STELLA: Towards a Framework for the Reproducibility of Online Search Experiments

Timo Breuer
Philipp Schaer
firstname.lastname@th-koeln.de
Technische Hochschule Köln
Cologne, Germany

Narges Tavakolpoursaleh
Johann Schaible
firstname.lastname@gesis.org
GESIS
Cologne, Germany

Benjamin Wolff
Bernd Müller
{wolff,muellerb}@zbmed.de
ZB MED - Information Centre for Life Sciences
Cologne, Germany

ABSTRACT

Reproducibility is a central aspect of offline as well as online evaluations, to validate the results of different teams and in different experimental setups. However, often it is difficult or not even possible to reproduce an online evaluation, as solely a few data providers give access to their system, and if they do, it is limited in time and typically only during an official challenge. To alleviate the situation, we propose STELLA: a living lab infrastructure with consistent access to a data provider's system, which can be used to train and evaluate search- and recommender algorithms. In this position paper, we align STELLA's architecture to the PRIMAD model and its six different components specifying reproducibility in online evaluations and illustrate two use cases with two academic search systems.

1 INTRODUCTION

Reproducibility[1] is still an open issue in TREC-style IR offline evaluations. Hanbury et al. [3] named this setting the *Data-to-Algorithms paradigm* where participants submit the output of their software when running on a pre-published test collection. Recently, the IR community extended this idea by thinking of Evaluation-as-a-Service (EaaS) that adopts the *Algorithms-to-Data paradigm*, i.e., in the form of living labs [1]. In living labs, relevance assessments are produced by actual users and instead resemble their satisfaction with the search system in contrast to the explicit relevance assessments in TREC-style test collections. To obtain the satisfaction rate relevance is measured by observing user behavior, e.g., navigation, click-through rates, and other metrics.

In theory EaaS might enhance reproducibility by keeping the data, algorithms, and results in a central infrastructure that is accessible through a standard API and allows for sharing open-source software components. However, the live environment for evaluating experimental systems typically has the consequence that the results are not reproducible since the users' subjective impression of relevance is very inconstant. This makes reproducibility of online experiments more complicated than their offline counterpart. To what extent online experiments in living labs can be made reproducible remains a central question. Although the user interactions and all rankings generated by the systems can be stored and used for subsequent calculations, there are no clear guidelines as to how the logged interaction data can contribute to a valid and reproducible evaluation result. The major problem is that the recorded interactions are authentic only for the particular situation in which they were recorded.

Conceptionally, Ferro et al. [2] introduce the PRIMAD model, which specifies reproducibility in several components: *Platform*, *Research goal*, *Implementation*, *Method*, *Actor*, and *Data*. PRIMAD is a conceptional framework for the assessment of reproducibility along the suggested components. Although PRIMAD explicitly discusses the application for both offline and online experiments, we see a gap when we apply it to living labs that involve the interaction of real users and real-time online platforms. The suggestion of thinking of users as "data generators" undervalues their role within the evaluations.

To overcome these issues, we introduce STELLA (InfraSTructurEs for Living LAbs) a living lab infrastructure. STELLA allows capturing document data, algorithms, and user interactions in an online evaluation setup that is based on Docker containers. We align the different components of the infrastructure to the PRIMAD model and discuss their match to the model. We see a particular need to pay attention to the *actor* and *data* components. These components represent the human-factors like users' interactions that affect the outcomes of online experiments and not only the experimenters' perspective on the experiment.

In the following paper, we present the design of the STELLA living lab Docker infrastructure (cf. Section 2). We align STELLA's components to the dimensions of the PRIMAD model and illustrate a use case with two academic search systems (cf. Section 3). Finally, we discuss the benefits and limitations of our work and conclude the paper (cf. Section 4).

