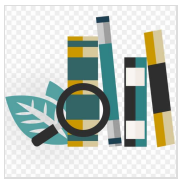


Feb 19, 2022

Systematic Literature Review about Software for References Extraction

Alessia Cioffi[1]

[1]University of Bologna

dx.doi.org/10.17504/protocols.io.buz9nx96

Alessia Cioffi

Converting unstructured data, i.e. data coded in a format which is not structured in a predefined way, such as PDF, into structured data, i.e. clearly defined types of data organised in a structure, has several advantages. One of the most positive effects of this conversion is that data becomes easier to search, both for humans and for algorithms. Even if there are many tools which have this objective, through a systematic review of the existing literature it is possible to understand whether there is a software whose features allow it to have better performances than the others in order to carry out a specific task in this context. This protocol shows the methodology followed in order to make a systematic review of the literature regarding the software dedicated to the extraction and manipulation of references from papers in PDF file format. Thus, the objective of this research, which is reflected on the flow of the literature review methodology, is to retrieve the most suitable software for the specified purpose, i.e.**retrieving and manipulating citations from PDF files**.

DOI

dx.doi.org/10.17504/protocols.io.buz9nx96

Alessia Cioffi 2022. Systematic Literature Review about Software for References Extraction. **protocols.io** https://dx.doi.org/10.17504/protocols.io.buz9nx96

protocol ,

May 14, 2021

Feb 19, 2022

49953

The following table synthetically shows the contents of the protocol:

Step	Explanation	output
Create the materials for the research	In this step we create the materials that will be used in order to transcribe the essential information of the paper which we will accept for the literature review and the ones which we will not consider.	“Papers_to_keep”, “Papers_to_discard”, “Papers_to_analyze”
Search with seed papers	In this step, after checking the validity of the seed papers, we search new literature through citations. We look for the papers which are cited by the seed papers and to the papers that cite the seed papers.	“Papers_to_keep”, “Papers_to_discard”, “Papers_to_analyze”
Search with keywords	The search is carried out with keywords, some selected before the beginning the research and others eventually required during the research itself. The research with keywords is performed through openly available research systems.	“Papers_to_keep”, “Papers_to_discard”
Quality assessment-full text reading	In this final step the full-text of the papers is read and it is decided whether to keep them in the review or not, on the basis of some specified criteria. So, first of all, two files are created, one for the papers which pass also the final check, "Papers", and another for listing the retrieved softwares called "Software".	“Papers_to_keep”, “Papers”, “Software”

Apart from the table reassuming the steps of the protocol, it is attached a workflow representing graphically the sequence of the steps and substeps reported in the protocol.

protocol_flowchart.pdf

Seed papers list.

The list of seed papers has been provided before the beginig of the research. Whether other researchers have another list of seed papers, they can use it instead of the one provided in this protocol.

[Waleed et al. Construction of the Literature Graph in Semantic Scholar,
Lopez P. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications,
Romary L. & Lopez P. GROBID - Information Extraction from Scientific Publications,
Körner M. et al. Evaluating Reference String Extraction Using Line-Based Conditional Random Fields: A Case Study with German Language Publications,
Hosseini A. et al. EXCITE - A toolchain to extract match and publish open literature references,
Boukhers Z. et al. An End-to-end Approach for Extracting and Segmenting High-Variance References from PDF Documents,
Ghavimi B. et al. An Evaluation of the Effect of Reference Strings and Segmentation on Citation Matching].

Keywords lists.

This section is dedicated to the lists of keywords that will be used during the research with keywords step. The lists should be provided before the beginning of the research and there should be a list for all the languages selected in the language parameter in the section ***General IC for papers*** in the Inclusion Criteria section. In my case there are two lists, one in english and one in italian. Whether other researchers want to perform their research in other languages they can update/change them.

