

Literature Search Strategy

In this document, we outline the steps we took to conduct our literature search. The search is conducted last on 28th October 2021 and included several steps.

Step 1. Determine the Search Topics and Synonyms / Related Keywords.

The study focuses on the decomposition of monolithic applications into microservices. Based on the research questions (see Table 1) and the area we selected, we identified three main topics of the research.

Table 1. Research Questions

No	Research Question
RQ1	What are the primary phases of monolith-to-microservices migration and the major constituent elements of those phases?
RQ2	Within that framework, what are the existing approaches, tools and methods observed in the decomposition of monolithic applications to microservices?
RQ3	What are the metrics, datasets, and benchmarks used for evaluating and validating the microservice extraction process and its results?
RQ4	What research gaps can be identified in the current literature?

Each topic was represented by a keyword and further expanded to include synonyms and related terms used in the field. Following the approach discussed in [1, 2], we identified the following three topics (see Table 2) and defined the keywords along with their synonyms and related terms.

Table 2. Topics, Keywords, and Synonyms/Related Terms

Topic	Keyword	Synonym/ Related Term
Topic 1	Monolith	monolith, existing, and legacy
Topic 2	Microservice	microservice and micro-service
Topic 3	Decomposition	decomposition, migration, identification, extraction, refactoring, modularisation, transformation, transition, and conversion

Preparing the Search Key

Then, we combined the keywords to construct our search key. We combined the three topics using Boolean “AND” and the synonyms under each topic using Boolean “OR” operators. We further use the wild card “*” to extract variants of the keywords. This is to ensure the coverage of studies by using the stem of the keywords. The final search string is represented as:

Search Key	(monolith* OR exist* OR legacy) AND (microservice* OR micro-service*) AND (decompos* OR migrat* OR identif* OR extract* OR refactor* OR modular* OR transform* OR transit* OR conver*)
------------	--

Step 2: Decide the Search Platforms

We included all the established publication platforms to gain as many related studies as possible. In addition to that, we also used Google Scholar and Publish or Perish¹ to harvest recent publications and much data as possible. We conducted the search on all available metadata on the platforms (refer Table 3).

Table 3. Search Platforms.

Search Platform
SpringerLink
IEEE Xplore
ACM Digital Library
Science Direct
Wiley Online
Scopus
Google Scholar
Publish OR Perish (PoP)

Step 3: Customise the Search String for the Platforms.

Some platforms do not support more than eight keywords when combined using the Boolean operators for an advanced search. Thus, we decomposed the final search key into three subsets of search strings (refer to Table 4).

Table 4. Search Keys

No	Search Key
Search Key 1	(monolith* OR existing OR legacy) AND (microservice* OR micro-service*) AND (decompos* OR migrat* OR identif*)
Search key 2	(monolith* OR existing OR legacy) AND (microservice* OR micro-service*) AND (extract* OR refactor* OR modular*)
Search Key 3	(monolith* OR existing OR legacy) AND (microservice* OR micro-service*) AND (transform* OR transit* OR conver*)

These three search keys are executed individually, and results are combined by removing duplicate entries. A record is considered duplicate if the record contains

- The same authors AND
- The same publication Year AND
- The same Title.

Step 4: Conduct Initial Search

We executed the above three search keys consecutively on the selected platforms step by step. We captured all the data that is returned in response to the search activity. A summary of the results is presented in Table 5. These results reflect the search hit when the above three search keys are executed, and the results combined. This includes duplicate publications found on different platforms

¹ <https://harzing.com/resources/publish-or-perish>

and incomplete data. From the 5022 search results, we removed duplicate studies, incomplete data, books, websites, and reports which resulted in 968 scientific publications.

Table 5. Summary of the Initial Search Result.

Platform	Initial Hit
SpringerLink	2247
EEE Explore	267
ACM Digital Library	1191
Science Direct	367
Wiley Online	461
Scopus	489
Total	5022

Refining the Search.

We followed several steps to refine the search. All the steps and the resulting number of studies qualifying for the next round are summarised below. The input for the refinement is the 968 papers.

The following inclusion and exclusion criteria (See Table 6) are prepared to refine the search results.

Table 6. Inclusion and Exclusion Criteria.

Inclusion Criteria	IC1: the primary objective of the studies should be the decomposition of monolithic applications into microservices,
	IC2: the studies should include structured and preferably automatic or semi-automatic decomposition approaches.
	IC3: the study should sufficiently describe the decomposition method, code, algorithm, and its evaluation (i.e., abstracts and extended abstracts are not included).
Exclusion Criteria	EC1: Duplicate Studies
	EC2: Books and Patents
	EC3: Not peer-reviewed Studies.
	EC4: Secondary Studies
	EC5: Studies written in languages other than English.
	EC6: Papers published before 2015.

