CRUISE–Screening: Living Literature Reviews Toolbox

Wojciech Kusa
TU Wien
Vienna, Austria
wojciech.kusa@tuwien.ac.at

Petr Knoth
The Open University
Milton Keynes, UK
petr.knoth@open.ac.uk

Allan Hanbury
TU Wien
Vienna, Austria
allan.hanbury@tuwien.ac.at

ABSTRACT
Keeping up with research and finding related work is still a time-consuming task for academics. Researchers sift through thousands of studies to identify a few relevant ones. Automation techniques can help by increasing the efficiency and effectiveness of this task. To this end, we developed CRUISE–Screening, a web-based application for conducting living literature reviews—a type of literature review that is continuously updated to reflect the latest research in a particular field. CRUISE–Screening is connected to several search engines via an API, which allows for updating the search results periodically. Moreover, it can facilitate the process of screening for relevant publications by using text classification and question answering models. CRUISE–Screening can be used both by researchers conducting literature reviews and by those working on automating the citation screening process to validate their algorithms. The application is open-source, and a demo is available under this URL: https://citation-screening.ec.tuwien.ac.at.

CCS CONCEPTS
• Information systems → Search interfaces; Clustering and classification; Document filtering; • Computing methodologies → Natural language processing.

KEYWORDS
information retrieval, natural language processing, literature reviews, living reviews, systematic reviews, citation screening

ACM Reference Format:

1 INTRODUCTION
Literature reviews are a critical component of all research domains. Two different types of reviewing the literature can be distinguished: systematic reviews and, more general, exploratory literature reviews. Systematic literature reviews follow strict criteria and are commonly used in healthcare and medical domains, as they provide the gold standard in evidence-based medicine [9, 29]. Such reviews are a tedious, recall-oriented and repetitive process; they typically involve several stages, including formulating a research question, defining inclusion and exclusion criteria, searching multiple databases, screening for relevant studies, assessing study quality, extracting data, and synthesizing the findings. In recent years, several tools and procedures have been developed to automate this process, making it more efficient and effective.

The process applied in systematic literature reviews in medicine was also transferred into other scientific domains, such as environmental sciences [4], software engineering [18], social sciences [33] and engineering [5]. There is an opportunity to reduce the human labour as well as the time it takes to deliver systematic reviews with the use of technology, specifically NLP and ML.

General literature reviews (usually more exploratory than “systematic”) are also conducted in the academic setting, often by PhD and Masters students [12, 38]. This process enables researchers to familiarise themselves with the current state of the art, theory and methods in their field. Overlapping reviews are very often repeated by different groups as there is no data sharing and exchange format that could enable reusing past reviews as it is in systematic reviews in medicine [16, 37]. Guidelines and methodologies also aim to improve this process but do not mention any automated approaches, and the search process itself is not very structured but, instead, exploratory [34]. There is substantial potential in developing standards and tools that academics could adopt for the purpose of literature reviews.

In recent years, there has been a growing interest in the concept of living literature reviews [10, 43]. These reviews are continuously updated to reflect the latest research in a particular field.

In this paper, we present CRUISE–Screening, a web-based tool for conducting living literature reviews. CRUISE–Screening, developed to improve the efficiency of the literature review process, is connected to several search engines via API and facilitates the process of screening for relevant publications using NLP and machine learning methods. We discuss the development and functionalities of CRUISE–Screening, as well as present the challenges in developing such a tool. The system has notable novelty as it integrates search and screening capabilities into a single application and can connect with several machine learning models. We foresee two use cases for our system: (1) primarily by researchers wanting to review the literature to locate the relevant work in their field of expertise; and (2) people developing automation models for literature reviews wanting to compare their approaches with others.

While most tools in this domain are developed specifically for systematic reviews, our system is among the first to apply systematic review concepts to general literature reviews. This sets our system apart from the rest of literature review tools, which are primarily recommendation systems for papers. Our system has the potential to promote collaboration and facilitate the exchange of ideas among researchers.

