements of Software and Data

Frank Krüger David Schindler

July 29, 2019

1 Structured Search

We performed a systematic literature review [1] to analyse existing work dealing with the identification and extraction of usage statements of research artefacts from scientific literature. To this end, we searched in public accessible databases, collected the results, and selected articles of interest. Figure 1 provides an overview of the workflow we followed during the literature retrieval.

Firstly, based on an initial set of relevant articles, the keywords for further search were identified in a manual process. Similarities between these articles of interest were then assessed and used to search and identify further relevant articles in a broad range of possible sources. Finally, an analysis of the articles with respect to technical approaches and the employed datasets was conducted. Below, we first describe the individual steps of the collection, then the selection and finally the analysis of the documents. The intermediate results of all steps, as well as the list of finally analysed documents are available at [2].

In order to establish a search query to retrieve the documents for further analysis, we initially performed a manual, unstructured search for relevant literature by use of Google Scholar (see Figure 1 **A1**). As a result we obtained 13 articles about the identification and extraction of usage statements for research artefacts (Figure 1 **A2**). These articles were used to 1. select suited literature databases and 2. create a query string based on keywords.

1.1 Selection of scientific literature databases

One reason to create the initial set of documents was to be able to estimate the coverage of the databases of interest. We considered the following databases as search targets: • Scopus • Science Direct • Google Scholar • ISI Web of Science • Microsoft Academic Search • Proquest • Pubmed • CochraneLibrary • Springer • JSTOR • PLoS • Wiley • IEEE • ACM • Europe PMC • Directory of Open Access Journals (DOAJ). In a previous step, three databases had to be excluded, because they do not support the intended query. Microsoft Academic and Semantic Scholar do not allow the direct input of logical queries and Google Scholar has a query length restriction of 256 characters in place.

For all remaining databases we manually assessed how many articles from our base set they contain (see Figure 1 **B**). We chose • Scopus, • ISI Web Of Science, • DOAJ and • Science Direct as our final search targets. Scopus and Web of Science were chosen because of the best coverage of 7 and 5 articles from the set. DOAJ only covers 2 articles, but was included because both articles were uniquely found in this database. Science Direct was included to fully cover the base of Elsevier even so only 1 article from the base set is available from there. From the excluded databases, Wiley includes 3 articles, Springer 2 and IEEE 1, all of which were also available from other databases.

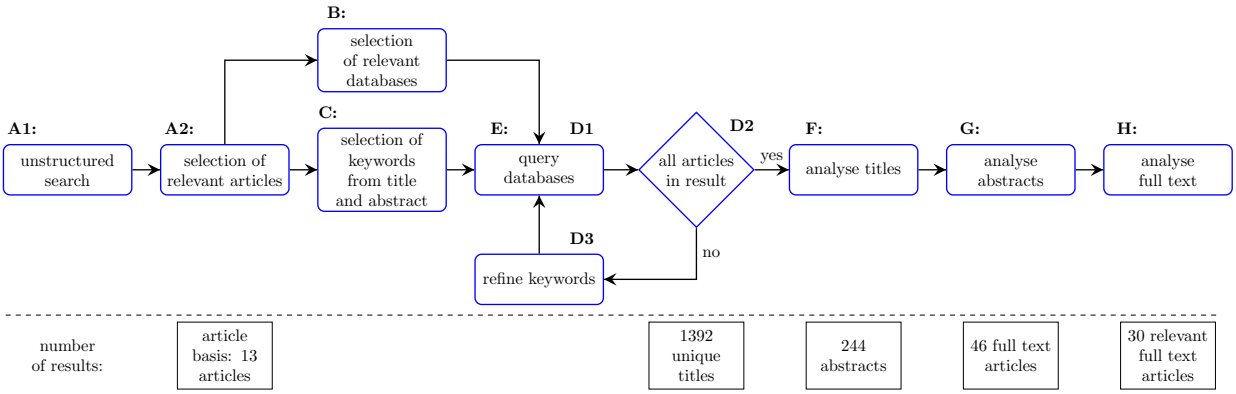

Figure 1: Workflow of the systematic review process. Blue nodes with rounded corners represent the basic steps in the workflow. The diamond node denotes a step including a decision with more than one possible outcome. Below the dashed line the resulting article sets of the corresponding steps above are given.

All remaining databases did not contain any of the base set articles. With the selected 4 databases we could extract 12 out of 13 articles, which is the best coverage achievable on the considered databases. The remaining article was only available from semantic scholar and from the publishing journal.

