

# Overview of LiLAS 2021 – Living Labs for Academic Search (Extended Overview)

Philipp Schaer<sup>1</sup>, Timo Breuer<sup>1</sup>, Leyla Jael Castro<sup>2</sup>, Benjamin Wolff<sup>2</sup>, Johann Schaible<sup>3</sup> and Narges Tavakolpoursaleh<sup>3</sup>

<sup>1</sup>TH Köln – University of Applied Sciences, Cologne, Germany

<sup>2</sup>ZB MED – Information Centre for Life Sciences, Cologne, Germany

<sup>3</sup>GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany

## Abstract

The Living Labs for Academic Search (LiLAS) lab aims to strengthen the concept of user-centric living labs for academic search. The methodological gap between real-world and lab-based evaluation should be bridged by allowing lab participants to evaluate their retrieval approaches in two real-world academic search systems from life sciences and social sciences. This overview paper outlines the two academic search systems LIVIVO and GESIS Search, and their corresponding tasks within LiLAS, which are ad-hoc retrieval and dataset recommendation. The lab is based on a new evaluation infrastructure named STELLA that allows participants to submit results corresponding to their experimental systems in the form of pre-computed runs and Docker containers that can be integrated into production systems and generate experimental results in real-time. Both submission types are interleaved with the results provided by the productive systems allowing for a seamless presentation and evaluation. The evaluation of results and a meta-analysis of the different tasks and submission types complement this overview.

## Keywords

Living labs, evaluation, academic search, dataset recommendation, ad-hoc retrieval, STELLA framework

## 1. Introduction

The Living Labs for Academic Search (LiLAS) lab aims to strengthen the concept of user-centric living labs for the domain of academic search. By allowing lab *participants* to evaluate their retrieval approaches in two real-world academic search portals (called *sites*) from life sciences and social sciences, the methodological gap between real-world and lab-based evaluations is effectively reduced.

This gap is based on the different opportunities available to researchers in academia and industry. While industry-based research in the field of information retrieval (IR) has the opportunity to conduct experiments in-vivo – thanks to the availability of large systems, with a wide range and correspondingly large user base – these opportunities usually remain closed to

---

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania


✉ philipp.schaer@th-koeln.de (P. Schaer); timo.breuer@th-koeln.de (T. Breuer); ljgarcia@zbmed.de (L. J. Castro); wolff@zbmed.de (B. Wolff); johann.schaible@gesis.org (J. Schaible); narges.tavakolpoursaleh@gesis.org (N. Tavakolpoursaleh)

🌐 <https://ir.web.th-koeln.de> (P. Schaer)

🆔 0000-0002-8817-4632 (P. Schaer); 0000-0002-1765-2449 (T. Breuer); 0000-0003-3986-0510 (L. J. Castro); 0000-0001-9345-8958 (B. Wolff); 0000-0002-5441-7640 (J. Schaible); 0000-0001-9324-3252 (N. Tavakolpoursaleh)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** Overview of the live evaluation pipeline.

academic research. In-vivo here describes the possibility to perform IR experiments integrated into real-world systems and to conduct experiments where the actual interaction with these systems takes place. It should be emphasized here that these are not classic user experiments in which the focus is on the individual interactions of users (e.g., to investigate questions of UI design), but rather aggregated usage data is collected in large quantities in order to generate reliable quantitative research results. The potential of living labs and real-world evaluation techniques has been shown in previous CLEF labs such as NewsREEL [1] and LL4IR [2], or TREC OpenSearch [3]. In a similar vein, LiLAS is designed around the living lab evaluation concept and introduces different use cases in the broader field of academic search. Academic search solutions, which have to deal with the phenomena around the exponential growing rate [4] of scientific information and knowledge, tend to fall behind the real-world requirements and demands. The vast amount of scientific information does not only include traditional journal publication, but also a constantly growing amount of pre-prints, research datasets, code, survey data, and many other research objects. This heterogeneity and mass of documents and datasets introduces new challenges to the disciplines of information retrieval, recommender systems, digital libraries, and related fields. Academic search is a conceptual umbrella to subsume all these different disciplines and is well-known through (mostly domain-specific) search systems and portals such as PubMed, arXiv.org, or dblp. While those three are examples of open-science-friendly systems as they allow re-use of metadata, usage data and/or access to fulltext data, other systems such as Google Scholar or ResearchGate. The later offer no access at all to their internal algorithms and data and are therefore representatives of a closed-science (and commercial) mindset.