- *English*: [“PDF extractor”, “PDF extractor software”, "software for data extraction from PDF", “extraction of information from PDF”, "retrieve information from PDF", "extraction of non structured data", "Grobid", "Excite"];
- *Italian*: ["estrazione dati da file PDF", "estrazione di dati non strutturati", "estrazione PDF", "software estrazione dati da PDF", "Grobid", "Excite"].

Free open platforms.

Citation: Alessia Cioffi Systematic Literature Review about Software for References Extraction
https://dx.doi.org/10.17504/protocols.io.buz9nx96

These platforms have been selected in order to make the research with keywords. The paremater which led to the selection of these specific platforms is that of the openness. If other researchers want to use other platforms they can add them to the list reported here above, or change it.

- Google Scholar,
- Lens,
- Microsoft Academic Graph,
- EBSCO,
- ProQuest,
- IEEE Xplore,
- Open citations,
- ACM Digital Library.

Platforms for forward search.

The forward search, the research performed on the articles which cite the article taken into consideration, can be done manually and with the help of some platforms which have that specific feature. The ones selected are open and free.

- Google Scholar,
- ISI Citation Index,
- Lens,
- Microsoft Academic Graph,
- OpenCitations.

Operating systems considered.

The list of the operating systems that will be used to test the softwares, detected through the literature review, is essential in order to classify the softwares as "valid" or "invalid". Indeed only the softwares which work on at least one of the selected operating systems will be accepted as "valid", and, therefore, the respective papers will be kept for the literature review.

macOS (the one used with preference if the available software is compatible);
Windows;
Linux.

Inclusion Criteria

Before starting the systematic review process, *Inclusion Criteria (IC)* have to be defined. The IC are the principles guiding the literature review process. Indeed, thy provide parameters on the basis of which it must be decided whether to keep or not the papers retrieved during the systematic literature review. Three different levels of criteria for inclusion have been defined. Each of them is used in a different step of the review process, from the first, more genrical, selection to the final, and more selective, one. The number of steps is not mandatory, and can be changed on the basis of the number of levels of analysis that are necessary in order to carry out different kinds of researches. For each of the three IC it has been presented an example extracted from one of the seed papers in **Seed papers list,** "Construction of the Literature Graph in Semantic Scholar", (Waleed et al. 2018)

1. General IC for papers. These inclusion criteria represent the first step to select the papers that will be considered in the systematic review. All of them are mandatory since they are very concrete and essential requirements for the papers to be selected. The ones I have identified for my review are the language and the publication date. The languages selected are the ones through which the researcher can have a complete understanding of the text, i.e. Italian and English; instead, since this kind of technology evolves rapidly, in the order of years, a limit for the publication date is required not to select too old papers which could convey very few or aged information with respect to researches made in the last 16 years. Whether other researchers require different criteria (e.g. version of the document) or other specifications (e.g. languages: en, fra), she is free to change them. The *General IC for papers* can be synthetised with:

Citation: Alessia Cioffi Systematic Literature Review about Software for References Extraction
https://dx.doi.org/10.17504/protocols.io.buz9nx96

- **languages**: en, it;
- **publication date**: from 2005 on.

An example regarding the date.

