Evaluation Infrastructures for Academic Shared Tasks: Requirements and Concept Design for Search and Recommendation Scenarios

Schaible, Johann; Breuer, Timo; Tavakolpoursaleh, Narges; Müller, Bernd; Wolff, Benjamin; Schaer, Philipp

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 407518790 / Funded by the German Research Foundation (DFG) - Project number 407518790

Empfohlene Zitierung / Suggested Citation:
Schaible, J., Breuer, T., Tavakolpoursaleh, N., Müller, B., Wolff, B., & Schaer, P. (2020). Evaluation Infrastructures for Academic Shared Tasks: Requirements and Concept Design for Search and Recommendation Scenarios. *Datenbank-Spektrum : Zeitschrift für Datenbanktechnologien und Information Retrieval*, *20*(1), 29-36. https://doi.org/10.1007/s13222-020-00335-x

Datenbank Spektrum (2020) 20:29–36
https://doi.org/10.1007/s13222-020-00335-x

SCHWERPUNKTBEITRAG

Evaluation Infrastructures for Academic Shared Tasks

Requirements and Concept Design for Search and Recommendation Scenarios

Johann Schaible[1] · Timo Breuer[3] · Narges Tavakolpoursaleh[1] · Bernd Müller[2] · Benjamin Wolff[2] · Philipp Schaer[3]

Received: 15 October 2019 / Accepted: 21 January 2020 / Published online: 7 February 2020

Abstract
Academic search systems aid users in finding information covering specific topics of scientific interest and have evolved from early catalog-based library systems to modern web-scale systems. However, evaluating the performance of the underlying retrieval approaches remains a challenge. An increasing amount of requirements for producing accurate retrieval results have to be considered, e.g., close integration of the system's users. Due to these requirements, small to mid-size academic search systems cannot evaluate their retrieval system in-house. Evaluation infrastructures for shared tasks alleviate this situation. They allow researchers to experiment with retrieval approaches in specific search and recommendation scenarios without building their own infrastructure. In this paper, we elaborate on the benefits and shortcomings of four state-of-the-art evaluation infrastructures on search and recommendation tasks concerning the following requirements: support for online and offline evaluations, domain specificity of shared tasks, and reproducibility of experiments and results. In addition, we introduce an evaluation infrastructure concept design aiming at reducing the shortcomings in shared tasks for search and recommender systems.

Keywords Evaluation Infrastructures · Domain-Specific Academic Search · Online/Offline Evaluation · Reproducibility

1 Introduction

Academic search systems help users to find scholarly resources relevant to specific scientific information needs. Users of academic search systems are typically (upcoming) scientists or domain experts whose search patterns are different from general search system customers. Web search users are often non-experts, whereas users of academic search systems (usual experts) have a higher level of domain knowledge. Therefore, they assess the quality of retrieved items differently by making connections, triangulating sources, or assessing their source [22]. Domain knowledge of expert users in academic search systems does have a positive impact on information search performance [18]. This suggests reasons to further look into the domain of academic search as a domain-specific, group-centric and highly specialized field of research. Large scale user studies on academic users showed how their search patterns are influenced by factors such as discipline or academic position [20]. Biologists are different from sociologists regarding their information needs, and professors have different search patterns than students. Thus, evaluating search and

✉ Johann Schaible
johann.schaible@gesis.org

Timo Breuer
timo.breuer@th-koeln.de

Narges Tavakolpoursaleh
narges.tavakolpoursaleh@gesis.org

Bernd Müller
muellerb@zbmed.de

Benjamin Wolff
wolff@zbmed.de

Philipp Schaer
philipp.schaer@th-koeln.de

[1] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany

[2] ZB MED – Information Centre for Life Sciences, Cologne, Germany

[3] TH Köln – University of Applied Sciences, Cologne, Germany

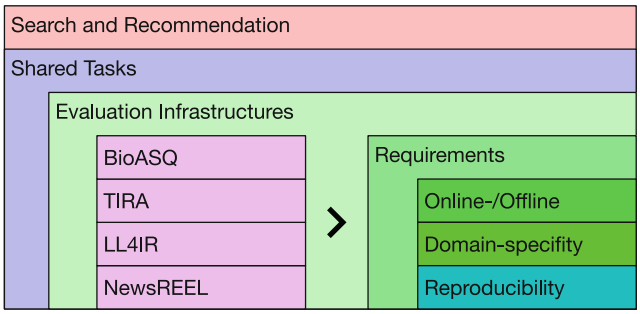

Fig. 1 Evaluation Infrastructures of shared tasks. Evaluating academic search with shared tasks, requires to investigate to which extent recent evaluation infrastructures employ online/offline evaluations, domain-specificity, and allow to reproduce an evaluation

recommendation task performance in a user-centric manner is a significant aspect in order to verify and improve retrieval results in a structured way. This leads to the need for rather sophisticated evaluation approaches for which specific evaluation infrastructures are beneficial [13, 8].