Progress in the field of academic search and its corresponding domains is usually evaluated by means of shared tasks that are based on the principles of Cranfield/TREC-style studies [5]. Typical shared tasks at the Conference and Labs of the Evaluation Forum (CLEF) and the Text Retrieval Conference (TREC) are based on the offline computation of results/runs missing a valuable link to real-world environments [6]. Most recently the TREC-COVID [7] evaluation campaign run by NIST attracted a high number of participants and showed the high impact of scientific retrieval tasks in the community. Within TREC-COVID a wide range of systems and retrieval approaches participated and generally showed the massive retrieval performance that recent BERT and other transformer-based machine learning approaches are capable of. However, classic vector-space retrieval was also highly successful using the well-known SMART system<sup>1</sup> and showed the limitations of the test collection-based evaluation approach of TREC-COVID and

<sup>1</sup><https://ir.nist.gov/covidSubmit/archive.html>

the general need for innovation in the field of academic search and IR. Meta-evaluation studies of system performances in TREC and CLEF showed a need for innovation in IR evaluation [8, 9]. The field of academic search is no exception to this. The central concern of academic search is finding both relevant and high-quality documents. The question of what constitutes relevance in academic search is multilayered [10] and an ongoing research area.

In 2020 we held a first iteration of LiLAS as a so-called workshop lab. This year we provide participants exclusive access to real-world systems, their document base (in our case a very heterogeneous set of research articles and research data including, for instance, surveys), and the actual interactions including the query string and the corresponding click data (see overview on the setup in Figure 1). To foster different experimental settings we compile a set of head queries and candidate documents to allow pre-computed submissions. Using the STELLA-infrastructure, we allow participants to easily integrate their approaches into the real-world systems using Docker containers and provide the possibility to compare different approaches at the same time.

This extended lab overview is a longer version of the condensed LNCS lab overview [11]. It is structured as follows: In Sections 2 and 3 we introduce the two main use cases of LiLAS which are bond to the sites granting us access to their retrieval systems: LIVIVO and GESIS Search. In these two sections the systems, the provided datasets, and task are described. In Section 4 we outline the evaluation setup and STELLA, our living lab evaluation framework, and the two submission types, namely pre-computed runs and Docker container submissions. Section 4 also includes the description of the evaluation metrics used with in the lab and a short overview on the organizational structure of the lab. In Section 5 we introduce the participating groups and approaches. We outline the results of the evaluation rounds in Section 6 and conclude in Section 7. In addition to the condensed LNCS overview we included some more textual details and additional tables and figures in Appendix A.

## 2. Ad-hoc Search in LIVIVO

### 2.1. LIVIVO Literature Search Portal

LIVIVO<sup>2</sup> [12] is a literature search portal developed and supported by ZB MED – Information Centre for Life Sciences. ZB MED is a non-profit organization providing specialized literature in Life Sciences at a national (German) and international level and hosting one of the largest stock of life science literature in Europe. Since 2015, ZB MED supports users including librarians, students, general practitioners and researchers with LIVIVO, a comprehensive and interdisciplinary search portal for Life Sciences.

LIVIVO integrates various literature resources from medicine, health, environment, agriculture and nutrition, covering a variety of scholarly publication types (e.g., conferences, preprints, peer-review journals). LIVIVO corpus includes about 80 million documents from more than 50 data sources in multiple languages (e.g., English, German, French). To better support its users, LIVIVO offers an end-user interface in English and German, an automatically and semantically enhanced search capability, and a subject-based categorization covering the different areas it supports (e.g., environment, agriculture, nutrition, medicine). Precision of search queries is

---

<sup>2</sup><https://www.livivo.de>

```

# Sample head query
{ "qid": 1001, "qstr": "integrierte AND versorgung", "freq": 12 }

# Sample documents
{ "DBRECORDID": "AGRISFR2016215853",
  "TITLE": ["Dissection ..."],
  "AUTHOR": ["Teyssèdre, Simon"],
  "SOURCE": ["Dissection ..."],
  "LANGUAGE": ["fra"],
  "DATABASE": ["AGRIS"] }

# Sample candidate list
{ "qid": 1001,
  "qstr": "integrierte AND versorgung",
  "candidates": ["C951899619", "C676171", "848078", "C765841" ... ] }