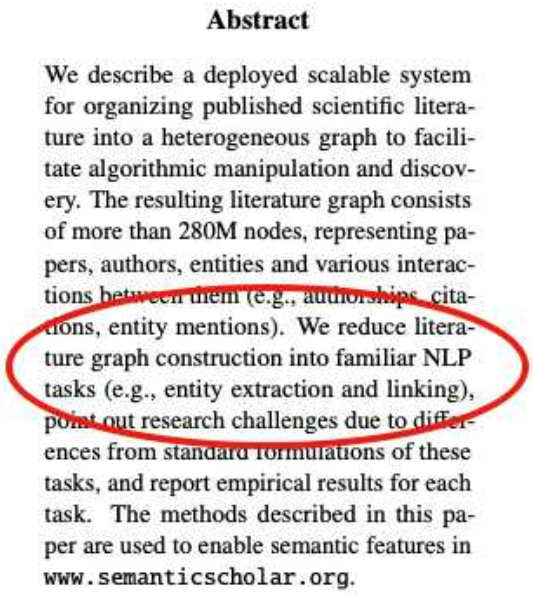
2. IC for screening procedure on papers. These criteria allow, when having a first look at the papers retrieved, to understand whether the topic outlined by the highlights of the papers is in line with the topic of the research. These highlights are are identified with the paper metadata like title, abstract and keywords. The title of a paper turns out to be really helpful in case the article is centered on one of the sub-topics of the research topic and therefore, since the aim of a title is presenting with a phrase the entire topic of a paper, that aspect should be present in the title itself. The keywords usually precede the abstract and, again, they have the objective of presenting the topics of the articles with single words, and also in this case it can be helpful if the article is centered on a topic regarding the research question. Finally, since the abstract reports in a short way the contents of a paper, the actual topics it is about are clearer than in the title and in the keywords. Also, since some sub-topic may be present in an article not being relevant for the discourse, it can happen that those contents are revealed in the abstract. **Whether other researchers want to use different parameters from the one presented in this section, they can.** Differently from the *General IC ,* one of these criteria is enough to accept the selected paper. However, the more criteria met, the surer it is that the paper's topic is correlated to the research one. This part should be still quite inclusive, and if the researcher is in doubt, she should keep the article (an even more precise investigation will be carried out by the full text reading).
Here the IC used in this research are reported. Nonetheless, whether different parameters are required, they can be modified on the basis of the specific needs:

Title (it should include a precise word or concept relating to the topic of the research):
1. Includes the name of a software for data extraction from PDF files;
2. Contains a procedure for data extraction for PDF;
3. Includes the sequence of the word 'PDF' and a noun/verb/adjective derived from 'extract'.

Keywords (one or more keywords should together compose the concept of the research question or similar ones):
1. A keyword expressing the concept of extraction of data from PDF file formats e.g. "PDF extraction";
2. Two or more keywords, one including the word "PDF" and the other expressing the concept of extraction e.g. "data extraction" + "PDF" or "data extractor" + "software" + "PDF".

Abstract (it should be present a concept or procedure similar or identical to the one required by the research topic):
1. Contains the concept of "data extraction (from PDF files)", also expressed in a different way (e.g. "This paper is concerned with the extrapolation of the information contained in the titles of files in PDF format.");
2. Includes the name of a software for data extraction from PDF files.

Citation: Alessia Cioffi Systematic Literature Review about Software for References Extraction
https://dx.doi.org/10.17504/protocols.io.buz9nx96



Abstract

We describe a deployed scalable system for organizing published scientific literature into a heterogeneous graph to facilitate algorithmic manipulation and discovery. The resulting literature graph consists of more than 280M nodes, representing papers, authors, entities and various interactions between them (e.g., authorships, citations, entity mentions). We reduce literature graph construction into familiar NLP tasks (e.g., entity extraction and linking), point out research challenges due to differences from standard formulations of these tasks, and report empirical results for each task. The methods described in this paper are used to enable semantic features in `www.semanticscholar.org`.

In this case it is not directly said that the extraction regards PDF, but if in doubt the paper should be kept for the full-text reading.

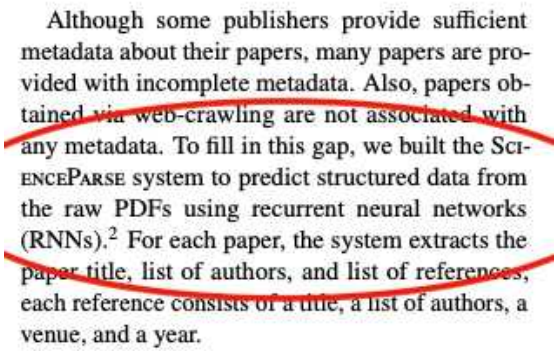
3. IC for the papers full text. These are the last criteria to apply in order to accept or reject the papers taken into consideration. Indeed these parameters regard the content of the papers full text. In this case two different parameter classes were needed, one defining the papers to take into consideration since they reference a valid software (see definition of valid at step 4) and the other identifying those papers that reference an invalid software (see definition of invalid at step 4). The papers which do not stick to these parameters are excluded from the review because their topic is actually different or non-relevant with respect to the research question purposes. Whether other researchers require different parameters in order to carry out their own research, they can change them, and the same is true for the specifications of the parameters outlined for this research on softwares for data extraction. Below the ***IC for the papers full text*** selected for this research are listed.