Evaluation infrastructures typically employ *shared tasks* giving researchers access to data or even to entire systems. Among others, four prominent state-of-the-art evaluation infrastructures are BioASQ [27], TIRA [11], the Living Labs for Information Retrieval (LL4IR and also known as the OpenSearch campaign) [17], and NewsREEL [15]. The former three correspond to search systems for academic domains, while the latter is a news articles recommendation system. Information Retrieval (IR) researchers (in the following also referred to as *participants*) use the infrastructures to obtain data from search systems (in the following referred to as *sites*) for developing their algorithms and submitting the retrieval results or entire algorithm implementations. Due to the importance of the academic domain, the close integration of the sites' users, and the validation of retrieval results, the IR community recently puts a focus on three requirements: (i) online/offline evaluations, (ii) domain specificity in IR, and (iii) the reproducibility of evaluation experiments and results. Fig. 1 illustrates the general stack of evaluating retrieval algorithms using shared tasks in academic search.

In this paper, we elaborate on evaluation infrastructures for search and recommendation tasks serving the academic community with respect to the requirements (i) to (iii). We focus on infrastructures for open academic search that have consistently been subject to open evaluation campaigns in the scientific community, such as those at the Conference and Labs of the Evaluation Forum (CLEF)[1] or the Text REtrieval Conference (TREC)[2]. We exclude web search systems, such as Google or Bing, or proprietary academic search systems, like Google Scholar or Mendeley. Therefore, we consider BioASQ, TIRA, LL4IR, and NewsREEL in our elaboration, as they have been subject to previous major open living labs campaigns.

We first discuss the requirements (i) to (iii) and their importance in retrieval evaluations (cf. Sect. 2), which is followed by a brief description of BioASQ, TIRA, LL4IR, and NewsREEL (cf. Sect. 3). A detailed comparison between these evaluation infrastructures and a discussion of their shortcomings with respect to the three requirements is provided in Sect. 4. Subsequently, we introduce a novel evaluation infrastructure concept design targeting to reduce existing shortcomings (cf. Sect. 5), before we conclude the paper (cf. Sect. 6).

2 Requirements

Online/Offline Evaluations. There are various possibilities to evaluate the performance of retrieval algorithms, comprising *offline evaluations* and *online evaluations*.[3] While offline evaluations use pre-defined datasets, so-called *test collections*, for measuring an algorithm's performance, online evaluations are performed on a live system with real users of the system. The experimental retrieval results are shown to the users and the system calculates the results' acceptance by observing the users' interaction with metrics, such as click-through rate (CTR) [14]. Evaluation infrastructures typically employ one of these two methods, and both of them have numerous advantages and shortcomings. On the one hand, offline evaluations can be performed quite fast and without any risks to the live system. However, they are limited in reflecting the current and specific information need as well as user behavior [13]. On the other hand, online evaluations precisely capture the users' information need as well as their interactions with the system. However, IR experts are needed to perform online evaluations, and typically, academic search system providers are not IR experts. Another problem can be posed by unsatisfying retrieval results, which might frustrate the users in a way that they decide to abandon the system. Despite these shortcomings of online evaluations, they receive an increasing interest from the IR community, as observing user interactions is more and more in focus for obtaining reliable retrieval results [3].

Domain Specificity in Evaluations. Open academic search systems exist for various domains, such as medicine and life sciences (e.g., LIVIVO[4]), the social sciences (e.g.,

[1] http://www.clef-initiative.eu/. This and all other online resources were checked for validity in January 2020.

[2] https://trec.nist.gov/.

[3] User studies, explicit feedback systems or surveys are other evaluation approaches, but are not part of this paper.

[4] https://www.livivo.de/.