```

**Figure 2:** Examples for head queries, documents, and candidate lists for the LIVIVO system.

improved by using descriptors with semantic support; in particular, LIVIVO uses three multi-lingual vocabularies to this end (Medical Subject Headings MeSH, UTHES, and AGROVOC). In addition to its search capabilities, LIVIVO also integrates functionality supporting inter-library loans at a national level in Germany. Since 2020, LIVIVO also offers a specialized collection on COVID-19<sup>3</sup>

## 2.2. LIVIVO Dataset

For the LiLAS challenge, we prepared training and test datasets comprising head queries together with 100-document candidate list. In Figure 2 we include an excerpt of the different elements included in the data. Data was formatted in JSON and presented as JSONL files to facilitate processing. Participating head queries were restricted to keywords-based search and keywords-based search plus AND, OR and NOT operators.

Head queries were assigned an identifier, namely *qid*, a query string, *qstr* and as an additional information the query frequency, *freq*. For each head query, a candidate list was also provided. Candidate lists include the query identifier as well as corresponding string, together with a list of 100 document identifiers (i.e. the native identifier used in the LIVIVO database).

In addition to head queries and candidate lists, we also provided a set of documents in LIVIVO corresponding to three of the major bibliographic scholarly databases so participants could create their own indexes. The document set contains metadata for approx. 35 million documents and is provided as a JSONL file. To reduce complexity and keep the data manageable, we decided to provide only the 6 most important data fields (DBRECORDID, TITLE, AUTHOR, SOURCE, LANGUAGE, DATABASE). Additional metadata and fulltext is mostly available from the original database curators. The aforementioned databases correspond to Medline, the National Library of Medicine's (NLM) bibliographic database for life sciences and biomedical information including about 20 million of abstracts; the NLM catalog, providing access to bibliographic

<sup>3</sup><https://www.livivo.de/covid19>.

data for over 1.4 million journals, books and similar data; and the Agricultural Science and Technology Information (AGRIS) database, a Food and Agriculture Organization of the United Nations initiative compiling information on agricultural research with 8.9 million structured bibliographical records on agricultural science and technology.

### **2.3. Task**

Finding the most relevant publications in relation to a head query remains a challenge in scholarly Information Retrieval systems. While most repositories or registries deal mostly with publications in English, LIVIVO, the production system used at LiLAS, supports multilingualism, adding an extra layer of complexity and presenting a challenge to participants.

The goal of this ad-hoc search task is supporting researchers to find the most relevant literature regarding a head query. Participants were asked to define and implement their ranking approach using as basis a multi-lingual candidate documents list. A good ranking should present users with the most relevant documents on top of the result set. An interesting aspect of this task is the multilingualism as multiple languages can be used to pose a query (e.g. English, German, French); however, regardless of the language used on the query, the retrieval can include documents in other languages as part of the result set.

## **3. Research Data Recommendations in GESIS-Search**

### **3.1. GESIS Search Portal**

GESIS Search<sup>4</sup> is a search portal for social science research data and open access publications developed and supported by GESIS - Leibniz Institute for the Social Sciences. GESIS is a member of the Leibniz Association with the purpose to promote social science research. It provides essential and internationally relevant research-based services for the social sciences, and as the largest European infrastructure institute for the social sciences, GESIS offers advice, expertise and services to scientists at all stages of their research projects.

GESIS Search aims at helping its users find appropriate scholarly information on the broad topic of social sciences [13]. To this end, it provides different types of information from the social sciences in multiple languages, comprising literature (114.7k publications), research data (84k), questions and variables (13.6k), as well as instruments and tools (440). A well-configured relevance ranking together with a well-defined structure and faceting mechanism allow to address the users' information needs, however, the most interesting aspect is the inclusion of scientific literature with research data. Typically, those types of information are accessible through different portals only, posing the problem of a lack of links between these two types of information. GESIS Search provides such an integrated access to research data as well as to publications. The information items are connected to each other based on links that are either manually created or automatically extracted by services that find data references in full texts. Such linking allows researchers to explore the connections between information items interactively.