Information about a valid software (one of them is enough):

- The article contains an explicit reference to a useful and valid software;
- The article contains a link to a useful and still valid software;
- The article describes a still valid software.

Information about an invalid software:

- The article contains the description or reference to a software which is not useful or no more valid.



Although some publishers provide sufficient metadata about their papers, many papers are provided with incomplete metadata. Also, papers obtained via web-crawling are not associated with any metadata. To fill in this gap, we built the SCIENCEPARSE system to predict structured data from the raw PDFs using recurrent neural networks (RNNs).[2] For each paper, the system extracts the paper title, list of authors, and list of references; each reference consists of a title, a list of authors, a venue, and a year.

The full-text parameters confirm the presence of a topic relevant with respect to the research question.

Citation: Alessia Cioffi Systematic Literature Review about Software for References Extraction
https://dx.doi.org/10.17504/protocols.io.buz9nx96

Create the materials for the research

1 The protocol reports the different and complementary steps required in order to make a systematic literature review. In this case, the topic of the research is focused on the retrieval of software for references extraction from PDF files. The structure of the review is split into two parts, search strategies and quality assessment. The search strategies are based one on the seed papers and the other on the keywords search; to each of both has been devoted a section. They are two complementary ways to maximize the retrieval of papers relevant with respect to the topic of the research question. Then, the quality assessment is positioned in the penultimate section. But before the beginning of the research there is the need to prepare the materials for the research. Indeed, this first step focuses on the creation of the materials that will be used later in the research process to keep track of the materials retrieved.

1.1 Create three files:

- "*Papers_to_keep*" will contain the papers compliant with the *General IC for papers* and with at least one of the points of the *IC for screening procedure on papers*. These papers will be further analyzed by reading the full text.
- "*Papers_to_discard*", instead, includes all the files excluded from the research because not compliant with at least one point of the *General IC for papers* or with neither of the points in the *IC for screening procedure on papers*. This file is useful to keep track of the discarded papers and to check whether a newly retrieved paper appears in that list, so that we already know it must be discarded.
- "*Papers_to_analyze*" can be considered as an intermediate file. Indeed, this file will be populated with the papers retrieved in each specific step of the research process. After the end of that research step, each of the papers listed in this file will be 1. moved to "*Papers_to_keep*" or to "*Papers_to_discard*" on the basis of some characteristics that will be defined later on; 2. deleted from "*Papers_to_analyze*". Therefore, since these two steps will be iterated on all the steps of the research process, at the end of the research "*Papers_to_analyze*" will be empty.

All these three files will have the same structure, i.e. a 8 columns table containing the information necessary to fully identify a paper: Title, DOI (or link if the DOI is not available), Author(s), Publication Date, Venue, Volume, Issue, Pages. All this information (or only some of them, depending on the kind of publication), are necessary in order to completely identify a production and to keep track of all the information that could be useful also in further steps of the review.

The following table is an example of how the files will look like, the examples are taken from "Construction of the Literature Graph in Semantic Scholar" by Waleed et al. from the '**Seed papers list'** (from a book) and "Guidance on Conducting a Systematic Literature Review" by Xiao and Watson (from a journal).