GESIS-wide Search[5]), and others. These systems comprise literature and research data mainly from its domain, including different domain-specific terminology. Thus, they also have distinct users with different information needs and behavior [12, 20], which poses problems for generalized across-domain IR approaches. A well-performing retrieval algorithm in one discipline does not mean, it is a well-performing algorithm in another domain. Evaluation infrastructures allowing to focus on a specific discipline alleviate this situation. For example, besides general tracks at TREC, there are several domain-specific sessions for the life sciences such as the TREC-Genomics, TREC-Chemistry, and TREC-Med. However, developing and maintaining a single evaluation infrastructure for each discipline is cumbersome, resulting in a high chance for such an infrastructure to disappear after a while. Thus, there is a need for a central evaluation infrastructure which can be used for various independent domain-specific evaluations.

Reproducibility. The requirement for reproducibility in evaluations of retrieval approaches gains growing attention in IR research, as it allows for validating results and comparisons between approaches [9]. At evaluation campaigns of scientific conferences, authors typically submit their results along with a scientific publication describing the retrieval approach. This is not ideal, as usually the approach's implementation as well as additional information is needed to facilitate reproducing the experiment and results. For example, in [6], it is reported that only 7 out of 18 major neural recommendation approaches published at top tier conferences were reproducible, as only limited information was shared by the authors. Following the PRIMAD model [4, 7] would resemble a large step towards reproducibility. It specifies the major components of an experiment (**P**latform, **R**esearch Goal, **I**mplementation, **M**ethod, **A**ctor, **D**ata) and in order to ensure reproducibility, participants should share as many PRIMAD components as possible. For example, standard evaluations campaigns generally specify a research goal, for which data is provided by the cooperating sites. Participants (the actors according to [7]) usually develop their methods with their implementation on their own platform. This poses a large problem for others to reproduce the entire setup. Supporting participants in sharing all components lies in the concept of the Evaluation as a Service (EaaS) [16] paradigm. Besides the research goal and data, this concept facilitates a common platform that participants can use to run the experiments for which they share their method implementation.

3 Evaluation Infrastructures

Evaluation infrastructures are typically a form of a cloud-based system allowing sites and participants for evaluating retrieval algorithms [16]. Sites are the cooperating search system, e.g., LIVIVO or GESIS-wide Search, and provide content and interaction data of their system. Participants, typically some IR research groups, use the sites' data to develop retrieval algorithms and then submit the implementations or just the retrieval result to the evaluation infrastructure.

BioASQ, TIRA, and initiatives concerned with living labs for IR (LL4IR) as well as recommender systems (NewsREEL) mark the current state-of-the-art with respect to the requirements (i)-(iii): online/offline evaluation, domain specificity, and reproducibility. These infrastructures comprise shared tasks that are open for participation during evaluation campaigns and conferences, such as CLEF and TREC. In the folliwing, we describe further details on these infrastructures.

BioASQ. The challenge on biomedical semantic indexing and question answering (BioASQ) [27] aims to improve the indexing process of PubMed[6] articles with the annotation of terms from the Medical Subject Headings (MeSH)[7]. In 2013, the BioASQ challenge was initiated in cooperation with the National Library of Medicine (NLM). BioASQ is split in two independent tasks. Task A (Biomedical Semantic Indexing) requires participants to annotate new PubMed articles with MESH-terms before curators annotate them manually. Task B (Biomedical Semantic QA) asks participants to respond to natural language questions (in English), which reflect real-life information needs, with relevant concepts, articles, or snippets from articles, RDF triples, and exact answers. We consider BioASQ in this work, as it is the leading example of an IR evaluation infrastructure in the field of life sciences, which at the same time is publicly available for participation.

TIRA. The TIRA (Testbed for Information Retrieval Algorithms) framework aims at sharing retrieval algorithms by leveraging the capabilities of the web. More concretely, by following the software as a service principle, TIRA requires its participants to upload not some retrieval results, but the retrieval algorithm itself. Furthermore, it facilitates the involvement of sites and their own shared tasks, i.e., participants and sites publish everything regarding an experiment via this platform. This way, everything regarding the evaluation challenge is in one place. Given one participant's opt-in, their algorithms, as well as the results get published. This makes the activities of the challenge highly visible and reproducible. It is thus a prime example of an

[5] https://search.gesis.org.

[6] https://www.ncbi.nlm.nih.gov/pubmed/.

[7] https://meshb.nlm.nih.gov/search.