---

<sup>4</sup><https://search.gesis.org/>

```

# Sample publication document
{ "id": "csa201419416",
  "title": "The Changing Value...",
  "abstract": "This article reviews...",
  "topic": [
    "Children",
    "Child Mortality",
    "Values" ] }

# Sample research dataset document
{ "id": "DA3433",
  "title": "Kindheit, Jugend und Erwachsenwerden...",
  "title_en": "Childhood, Adolencence, and Becoming an Adult...",
  "abstract": "Die Hauptthemen der Studie...",
  "abstract_en": "The primary topics of the study...",
  "topic": ["Familie und Ehe", "Kinder"],
  "topic_en": ["Family life and marriage", "Children" ] }

# Sample candidate list
{ "s_id": "gesis-ssoar-62031",
  "candidate_docs": {
    "ZA6752": 0.1856689453125,
    "ZA6751": 0.183837890625,
    "ZA6749": 0.181396484375,
    "ZA6782": 0.1795654296875 } }

```

**Figure 3:** Examples for publication documents, research dataset documents, and candidate lists for the GESIS Search system.

### 3.2. GESIS Search Dataset

For LiLAS, we focus on all publications and research data comprised by GESIS Search. The publications are mostly in English and German, and are annotated with further textual metadata including title, abstract, topic, persons, and others. Metadata on research data comprises (among others) a title, topics, datatype, abstract, collection method, primary investigators, and contributors in English and/or German.

The data provided to participants comprises the mentioned metadata on social science literature and research data on social science topics comprised in the GESIS Search. In Figure 3 we include an excerpt of the different elements included in the data. For the dataset recommendation task with pre-computed results (see details in Section 3.3), in addition, the participants were given the set of research data candidates that are recommended for each publication. This candidate set is computed based on context similarity between publications and research data. It is created by applying the TF-IDF score to vectorize the combination of title, abstract, and topics for each document type and computing the cosine similarities between cross-data types. It contains a list of research data for each publication with the highest similarities to the publication among other research data in the corpus.

### 3.3. Task

Research data is of high importance in scientific research, especially when making progress in experimental investigations. However, finding useful research data can be difficult and cumbersome, even if using dataset search engines, such as Google Dataset Search<sup>5</sup>. Another approach is scanning scientific publication for utilized or mentioned research data; however, this allows to find explicitly stated research data and not other research data relevant to the subject. To alleviate the situation, we aim at evolving the recommendation of appropriate research data beyond explicitly mentioned or cited research data. To this end, we propose to recommend research data based on publications of the user’s interest between a scientific publication and possible research data candidates.

The main task is: given a seed-document, participants are asked to calculate the best fitting research data recommendations with regards to the seed-document. This resembles the use case of providing highly useful recommendations of research data relevant to the publication that the user is currently viewing. For example, the user is interested in the impact of religion on political elections. She finds a publication regarding that topic, which has a set of research data candidates covering the same topic.

The participants were allowed to submit pre-computed and live runs (see section 4.2 for more details). For submitting the pre-computed run, the participants also received a first candidate list comprising 1k publication each having a list of recommended research data. The task here was to re-rank this candidate list. On the contrary, for submitting the live runs, such a candidate list was not needed, as the recommended candidates needed to be calculated first. To do so, participants are provided metadata on publications as well as on the research data comprised in GESIS Search (see Section 3.1 for more details on the provided data).

