



# 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,<sup>1</sup> 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:

Wojciech Kusa, Petr Knoth, and Allan Hanbury. 2023. CRUISE–Screening: Living Literature Reviews Toolbox. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, October 21–25, 2023, Birmingham, United Kingdom. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3583780.3614736>

## 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

<sup>1</sup><https://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).  
ACM ISBN 979-8-4007-0124-5/23/10.  
<https://doi.org/10.1145/3583780.3614736>

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.

## 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<sup>2</sup>, Semantic Scholar [2], CORE [20], OpenAlex [35] and PubMed<sup>3</sup> 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<sup>4</sup> enumerates 45 tools helping users during the screening phase, whereas Harrison et al. [14] 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 [28, 30].

In addition to the commercial tools, a plethora of free or open-source tools is available, usually created by academics. These tools, such as Abstrakt [42], Rayyan [11], or ASReview [41] 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 [32]. Automated citation screening is an umbrella term for using NLP, machine learning and information retrieval (IR) 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 [23, 24].

Previous approaches ranged from statistical models like naïve Bayes classification [3, 27], support vector machine (SVM) [7, 8, 15, 26], voting perceptron [6] and random forest [19] to neural networks [21, 22]. 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 [40]. 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 [25].

<sup>2</sup><https://www.scholar.google.com/>

<sup>3</sup><https://www.ncbi.nlm.nih.gov/pubmed/>

<sup>4</sup><http://www.systematicreviewtools.com>

## 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. [31] 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 API<sup>5</sup>, CORE API<sup>6</sup>, PubMed via ENTREZ API<sup>7</sup> 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<sup>8</sup> 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 ❶. The protocol (Figure 2) 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 ❷. Merged and deduplicated search results are stored in a PostgreSQL database ❸. 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.

<sup>5</sup><https://www.semanticscholar.org/product/api>

<sup>6</sup><https://core.ac.uk/services/api>

<sup>7</sup><https://www.ncbi.nlm.nih.gov/search/>

<sup>8</sup><https://www.aminer.cn/citation>

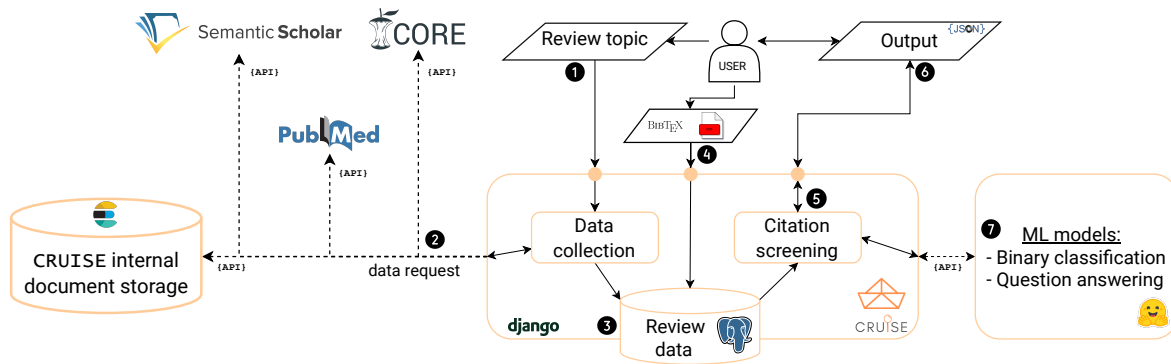


Figure 1: Overview of the CRUISE–Screening architecture.

Literature review title:

How is the knowledge gap modelled in information retrieval?

Literature review description:

I would like to know if there are search engines that use a representation of the user's knowledge to measure their knowledge gap with their target information need. Therefore, I would like to know what techniques are used to model users' knowledge and the target knowledge and how is a gap calculated based on these two knowledge representatives. How and how often is the knowledge gap updated during the search and how is the knowledge gap used to improve the exploratory search experience.

Type in your search queries, each query on a new line:

knowledge gap models in information retrieval  
 knowledge gap in information retrieval  
 model of knowledge gap in IR  
 knowledge gap model in information retrieval

Type in your inclusion criteria, each one on a new line:

Paper about information retrieval  
 Paper about knowledge gap  
 Paper including a model of knowledge gap  
 Paper about knowledge delta

Type in your exclusion criteria, each one on a new line:

Paper written in language other than English  
 Only title is available  
 Paper from other domains than Information Retrieval  
 Paper older than 2000

Figure 2: Example literature review protocol containing review title, description, search queries and criteria for inclusion and exclusion.

