Data for “Composition of Heterogeneous Web Services: A Systematic Review”

Alexis Huf and Frank Siqueira  
Graduate Program in Computer Science  
Federal University of Santa Catarina  
alexis.huf@posgrad.ufsc.br, frank.siqueira@ufsc.br  
January 3, 2019

1 Introduction

This document describes the data files for a Systematic Literature Review (SLR) on heterogeneous web service composition in detail. For context, motivation, goals and a general overview of the review protocol (in addition to results), see the corresponding publication, linked in the mendeley data page. To avoid copyright issues, this document mostly discusses details omitted from the main publication, assuming its readers have read the associated publication.

This data repository contains has the following structure:

- works.csv: Result file of the automated selection phases. Includes all works that remained after stage 2 (Section 2). Documents excluded by stage 3 are included, but have are marked on the relevant columns for the exclusion criteria they met;
- manual.ods: A LibreOffice spreadsheet with the results of manual selection (stages 4 and 5, see Section 2) and of data extraction. Includes also a sheet with the validation runs of stages 4 and 5 and an analysis sheet;
- manual.csv: The main sheet from manual.ods, containing selection decision and extracted data, in a machine readable format;
  - *.csv,bib: Result files. IEEE results are saved as Comma Separated Values (CSVs). Scopus results are saved as BibTeX files with abstracts and keywords included;
  - *.qry: Actual query submitted to obtain the corresponding results.
- rest-terms: Queries and results that measure occurrences of RESTful web services and variants against RESTful services and RESTful web APIs;
- Makefile: Executes the automated selection procedure;
- slrpk: Sources and binaries of a software used to filter results;

1https://github.com/alexishuf/slrpk
expressions: Expression files interpreted by slrpk
stage: Intermediary results of the selection pipeline;

This document is structured as follows. Section 2 gives a brief overview on the selection process. Section 3 describes the contents of manual.ods in detail. Subsection 3.1 has instructions on how to execute the automated portion of document selection.

2 Selection procedure overview

The search strategy had two sources: IEEE Xplore\(^2\) and Scopus\(^3\). Search strings used can be found under \texttt{qry-2018-07-22}, with .qry file extensions. For Scopus, the “Advanced” query interface must be used. For IEEE, the “Command Search” must be used, with care to not leave blank lines at the end of the query submission form.

Scopus queries are the authoritative queries, derived from the research question and inclusion criteria discussed in the associated publication. Two queries are submitted to Scopus:

1. Types query: Searches for two or more service types, and composition-related keywords. Identified in data files by the column \texttt{qry-n};
2. Heterogeneity query: Searches for heterogeneity of interaction paradigm, communication protocol and other variations. Identified in data files by the column \texttt{qry-h}.

For IEEE Xplore, limitations on the search interface require the first search string to be divided into three queries. These sub-queries must also be relaxed, due to a limitation on the number of keywords in a single query. Therefore, these queries are less selective than their Scopus counterpart.

The selection process had 5 stages, of which the first 3 are automated:

1. Revert query relaxation applied to IEEE Xplore;
2. Require For each query type, require the intervention and population keywords to occur on the same sentence;
3. Apply automated exclusion criteria;
4. Manually apply exclusion criteria to Title, Abstract or Keyword (TAK) only;
5. Manually apply exclusion criteria to full-text.

There was a validation run of stages 4 and 5. For this, the reviewers had only Title, Abstract and Keywords visible (i.e., no previous decision status was shown). File \texttt{manual.ods} has a separate sheet for the validation run. The errors identified were fixed in the main sheet, and are only visible within the validation sheet itself. The criteria that justifies exclusions is also on this validation sheet.