## 4. Evaluation Setup

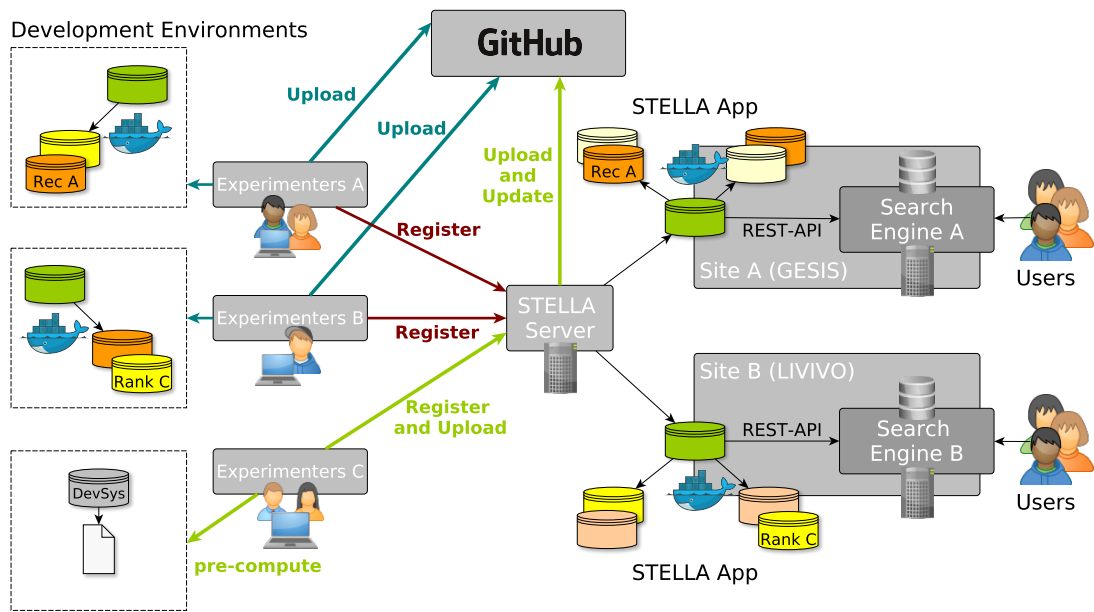
### 4.1. STELLA Infrastructure

The technical infrastructure and platform was provided by our evaluation service called STELLA (as illustrated in Figure 4). It complements existing shared task platforms by allowing experimental ranking and recommendation systems to be fully integrated into an evaluation environment, with no interference in the interaction between the users and the system as the whole process is transparent for users. Besides transparency and reproducibility, one of the STELLA main principles is the integration of experimental systems as micro-services. More specifically, lab participants package their single systems as Docker containers that are bundled in a multi-container application (MCA). Providers of academic research infrastructures deploy the MCA in their back-end and use the REST-API either to get ranking and recommendations or to post the corresponding user feedback that is mainly used for our evaluations. Intermediate evaluation results are available through a public dashboard service that is hosted on a central server, also part of the STELLA infrastructure. After authentication, participants can register experimental systems at this central instance and access feedback data that can be used to optimize their systems. In the following, each component of the infrastructure is briefly described to give

---

<sup>5</sup><https://datasetsearch.research.google.com/>





**Figure 4:** Overview of the STELLA infrastructure

the reader a better idea on how STELLA serves as a proxy for user-oriented experiments with ranking and recommendation systems.

#### 4.1.1. Micro-services

As pointed out before, we request our lab participants to package their systems with Docker. For the sake of compatibility, we provide templates for these micro-services to implement minimal REST-based web services. Participants can adapt their systems to these templates as they see it fits as long as the pre-defined REST endpoints deliver technically correct responses. The templates can be retrieved from GitHub<sup>6</sup> that is fundamental to our infrastructure. Not only the templates, but also the participant systems should be hosted in a public Git repository in order to be integrated into the MCA. As soon as the developments are done, the participants register their Git(Hub) URL at the central dashboard service of the infrastructure.

#### 4.1.2. Multi-container Application (MCA)

Once the experimental systems pass technical unit tests and sanity checks for selected queries and target items, they are ready to be deployed and evaluated via user interactions. To reduce the deployments costs for the site providers, the single experimental systems are bundled into an MCA which serves as the entry point to the infrastructure. The MCA handles the query distribution among the experimental systems and also sends user feedback data to the central server at regular intervals. After the REST-API corresponding to the MCA is connected to the

<sup>6</sup><https://github.com/stella-project/stella-micro-template>



search interface, the user traffic can be redirected to the MCA which will actually deliver the experimental results. We then interleave results of single experimental systems with those from the baseline system by using a Team-Draft-Interleaving (TDI) approach. This results in two benefits: 1) we prevent users from subpar retrieval results that also might affect the site’s reputation, and 2), as shown before, interleaved results can be used to infer statistically significant results with less user data as compared to conventional A/B tests. The site providers rely on their own logging tools. STELLA expects a minimal set of information required when sending feedback; however, sites are free to add any additional JSON-formatted feedback information and interactions to the data payload, for instance logged clicks on site-specific SERP elements. The underlying source code of the MCA is hosted in a public GitHub repository<sup>7</sup>.