CRUISE–Screening also allows for the additional direct import of data for screening from two sources 4:

- 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 5 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 6. 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 7. 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 re-trained 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

Back Review title: How is the knowledge gap modelled in information retrieval? All Reviews

## Accounting for User's Knowledge and Search Goals in Information Retrieval Evaluation - Extended Abstract

**Abstract:** Accounting for the user's cognitive aspects in the information retrieval field is still considered a challenge up until our days. Knowing that recent frameworks are trying to fill this gap, the bigger challenge remains to evaluate those frameworks and to measure the results' relevance in view of the user cognition. The majority of existing evaluation measures often consider isolated document–query environments. Traditional evaluation measures, for example, precision and recall, are not suitable to evaluate the quality of such IR algorithms. Goffman et al. recognised that the relevance of a document must be determined with respect to the documents appearing before it while Boyce et al. claimed that the change a document makes in the knowledge state must be reflected in the choice of document for the second position. The few measures that account for the user's cognitive aspects when evaluating the "relevance" of a result or ranking are limited to one search session, one query, or one search goal. The evaluation metric proposed by Clarke et al. for example systematically... [Show full abstract](#)

Dima El Zein, C. Pereira

2022 — CIRCLE

**1. Relevance \***

Domain relevance:  very relevant  somewhat relevant  not relevant

Topic relevance:  very relevant  somewhat relevant  not relevant

**2. Inclusion criteria**

Paper about information retrieval:  Yes  Not sure  No

Paper about knowledge gap:  Yes  Not sure  No

Paper including a model of knowledge gap:  Yes  Not sure  No

Paper about knowledge delta:  Yes  Not sure  No

**3. Exclusion criteria \***

Paper written in language other than English:  Yes  Not sure  No

Only title is available:  Yes  Not sure  No

Paper from other domains than IR:  Yes  Not sure  No

Paper older than 2000:  Yes  Not sure  No

**4. Descriptive reason**

Paper is not modelling user's knowledge gap but still is relevant

**5. Decision based on title and abstract \***  Include  Not sure  Exclude

Figure 3: Example screening interface in CRUISE–Screening presenting single paper with answered questions.

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.