\(^2\)https://ieeexplore.ieee.org/
\(^3\)https://scopus.com
<table>
<thead>
<tr>
<th>Query Date</th>
<th>Documents in Stage 3</th>
<th>Stage 3 errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>November 19, 2017</td>
<td>473</td>
<td>0 (random sample of 100)</td>
</tr>
<tr>
<td>July 7, 2018</td>
<td>524</td>
<td>3 (full validation)</td>
</tr>
<tr>
<td>July 22, 2018</td>
<td>546</td>
<td>+1</td>
</tr>
</tbody>
</table>

Table 1: Previous runs of the SLR protocol

2.1 Multiple runs of the protocol

The tooling used on this SLR allows incremental updates (see Subsection 3.1). In addition to pilot runs of the protocol, there were three major runs of the protocol, with minor adjustments, that led to this dataset. Table 1 shows the dates in which the sources were queried for each of these three runs of the protocol. The validation runs of stages 4 and 5 was only executed no the last two runs of the protocol, totaling 4 false exclusion errors on stage 4.

3 Documents Spreadsheet

The manual.ods contains the following sheets:

- **analysis**: Summary information and general counts over the extracted data;
- **manual**: Selection decisions and extracted data;
- **abs_reeval_full**: Full re-evaluation of stages 4 and 5;
- **ex_crit**: Exclusion criteria code names used in abs_reeval_full;
- **works**: This sheet is linked to the works.csv file
- other helper sheets, used by formulas

The works sheet contains the results of stage 3, produced by the Makefile. The columns can be summarized as follows:

- Columns converted from Scopus BibTeX or IEEE CSV;
- **qry-n**: 1 if the document was found by the queries that test service type combinations
- **qry-h**: 1 if the document was found by the query that tests heterogeneity claims
- **qry-ieee-**: 1 if the document was found by the corresponding IEEE query;
- **tiabskw**: Concatenation of title, abstract and keywords, used by some expressions during automated selection;
- **nprv-book, nprv-booklet, nprv-manual, nprv-techreport, nprv-thesis**: 1 if the document matched some of these BibTeX types;
- **nprv-editorial, nprv-tutorial, nprv-keynote**: 1 if the Title contains these words
- **tp-proc, tp-survey**: 1 of the document has proceedings or survey keywords in the title;
• **tak-event-oos**: 1 if the document TAK only has bogus occurrences of event keywords and does not satisfy the automated inclusion criteria without the event-related keyword;

• **tak-tps-oos**: 1 if the document was found by a query that tested service types co-occurrence, but the co-occurrence was not found.

The specific rules on how the above columns are computed can be found in the Makefile, in the recipes for the targets under `auto-select/`. The last two columns are computed using expression files in the `expressions/` directory.

The **manual** sheet contains all selection information and extracted data. In addition to some columns equivalent to the above, but manually filled, the selection-related columns are:

• **excluded**: Automatic exclusion in stage 3 (columns above);

• **manual-included, manual-excluded**: Decision for stage 4;

• **cnt-excluded**: 1 if excluded by full-text (stage 5);

• **cnt-na, cnt-ext, cnt-oos**: reason for exclusion in stage 5: not available, extension selected or out of scope;

For data extraction, there are some columns where multiple values in a single cell are expected. In these cases, values are separated by semicolons (;). When categorizations are expected, additional details for a particular classification are provided inside parenthesis (e.g., “eos(Feeds)”). The data primary data columns, related to the document text, to composition and to the heterogeneity support method are the following:

• **comp-term-artifact**: The term composition is predominantly used to mean an action or an artifact;

• **comp-term-both**: The term composition is used with both meanings;

• **comp-executes**: Whether the document (implicit) position is that composition includes execution of the composite service;

• **rest-is-web-service**: Whether RESTful services are considered web services;

• **supp**: Supported service types, semi-colon separated;

• **auto-synt**: Whether a composite service specification is generated by an algorithm;

• **dyn-func-comp**: Whether the previous occurs during execution;

• **dyn-proc**: Whether the composite service changes during execution (this may occur in some event-oriented composition engines and is always the case for `dyn-func-comp`);

• **comp-dyn-binding**: Whether service selection occurs during execution;

• **rmm**: Richardson Maturity Model (RMM) [Webber et al., 2010, p. 18] maturity level;

• **hateoas**: Reason for failure in supporting Hypermedia As The Engine Of Application State (HATEOAS);

• **ct-description, ct-language, ct-composition**: Whether the document proposes a service description, a composite service specification language, or a service composition method/approach/algorithm;
• **comp-style**: Composition style;
• **category** The categories of heterogeneity support methods used in the document (see the published article);
• **comment** A summary of the articles contribution, of heterogeneity support methods used, and any justifications about the value of other columns. For documents excluded by full-text, this contains a reason for the exclusion.