2 ONLINE EVALUATION WITH STELLA

The STELLA infrastructure allows researchers to evaluate search and recommender algorithms in an online environment, i.e., within a real-world system with real users. When using STELLA, the researchers' primary goal is to introduce ranking models for search results and recommendations that outperform the existing baseline. In the following, we describe STELLA's workflow as well as its technical infrastructure. We adhere to the wording of TREC OpenSearch [5] with regards to several components of the living lab infrastructure. Providers of search engines and corresponding web interfaces are referred to as *sites*. Research groups that

OSIRRC 2019 co-located with SIGIR 2019, 25 July 2019, Paris, France.

[1] One can differentiate between repeatability (same team, same experimental setup), replicability (different team, same setup), and reproducibility (different team, different setup). For the sake of simplicity, we use the term *reproducibility* to refer to all of these three types.

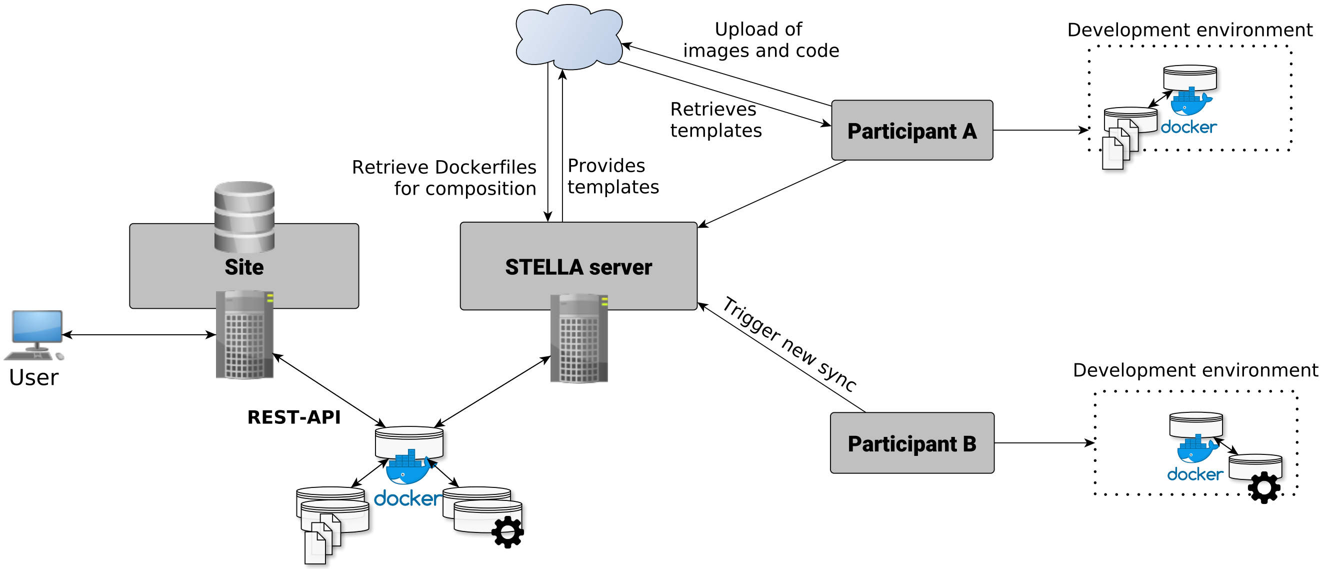

Figure 1: Living lab infrastructure based on Docker. Participants contribute their experimental retrieval and recommender systems by uploading Dockerfiles and source code. The STELLA server composes a multi-container application out of single experimental systems. This application is deployed locally at the sites. Queries are forwarded to this application, which delivers results from the experimental systems in return. User feedback is sent to the STELLA server and stored within the Docker container.

contribute experimental retrieval and recommender systems are referred to as *participants*.

STELLA's main component is the central living lab API. It connects data and content providers (*sites*) with researchers in the fields of information retrieval and recommender systems (*participants*). When linked to the API, the sites provide data that can be used by participants for implementing search and recommendation algorithms, e.g., metadata about items, users, session logs, and click paths. Participants use the API in order to obtain this data from the sites and enhance their experimental system, e.g., by delivering personalized search results. Subsequently, this experimental system is integrated into the living lab infrastructure and made accessible to site users. Within the site, the submitted system is implemented automatically. The system is used within an A/B-testing or interleaving scenario, such that the system's results are presented to the real users of the site. The users' actions, e.g., clicking or selecting a ranked item which determines the click-through rate, is recorded and send back to the central living labs API. There, this data is aggregated over time in order to produce reliable results to determine the system's usefulness for that specific site.