Table 1 Evaluation infrastructures with respect to online/offline evaluations, domain specificity, and reproducibility

Infrastructure	Description	Evaluation		Domains	Reproducibility					
		Online	Offline		P	R	I	M	A	D
BioASQ	BioASQ is an initiative for semantic indexing and question answering [27]	◐	●	Biomedical text indexation and Question Answering	○	●	○	●	●	●
TIRA	Underlying platform of the PAN lab dedicated to digital text forensics [23]	○	●	Multi-domain, e.g., text forensics (author profiling), clickbait detection	●	●	●	●	◐	●
LL4IR	Living labs infrastructure for ad-hoc search using interleaving [17]	●	○	Multi-domain, e.g., online shop, academic repositories and search systems	○	●	○	●	●	●
NewsREEL	Living lab infrastructure to evaluate news article recommenders [15]	◐	●	Commercial news articles recommendation	◐	●	○	◐	◐	●

● requirement not fulfilled, ◐ requirement partially fulfilled, ○ requirement fulfilled

EaaS-infrastructure and has been used in various workshops and IR challenges. It has been used for several shared tasks at the CLEF or CoNLL conference. For technical details on the architecture and the underlying concept, we refer the reader to the reports by TIRA's developers [10, 11]. We chose to consider TIRA in our overview, as it marks a milestone in modern evaluation campaigns encouraging open science and reproducibility.

LL4IR. The Living Lab for IR (LL4IR) evaluation represents a rather user-centric study methodology for participants to evaluate their ranking approaches. The living labs paradigm provides a live setting with shared data for participants and enables sites to observe their real users' interactions with the experimental retrieval algorithms. In detail, participants gain access to a site's usage data, like click data and query logs, to develop their retrieval algorithms locally. LL4IR focuses on so-called "head queries", which resemble the most queried terms by a site's users [1]. Participants thus submit the rankings for solely the head queries, which are incorporated in the site's live system. This forms the basis for the evaluation. The retrieval algorithm (i.e., implementation) itself is not submitted. LL4IR was first organized as a lab at CLEF 2015 and continued with TREC OpenSearch 2016 and 2017 [2]. We chose to include LL4IR, since it is the first fully-integrated living lab-based evaluation infrastructure to facilitate online retrieval experiments for academia with real users.

NewsREEL. The News REcommendation Evaluation Lab (NewsREEL) is an evaluation infrastructure for experimenting with news article recommendations. It is designed as a living lab, where participants of a campaign use requested articles to recommend similar articles. The Open Recommendation Platform (ORP) [5], provided by plista, allows the participants to register different recommendation algorithms and evaluate their performance. Other evaluation infrastructures used at campaigns and challenges, such as the ACM RecSys Challenge 2017[8] do also exist, but the sites specifying the challenge objective change almost every year comprising different requirements: XING organized it in 2017, Spotify in 2018, and trivago in 2019. Furthermore, evaluation infrastructures from such companies are not open, lacking transparency of the campaign results. NewsREEL, however, has started in 2013 as the international news recommender systems workshop and challenge at the ACM RecSys conference and has evolved into an evaluation lab of CLEF since then. Although NewsREEL is not per se an evaluation infrastructure for academic search, it is the most open one to academic researchers and the only infrastructure that is consistent in its campaigns and requirements. Thus, we include NewsREEL in our work.

4 Requirement-based Comparison

As mentioned, BioASQ, TIRA, LL4IR, and NewsREEL resemble the current state-of-the-art in evaluating retrieval systems. In the following, we will provide a detailed discussion to which extent they comply to the three requirements (i) online/offline evaluations, (ii) domain specificity, and (iii) the reproducibility of evaluation results. Table 1 provides an overview of this discussion.

4.1 BioASQ

Online/offline evaluation: BioASQ comprises a typical offline evaluation with predicting held out information. Win-

[8] http://www.recsyschallenge.com/2017/.

ning systems from 2014 to 2017 have been made available online, comprising a user interface [21]. There, articles can be further manually labeled, allowing submissions by participants to be evaluated if new annotations become available. This feedback loop allows ongoing evaluations as more assessments by human evaluators become available, which resembles a sort of online evaluation.

Domain specificity: BioASQ is dedicated to the question of answering and annotation of documents from the biomedical domain. Documents have to be indexed with regard to a biomedical ontology concept. The Question-Answering campaign requires to provide domain-specific answers to domain-specific natural language questions. Hence, annotating documents, classifying questions, retrieving relevant text passages, or summarizing contents are all adjusted to the needs in biomedical indexing and QA. This also applies to the utilized datasets in BioASQ as well as to its evaluation tasks.

Reproducibility: BioASQ does not address the reproducibility of experiments, as participants submit their results only, followed by an adjunct publication comprising the method description. Simple and direct reuse of the method and implementations is thus not possible. Regarding PRIMAD, the research goal, the actors (participants and users), the method (via a publication), and the data are given. The platform is not given as well, as there is no need to specify which platform the implementation runs on, and there is no EaaS-platform for the implementation.