Citation: Alessia Cioffi Systematic Literature Review about Software for References Extraction
https://dx.doi.org/10.17504/protocols.io.buz9nx96

Title	DOI/Link	Author(s)	Pub Date	Venue	Volume	Issue	Page(s)
Construction of the Literature Graph in Semantic Scholar	10.18653/v1/N18-3011	Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni	06/2018	Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)	N/A	N/A	84–91
Guidance on Conducting a Systematic Literature Review	10.1177/0739456X17723971	Yu Xiao, Maria Watson	2019	Journal of Planning Education and Research	39	1	93-112

Search with seed papers

2 The second step of the research is based on the **seed papers**, i.e. papers identified before the beginning of the research. These papers give place to the first typology of research, based on an iterative research performed on these initial papers and on the successively retrieved ones. Once verified whether the seed papers meet some

Citation: Alessia Cioffi Systematic Literature Review about Software for References Extraction
https://dx.doi.org/10.17504/protocols.io.buz9nx96

specified requirements, the focus will be shifted to the research and analysis of the papers where the seed papers are cited (forward) and of the papers that the seed papers cite (backward) to retrieve useful papers for the research.

FASTEN UP THE REJECTION PROCESS

- A technique that can be used at any stage of the research is looking in an automatic way to the abstract of the papers with a DOI on the Crossref API. This can be done through the use of a keywords list: whether one or more of the keywords in the list are found the paper shall be added to ***papers_to_analyze***, otherwise, if none of the selected keywords are retrieved in the abstract, the title shall be added to ***papers_to_discard***. In this case the selected keywords were: **keywords_l** = ['reference', 'references', 'reference strings', 'bibliographic', 'bibliography', 'citation', 'citations', 'extract', 'extraction', 'extractions', 'mining', 'mine', 'retrieve', 'CERMINE', 'GROBID', 'ParsCit'].

- In case the length of the papers list in ***papers_to_analyze*** becomes quite long (from around 300 papers on), it is possible to use a keywords technique to fasten up the process of rejection of unuseful papers. This technique is based on the fact that there could be some words, in the title of the papers relevant to understand the content of the paper itself. From this assumption, derives the fact that some words could be relevant in order to identify some (or many) papers which probably are not connected to the research topic. By selecting these keywords and making a targeted search, a huge number of unuseful papers can be fastly discarded. In this case some of the most useful keywords were: "knowledge" (connected to the creation/ use of graphs rather than to citation extraction), "semantic" (word usually not related to the extraction but the classification of the citations), "keyword" (related to keywords extraction), "covid" (and other topics related to medicine where the focus is, in most cases, on the content rather than on the citations), "measur" (often used to identify the citation index, which does not include the extraction phase).

2.1 **Iteration on the seed papers.** For each of the seed papers listed in '**Seed papers list'** (in the section "Materials") we make an analysis of the seed paper: we check whether the selected paper meets all the requirements pointed out in *General IC for papers.* There are two possibilities:

- If the work meets the specified criteria **we keep the selected seed paper** and we create a new row in "*Papers_to_analyze*" and add there the information about the selected article, as shown in the table at Step 1 **go to step #1.1** .
- Instead, if the paper does not meet the specified criteria, **we do not keep the selected paper** and we create a new row in "Papers_to_discard" and add there the information of the selected article, as shown in the table at Step 1 **go to step #1.1** .

Then, in both cases, we jump to the next paper in '**Seed papers list'** or, if we were analysing the last item of the list, we can procede with the **iteration on "*Papers_to_analyze*"**, in the next step.

2.2 **Iteration on "Papers_to_analyze"**. For each of the papers listed in "Papers_to_analyze" we make:

1. Analysis of the paper: check its title, keywords and abstract. In case the article seems useful for the purposes of the research (i.e. it meets at least one of the requirements listed in *IC* for screening procedure on papers):

- create a new row in "Papers_to_keep" and add there the information of the selected article, as shown in the table at Step 1 **go to step #1.1** ;

Citation: Alessia Cioffi Systematic Literature Review about Software for References Extraction
https://dx.doi.org/10.17504/protocols.io.buz9nx96

- remove the referring row in "Paper*s_to_analyze*";
- proceed with the backward and forward search (points 2 and 3).

