UP2DATE: Shared Task on Systematic Review Updates

Pierre Achkar[1], Tim Gollub[2], Martin Potthast[3], Carsten Eickhoff[4] and Harrisen Scells[4]

[1]Leipzig University and Fraunhofer ISI
[2]Bauhaus-Universität Weimar
[3]University of Kassel, hessian.AI, and ScaDS.AI
[4]University of Tübingen

Abstract

In all areas of science, the volume of literature has become overwhelming, and staying current with the bleeding edge of research is increasingly unmanageable. This problem is particularly acute in medicine, where studies are published at a rate of approximately two per minute, and up-to-date information is crucial for decision-making processes. Maintaining the evidence base by updating systematic reviews is a costly and time-consuming endeavour, yet there is comparatively little research that tackles this important problem via computational means. We propose a shared task that addresses this challenge in (biomedical) scholarly information access. Specifically, we propose three tasks involved in updating systematic reviews: determining when a review should be updated, retrieving literature for the review update, and classifying whether each piece of retrieved literature should be included in the review. These are difficult problems that have received limited research attention, and improved solutions could have considerable impact.

Keywords

Systematic Reviews, Information Access, Scientific Literature Retrieval

1. Background

Systematic reviews, i.e., comprehensive literature reviews on narrowly focused questions, constitute the highest form of evidence in medicine. They are central to decision-making across health and medicine, including the regulatory approval of new treatments, the development of clinical guidelines, and the formulation of institutional and governmental health policies. However, creating a systematic review cost upwards of €130,000 and may take over a year to complete [1].

Unfortunately, it is exceedingly common for a systematic review to be outdated by the time it is published. As medical studies appear at a rate of roughly two per minute [2], the rigorous procedures that ensure systematic reviews' comprehensiveness and accuracy prevent medical experts from matching this pace. To foster automation research, the CLEF Technology Assisted Reviews in Empirical Medicine track (CLEF TAR) introduced shared tasks on systematic review creation [3, 4, 5], primarily targeting the initial version of a review. However, since medical experts ultimately bear responsibility for its accuracy, and automatically created systematic review has to be thoroughly validated to ensure that no relevant evidence has been overlooked nor misrepresented.

With our shared task, we propose a different approach by focusing on *updating* existing systematic reviews. Searching for related work, we found only a single study [6] that uses the only available dataset for systematic review updates [7]. That study investigates how to update the Boolean query used for study retrieval to include relevant studies missed by the original query. In addition to this task, our shared task also considers identifying *when* a review should be updated and classifying if a retrieved study should be included. These three tasks align well with the topics of the SCOLIA workshop, and we have designed them to cater to the varying interests of the participants of SCOLIA; not just in terms of natural language processing and information retrieval, but scientometrics/bibliometrics in general.

SCOLIA 2026: The Second International Workshop on Scholarly Information Access
pierre.achkar@uni-leipzig.de (P. Achkar); tim.gollub@uni-weimar.de (T. Gollub); martin.potthast@uni-kassel.de (M. Potthast); carsten.eickhoff@uni-tuebingen.de (C. Eickhoff); harrisen.scells@uni-tuebingen.de (H. Scells)

2. Task Overview

Task 1: Review Update Prediction The first task concerns deciding *when* to update a review. This decision is typically just time-based [8] (i.e., after n months). Recommended update intervals depend on several factors, including the stakeholders affected by the review, the medical discipline, the degree of uncertainty in the evidence, and the urgency of the clinical question [9]. This approach risks choosing intervals that are either too short, wasting effort when no relevant new studies exist, or too long, allowing the evidence base to become outdated. Figure 1 illustrates the acuteness of this problem: many studies are published long before a review is updated.

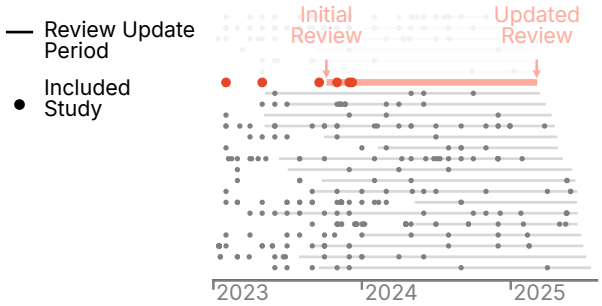

Figure 1: A selection of systematic reviews from the task dataset, highlighting the time between successive updates and the publication dates of studies included in each update.