### 4.1.3. Central Server

The central server instance of the infrastructure fulfills four functionalities: 1) participants, sites and administrators visit the server to register user accounts and systems; 2) a dashboard service provides visual analytics and first insights about the performance of experimental systems; 3) likewise, feedback data in the form of user interactions is stored in a database that can be downloaded for system optimizations and further evaluations; and 4) the server implements an automated update job of the MCA in order to integrate newly submitted systems if suitable.

Each MCA that is instantiated with legitimate credentials posts the logged user feedback to the central infrastructure server. Even though the infrastructure would allow continuous integration of newly submitted systems, we stuck to the official dates of round 1 and 2 when updating the MCAs at the sites. Due to moderate traffic, we run the central server on a lightweight single core virtual machine with 2GB RAM and 50GB storage capacity<sup>8</sup>. More technical details about the implementations can be found in the public GitHub repository<sup>9</sup>.

## 4.2. Submission Types

Participants can choose between two different submission types for both tasks (i.e. ad-hoc search and dataset recommendation). Similar to previous living labs, **Type A** are pre-computed runs that contain rankings and recommendations of the most frequent queries and the most frequently viewed document, respectively for reach task. Alternatively, it is possible to integrate the entire experimental system as a micro-service as part of a **Type B** submission. Both submission types have their own distinct merits as described below.

### 4.2.1. Type A - Pre-computed Runs

Even though the primary goal of the STELLA framework is the integration of entire systems as micro-services, we offer the possibility to participate in the experiments by submitting system outputs, i.e. in the form of pre-computed rankings and recommendations. We do so for two reasons. First, the Type A submissions resemble those of previous living labs and serve as the baseline in order to evaluate the feasibility of our new infrastructure design. Second, we hope

---

<sup>7</sup><https://github.com/stella-project/stella-app>

<sup>8</sup><https://lilas.stella-project.org/>

<sup>9</sup><https://github.com/stella-project/stella-server>

to lower technical barriers for some participants that want to submit the system outputs only. To make it easier for participants, we follow the familiar TREC run file syntax.

Depending on the chosen task, for each of the selected top-k queries or target items (identified by `<qid>`) a ranking or recommendation has to be computed in advance and then uploaded to the dashboard service. The upload process is tightly integrated into the GitHub ecosystem. Once the run file is uploaded, a new repository is automatically created from the previously described micro-template to which the uploaded run is committed. This is made possible thanks to GitHub API and access tokens. The run file itself is loaded as a pandas `DataFrame` into the running micro-service when the *indexing* endpoint is called. Upon request, the queries and target items are translated into the corresponding `<qid>` to filter the `DataFrame`. Due to manageable sizes of top-k queries and target items, the entire (compressed) run file can be uploaded to the repository and can be kept in memory after it is indexed as a `DataFrame`. As a technical safety check, we also integrate a dedicated verification tool<sup>10</sup> in combination with GitHub Actions to verify that the uploaded files follow the correct syntax.

#### 4.2.2. Type B - Docker Containers

Running fully-fledged ranking and recommendation systems as micro-services overcomes the restrictions of responses that are limited to top-k queries and target items. Therefore, we offer the possibility to integrate the entire systems as a Docker container into the STELLA infrastructure as part of Type B submissions. As pointed out earlier, participants fork the template of the micro-services and adapt it to their experimental system. While Docker and the implementation of pre-defined REST endpoints are hard requirements, participants have total freedom w.r.t. the implementation and tools they use within their container, i.e., they do not even have to build up on the Python web application that is provided in the template. Solely, the *index* endpoint and, depending on the chosen task, either the *ranking* or *recommendation* endpoint have to deliver technically correct results. For this purpose, we include unit tests in the template repository that can be run in order to verify that the Docker containers can be properly integrated. If these unit tests pass, the participants register the URL of the corresponding Git repository at the dashboard service. Later on, the system URL is added to the build file of the MCA when an update process is invoked. If the MCA is updated at the sites, newly submitted experimental systems are build from the Dockerfiles in the specified repositories.

