Living Labs for Academic Search at CLEF 2020

Philipp Schaer[1(✉)], Johann Schaible[2], and Bernd Müller[3]

[1] TH Köln - University of Applied Sciences, Cologne, Germany
philipp.schaer@th-koeln.de
[2] GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany
johann.schaible@gesis.org
[3] ZB MED - Information Centre for Life Sciences, Cologne, Germany
muellerb@zbmed.de

Abstract. The need for innovation in the field of academic search and IR, in general, is shown by the stagnating system performance in controlled evaluation campaigns, as demonstrated in TREC and CLEF meta-evaluation studies, as well as user studies in real systems of scientific information and digital libraries. The question of what constitutes relevance in academic search is multi-layered and a topic that drives research communities for years. The Living Labs for Academic Search (LiLAS) workshop has the goal to inspire the discussion on research and evaluation of academic search systems by strengthening the concept of living labs to the domain of academic search. We want to bring together IR researchers interested in online evaluations of academic search systems and foster knowledge on improving the search for academic resources like literature, research data, and the interlinking between these resources. The employed online evaluation approach based on a living lab infrastructure allows the direct connection to real-world academic search systems from the life sciences and the social sciences.

Keywords: Evaluation · Living labs · Academic search · CLEF

1 Introduction

The search for scientific information is a challenge that the field of Information Retrieval has been dealing with for many years in the unique environment of digital libraries or academic search systems. Its roots go back to the fundamental papers of IR research, such as the early Cranfield studies where Cleverdon et al. conducted their experiments in the context of scientific document retrieval. Even today, the search for academic material is a broad field of research with international conferences such as the ACM/IEEE Joint Conference on Digital Libraries (JCDL) or the Theories and Practices in Digital Libraries (TPDL), which always have a strong IR-specific background.

The need for innovation in this field of academic search and IR, in general, is shown by the stagnating system performance in controlled evaluation campaigns, as demonstrated in TREC and CLEF meta-evaluation studies [1,11], as

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 580–586, 2020.
https://doi.org/10.1007/978-3-030-45442-5_75

well as user studies in real systems of scientific information and digital libraries. Even though large amounts of data are available in highly specialized subject databases (such as ArXiV or PubMed), digital libraries (like the ACM Digital Library), or web search engines (Google Scholar or SemanticScholar), many user needs and requirements remain unsatisfied. The central concern is to find both relevant and high-quality documents - if possible, directly on the first result page - for the actual user of an academic search system. The question of what constitutes relevance in academic search is multi-layered [4] and a topic that drives research communities like the Bibliometrics-enhanced Information Retrieval (BIR) workshops [10].

The Living Labs for Academic Search (LiLAS) workshop would like to inspire the discussion, research, and evaluation of academic search systems by strengthening the concept of living labs to the underrepresented domain of academic search. To do so, the CLEF 2020 workshop lab is about to bring together IR researchers interested in the online evaluation of academic search systems. The goal is to foster knowledge on improving the search for academic resources like literature (ranging from short bibliographic records to full-text papers), research data, and the interlinking between these resources. The employed online evaluation approach based on a living lab infrastructure allows the direct connection to real-world academic search systems from the Life Sciences and the Social Sciences.

The unique feature of the proposed living lab [6] is that the evaluations can be carried out in a productive online system. Contrary to the usual TREC tasks, no expert judgments are used for the evaluation of the information systems and the underlying retrieval systems. The lab employs A/B tests or more complex interleaving procedures to present the experimental systems to the users. However, this requires that the developers of the experimental systems also have access to the production systems, which is rarely the case. Living labs make this possible by bringing platform operators and researchers together. At the same time, they provide a methodological and technical framework for online experiments.

From a broader perspective, the motivation behind this lab is to (a) bring together interested researchers, (b) advertise the online evaluation campaign idea, and (c) develop ideas, best practices, and guidelines for a full online evaluation campaign at CLEF 2021.

2 Background of the LiLAS Workshop

We describe two aspects that form the basis of the LiLAS workshop at CLEF: (1) Evaluating IR systems based on the so-called living labs paradigm and (2) the focus on academic search.