This task corresponds to the phase of the systematic review process in which reviewers decide *when* an update is warranted. Since there is little methodological research to guide update timing, methods developed in this shared task have a good chance to impact future guidelines for systematic review updates. Participants will develop approaches that, given a systematic review topic and a collection of studies, predict a point in time such that all new studies would be captured by an updated review.

Task 2: Study Retrieval The second task concerns retrieving relevant studies for a review update. Typically, this work is delegated to an information specialist, namely a librarian who collaborates with the review team, who crafts a complex Boolean query. Revising the previously used query when updating a review has been emphasized as an important methodological practice [10], and our preliminary analysis in Figure 1 empirically supports this point: many studies published before the time periods marked by the grey horizontal lines were not retrievable with the previous version of the search strategy. In these cases, updating the query refocused the review topic to enable the retrieval of the new studies not captures by the previously used query. As already mentioned, methodological research on revising reviews remains scarce [11], and only a *single study* has explored computational approaches to query updating [6].

This task corresponds to the study retrieval phase of a systematic review. Since the CLEF TAR tracks featured a related retrieval task, participants familiar with those tracks should have an easy entry point. However, our setting specifically targets retrieving the additional studies that should be included in a review *update*, allowing participants to exploit knowledge about the studies have been included in a review's initial version. Participants will develop Boolean queries, either manually or via automatic approaches [12], that, given a systematic review topic and a collection of studies, maximize recall and precision, ideally retrieving exactly the studies that are ultimately included.

Task 3: Study Classification The third task involves classifying retrieved studies as to-be-included in the review update. In current practice, determining which newly retrieved studies should be included into an updated review is typically done either through manual screening [13] or via bespoke machine learning methods trained on a per-review basis [14].

This task corresponds to the study screening phase of a systematic review. Since the CLEF TAR tracks also included a screening task, participants familiar with those tracks should find the setup familiar. Our variant, however, explicitly allows participants to exploit information about studies included in the initial version of the review. Participants will develop classification systems, either human-in-the-loop or fully automatic, that predict whether a retrieved study should be included in the updated review.

2.1. Connection to SCOLIA

This shared task fits neatly with the goals of the SCOLIA workshop, targeting practitioners in natural language processing and information retrieval who work on the analysis of scientific (biomedical) documents. It also addresses problems in academic literature processing that have received little research attention. We expect participants will find the tasks engaging, while also contributing new methods toward open challenges in the field.

In particular, this shared task directly targets several topics SCOLIA aims to address, including the construction of scholarly information access systems, user models and collections, and quality assurance and scientific integrity (i.e., filtering high-quality research papers for inclusion in reviews).

2.2. Artefacts Expected by Participants

The following artefacts that we expect participants to submit will be collected via the TIRA platform [15]. TIRA also enables participants to submit their code and working systems in addition to runs, supporting reproducibility.

	Task	Artefacts	Notes
■	Review Update Prediction	Plain text file containing a single date in the form `YYYY-MM-DD`.	We will use the predicted date to retrieve all studies published between the initial review and the update by applying a single date filter to the document collection.
■	Study Retrieval	Query used for retrieval and TREC run file containing retrieval results.	We will provide a tool to validate that the submitted queries correspond to the submitted TREC run files.
■	Study Classification	Plain text file where each line in the file corresponds to a study classified as included to the review.	We will provide a tool to validate that submitted results are compatible with our evaluation setup.

2.3. Evaluation Setup

Datasets For this task, we acquired an open-access subset of Cochrane systematic reviews that includes review updates. While we cannot distribute the raw data due to licensing constraints, we can use the reviews' associated metadata (e.g., bibliographic information) to construct an open test collection tailored to this task.

In total, the dataset comprises 35 systematic review updates (10 more topics than the next-largest systematic review update dataset [7]).[1] This scale is comparable to that of the CLEF TAR tracks. We will provide topic data similar to CLEF TAR (e.g., title, query, included docids). In our setting, each topic is split into two versions such that the docids of the updated version are withheld from participants initially. This design allows participants to leverage information about the studies included in the initial version when tackling each task.