The data associated with these different steps are included in the *Literature_Search_Data* folder in the CSV files identified in the "Associated File" column of Table 7. Finally, we applied a snowballing process to further extract additional papers that we might have missed during the search or those papers that are not accessible. The snowballing process is discussed in detail in the next section.

Table 7. Refinement Steps and Results.

Step	Refinement Strategy	Metadata Fields	Studies	Associated File
1	Data Cleaning	All	968	<i>1st_Round_Filtering_968.csv</i>
2	Apply Exclusion criteria.	Title	126	<i>2nd_Round_Filtering_126.csv</i>
3	Apply Inclusion and Exclusion criteria.	Title, Abstract, and Conclusion	92	<i>3rd_Round_Filtering_92.csv</i>
4	Apply Inclusion and Exclusion criteria.	Full Text	33	<i>4th_Round_Filtering_33.csv</i>

Snowballing Strategy.

We have applied both backward snowballing and forward snowballing methods to extract additional references. We extracted the references from the pdf of the papers for backward snowballing. For forward snowballing citations of the 33 papers are reviewed after we did the backward snowballing. We followed the same procedure for the forward snowballing process. Because of the similarities between the backward snowballing and forward snowballing approaches, only one (the backward snowballing approach) is discussed in detail.

Step 1: Backward Snowballing.

This approach uses the following steps to conduct the backward snowballing.

We extracted all the references from the 33 selected papers. The automatic reference extractor tool we used is Scholarcy API². The tool extracts all the references except a few exceptions that introduced some noisy data. We manually fixed those errors and made them suitable for the analysis. Note that we are mostly interested in the title, authors and year. The 941 papers extracted following the backward snowballing approach are included in the Snowballing_data folder in the CSV files identified in the associated file column of Table 8.

Table 8. Snowballing Step 1.

Papers included	References Extracted	Metadata Fields	Associated File
33	941	Item Type, Author, Title, Publication Year, ...	<i>Snowballing_step_1.csv</i>
33	212	Item Type, Author, Title, Publication Year, ...	<i>Snowballing_step_2.csv</i>
33	2	Item Type, Author, Title, Publication Year, Found During Initial Search, Exclusion Criteria, Unsatisfied Inclusion Criteria, Comment	<i>Snowballing_step_3.csv</i>

We cleaned the data by fixing errors in the author, year and title fields. Further, we removed duplicate references (based on title, author, year combinations and title and author as well as title and year combinations). We further removed blank titles, years, and author records. We also removed any book, literature review, systematic mapping study etc. Finally, we extracted 212 unique references, as documented in the second row of Table 8.

Then, we compared the snowballing result with the search result and removed the publications that are already included in the search. Then we applied title, abstract, and full paper revision using inclusion and exclusion criteria. This process is documented in the third row of Table 8 under Snowballing_step_3.csv.

² <https://ref.scholarcy.com/api/>

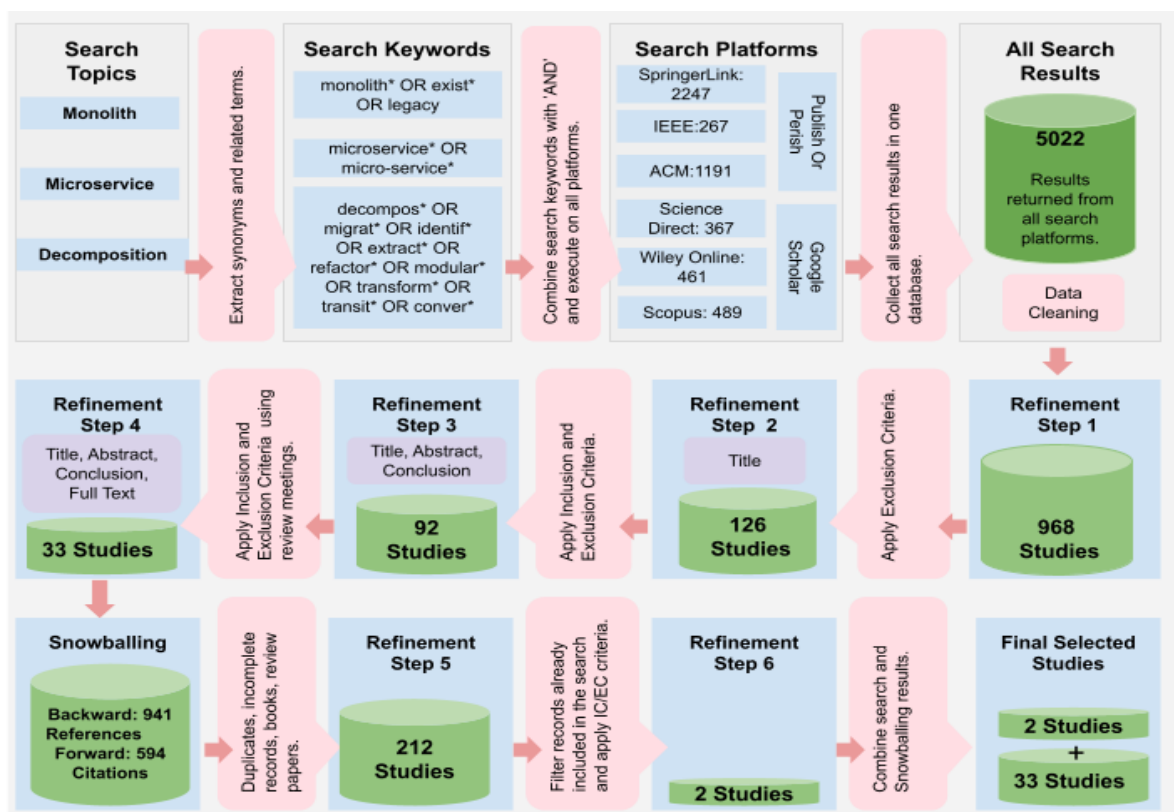
Finally, the two papers that were not found in the initial search, that were not excluded by the exclusion criteria and that satisfied the inclusion criteria are included. The two papers included are listed in the file *Snowballing_step_3.csv*.