1.2 Formalization of Keywords

The set of initial articles deals with different research questions but generally follow the topic: "the extraction of software and data usage statements". Our goal is to identify the keywords and phrases used to describe this topic and formalize them (see Figure 1 **C**). Purpose of creating the formalized description of the topic was to systematically match further articles in the selected databases. The formalization of choice is a database query which can be used to retrieve literature from scientific databases. Since the topic of an article is summarized in its title and abstract, we only considered those parts to create the query. During the initial analysis of the base set we identified four parts for the potential query:

Resource Type: scientific resource investigated in the work (e.g. software or data)
Reference: term used to state the usage (e.g. reference or citation)
Information Extraction: term used to describe the extracting of information from an article (e.g. identification or extraction)
Source: text source used for the analysis (e.g. publication or literature)

Table 1 contains the particular terms that were identified. The terms were selected over multiple iterations in which we applied the query on the selected databases and tested the coverage (see Figure 1 cycle **D1**–**D3**).

The extracted keywords are connected through the logic operators *AND* and *OR* to form the query. As it is not of interest which particular keyword of a category appears, the rows of each column were connected by *OR*. The columns were connected by *AND* because each of the four components had to appear once. In the source component a nested logic expression was used between the two sub-columns that was connected in the same manner.

The query further distinguishes whether a keyword appears in the *title* (T) or in *title and abstract* (TA) (see Table 1). This distinction mainly influences the extraction target, which we found to be commonly mentioned in the titles of the base set articles. All other parts of the query are searched in title and abstract. The resource name "scientific software" is searched in title and abstract, because some work focusses on specific scientific software which is only mentioned by name in the

Table 1: Explicit keyword for all components of the query. (T) states that the term is searched in the title only, (TA) keywords are searched in title and abstract.

Extraction Target	Reference	Information Extraction	Source	
data (T)	citation (TA)	identify (TA)	publication (TA)	
dataset (T)	reference (TA)	extraction (TA)	full text (TA)	
software (T)	statement (TA)	detecting (TA)	academic (TA)	text (TA)
scientific software (TA)	mentioned (TA)	content analysis (TA)	research (TA)	document (TA)
	mentions (TA)	examining (TA)	scientific (TA)	corpus (TA)
	mentioning (TA)	tracking (TA)	journal (TA)	article (TA)
	association (TA)		published (TA)	literature (TA)
				paper (TA)

title. Therefore, we broadened the search to the abstract in this case. To reduce the number of overall retrieved articles we restricted the scope of the search to articles with a topic of computer science and main language English.

The query was further adjusted to the individual search characteristics of the selected databases (see Section 1.1). Varying features we considered are: • plural extension, • stemming and automatic extension of words, • hyphen replacement (full-text (TA)), • precedence of logic operators and • available search fields. The main intention was to keep the query's function consistent across all databases it is applied on.

1.3 Querying literature databases

To retrieve the results we applied the adjusted query onto all 4 selected databases (see Figure 1 **E**). The articles were either retrieved by a custom python script over provided API functions or retrieved manually. The results were summarized into a CSV file and duplicates were eliminated based on DOIs and titles. For the purpose of query construction we concentrated on the DOI and the title in order to test the coverage. For the final results we extracted • DOI, • title, • abstract and • authors. The execution of the final query (see Table 1) in all databases resulted in 1392 unique and potentially relevant candidate articles. In the following, the selection of literature of actual interest is described.

2 Selection of appropriate Literature

2.1 Title analysis

Because the number of articles (n=1392) retrieved in the previous step was to high for a manual full text examination, it was reduced through a title analysis (see Figure 1 **F**). All titles were manually analysed by both authors and explicitly removed from the set if they were determined as not relevant with the objective to only exclude articles clearly dealing with a different topic. Inter-rater agreement for this task was assessed for both authors for a sample of 20 articles, for 17 of which they agreed on. For the remaining three articles the issues could be resolved and one of the authors coded the entire set of articles. The set of candidates was reduced to 244 articles in this step.

2.2 Abstract analysis

To further reduce the number of articles (244), the topics of the articles were analysed by reading their abstracts (see Figure 1 **G**). For this task the inter-rater agreement was tested on a set of 7

abstracts for which the authors agreed on all. One author then examined all abstracts. This step reduced the article set to 46 candidates.

2.3 Full text analysis

For all remaining 46 articles we tried to retrieved or requested access to full text versions. The full text versions were read by both authors and unsuited articles removed under agreement (see Figure 1 **H**). After this step 30 articles remained that were determined as relevant to the topic. In this step 4 articles were added to the set, which were referenced in the relevant literature, but not located through the systematic review. The articles could not be localized through the review, because they were not tagged with the 'computer science' research tag. We decided to retain the tag, because the number of articles would have increased considerably by removing it. The final set of articles was then carefully studied and analysed with respect to the employed technical approach, the dataset and the target research domain. The results of this analysis are presented in the following.

References

[1] Barbara Kitchenham. Procedures for performing systematic reviews. 33, 08 2004.

[2] Frank Krüger and David Schindler. Data for: A systematic literature review on the extraction of usage statements of scientific artefacts, May 2019.