The document collection will be based on a baseline PubMed dump that we will index for participants. The index can be used locally via pybool_ir [16], which is available as a Docker image and requires minimal setup. The same index will also be accessible through an API, allowing participants to retrieve documents and produce TREC runs directly. If participants prefer not to use pybool_ir or the provided API, the index is a standard Lucene index that can be accessed through many other toolchains, e.g.,

[1] We will actively investigate if frontier models have memorized our data and develop corresponding baselines.

PyTerrier.[2] Constraining the document collection to this index ensures reproducibility and enables others to continue working on the tasks after the shared task has concluded.

Measures

- Task 1: A TREC run file derived from the submission and evaluated in terms of recall and precision.
- Task 2: The submitted TREC run file will be evalauted in terms of recall and precision.
- Task 3: The result file will be evaluated in terms of accuracy and micro/macro F1

2.4. Preliminary Timeline

We plan to run the shared task in the second half of 2026, starting in late July and ending in late October. The three tasks will be released in phases, each running for approximately one month. A visual timeline of this plan is shown in Figure 2. This phased setup is intended to let participants incrementally build their systems over the course of the shared task. With the data release for Task 1, we will provide the topics with their initial review versions, along with metadata such as review titles and the list of included studies. For Task 2, we will additionally release the original Boolean queries used to retrieve studies for the initial review. For Task 3, we will provide the set of documents retrieved by the Boolean query for the updated review. Each task can be attempted independently, with no dependencies on earlier phases, so participants may choose to tackle any subset of the tasks.

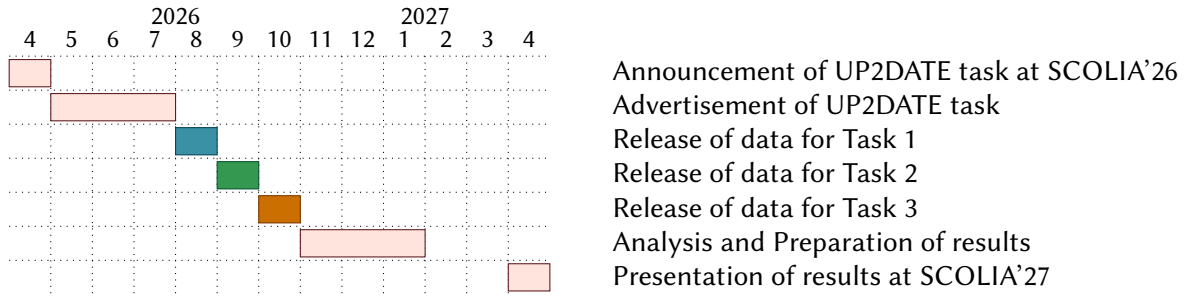

Figure 2: Preliminary timeline of the shared task.

3. Details on Organisers

The organizing committee, listed below, comprises experienced members, most of whom have co-organized multiple shared tasks in the past, including the PAN workshop series since 2007 and Touché since 2020. Other co-organized tasks include the SCAI'21 task on Conversational Question Answering, the SemEval-2019 task on Hyperpartisan News Detection, and the SemEval-2023 task on Human Value Detection. All members are familiar with, or actively working on, challenges in systematic reviews and information access for scientific literature, and have experience developing test collections for information retrieval and natural language processing.

Pierre Achkar is a PhD student at Leipzig University and Fraunhofer ISI. His research focuses on information access in scientific literature.

Tim Gollub is a postdoctoral researcher at Bauhaus-Universität Weimar. He is one of the main developers of the IR-Anthology project and has co-organized several shared tasks in the past, including at SemEval and CLEF.

Martin Potthast is a Professor at the Unviersity of Kassel. He has a wealth of experience organising numerous IR and NLP shared tasks at CLEF and SemEval.

[2] https://pyterrier.readthedocs.io/en/latest/ext/pyterrier-anserini/index.html

Carsten Eickhoff is a Professor at the University of Tübingen. He is an expert in biomedical IR and NLP and has experience organising shared tasks at venues like CLEF.

Harrisen Scells is an Assistant Professor at the University of Tübingen. He has a great amount of experience working on problems at the intersection of systematic reviews and IR and has experience organising shared tasks at venues like CLEF.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check, paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Michelson, K. Reuter, The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials, Contemp Clin Trials Commun 16 (2019) 100443. doi:10.1016/j.conctc.2019.100443.

[2] J. Novoa, M. Chagoyen, C. Benito, F. J. Moreno, F. Pazos, Pmidigest: Interactive review of large collections of pubmed entries to distill relevant information, Genes 14 (2023). doi:10.3390/genes14040942.