Otherwise:

- create a new row in "Papers_to_discard" and add there the information of the selected article, as shown in the table at Step 1 go to step #1.1 ;
- remove the referring row in "*Paper*s_to_analyze";
- jump to the next paper in "*Papers_to_analyze*" (restart from point 1 with the next paper).

2. Backward search. Check the articles cited by the reference papers by looking at the list of references at the end of the article. Then, for each of the articles retrieved with the backward search, if its title is not in "Papers_to_keep", "Papers_to_discard" or "*Papers_to_analyze*" and if it meets all the requirements listed in General I*C for papers*, add it to "*Papers_to_analyze*".

3. Forward search. The forward search is carried out both manually and with the help of some platforms which have that specific functionality (see "**Free open platforms**" in the section "Materials"). Then, for each of the articles retrieved with the forward search, if its title is not in "Papers_to_keep", "Papers_to_discard" or "*Papers_to_analyze*" and if it meets all the requirements listed in General I*C for papers*, add it to "*Papers_to_analyze*".

2.3 When no new article is retrieved with this search method or when there are no more rows in the table of "*Papers_to_analyze*", we can consider finished the research through seed papers and we can skip to the next section, "Search with Keywords".

Search with keywords (OPTIONAL)

3 This step, based on a **search with keywords** through the most used openly available research platforms, directly follows the previous one, based on the search with seed papers, with a complementary perspective. **go to step #2 Research with seed papers** . Indeed, this approach is able to fill an eventual gap of papers and publications left by the previous step. At the end of this section, all the available documents should be retrieved.

ONGOING CHANGE OF KEYWORDS

Because of practical reasons, apart from the keywords selected before the beginning of the research of the literature, new keywords can be identified during the research itself because new information becomes available from the reading of some papers. For instance, this can happen after the discovery of new software or methodologies. In this case, they may require to be further searched in order to get more information about them. Thus, every time a new keyword is required, this should be added to the original keywords list. That is why this step can be considered both an iterative step, since if at the end some information is missing the process restarts with the new keywords, and a step iteratively recallable from the fourth one ***"Quality assessment-full text reading"***: Indeed, whether from the full-text reading it comes out that there are no enough information about a specific topic or software it can be necessary to come back to the search with keywords and specifically analyze that specific sub-topic.

3.1 This passage is about the further keywords that will be added to the original list of keywords during the research process. Indeed, there is a possibility that the keywords identified before the beginning of the research won't be enough in order to carry out all the possible information necessary in order to completely review all the software or the subtopics identified throught the retrieved papers. Therefore there could be the necessity to create new keywords, in order to search for a specific topic or to answer to a specific need of the research, and those keywords

Citation: Alessia Cioffi Systematic Literature Review about Software for References Extraction
https://dx.doi.org/10.17504/protocols.io.buz9nx96

should be added to the original list of keywords in order to keep track of all the keywords used. Since in some cases this will be an iterative step, there are two possible cases:

1. **It is the first iteration on this step, there is no new keyword to add**: just skip the passage and go to the next step, 3.2;
2. **It is the second, or more, time in which we go through this step**: we need to add the further keywords which we want to use in the keyword search to the keyword list.

3.2 **Iteration on the keywords**. For each of the platforms under the voice "**Free open platforms**" in the section "Materials":

Research: Perform on that platform a research for each of the keywords listed in **"Keywords list"** (in the section "Materials"). For each of the articles retrieved through each keyword, if its title is not in "Papers_to_keep", "Papers_to_discard" or "*Papers_to_analyze*", and it meets all the requirements listed in *General IC for papers*, add it to "*Papers_to_analyze*".