2.1 Living Lab Evaluations at CLEF and TREC

The Living Labs for Information Retrieval (LL4IR) and Open Search initiatives were launched as part of the TREC and CLEF evaluation campaigns to promote

cooperation between industrial and academic research in the field of living labs. As part of these initiatives, an initial evaluation infrastructure was created in the form of an API[1] that allows academic researchers to access the search systems of other platforms. A hard requirement is that the platforms actively participate in the initiative and have implemented the corresponding API for their systems. The API developed under a free license is publicly available and represents the greatest success of the campaign to date. Within CLEF, LL4IR was first organized as a lab in the form of a challenge at the CLEF 2015 conference series and continued with TREC OpenSearch 2016 and 2017 [2].

Another living lab campaign that ran at CLEF was NewsREEL that provides the possibility to predict user interactions in an offline dataset as well as to recommend news articles in real-time. Researchers could evaluate their models either with a test collection (offline) or by delivering the recommended content to real users of the news publishing platforms (online) via the Open Recommendation Platform (ORP)[2]. Since NewsREEL joined the MediaEval Benchmarking Initiative for Multimedia Evaluation[3] in 2018, the evaluation setup comprises offline evaluations only [8].

In a nutshell, the living lab evaluation paradigm represents a user-centric study methodology for researchers to evaluate the performance of retrieval systems within real-world applications. Thus, it offers a more realistic experiment and evaluation environment as offline test collections, and therefore should be further investigated to raise IR-evaluation to the next level.

2.2 Academic Search Evaluation

Since the early days of the Cranfield experiments, scientific retrieval has developed and improved relatively little in comparison and contrast to everyday web search, commercial search, or social network search. The common interest in academic search seems small compared to these scenarios. Many information systems are based on bibliographical metadata (e.g., WoS, Scopus, PubMed) and do not make use of full-text information, which in most cases is not available due to licensing hiccups of the academic publishing system. Even when full-text is searchable, the characteristics of scientific communication are not thoroughly operationalized and are even ignored. This leads to simple search systems that still heavenly rely on Boolean Retrieval techniques to get the most out of the available (meta-)data. Although during the last years, the BIR community added much traction for the overall topic of academic search, it is not focusing on the rigorous evaluation of proposed methods. The lack of an overall evaluation scenario leads to little understanding of the consequences for the different domains and fields of science.

[1] https://bitbucket.org/living-labs/ll-api, all online resources were checked for validity in January 2020.
[2] https://orp.plista.com/.
[3] http://www.multimediaeval.org/mediaeval2018/newsreelmm/.

Within TREC and CLEF, only very few tracks or labs focused on the evaluation of academic search systems, although some made use of scientific documents or use-cases to generate test collections. An example might be the CLEF Domain-specific track [7] (an offline evaluation lab). In this track, a corpus of bibliographic records and research project descriptions from the Social Sciences was created to test the special needs of scientific retrieval tasks. As stated by the organizers: "General-purpose news documents require very different search criteria than those used for reference retrieval in databases of scientific literature items, and also offer no possibility for comparable test runs with domain-specific terminology." The DS track ran for many years and developed a total of four versions of the GIRT test collection of scientific reference documents. More recently, the TREC Precision Medicine/Clinical Decision Support Track released a large test collection in 2016 based on 1.2 million open access full-text documents from PubMedCentral. Another test collection from the sciences is iSearch [9]. Both collections contain citation data.

3 Outline of the LiLAS Workshop

3.1 Roadmap for CLEF 2020 and Beyond

The timeline for LiLAS starts with a workshop lab at CLEF 2020 to kick-off the evaluation lab that is planned for CLEF 2021. The concrete details of the 2021 online experiments, like tasks, metrics, and technologies, are to be discussed at the workshop, but right now, we favor a Docker-based container infrastructure setup that is briefly described in Sect. 4.

The LiLAS workshop is designed to be a blend of the most successful parts of LL4IR, NewsREEL, and the DS track. The DS track had a strong focus on scientific search, thesauri, and multilingual search. NewsREEL had an active technological component, and LL4IR turned from product search to academic search but was not able to implement the scientific focus into the last iteration. There is much potential that is still not used in the question of how to online evaluate scientific search platforms.