## REFERENCES

- [1] 2008–2022. Grobid. <https://github.com/kermitt2/grobid>.
- [2] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*.
- [3] Tanja Bekhuis, Eugene Tseytlin, Kevin J Mitchell, and Dina Demner-Fushman. 2014. Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *PLoS one*, 9(1):e86277.
- [4] Gary S Bilotta, Alice M Milner, and Ian Boyd. 2014. On the use of systematic reviews to inform environmental policies. *Environmental Science & Policy*, 42:67–77.
- [5] Stefanie Castillo and Petar Grbovic. 2022. The apisser methodology for systematic literature reviews in engineering. *IEEE Access*, 10:23700–23707.
- [6] A. M. Cohen, W. R. Hersh, K. Peterson, and Po Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219.
- [7] Aaron M. Cohen. 2008. Optimizing Feature Representation for Automated Systematic Review Work Prioritization. *AMA Annual Symposium Proceedings*, 2008:121.
- [8] Aaron M Cohen. 2011. Letter: Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure. *Journal of the American Medical Informatics Association : JAMIA*, 18(1):104.
- [9] Deborah J Cook, Cynthia D Mulrow, and R Brian Haynes. 1997. Systematic reviews: synthesis of best evidence for clinical decisions. *Annals of internal medicine*, 126(5):376–380.
- [10] Julian H. Elliott, Tari Turner, Ornella Clavisi, James Thomas, Julian P.T. Higgins, Chris Mavergames, and Russell L. Gruen. 2014. Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. *PLoS Medicine*, 11(2).
- [11] A Elmagarmid, Z Fedorowicz, H Hammady, I Ilyas, M Khabsa, and M Ouzzani. 2014. Rayyan: a systematic reviews web app for exploring and filtering searches for eligible studies for cochrane reviews. In *Evidence-Informed Public Health: Opportunities and Challenges. Abstracts of the 22nd Cochrane Colloquium*, pages 21–26. John Wiley & Sons Hyderabad, India, India.
- [12] Rosemary Green. 2009. *American and Australian doctoral literature reviewing practices and pedagogies*. Deakin University (Australia).
- [13] Nathalia Sernizon Guimarães, Andréa JF Ferreira, Rita de Cássia Ribeiro Silva, Adelzon Assis de Paula, Cinthia Soares Lisboa, Laio Magno, Maria Yury Ichiara, and Maurício Lima Barreto. 2022. Deduplicating records in systematic reviews: There are free, accurate automated ways to do so. *Journal of Clinical Epidemiology*, 152:110–115.
- [14] Hannah Harrison, Simon J Griffin, Isla Kuhn, and Juliet A Usher-Smith. 2020. Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. *BMC medical research methodology*, 20(1):1–12.
- [15] Brian E. Howard, Jason Phillips, Kyle Miller, Arpit Tandon, Deepak Mav, Mihir R. Shah, Stephanie Holmgren, Katherine E. Pelch, Vickie Walker, Andrew A. Rooney, Malcolm Macleod, Ruchir R. Shah, and Kristina Thayer. 2016. SWIFT-Review: A text-mining workbench for systematic review. *Systematic Reviews*, 5(1):1–16.
- [16] John PA Ioannidis. 2016. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, 94(3):485–514.
- [17] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, volume 2, pages 427–431.
- [18] Staffs Keele et al. 2007. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, ver. 2.3 ebse technical report. ebse.
- [19] Madian Khabsa, Ahmed Elmagarmid, Ihab Ilyas, Hossam Hammady, and Mourad Ouzzani. 2016. Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning*, 102(3):465–482.
- [20] Petr Knuth and Zdenek Zdrahal. 2012. CORE: three access levels to underpin open access. *D-Lib Magazine*, 18(11/12):1–13.
- [21] Georgios Kontonatsios, Sally Spencer, Peter Matthew, and Ioannis Korkontzelos. 2020. Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. *Expert Systems with Applications: X*, 6:100030.
- [22] Wojciech Kusa, Allan Hanbury, and Petr Knuth. 2022. Automation of Citation Screening for Systematic Literature Reviews using Neural Networks: A Replicability Study. In *Advances in Information Retrieval, 44th European Conference on IR Research, ECIR 2022*.
- [23] Wojciech Kusa, Aldo Lipani, Petr Knuth, and Allan Hanbury. 2023. An Analysis of Work Saved over Sampling in the Evaluation of Automated Citation Screening in Systematic Literature Reviews. *Intelligent Systems with Applications*, 18:200193.
- [24] Wojciech Kusa, Aldo Lipani, Petr Knuth, and Allan Hanbury. 2023. Vombat: A tool for visualising evaluation measure behaviour in high-recall search tasks. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*, page 5, Taipei, Taiwan. ACM.
- [25] Wojciech Kusa, Guido Zuccon, Petr Knuth, and Allan Hanbury. 2023. Outcome-based evaluation of systematic review automation. In *Proceedings of the 2023 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '23)*, page 9, Taipei, Taiwan. ACM.
- [26] David Martinez, Sarvnaz Karimi, Lawrence Cavedon, and Timothy Baldwin. 2008. Facilitating biomedical systematic reviews using ranked text retrieval and classification. In *Australasian Document Computing Symposium (ADCS)*, pages 53–60.
- [27] Stan Matwin, Alexandre Kouznetsov, Diana Inkpen, Oana Frunza, and Peter O’Blenis. 2010. A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association*, 17(4):446–453.
- [28] Cristhian D Morales-Plaza, David A Forero-Peña, and Fhabián S Carrión-Nessi. 2022. Resource use during systematic review production varies widely: a scoping review: response to nussbaumer-streit et al. *Journal of Clinical Epidemiology*, 142:319–320.
- [29] M Hassan Murad, Noor Asi, Mouaz Alsawas, and Fares Alahdab. 2016. New evidence pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127.
- [30] B Nussbaumer-Streit, LE Ziganshina, M Mahmić-Kaknjó, G Gartlehner, R Sfetcu, and H Lund. 2022. Resource use during systematic review production varies widely: a scoping review: authors’ reply. *Journal of Clinical Epidemiology*, 142:321–322.
- [31] Barbara Nussbaumer-Streit, Irma Klerings, Gernot Wagner, Thomas L. Heise, Andrea I. Dobrescu, Susan Armijo-Olivo, Jan M. Stratil, Emma Persad, Stefan K. Lhachimi, Megan G. Van Noord, Tarquin Mittermayr, Hajo Zeeb, Lars Hemkens, and Gerald Gartlehner. 2018. Abbreviated literature searches were viable alternatives to comprehensive searches: a meta-epidemiological study. *Journal of Clinical Epidemiology*, 102:1–11.
- [32] Alison O’Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, 4(1):5.
- [33] Mark Petticrew and Helen Roberts. 2008. *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons.
- [34] Catherine Pickering and Jason Byrne. 2014. The benefits of publishing systematic quantitative literature reviews for PhD candidates and other early-career researchers. *Higher Education Research & Development*, 33(3):534–548.
- [35] Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.
- [36] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.
- [37] Konstantinos C Siontis, Tina Hernandez-Boussard, and John PA Ioannidis. 2013. Overlapping meta-analyses on the same topic: survey of published studies. *Bmj*, 347.
- [38] Ayah Soufan, Ian Ruthven, and Leif Azzopardi. 2022. Searching the literature: an analysis of an exploratory search task. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 146–157.
- [39] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: Extraction and mining of academic social networks. In *KDD’08*, pages 990–998.
- [40] Guy Tsafnat, Paul Glasziou, George Karystianis, and Enrico Coiera. 2018. Automated screening of research studies for systematic reviews using study characteristics. *Systematic Reviews 2018 7:1*, 7(1):1–9.
- [41] Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix Weijdem, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, et al. 2021. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133.
- [42] Byron C. Wallace, Kevin Small, Carla E. Brodley, Joseph Lau, and Thomas A. Trikalinos. 2012. Deploying an interactive machine learning system in an Evidence-based Practice Center: Abstract. *IHI’12 - Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 819–823.
- [43] Michel Wijkstra, Timo Lek, Tobias Kuhn, Kasper Welbers, and Mickey Steijaert. 2021. Living literature reviews. In *Proceedings of the 11th on Knowledge Capture Conference*, pages 241–248.
- [44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.