To support reproducibility, we employ the Docker technology in order to keep as many components as possible the way they were in the first experiment. This includes reusing the utilized software, tools being used to develop the method, and in a user-oriented study, usage data generated by the users. Within the framework, experimental retrieval and recommender systems are distributed with the help of Docker images. Sites deploy local multi-container applications running these images. Participants contribute their experimental systems by extending prepared Dockerfiles and code templates. The underlying infrastructure will assure the synchronization and comparability of experimental retrieval and recommender systems across different sites. See figure 1 for a schematic visualization of the framework.

Sites deploy a multi-container environment that contains the experimental systems. Queries by site users are forwarded to these systems. A scheduling mechanism assures even distribution of queries among the participants' systems. The multi-container environment logs user feedback and forwards this usage data to the experimental systems. Likewise logged user feedback is sent to the central STELLA server where it is filed, and overall statistics can be calculated.

The main functionalities of the STELLA server are the administration and synchronization of infrastructural components. After the submission of extended template files, the STELLA server will initiate the build process of Docker images. The build process is triggered by new contributions or changes within the experimental systems. Participants and sites will be able to self-administrate configurations and have insights into the evaluation outcomes by visiting a dashboard service.

In previous living lab campaigns [5], sites had to implement a REST-API that redirected queries to the central living lab server and an interleaving mechanism to generate the final result list. In our case, sites can entirely rely on the Docker applications. The site's original search system can – but do not have to – be integrated as an additional container. Optionally the REST-API can be reached by the conventional way of previous living lab campaigns [1]. The Docker application can also be deployed on the STELLA server from where it can be reached over the internet. This may be beneficial

PRIMAD variable	Instance
Platform	Docker-based framework
Research goal	Retrieval effectiveness
Implementation	Chosen by participants
Method	Chosen by participants
Actor	Participants, site users
Data	Domain/site specific

Table 1: Alignment of STELLA components to the dimensions of the PRIMAD model

for those sites, which want to participate but do not have sufficient hardware capacities for the Docker environment.

Participants develop their systems in local environments and submit their systems upon completion. They contribute their systems by extending Dockerfile templates and providing the necessary source code. The submission of systems can be realized with already existing infrastructures like online version control services in combination with the Docker Hub.

3 RELATION TO PRIMAD

The PRIMAD model offers orientation to what extent reproducibility in IR experiments can be achieved. Table 1 provides an overview of the PRIMAD model and corresponding components in the STELLA infrastructure. The *platform* is provided by our framework, which mainly relies on Docker and its containerization technology. Increasing retrieval effectiveness is the primary *research goal.* Sites, e.g., a digital library, and participants, i.e., the external researchers, benefit from the cooperation with each other using STELLA. Sites, for instance, may be interested in finding adequate retrieval and recommender algorithms, whereas participants get access to data from real-world user interactions. Both the *implementation* and the *method* is chosen by the participants. In ad-hoc retrieval experiments, solely the researcher would be considered an *actor.* In a living lab scenario, this group is extended by site users who affect the outcome of experiments. Finally, *data* will consist of logged user interaction as well as domain-specific text collections.

In the following, we present use cases of how the STELLA infrastructure is concretely aligned to the PRIMAD components. Two early adopters from the domain of academic search implement our framework such that they are part of the STELLA infrastructure, LIVIVO[2] [7] and the GESIS-wide Search[3] [4]. LIVIVO is an interdisciplinary search engine and contains metadata on scientific literature for medicine, health, nutrition, and environmental and agricultural sciences. The GESIS-wide Search is a scholarly search system where one can find information about social science research data, instruments and scales, as well as open access publications.