SELECTION OF THE NUMBER OF RESULTS

For the purposes of the keyword research in this protocol it is kept into consideration max 20 pages (Google Scholar)/200 results. These numbers have been selected because they present the most relevant results with respect to the keywords (which are not too generic and therefore the relevant results should stick around these numbers). This selection has been based on the reports:
Ranking #1 on Google Is Overrated (Ahrefs' Study of 100k Keywords), Guidance on Conducting a Systematic Literature Review.
Nonetheless, this is not a mandatory number, and whether someone else wants to change the amount of results considered, either diminuishing or extending them, he/she can freely do it.

3.3 **Iteration on the retrieved papers.** For all the publications listed in "*Papers_to_analyze*":

1. Analysis of retrieved articles. Check the title, keywords and abstract of the selected article.
If the article meets at least one of the requirements listed in *IC for screening procedure on papers*:

- create a new row in "Papers_to_keep" and add there the information of the selected article, as shown in the table at Step 1 **go to step #1.1** ;
- remove the referring row in "*Papers_to_analyze*";
- proceed with the backward and forward search (points 2 and 3).

Otherwise, if the papers does not meet any of the requirements in *IC for screening procedure on papers*:

- create a new row in "Papers_to_discard" and add there the information of the selected article, as shown in the table at Step 1 **go to step #1.1** .
- remove the referring row in "*Papers_to_analyze*";
- jump to the next paper in "*Papers_to_analyze*" (restart from point 1 with the next paper).

2. Backward search. Look at the articles cited by the reference paper by looking at the list of references at the end of the article. Then, for each of the articles retrieved with the backward research, if its title is not in "Papers_to_keep", "Papers_to_discard" or "*Papers_to_analyze*" and if

Citation: Alessia Cioffi Systematic Literature Review about Software for References Extraction
https://dx.doi.org/10.17504/protocols.io.buz9nx96

it meets all the requirements listed in *General IC for papers*, add it to "*Papers_to_analyze*".

3. Forward search. The forward search is carried out both manually and with the help of some platforms which have that specific functionality (see "**Free open platforms**" in the section "Materials"). Then, for each of the articles retrieved with the forward search, if its title is not in "Papers_to_keep", "Papers_to_discard" or "Papers_to_analyze" and if it meets all the requirements listed in *General IC for papers*, add it to "Papers_to_analyze".

3.4 When no new keyword leads to new results or when there are no more rows in the table of "Papers_to_analyze", we can consider finished the research. We can say that almost all the articles regarding this topic have been found and we can proceed by reading the full text and extracting the information useful to retrieve the software and select the best one with respect to the purposes of the research question.

Quality assessment-full text reading

4 The step of the quality assessment takes as input the two files produced in the previous step: "Papers_to_keep" and "Papers_to_discard". The output file will be these two, with some eventual modifications due to the identification or not of parameters defined in section 1.3 (*IC for the papers full text*) in the full text of the articles accepted in the previous passage, and other two files ("papers", "software") which allow a further distinction on the basis of the content of the articles.
This step can be considered iterative with respect to the previous one,
go to step #3 Search with keywords . Indeed, since only in this section it is fully clear the content of the articles identified during the research process, thanks to the full-text reading, it may result in the necessity of providing more information on a topic or software about which only few or not enough information was retrieved in the previous research steps.

VALID

The definition of validity can be associated to that of usable. A valid software is, indeed, a software who can be used in the current moment for working on a specific topic. In order to be considered so it must meet the following requirements:

- it is available (better if in an updated version);
- it is compatible with at least one of the major operating systems (see the Materials section for more);
- it is a free open source software or, at least, freely available.

INVALID

A software is classified as invalid (decayed) if, for any reason, it cannot be used for the purposes of the research. In practical terms, a decayed software can be identified through the following parameters:

- it is not available;
- it is not compatible with the operating systems we want to use;
- it is available on payment.