1https://github.com/ProjectDoSSIER/cruise-screening

This work is licensed under a Creative Commons Attribution International 4.0 License

CIKM ’23, October 21–25, 2023, Birmingham, United Kingdom
© 2023 Copyright held by the owner/author(s).
https://doi.org/10.1145/3585780.3614736

5071
2 RELATED WORK

2.1 Academic Search Engines

Private academic search engines, citation indices and paywalled collections such as ScienceDirect and Web of Science are one source of finding publications. Public search engines and publication aggregators such as Google Scholar, Semantic Scholar, CORE, OpenAlex and PubMed are becoming increasingly popular for allowing researchers to freely access the latest publications. Their main goal is creating a citation network, and their support for conducting systematic literature reviews is often minimal. Moreover, only a few of these tools provide an API, and none of them allow for a traditional systematic literature review workflow.

2.2 Systematic Review Toolboxes

There is already a number of tools helping researchers conduct systematic literature reviews. An online catalogue enumerates 45 tools helping users during the screening phase, whereas Harrison et al. found 15 of them accessible and available without specific computing infrastructure for a title and abstract screening step. Dedicated commercial tools exist to help medical researchers conduct systematic literature reviews. They are usually customised to medical reviews and require purchasing a subscription which can be a bottleneck to academic researchers from lower- and lower-middle income countries.

In addition to the commercial tools, a plethora of free or open-source tools is available, usually created by academics. These tools, such as Abstrakr, Rayyan, or ASReview usually support only one of the systematic review stages.

2.3 Automated Citation Screening

All the documents retrieved from the search step constitute the input to the citation screening step. In a manual screening scenario, reviewers read all these documents to select only the fraction relevant to the systematic review. Because the total number of retrieved studies can go into tens of thousands, it is essential to find a way of improving this process. Automated citation screening is an umbrella term for using NLP, machine learning and information retrieval techniques with the goal of decreasing the time spent on manual screening. Classification approaches train a supervised model on an annotated dataset to determine whether a paper should be included or excluded from the review.

Previous approaches ranged from statistical models like naïve Bayes classification, support vector machine (SVM), voting perceptron and random forest to neural networks. A significant limitation of all these approaches is the need for a large number of manual annotations that must be completed before developing a reliable model for every new systematic review. Moreover, the majority of the classification models are evaluated only retrospectively which might raise questions of data leakage when considering large amounts of data used for pretraining language models.

3 CRUISE-SCREENING

Figure 1 shows the architecture of CRUISE-SCREENING, which is built with Python 3.9, Django 4, Bulma and AlpineJS frameworks.

3.1 Data Resources

Good quality input data covering multiple domains is the crucial ingredient of a successful literature review. Nussbaumer-Streit et al. found that combining two separate databases may suffice to reliably determine the conclusions of a systematic review in medicine. Therefore, CRUISE-SCREENING was designed to use multiple data sources and to allow for extending them when needed. Currently, it supports the following four search engines as data sources: Semantic Scholar, CORE, PubMed via ENTREZ API and internal document storage.

The first three APIs call search engines that are used as primary data sources when searching for documents. Using three different search engines with contrasting scopes and content enables good search results coverage.

The tool also allows for indexing documents in the internal database. It is implemented using Elasticsearch and communicates with the main application using the API. It can be used, for example, when one wants to store private documents or content not covered by other search engines. For this demo, we index the DBLP-Citation-network Version 13 collection created by Tang et al. [39].

The system could be expanded to connect to other search engines offering API access. As the system is a meta-search engine, we use a script to deduplicate the search results based on papers’ metadata.

3.2 Screening Workflow

The typical screening workflow for systematic literature reviews consists of two stages. In the first stage, the researcher searches for documents potentially related to the research topic. In the second stage, the documents are screened for relevance. We have implemented these two stages inside CRUISE-SCREENING.

Search for relevant items. First, the user creates a new literature review by defining the research protocol. The protocol consists of the review’s title, description, at least one search query and a set of inclusion and exclusion criteria (eligibility criteria). The tool allows for specifying search engines, by default selecting all four available sources described in Section 3.1. The search can be limited to only the first N results if the reviewer is not interested in a comprehensive literature review.