Platform: In STELLA the platform is implemented via a Docker-based framework as shown in Figure 1. It connects the sites with the participants and assures (i) the sites' flow of data for computing IR-models, (ii) deploying the participants' IR-models on the sites, and (iii) obtaining the usefulness of the deployed IR-model, e.g., via accumulated click-through rates. Both sites are configured in a way that they can interact with the central STELLA server, i.e., provide data on the sites' content and users as well as their queries and interactions.

Research Goal: The research goal is the retrieval of scientific datasets and literature which satisfy a user's information need. This includes retrieval using string-based queries as well as recommending further information using item-based queries. In LIVIVO, the research goal is finding appropriate domain-specific scientific literature in medicine, health, nutrition, and environmental and agricultural sciences. Besides scientific publications in the social sciences, the GESIS-wide Search offers to search for research data, scales, and other information. The research goal also includes finding appropriate cross-item recommendations as recommending research data based on a currently viewed publication.

Method and Implementation: The participant chooses both the method and implementation. With the help of the Docker-based framework, participants are free to choose which methods for retrieval to use as well as the methods' implementation as long as the interface guidelines between the sites and the Docker images are respected.

Actor: The actors in online evaluations are (i) the potential site users and (ii) the participants that develop an experimental system to be evaluated on a site. In both LIVIVO and GESIS-wide Search, the site users range from students to scientists as well as librarians. Specifically, the users of both sites differ by their domain of interest (medicine and health vs. social sciences), their experience, and the granularity of their information need, which can be trivial, but also very complex. In any case, the site users' interactions are captured by STELLA, which allows observing differences in their behavior. Participants submit their experimental systems using the Docker-based framework and receive the evaluation results from STELLA after some evaluation period. This way, they can adapt their systems and re-submit them whenever possible.

Data: The data comprises all information the sites are willing to provide to participants for developing their systems. It usually includes some structured data, such as database records containing metadata on research data and publications, as well as unstructured data like full texts or abstracts. LIVIVO uses the ZB MED Knowledge Environment (ZB MED KE) as a data layer that semantically enriches (by annotating the metadata with concepts from life sciences ontologies) the textual content of metadata from about 50 different literature resources with a total of about 55 Million citations. Additionally, LIVIVO allows users to register and maintain a personalized profile, including a watch list. The GESIS-wide Search integrates metadata from different portals into a central search index that uses a specific metadata schema based on Dublin Core. Additionally, each data record is enriched with explicit links between different information items like links between a publication and a dataset. This link specifies that the publication uses that particular dataset. As there is no option for users to register, GESIS-wide Search provides users' session logs and click paths but no user profiles. The

[2] https://www.livivo.de, Accessed June 2019

[3] https://www.gesis.org/en/home/, Accessed June 2019

entire document corpora of LIVIVO and GESIS-wide Search is made available for participants.

In this living lab scenario P, R, I, and M would be conserved within the Docker container and the STELLA infrastructure. $A \rightarrow A'$ and $D \rightarrow D'$ would remain as components that change in another system and another point in time. By recording the interaction and click data, we can try to preserve parts of A and D.

4 DISCUSSION

Compared to previous living lab campaigns, the Docker-based infrastructure results in several advances towards enhancing reproducibility in online evaluations. The most important advances are:

Transparency: Specifying retrieval and recommender systems, as well as their specific requirements in a standardized way, may contribute to the transparency of these systems.

No limitation to head queries: Pre-computed rankings were limited to the top-k queries. Even though this is an elegant solution, it may influence evaluation outcomes. Using locally deployed Docker applications, there is no need for this restriction anymore. Rankings and recommendations can be determined based on the complete corpus.

Avoidance of network latencies: Network latencies after the retrieval of rankings or recommendations might affect user behavior. Also, implementing workarounds like timeouts resulted in additional implementation effort for sites. By deploying local Docker images, these latencies are eliminated.