In retrospective, the entrance barriers for LL4IR and NewsREEL for new participants were quite high. This was due to the unusual evaluation schema that raised many questions and might have put off some interested groups. To surpass these issues, we would like to offer different ways to participate in LiLAS: Simple non-technical submissions of pre-computed results and Docker-based containers that offer more evaluation possibilities but require the participants to develop a good-performing system. Our blueprint for this approach is TIRA (short for Testbed for Information Retrieval Algorithms), as it has proven itself to be a reliable infrastructure, although it does not support living lab functionalities.

As the lab specifications for 2021 are still open for discussion, we would like to use the workshop to address some technical and conceptional ideas for this. Questions are how to allow a low barrier entrance to this campaign (e.g., in the form of ready to use templates, reusing Docker containers from campaigns like OSIRRC, or offering tutorial sessions at ECIR or other venues). The unique

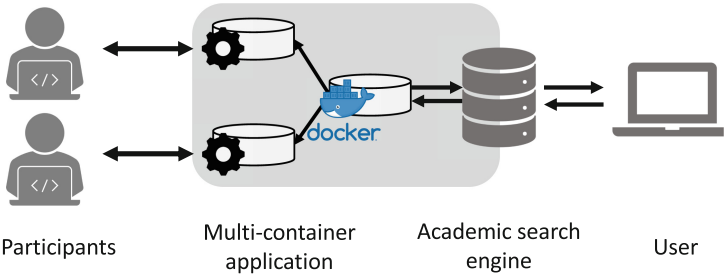

Fig. 1. Infrastructure design for an online living lab evaluation with experimental academic search systems: Participants package their systems with the help of Docker containers that are deployed in the backend of academic search engines. Users retrieve results from these systems and deliver feedback for later evaluation.

selling points of this lab should be advertised and communicated very clearly so that new participants see the potential both in the available data sets (literature *and* research data) and the possibility to have an online evaluation for their retrieval and recommendation approaches.

To prepare the workshop, we will release some sample data sets from the scientific search systems LIVIVO[4] and GESIS-wide Search [5][5] and some Docker templates to allow early adopters to implement first prototypes. At the workshop, we would like to have these early adopters who took part in this *open beta phase* to present their first-hand experiences to lay a foundation for a full-size campaign in 2021.

4 Evaluation Infrastructure

The evaluation infrastructure for this lab is called STELLA [3]. Figure 1 shows the general structure behind it. It would incorporate usage feedback like click-through rates and would allow participating research groups to package their systems with the help of Docker containers. These containers are deployed in the backend of academic search engines. Users search and interact with these systems and deliver feedback for later evaluation.

The infrastructure is based on the container virtualization Docker, which allows us to run and test different research algorithms on productive systems in a so-called multi-container environment. STELLA's main component is a central Living Lab API that connects data and content providers (sites) with research prototypes (participants) encapsulated in Docker containers. In addition to embedding in production systems and accessing their data, this solution also allows the backup of usage data (e.g., clickthrough or download data), which are necessary for an evaluation of the systems.

[4] https://livivo.de.
[5] https://search.gesis.org/.

Although STELLA and its predecessors, like the LL4IR API, have already been successfully evaluated, the software is far from being a fully-fledged production system. Through the discussion with future participants and other actors in the field, like the TIRA maintainers, we hope to achieve a more stable code base. Another practical benefit will result from the common requirements but different areas of application (online vs. offline evaluation). Future integration of the systems provides a unique opportunity to evaluate in two different stages of the systems: (1) Systems could measure themselves against static test collections, and later (2) prove themselves in online scenarios according to their basic performance. Finally, this would allow generating new components for future test collections through the recorded usage data.

This approach would surpass some of the shortcomings of the previous labs' implementations. The main advantages would be:

- No focus on head queries (the system's top-k most common queries)
- Retrieval and recommendation tasks within the same platform
- A technological platform that is built on open-source frameworks like Docker that allows easy distribution of implementation templates (like in the SIGIR 2019 OSIRRC workshop[6] that would lower the entrance barrier and foster reproducibility).