CRUISE-SCREENING sends API requests to selected search engines and gathers all responses. Merged and deduplicated search results are stored in a PostgreSQL database. In order to support living reviews, the user can re-run the search function periodically to update the list of references. However, since search engines only allow filtering by publication year and not month or day, the tool removes publications older than the year of the previous search during updates. The tool then relies on deduplication to ensure that new publications are not mistakenly added twice.
CRUISE–Screening also allows for the additional direct import of data for screening from two sources:

- **Bulk upload from reference files** – Currently, the tool supports .bib and .ris file extensions. These publications are imported to the PostgreSQL database.
- **Full text .pdf files** – These files are processed using GROBID [1] and then added to the database. Documents added this way can also be marked as seed studies. This way, these new documents are labelled as **relevant**, which can speed up the process of automated screening.

**Citation screening.** Currently, CRUISE–Screening implements the title and abstract screening while providing external URLs to full text articles whenever available. Figure 3 presents an example screening interface. From the top, it contains the title, abstract, authors, publication venue and year and the link to the full text of the screened paper. Below are two sections with eligibility criteria questions and a main inclusion question.

There are two screening workflows in CRUISE–Screening: strict and relaxed. Strict screening requires the annotators to conduct the process by manually answering every eligibility question. It mimics the citation screening process of systematic reviews. This mode could be used for in-depth systematic reviews or gathering manual annotations for training machine learning models.

The relaxed mode does not impose any requirements on which questions the annotator should answer except for the main **include**/*maybe*/**exclude** decision. There are optional questions about the reviewer’s prior knowledge of the paper and authors, which reviewers can turn on to control for the selection bias.

The output of the literature review can be exported in a json format. It contains the literature review protocol and all identified studies with corresponding automatic and manual decisions.

**3.3 Automation Methods**

Except for the fully manual workflow, CRUISE–Screening implements automation methods to increase the speed and coverage of the literature review. Implemented approaches include supervised text classification and zero-shot question-answering models. The tool connects to them using an API, which allows for extending the list of supported models.

**Text classification.** We implemented two examples of supervised classifiers based on previous literature: a logistic regression model using tf-idf text representation and a fastText classifier [17]. These models provide a single **yes/no** decision for each paper (corresponding to the main eligibility question from the manual workflow). Reviewers need to annotate a “training set” of at least three included and three excluded papers before using the models. When the reviewer annotates more publications, the models can be retrained to make an updated prediction on the remaining documents.

**Question answering.** In addition to supervised text classification, CRUISE–Screening enables users to conduct automatic screening.
using prompt-based language models with a question answering approach. The advantage of this method is that it does not require pre-labelled data and can make predictions for all inclusion and exclusion criteria. However, it can be computationally intensive and sensitive to the quality of input questions. The API is designed to support any Text2TextGeneration model implemented in the HuggingFace Transformers [44] library. We used the T0_3B and T0 models [36]. The example prompt consists of a single eligibility question and the same paper data as available during manual screening (Figure 3), namely the title, abstract, authors, journal name and publication year.

4 DISCUSSION AND CONCLUSION

The merging of results from multiple sources can present significant challenges in the context of scientific publications. Although using multiple data sources can lead to better coverage of relevant studies, combining the results from different sources is not a trivial task. The data can have different formats, fields, and identifiers, which require significant effort to reconcile. Additionally, data quality can be poor in some cases, which can further complicate the merging process. Therefore, careful consideration must be given to the merging process, and the use of automated tools can help improve the accuracy and efficiency of the process. Nevertheless, human intervention may still be required to resolve any inconsistencies or errors in the data [13].

The evaluation of models is an essential aspect of our tool, as it allows researchers to evaluate their models without the risk of data leakage. Due to the vast amount of training data used by large language models, it can be difficult to detect data leakage when retrospectively evaluating on common benchmarks. Therefore, our tool provides a solution by enabling researchers to make predictions with several models before starting a new review, storing the results, and evaluating the models after the manual review is conducted, without interfering with the manual workflow.

Compared to other tools, CRUISE–Screening combines search and screening stages into one workflow. Thanks to this, researchers can use the tool as an information system to systematise, manage and record their literature review workflows.

We note that the approach based on prompting large language models can generate non-reliable predictions. We added warnings to the user interface so the user knows these predictions can contain hallucinations. In future work, we plan to present the user with a predicted performance on this particular criterion if the same question was asked in the previous reviews and there was enough data on its accuracy against the ground truth.

ACKNOWLEDGEMENTS

This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval – DoSSIER (H2020-EU.1.3.1., ID: 860721). We thank all the members of the DoSSIER Project who contributed to the CRUISE workshop.