Lower entrance barrier for participation: Participants can contribute already existing systems to the STELLA infrastructure by simply dockerizing them. By specifying the required components and parameters, the deployment procedure is less error-prone. Researchers can use software and programming languages of their choice. Sites solely need to implement the REST-API and set up the local instance of the Docker application. Letting Docker deploy the application, human errors are avoided, and efforts are reduced.

The benefits mentioned above come at a cost. Especially the following limitations have to be considered:

Central server: The proposed infrastructure relies on a central server. This vulnerability might be a target for malicious intents and is generally a single point of failure.

Hardware limitations: The experimental systems will be deployed at sites. Available hardware capacities may vary across different sites. Furthermore, the hardware requirements of experimental systems should be following the available resources of sites. For instance, participants contributing machine/deep learning systems should not outsource training routines to external servers. In the first place, we will focus on lightweight experiments in order to keep the entrance barrier for participation low.

User interaction data: While all interaction data within the STELLA infrastructure can be logged, a reuse-setup is not outlined yet and must remain as future work.

Recording usage and interaction data for later reuse are not novel. An example of interaction data recorded to allow later verification and simulation was the NewsREEL lab at CLEF 2017 [6]. In NewsREEL Replay participants had access to a dataset comprising a collection of log messages analogous to NewsREEL Live. Analogous to the online evaluation, participants had to find the configuration with the highest click-trough-rate. By using the recorded data, participants experimented on a reasonable trade-off amid prediction accuracy and response time.

5 CONCLUSION

We present a living lab platform for the evaluation of online experiments and a Docker-based infrastructure which bridges the gap between experimental systems and real user interactions. Concerning the PRIMAD model, it is possible to assess reproducibility and corresponding components in our infrastructure proposal. Participants are free to choose which method and implementation to use and can rely on adequate deployment and environments. User interaction data will be logged and is accessible for optimizing systems and future applications.

REFERENCES

[1] Krisztian Balog, Anne Schuth, Peter Dekker, Philipp Schaer, Narges Tavakolpoursaleh, and Po-Yu Chuang. 2016. Overview of the TREC 2016 Open Search Track. In *Proceedings of the Twenty-Fifth Text REtrieval Conference (TREC 2016). NIST.*

[2] Nicola Ferro, Norbert Fuhr, Kalervo Järvelin, Noriko Kando, Matthias Lippold, and Justin Zobel. 2016. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". *SIGIR Forum* 50, 1 (June 2016), 68–82. https://doi.org/10.1145/2964797.2964808

[3] Allan Hanbury, Henning Müller, Krisztian Balog, Torben Brodt, Gordon V. Cormack, Ivan Eggel, Tim Gollub, Frank Hopfgartner, Jayashree Kalpathy-Cramer, Noriko Kando, Anastasia Krithara, Jimmy Lin, Simon Mercer, and Martin Potthast. 2015. Evaluation-as-a-Service: Overview and Outlook. *ArXiv e-prints* (Dec. 2015). http://arxiv.org/abs/1512.07454

[4] Daniel Hienert, Dagmar Kern, Katarina Boland, Benjamin Zapilko, and Peter Mutschke. 2019. A Digital Library for Research Data and Related Information in the Social Sciences. In *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) (forthcoming).*

[5] Rolf Jagerman, Krisztian Balog, and Maarten de Rijke. 2018. OpenSearch: Lessons Learned from an Online Evaluation Campaign. *J. Data and Information Quality* 10 (2018), 13:1–13:15.

[6] Benjamin Kille, Andreas Lommatzsch, Frank Hopfgartner, Martha Larson, and Torben Brodt. 2017. CLEF 2017 NewsREEL Overview: Offline and Online Evaluation of Stream-based News Recommender Systems. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. (CEUR Workshop Proceedings)*, Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl (Eds.), Vol. 1866. CEUR-WS.org. http://ceur-ws.org/Vol-1866/invited_paper_17.pdf

[7] Bernd Müller, Christoph Poley, Jana Pössel, Alexandra Hagelstein, and Thomas Gübitz. 2017. Livivo–the vertical search engine for life sciences. *Datenbank-Spektrum* 17, 1 (2017), 29–34.