4.2 TIRA

Online/Offline evaluation: Evaluations are run in an offline test environment with a "data lock"-mechanism, to prevent a leakage of test data outside the evaluation infrastructure. This encourages sites providing sensitive and/or proprietary data. Conceptually, TIRA does not incorporate online evaluations at the time of writing.

Domain specificity: With regards to domain specificity, TIRA has been used for shared tasks in various domains at conferences, such as CLEF, CoNLL and WSDM [23]. For example, the PAN lab [24] at CLEF is dedicated to digital text forensics ranging from authorship profiling and identification over to style change detection, to name some of the specific tasks.

Reproducibility: TIRA makes it possible to archive and re-execute submitted retrieval algorithms. With the use of virtual machines, participants are not confronted with technical barriers, but can choose the implementation and platform of their method. The algorithm implementation as well as its configuration is submitted by the participants. Therefore, TIRA's design strongly emphasizes the reproducibility of submitted retrieval systems, with the same test collection or another one, and thereby satisfies all PRIMAD components. With the use of VMs, the platform, implementation and method are replicable. Depending on the context, these systems can be rerun with different research goals, data and by different actors. Only the users, as part of the actors-component, is not satisfied, as TIRA does not contain an online evaluation.

4.3 Living Labs for IR Evaluation

Online/offline evaluation: Living labs allow for a realistic online evaluation and form an alternative to offline evaluations with test collections. Sites share their head queries and the corresponding user interactions with the system. The experimental rankings for the head queries are incorporated into the site by the the LL4IR central platform. An offline evaluation is not provided, such that the algorithm cannot be validated before it is tested directly with real users. However, as only head queries comprise the experimental ranking, the risk of frustrating users with unsatisfying ranking is minimized.

Domain specificity: Living Labs for IR has been applied to tasks in different domains: LL4IR at CLEF targeted product search on the Hungarian e-commerce site REGIO Játék. TREC OpenSearch 2016-2017 focused on ad-hoc document retrieval within the social sciences open access repository SSOAR and the academic search engine CiteSeerX [2]. Thereby, participants can evaluate their algorithms in academic search as well as in commercial product search.

Reproducibility: As mentioned, only the rankings for the head queries are submitted. Although all ranking results are saved and can also be compared afterward, the utilized ranking algorithms are only described within the accompanying papers. Implementations are not submitted. Thus, besides the research goal and data, only the method (via a publication) and the actors (participants and users) are the fulfilled PRIMAD components.

4.4 NewsREEL

Online/Offline evaluation: Starting from 2014, NewsREEL provides both an offline and an online evaluation. Participants use a dataset comprising recorded events of news publishing platforms to predict other articles the user might be interested in. Evaluation of the models is performed either with a test collection (offline) or by delivering the recommended content to real users of the news publishing platforms (online) via the the Open Recommendation Platform

(ORP). However, after joining the MediaEval Benchmarking Initiative for Multimedia Evaluation[9] in 2018, NewsREEL comprises offline evaluations only [19].

Domain specificity: As the name "News REcommendation Evaluation Lab" suggests, it is a project for evaluating recommendation algorithms for news articles. The infrastructure could allow other domains to participate, but plista, who provides this infrastructure, is a company for recommendation services for online publishers. It is thus unlikely that they open up their infrastructure for academic search.

Reproducibility: At CLEF 2017, NewsREEL offered two tasks, *NewsREEL Live* and *NewsREEL Replay*. In NewsREEL Live, participants can use plista's VMs for running their recommendation algorithms, which serve as a platform for the implementation. However, this is not mandatory. NewsREEL Replay aims at reproducing the click-through-rates of news recommendations by simulating requests of articles. NewsREEL thus provides a research goal with recommending relevant news articles and the according data. Providing the implementation is not required unless participants want to use plista's VMs, and even for the method, NewsREEL merely asks for working notes only. Since joining MediaEval, users are not part of the Actor-component.

4.5 Summarization and Discussion

In general, we can observe that no evaluation infrastructures offers the possibility for an offline as well as for an online evaluation. NewsREEL used to have both, but ever since being part of the MediaEval campaign, online evaluations are no longer part of NewsREEL. BioASQ does not offer a direct online evaluation but instead employs the winning algorithm into the live system to give the participants a chance to re-adjust the algorithm. This poses the problem that non-winning systems still might perform better than the winning systems in an online evaluation taking real users into account [3].