#### 4.3. Baseline Systems

LIVIVO baseline system for ranking is built on Apache Solr and Apache Lucene. The index contains about 80 million documents from more than 50 data sources in multiple languages and about 120 searchable fields ranging from basic data such as Title, Abstract, Authors to more specific such as MeSH-Terms, availability or OCR-Data. For ranking, LIVIVO uses the Lucene default ranker which is a variant of TF-IDF; on top of it, a custom boosting is added. Newer documents as well as search queries occurring in title or author fields are boosted. An exact match of search phrases in title-field results in a very high boosting. Moreover LIVIVO uses a Lucene-based plugin which executes NLP-tasks like stemming, lemmatization, multilingual

---

<sup>10</sup>[https://github.com/stella-project/syntax\\_checker\\_CLI](https://github.com/stella-project/syntax_checker_CLI)

search; it also makes use of semantic technologies, mainly based on the Medical Subject Headings (MeSH) vocabulary.

The baseline system for recommendation of research data based on publications in Gesis Search utilizes Pyserini, a Python interface to the IR toolkit built on Lucene designed to support reproducible IR research. The baseline system for recommendation applies the SimpleSearcher of Pyserini that provides the entry point for sparse retrieval BM25 ranking using bag-of-words representations. The Lucene-based index contains abstracts and titles of all research data. The publication identifier (target item of the recommendation) is translated into the publication title, which, in turn, is used to query the index with a BM25 algorithm. Accordingly, the research data recommendations are based on the title and abstracts of the research data and queries made from the publication titles.

#### 4.4. Evaluation Metrics

Our logging infrastructure allows us to track search sessions and the corresponding interactions made by users. Each session comprises a specific site user, multiple queries (or target items) as well as the corresponding results and feedback data in the form of user interactions, primarily logged as clicks with timestamps.

Similar to previous living lab initiatives, we design our user-oriented experiments with interleaved result lists. Given a list with interleaved results and the corresponding clicks of users, we determine *Wins*, *Losses*, *Ties*, and the derived *Outcomes* for relative comparisons of the experimental and baseline systems [2]. Following previous living lab experiments, we implement the interleaving method by the *Team-Draft-Interleaving* algorithm [14]. More specifically, we refactored exactly the same implementation<sup>11</sup> for the highest degree of comparability.

Furthermore we follow Gingstad et al.’s proposal of a weighted score based on click events [15] and define the *Reward* as

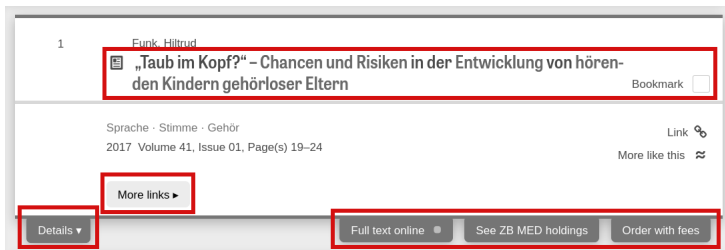
$$Reward = \sum_{s \in S} w_s c_s \quad (1)$$

where  $S$  denotes the set of all elements on a search engine result page (SERP) for which clicks are considered,  $w_s$  denotes the corresponding weight of the SERP element  $s$  that was clicked, and  $c_s$  denotes the total number of clicks on the SERP element  $s$ . The *Normalized Reward* is defined as

$$nReward = \frac{Reward_{exp}}{Reward_{exp} + Reward_{base}} \quad (2)$$

that is the sum of all weighted clicks on experimental results ( $Reward_{exp}$ ) normalized by the total *Reward* given by  $Reward_{exp} + Reward_{base}$ . Note that, only those clicks from the experimental systems where rankings were interleaved with results of the two compared systems are considered. Figure 5 shows the SERP elements that were logged at LIVIVO and the corresponding weights for our evaluations. We do not implement the *Mean Normalized Reward* proposed by Gingstad et al. due to a different evaluation setup. Our lab is organized in rounds

<sup>11</sup><https://bitbucket.org/living-labs/ll-api/src/master/ll/core/interleave.py>



SERP Element	$w_s$
Bookmark	10
Order	10
Fulltext	8
In Stock	8
More Links	2
Title	1
Details	1

**Figure 5 & Table 1:** Example illustrating the SERP elements for that clicks were logged at LIVIVO and the corresponding weights  $w_s$  according to Equation 1.