[3] E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, Clef 2017 technologically assisted reviews in empirical medicine overview., in: CLEF, 2017. URL: https://ceur-ws.org/Vol-1866/invited_paper_12.pdf.

[4] E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, Clef 2018 technologically assisted reviews in empirical medicine overview., in: CLEF, 2018. URL: https://ceur-ws.org/Vol-2125/invited_paper_6.pdf.

[5] E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, Clef 2019 technology assisted reviews in empirical medicine overview., in: CLEF, 2019. URL: https://ceur-ws.org/Vol-2380/paper_250.pdf.

[6] A. Alharbi, M. Stevenson, Refining boolean queries to identify relevant studies for systematic review updates., J. Am. Medical Informatics Assoc. (2020) 1658–1666. URL: https://doi.org/10.1093/jamia/ocaa148. doi:10.1093/JAMIA/OCAA148.

[7] A. Alharbi, M. Stevenson, A dataset of systematic review updates., in: SIGIR, 2019, pp. 1257–1260. URL: https://doi.org/10.1145/3331184.3331358. doi:10.1145/3331184.3331358.

[8] J. Elliott, A. Synnot, T. Turner, M. Simmonds, E. A. Akl, S. McDonald, G. S. *et al.*, Living systematic review: 1. introduction? the why, what, when, and how, Journal of Clinical Epidemiology 91 (2017) 23–30. URL: https://www.sciencedirect.com/science/article/pii/S0895435617306364. doi:https://doi.org/10.1016/j.jclinepi.2017.08.010.

[9] S. McDonald, S. Sharp, R. L. Morgan, M. H. Murad, D. Fraile Navarro, Methods for living guidelines: early guidance based on practical experience. paper 4: search methods and approaches for living guidelines, Journal of Clinical Epidemiology 155 (2023) 108–117. URL: https://www.sciencedirect.com/science/article/pii/S0895435622003481. doi:https://doi.org/10.1016/j.jclinepi.2022.12.023.

[10] P. Garner, S. Hopewell, J. Chandler, H. MacLehose, E. A. Akl, J. Beyene, S. Chang, R. Churchill, K. Dearness, G. Guyatt, C. Lefebvre, B. Liles, R. Marshall, L. Martínez García, C. Mavergames, M. Nasser, A. Qaseem, M. Sampson, K. Soares-Weiser, Y. Takwoingi, L. Thabane, M. Trivella, P. Tugwell, E. Welsh, E. C. Wilson, H. J. Schünemann, When and how to update systematic reviews: consensus and checklist, BMJ (2016). URL: https://www.bmj.com/content/354/bmj.i3507. doi:10.1136/bmj.i3507.

[11] M. Sampson, J. McGowan, Inquisitio validus index medicus: A simple method of validating medline systematic review searches, Research Synthesis Methods 2 (2011) 103–109. doi:10.1002/jrsm.40.

[12] S. Wang, H. Scells, B. Koopman, G. Zuccon, Reassessing large language model boolean query generation for systematic reviews., in: SIGIR, 2025, pp. 3296–3305. URL: https://doi.org/10.1145/3726302.3730329. doi:10.1145/3726302.3730329.

[13] J. Thomas, A. Noel-Storr, I. Marshall, B. Wallace, S. McDonald, C. Mavergames, P. G. *et al.*, Living systematic reviews: 2. combining human and machine effort, Journal of Clinical Epidemiology 91 (2017) 31–37. URL: https://www.sciencedirect.com/science/article/pii/S0895435617306042. doi:https://doi.org/10.1016/j.jclinepi.2017.08.011.

[14] M. Sood, S. Sharp, E. McFarlane, R. Willans, K. Hopkins, J. Karpusheff, F. Glen, Managing the evidence infodemic: automation approaches used for developing nice covid-19 living guidelines, medRxiv (2022) 2022–06. doi:10.1101/2022.06.13.22276242.

[15] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous integration for reproducible shared tasks with tira.io., in: ECIR, 2023, pp. 236–241. URL: https://doi.org/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[16] H. Scells, M. Potthast, pybool_ir: A toolkit for domain-specific search experiments., in: SIGIR, 2023, pp. 3190–3194. URL: https://doi.org/10.1145/3539618.3591819. doi:10.1145/3539618.3591819.