With respect to evaluation quality and study quality, the extracted data columns are:

• **heter-eval**: The evaluation methods used, according to Hevner et al. [2004] and separated by semicolons;
• **heter-eval-quality**: Quality of evaluation, graded as shallow, mostly_credible and credible;
• **perf-eval**: Evaluation method used for evaluating performance;
• **comment-eval**: Summary of evaluation and any justification deemed useful by the reviewer;
• **issn**: Internation Standard Serial Number (ISSN), for journal publications only;
• **year**: Year of publication;
• **sjr-**: SCImago Journal Rank for the given year (obtained from Scopus);
• **snip-**: Source Normalized Impact per Paper for the given year (obtained from Scopus);
• **c/d(2yrs)-**: Citations per document for the two years before the given year (obtained from SCImago);
• ***-pub**: The metric at the year of publication;
• **h5-index, h5-median**: H-Index for publications in the last 5 years and the median of citations per documents that compose the H-Index;
• **qualis-2016**: Venue classification in the QUALIS for computer science, a ranking of conferences and journals maintained by CAPES, a Brazilian funding agency;
• **cit-scopus**: Citations identified by Scopus;
• **cit-scholar**: Citations identified by Google scholar;
• **comp-style**: Composition styles, separated by “;” that the document adopts or assumes;
• **application-field**: Application field that the document targets, this is document-wise;
• **case-field**: Application field used only for a particular use case or scenario;
• **publisher**: Which company published the document.

The columns in abs_reeval_full have the following meanings:

• **In, Ex, Ex crit**: inclusion/exclusion decision and exclusion criteria code name;
• **amb-in**: 1 if the abstract is ambiguous or suspicious, and the inclusion decision was done to clear the doubt;
• **n-abs-ex**: 1 if the document was not excluded in stage 4 and 5;
3.1 Running the automated selection

The automated stages of selection can be executed using the Makefile provided. The software requirements are the following:

- make (tested on GNU Make) and coreutils;
- sed (tested on GNU sed 4.5);
- JDK 8
- Maven 3+ (tested on 3.5)

Go into a local copy of the data repository and run `make`. Based on timestamps, required targets will be rebuilt, up to the `works.csv` file. To force a rebuild, remove intermediary files (`rm stage/*`) re-run make. The Makefile reuses the `works.csv` file to maintain ids consistent. Therefore, removing `works.csv` file will cause already assigned IDS to change, leading to inconsistencies in several sheets of `manual.ods`.

In order to download new result files, create a new `qry-$DATE` directory, copying the `.qry` and `.sh` files from the current `qry-` directory into the new one. Change the queries, if needed, and submit them to IEEE Xplore “Command Search” and to Scopus “Advanced Search”. For IEEE, download CSV results, the `ieee-hdr-swap.sh` file performs transformations on these files to make them compatible with the rest of the process and will be automatically called by the Makefile. For Scopus, A BibTex export with Abstract and keyword data must be done. To use the new set of results, change the variable `BIBSDIR` in the Makefile or override it from the command-line.

The `manual.ods` file must be manually updated after changes in `works.csv`. The `manual` sheet must be reviewed for documents that have gone missing from `works.csv` (column `has-id`). New documents can be found in the `missing` column of `works_helper`. These must have their ids copied into `manual`, and the columns must be filled, either using formula auto-complete, or manually (according to the SLR protocol).

References

References