4.1 Before the beginning of the research we need to create the materials which will be necessary in order to carry out the task of the annotations of the full-text reading. First of all we have to create two files:

Citation: Alessia Cioffi Systematic Literature Review about Software for References Extraction
https://dx.doi.org/10.17504/protocols.io.buz9nx96

- “Papers”: this file will include all the articles which, after the first selection in the research process, will stick also to the full-text parameters, and therefore can be considered useful in order to answer the research question. The format will be again that of the table, in this case of a three columns table, where each column will correspond to the following values: topic, explicit mention of a relevant software, explicit mention of other software. This file will be particularly useful after the end of the research to connect the retrieved papers to their referring resources, in order to investigate the entirety of the software and to collect the previous and relative works.

Title	Topic	Relevant software	Other software
Costruction of the Literature Graph in Semantic Scholar	Use of the software science parse in order to retrieve information from PDF files: title, abstract and citations.	ScienceParse	Apache PDF box

- “Software”: it is a file aimed at collecting the software retrieved through the literature review. It is a 2 columns table which contains the information of the name of the software and of its validity status. In this way it is possible to collect all the software and, together with the information about its validity, analyse and classify them after the end of the research.

Name	Repository link	Validity status
ScienceParse	https://github.com/allenai/science-parse	Valid

Also, both the files must be created on a digital file so that clusters of topics can be created. Indeed this action has a different objective for the two files:

1. For the first file, "Papers", the reason for creating clusters is that of ordering the articles on the basis of their topic (i.e. the software, in the third column), so that in a following step they can by taken into consideration in groups defined by the topic and, also, it is easier to find the articles describing a specific software by grouping them.
2. Instead, in the case of the file "Software", it is used in order to distinguish the valid software from the invalid ones.

4.2 For each of the papers in “Papers_to_keep” read its full text. We outline three possible cases derived from the possible outcomes of the verification whether the articles include the points listed in *IC for the papers full text* or not.

Discard the paper. If the selected paper does not meet any of the requirements listed in *IC for the papers full text:*

1. remove it from “Papers_to_keep” and add it to “Papers_to_discard”. Indeed, the parameters listed in *IC for the papers full text* are still necessary in order to distinguish between the papers useful for the purposes of the research and the ones that are not so, even if they could stick to the inclusion criteria of previous steps;
2. do not add anything to the file "Papers" nor to "Software". Since the file does not contain useful information for the research, we must not take it into account in the list of the useful papers and software.

Add the papers to "Papers" and "Software" (invalid status). If the paper sticks to the second point in *IC for the papers full text:*

1. create a new row in the file “Papers” and add there its title and information. Indeed, even if it is about an invalid software it is to be considered in order to make a complete review of the

Citation: Alessia Cioffi Systematic Literature Review about Software for References Extraction
https://dx.doi.org/10.17504/protocols.io.buz9nx96

software which have been used and eventually investigating the reasons why it cannot be used anymore;
2. if the software name is not already in the file "Software", create a new row, add the software name in the first cell and "**invalid**" in the second cell, i.e. the one of the status of the software. If the software is already in, skip this step.

Add the papers to "Papers" and "Software" (valid status). If the paper contents meet at least one of the requirements of the first point in *IC for the papers full text:*
1. create a new row in the file “Papers” and add there its title and information;
2. if the software name is not already in the file "Software", create a new row, add the software name in the first cell and "**valid**" in the second cell, i.e. the one of the status of the software. If the software is already in, skip this step.

4.3 If at the end of the step 4.2 **go to step #4.2** , there are some information missing about one or more software, we have to go back to step 3 **go to step #3 Search with Keywords** and repeat the search with keywords, using the name or a concept related to the specific software of interest.
Otherwise, if no new information is required we can consider the research and selection of the literature over. At the end of the research and evaluation process we will return three non-empty output files: “Papers_to_keep”, "Papers" and "Software".

Citation: Alessia Cioffi Systematic Literature Review about Software for References Extraction
https://dx.doi.org/10.17504/protocols.io.buz9nx96