Step 2: Forward Snowballing.

The forward snowballing process followed a similar approach. We collected 594 citations from the 33 selected papers. We followed a similar procedure of filtering the citing papers. However, this resulted in no additional papers except those that are already included in the backward snowballing process.

Step 3: Final Result.

The final result combines the forward and backward snowballing process. Finally, we extracted the final two papers. The majority of research papers extracted during the snowballing process are also identified during the initial search (Table 8). A diagram summarising all the steps is presented in Figure 1. below.



M2MF Data Collection and Analysis

We extracted the data required to formulate the M2MF and answer RQ1 directly from the full text of the papers. The data was analysed using a light form of Content Analysis, or by following several

steps from Ground Theory (GT), dependent upon the contents of the paper (see below). We extracted the monolith-to-miscoservices migration phases and sub-categories from the introduction, methods, experiment, and evaluation sections of the papers.

Initially, the first author extracted the raw data from textual descriptions of the method, the frameworks and diagrams illustrating the steps, experimental setups, and evaluation procedures. Next, if phases or sub-categories were explicitly identified in the paper they were extracted using a light form of Content Analysis, and repeatedly applying constant comparison with previously identified phases and categories. That is, at the end of each full-text review, the extracted information was analysed, compared and assigned a category if the description fitted with previously identified phases and sub-categories; otherwise a new phase/sub-category was assigned. In studies where the information was not explicitly presented in the paper, we synthesised the appropriate phases/sub-categories, based on the GT components of memoing, open-coding, constant comparison and theoretical sorting, the latter employing an analysis lens of phases and approaches-within-phases, in the organisational spirit of [3]’s approach to GT. Again, constant comparison was employed after each paper was reviewed. Several weekly review meetings in 2020 and 2021 were held to present the results of the analysis at different times led by the first author and involving the third, fifth, sixth and seventh authors. A summary of the data that was extracted from the papers and the resulting encodings are also included in the DataExtraction.xlsx file under the *SLR_GT_Data* folder of the replication package.

Detailed Extraction and Analysis of the Literature

The rest of the data used to answer RQ2, RQ3, and RQ4 is extracted by the first and the fifth authors jointly, in accordance with a template (DataExtractionTemplate.xltx) structured to capture all the relevant data for those questions. This template was based, in part, on the emerging findings from the M2MF framework analysis. After the application of this template to individual sets of papers, several literature review meetings were held, presenting the results of the data extraction process for discussion and refinement. Additionally, the third, sixth, and seventh authors critically evaluated the presented results to review the relevance of the analysis and to evaluate the accuracy of the extraction, conceptualisation, and representation of the extracted data. The full-text versions of all the studies were shared among all authors enabling them to verify the accuracy of the data extraction process.

Using the data collected from the data extraction stage, the data synthesis stage focused on comparing, organising and re-organising the extracted data in a format that answers the research questions. It is important to note that this process was also revisited by the groups, every time a paper was reviewed. After several revisions conducted during the literature review presentation meetings, we reached the final analysis result. This final result is included in the folder under *SLR_GT_Data* folder.

Reference

- [1] B. Kitchenham, 'Procedures for Performing Systematic Reviews', *Keele, UK, Keele Univ.*, vol. 33, Aug. 2004.
- [2] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, 'Lessons from applying the systematic literature review process within the software engineering domain', *Journal of Systems and Software*, vol. 80, no. 4, pp. 571–583, Apr. 2007, doi: 10.1016/j.jss.2006.07.009.
- [3] A. Strauss and J. Corbin, *Basics of qualitative research*. Sage publications, 1990.