5 Outlook

We see academic search as a broader term for scientific and especially domain-specific retrieval tasks, which comprises document, dataset as well as bibliometric-enhanced retrieval. As web-search platforms like Google Scholar (or Google Dataset Search) or proprietary digital libraries like the ACM Digital Library are not open to public research and do not offer any domain-specific features, we focus on mid-size scientific search systems that offer domain-specific resources and use-cases. This focus allows for using many specific information types like bibliographic metadata, usage data, download rates, or citations in order to develop and evaluate innovative search applications. We see a clear connection to the Bibliographic-enhanced Information Retrieval (BIR) community that runs a series of successful workshops since 2013 at venues like ECIR or SIGIR. A proposed SMART lab on "Scientific mining and retrieval"[7] might be a good starting point to join forces. The common point of interest might be to create an integrated test collection for academic search that contains academic material and their citations, is useful for the scientific retrieval process, and is suitable for offline and online evaluation as well.

[6] https://github.com/osirrc.
[7] https://www.scientometrics-school.eu/images/esss_Programme2019.pdf.

References

1. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: Improvements that don't add up: ad-hoc retrieval results since 1998. In: Proceeding of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, pp. 601–610. ACM (2009). https://doi.org/10.1145/1645953.1646031
2. Balog, K., Schuth, A., Dekker, P., Schaer, P., Tavakolpoursaleh, N., Chuang, P.Y.: Overview of the TREC 2016 open search track. In: Proceedings of the Twenty-Fifth Text REtrieval Conference (TREC 2016). NIST (2016)
3. Breuer, T., Schaer, P., Tavalkolpoursaleh, N., Schaible, J., Wolff, B., Müller, B.: STELLA: towards a framework for the reproducibility of online search experiments. In: Proceedings of The Open-Source IR Replicability Challenge (OSIRRC) @ SIGIR (2019)
4. Carevic, Z., Schaer, P.: On the connection between citation-based and topical relevance ranking: results of a pretest using iSearch. In: Proceedings of the First Workshop on Bibliometric-Enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval (ECIR 2014), Amsterdam, The Netherlands, 13 April 2014. CEUR Workshop Proceedings, vol. 1143, pp. 37–44. CEUR-WS.org (2014). http://ceur-ws.org/Vol-1143/paper5.pdf
5. Hienert, D., Kern, D., Boland, K., Zapilko, B., Mutschke, P.: A digital library for research data and related information in the social sciences. In: 19th ACM/IEEE Joint Conference on Digital Libraries, JCDL 2019, Champaign, IL, USA, 2–6 June 2019, pp. 148–157. IEEE (2019). https://doi.org/10.1109/JCDL.2019.00030
6. Kelly, D., Dumais, S., Pedersen, J.O.: Evaluation challenges and directions for information-seeking support systems. Computer **42**(3), 60–66 (2009)
7. Kluck, M., Gey, F.C.: The domain-specific task of CLEF - specific evaluation strategies in cross-language information retrieval. In: Peters, C. (ed.) CLEF 2000. LNCS, vol. 2069, pp. 48–56. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44645-1_5
8. Lommatzsch, A., Kille, B., Hopfgartner, F., Ramming, L.: Newsreel multimedia at mediaeval 2018: news recommendation with image and text content. In: Working Notes Proceedings of the MediaEval 2018 Workshop. CEUR-WS (2018)
9. Lykke, M., Larsen, B., Lund, H., Ingwersen, P.: Developing a test collection for the evaluation of integrated search. In: Gurrin, C., et al. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 627–630. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12275-0_63
10. Mayr, P., Scharnhorst, A., Larsen, B., Schaer, P., Mutschke, P.: Bibliometric-enhanced information retrieval. In: de Rijke, M., et al. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 798–801. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06028-6_99
11. Yang, W., Lu, K., Yang, P., Lin, J.: Critically examining the "Neural Hype": weak baselines and the additivity of effectiveness gains from neural ranking models. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 2019, Paris, France, pp. 1129–1132. ACM Press (2019). https://doi.org/10.1145/3331184.3331340