NewsREEL and BioASQ are limited to specific domains: news articles and the biomedical domain, respectively. They are adjusted to the needs within their domain, such it is either difficult to open the infrastructures to other domain. In the case of plista's platform for news recommendations, it might not even be wanted due to their business model and specialization. Furthermore, recommendations in academia are very likely to be different from news article recommendations, as people read news articles quite differently compared to scientists reading scientific papers. Also, there is a vast amount of new news articles every day that need to be encountered by the employed recommendations system. In an academic search system, such phenomena do not occur. LL4IR and TIRA are open to various domains, depending on which sites take part. For TIRA in specific, there is also no limitation to domains with available open content only, as TIRA's evaluation relies on the previously described "data lock"-mechanism. In the case of LL4IR, their online evaluation requires quite an effort from the sites to provide all needed information and ensure the workflow of head queries and interleaving approaches.

Considering the PRIMAD model, we can observe that TIRA is the only evaluation infrastructure committed to addressing all components, only excluding the user out of the Actor-component, as no online evaluation is provided. Most importantly, all other infrastructures do not encourage submitting the implementation of a retrieval algorithm. This poses an enormous problem for the ability to reproduce entire experiments and retrieval results, as other teams must re-implement the algorithms based on the description in the paper.

5 A Novel Concept for Evaluation Infrastructures

Based on the three requirements (i) providing online and offline evaluations, (ii) enabling domain-specific evaluations, and (iii) allowing for reproducibility, we illustrate a novel concept for evaluation infrastructures that fulfills these three requirements. The concept involves a platform that makes it possible to run offline evaluations with test collections (if available) as well as online evaluations on sites for different domains and makes it possible to reproduce the experiments. Inspired by other initiatives, such as TIRA and LL4IR, the concept facilitates experimenting with submitted implementations in environments with real-world user feedback. The process is depicted in Fig. 2.

At the very core, the concept relies on Docker and its containerization technology. Participants package their experimental search and recommender algorithms with the help of clearly specified Dockerfiles. Single Docker images are integrated into a multi-container application (MCA) that is composed of several experimental systems. Keeping deployment effort low, sites only need to set up the MCA on local servers and integrate a REST-API to retrieve search results and recommendations. In return, the sites have to provide data collections to the MCA. Therefore it is necessary to agree upon a common document structure. Both the sites and the participants must adapt their collections or indexing routines to this common document structure. If necessary, re-indexing can be invoked by an API call that forces systems to update the indices. Eventually, sites send feedback data to the MCA. Feedback has to be logged in

[9] http://www.multimediaeval.org/mediaeval2018/newsreelmm/.

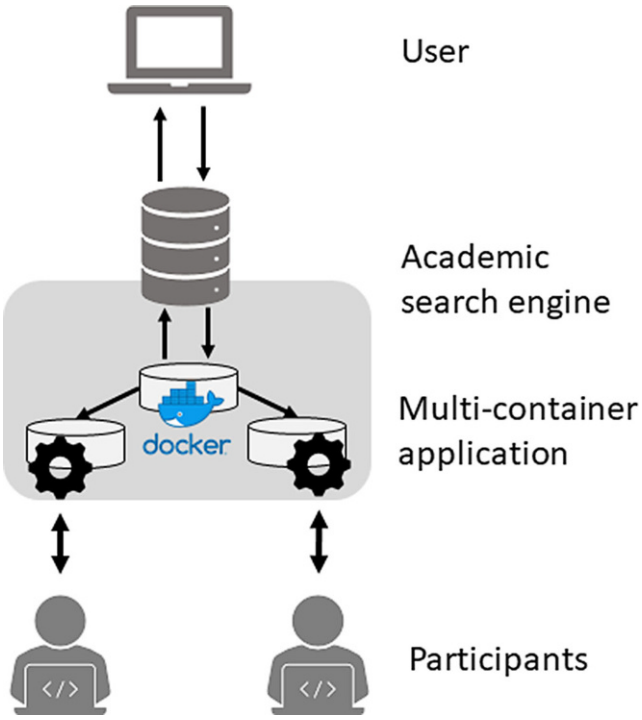

Fig. 2 Infrastructure design concept for online evaluations with experimental search and recommender systems. Participants package their systems with the help of Docker containers that are deployed in the backend of academic search engines. Users retrieve results from these systems and deliver feedback for later evaluation

the conventional form of query results, recommendations, session lengths, clicks, and reformulations.

The user feedback of different sites is aggregated on a central server by the provider of such an evaluation infrastructure. Subsequently, participants interact with this provider by visiting a dashboard-service that gives quantitative insights into the performance of their experimental algorithms. Moreover, it makes it possible to compare the retrieval algorithms from different participants. Besides hosting the dashboard service and managing participants, the primary task of the central server is the automatic composition of the MCA. It is made possible by templates that have to be adapted by participants. The templates include scripts with pre-defined names for indexing, searching, and recommending. Later on, these scripts are invoked by the MCA that assigns queries equally to the different systems on a least-served-basis.

Online/Offline Evaluation. Not only online evaluation in real-life systems with actual users can be employed, but also offline evaluations with previous data from academic search systems, in order to make retrieval algorithms generally applicable. This can be used to reduce the risk of frustrating users with poor-performing retrieval algorithms. Once an algorithm performs quite well in an offline evaluation, it can be dockerized and sent to the central server, where it will be part of an online evaluation.

Domain specificity. Domain specificity is ensured, as sites from different domains can be included in this evaluation infrastructure, as it is possible in LL4IR. Participants can evaluate their experimental retrieval algorithm within various academic search engines from different domains and compare their results with other participants but also across domains. This enables investigating which domain has which *key-performance-indicator* for high quality and reliable retrieval results.

Reproducibility. Reproducibility is ensured via the Docker framework. Contributions by participants are deployed as intended. Not only deployment effort but also error-proneness is lowered since no configurations or adaptions need to be made by the sites. This way, using one retrieval algorithm with exactly the same settings in combination with two different document collections, helps to understand if the experimental outcomes are not limited to specific constellations and transferable across different domains. Thus, all PRIMAD components are provided, as (a) the Docker framework serves as a EaaS-infrastructure comprising the **P**latform, **M**ethod, and **I**mplementation, (b) the sites specify the **R**esearch Goal and the **D**ata, and (c) the **A**ctors comprise the participants and the sites' users.

Such a concept is currently under implementation. The STELLA project [4] does not restrict search experiments to a specific domain, but rather encourages the same experimental systems in different domains. In cooperation with the early adopters GESIS-wide Search and LIVIVO, STELLA's initial evaluations are conducted in the domains of social and life sciences. It is integrated in these online academic search engines, which enables participants to evaluate their experimental retrieval systems (e.g., domain-specific research data recommendations based on publications [26]) with real users. STELLA will be part of the Living Labs for Academic Search (LiLAS) lab at CLEF 2020 [25][10].

6 Conclusion

In this paper, we elaborated on different evaluation infrastructures for shared tasks in evaluation campaigns for academic search systems concerning three requirements: online/offline evaluation, domain specificity, and reproducibility. We showed that current evaluation infrastructures do not fulfill all of the three requirements. Either the option for an online evaluation or the ability for reproducibility was not given, which poses a problem for validating user-centric experiment results. To approach this problem, we illustrated

[10] https://clef-lilas.github.io.

a concept design for an evaluation infrastructure that can fulfill these three requirements by providing an online evaluation in domain-specific academic search systems by using a dockerized framework for ensuring reproducibility.

Acknowledgements The STELLA project is funded by the Deutsche Forschungsgemeinschaft (DFG) – Project number 407518790.

Funding Open Access funding provided by Projekt DEAL.

References

1. Balog K, Kelly L, Schuth A (2014) Head first: living labs for ad-hoc search evaluation. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, New York
2. Balog K, Schuth A, Dekker P, Schaer P, Tavakolpoursaleh N, Chuang PY (2016) Overview of the trec 2016 open search track. Proceedings of the 25th Text REtrieval Conference 2016. Gaithersburg, NIST
3. Beel J, Genzmehr M, Langer S, Nürnberger A, Gipp B (2013) A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. Proceedings of the international workshop on reproducibility and replication in recommender systems evaluation. ACM, New York
4. Breuer T, Schaer P, Tavakolpoursaleh N, Schaible J, Wolff B, Müller B (2019) STELLA: towards a framework for the reproducibility of online search experiments. Proceedings of the Open-Source IR Replicability Challenge co-located with SIGIR, OSIRRC@SIGIR.
5. Brodt T, Hopfgartner F (2014) Shedding light on a living lab: The clef newsreel open recommendation platform. IIiX'14: Proceedings of the Information Interaction in Context Conference. ACM, New York
6. Dacrema MF, Cremonesi P, Jannach D (2019) Are we really making much progress? a worrying analysis of recent neural recommendation approaches. Proceedings of the 13th ACM Conference on Recommender Systems, RecSys '19. ACM, New York
7. Ferro N, Fuhr N, Järvelin K, Kando N, Lippold M, Zobel J (2016) Increasing reproducibility in IR: Findings from the dagstuhl seminar on "reproducibility of data-oriented experiments in e-science". ACM SIGIR Forum 50(1):68–82
8. Ferro N, Peters C (2019) From multilingual to multimodal: The evolution of CLEF over two decades. In: Information retrieval evaluation in a changing world – lessons learned from 20 years of CLEF
9. Fuhr N (2019) Reproducibility and validity in CLEF. In: Information retrieval evaluation in a changing world – lessons learned from 20 years of CLEF
10. Gollub T, Stein B, Burrows S (2012) Ousting ivory tower research: Towards a web framework for providing experiments as a service. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12. ACM, New York
11. Gollub T, Stein B, Burrows S, Hoppe D (2012) TIRA: configuring, executing, and disseminating information retrieval experiments. 9th International Workshop on Text-based Information Retrieval (TIR 2012) at DEXA. IEEE, Los Alamitos
12. Gregory K, Groth P, Cousijn H, Scharnhorst A, Wyatt S (2019) Searching data: a review of observational data retrieval practices in selected disciplines. J Assn Inf Sci Tec 70(5):419–432
13. Gunawardana A, Shani G (2015) Evaluating recommender systems. In: Recommender systems handbook. Springer, Heidelberg, Berlin, New York, pp 265–308
14. Hofmann K, Li L, Radlinski F et al (2016) Online evaluation for information retrieval. Found Trends Inf Retr 10(1):1–117
15. Hopfgartner F, Brodt T, Seiler J, Kille B, Lommatzsch A, Larson M, Turrin R, Serény A (2015) Benchmarking news recommendations: the CLEF newsreel use case. ACM SIGIR Forum 49(2):129–136
16. Hopfgartner F, Hanbury A, Müller H, Kando N, Mercer S, Kalpathy-Cramer J, Potthast M, Gollub T, Krithara A, Lin J, Balog K, Eggel I (2015) Report on the evaluation-as-a-service (eaas) expert workshop. ACM SIGIR Forum 49(1):57–65
17. Jagerman R, Balog K, de Rijke M (2018) Opensearch: lessons learned from an online evaluation campaign. J Data Inf Qual 10(13):1–15
18. Karanam S, Jorge-Botana G, Olmos R, van Oostendorp H (2017) The role of domain knowledge in cognitive modeling of information search. Inf Retr J 20(5):456–479
19. Lommatzsch A, Kille B, Hopfgartner F, Ramming L (2018) NewsREEL multimedia at mediaeval 2018: news recommendation with image and text content. In: Working notes Proceedings of the MediaEval workshop
20. Niu X, Hemminger BM (2012) A study of factors that affect the information-seeking behavior of academic scientists. J Am Soc Inf Sci 63(2):336–353
21. Peng S, Mamitsuka H, Zhu S (2018) MeSHLabeler and DeepMeSH: recent progress in large-scale MeSH indexing. Methods Mol Biol 1807:203–209
22. Pontis S, Kefalidou G, Blandford A, Forth J, Makri S, Sharples S, Wiggins G, Woods M (2016) Academics' responses to encountered information: context matters. J Assn Inf Sci Tec 67(8):1883–1903
23. Potthast M, Gollub T, Wiegmann M, Stein B (2019) TIRA integrated research architecture. In: Information retrieval evaluation in a changing world – lessons learned from 20 years of CLEF
24. Rosso P, Potthast M, Stein B, Stamatatos E, Pardo FMR, Daelemans W (2019) Evolution of the PAN lab on digital text forensics. In: Information retrieval evaluation in a changing world – lessons learned from 20 years of CLEF
25. Schaer P, Schaible J, Müller B (2020) Living labs for academic search at clef 2020. In: Advances in information retrieval – 42nd European Conference on IR Research, ECIR 2020. Lecture notes in computer science. Springer, Heidelberg, Berlin, New York
26. Tavakolpoursaleh N, Schaible J, Dietze S (2019) Using word embeddings for recommending datasets based on scientific publications. Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen", Berlin
27. Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, Weissenborn D, Krithara A, Petridis S, Polychronopoulos D, Almirantis Y, Pavlopoulos J, Baskiotis N, Gallinari P, Artières T, Ngomo AN, Heino N, Gaussier É, Barrio-Alvers L, Schroeder M, Androutsopoulos I, Paliouras G (2015) An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics 16(138):1–28