**Table 2**

Schedule for the LiLAS lab 2021

Event	Date
Data set release - documents for LIVIVO and GESIS Search	14 December 2020
Training phase 1	January + February 2021
Release of code tutorial for the living lab component	14 January 2021
Round 1 for GESIS Search	1 March - 28 March 2021
Round 1 for LIVIVO	5 March - 28 March 2021
Release of feedback data for round 1	29 March 2021
Training phase 2	30 March - 11 April 2021
Round 2 for GESIS Search	12 April - 24 May 2021
Round 2 for LIVIVO	19 April - 24 May 2021
Release of feedback data for round 2	11 and 18 May 2021

during which the systems as well as the underlying document collections are not modified and we already determine the *Normalized Reward* over all aggregated clicks of a specific round.

#### 4.5. Lab Rounds and Overall Lab Schedule

The lab was originally split in two separated rounds of 4 weeks each. Due to technical issues for LIVIVO round 1 was four days shorter and round 2 started one week later as planned. To compensate this, we decided to let round 2 last until 24 May 2021, so in total round 2 lasted nearly six instead of four weeks. An overview of the general LiLAS 2021 schedule is given in Table 2. Each participating groups received a set of feedback data after each round; the feedback was also made publicly available on the lab website<sup>12</sup>. Before each round a training phase was offered to allow the participants to build or adapt their systems to the new datasets or click feedback data.

<sup>12</sup><https://th-koeln.sciebo.de/s/OBm0NLEwz1RYI9N>

## 5. Participation

### 5.1. Team lemuren

Team lemuren participated in both rounds with pre-computed results and dockerized systems for the ad-hoc search task at LIVIVO [16]. For both rounds, they submitted two different approaches.

The pre-computed ranking results of `lemuren_elk` are based on built-in functions of Elasticsearch. This system uses a combination between the divergence from randomness model and the Jelinek-Mercer smoothing method for re-ranking candidate documents. The preprocessing pipeline implements stop-word removal, stemming and considers synonyms for medical and COVID19-related terms. The system was tuned only to the results in English.

`save_fami` is another pre-computed system. It also uses Elasticsearch combined with natural language processing (NLP) modules implemented with the Python package spaCy. Similar to the second submission for the pre-computed round, this dockerized system is build on top of Elasticsearch and spaCy. The indexing pipeline follows a multilingual approach supporting English and German languages. For both languages the system implements full solutions available in spaCy, either by the models `en_core_sci_lg` (English biomedical texts) or `de_core_news_lg` (general German texts). The system uses the Google Translator API<sup>13</sup> for language detection and automatic translating of incoming queries (from German to English and vice versa). For indexing and document-retrieval Elasticsearch was used with a custom boosting for MeSH and Chemical-tokens. `lemuren_elastic_only` (LEO) is the second dockerized system by this team which, different from LEPREP, relies only on Elasticsearchs built-in tools for indexing documents and processing queries. For indexing documents a custom ingestion pipeline is used to detect the documents language (English or German) and creating the corresponding language fields. Handling of basic acronyms was modeled by using the built-in word-delimiter function. Similar to LEPREP-System, LEO uses Google Translator API for automatic query translation. The system is complemented by a fuzzy match and fuzzy query-expansion to obtain better results for mistyped queries. Like `lemuren_elk` in round one, LEO also uses DFR and LMJelinekMercer to calculate a score and a similarity distance.

### 5.2. Team tekma

Team tekma contributed experiments to both rounds. In the first round, they submitted the pre-computed results of the system `tekma_s` for the ad-hoc search task at LIVIVO [17]. In the second round, they submitted pre-computed recommendations (covering the entire volume of publications) for the corresponding task at GESIS. Both systems are described below.

`tekma_s` used Apache Solr to index the document and used pseudo-relevance feedback to extend the queries for the ad-hoc search task. The system only considers documents in English. The system got few impressions and clicks in comparison to the baseline system. `tekma_n` participated in the second round producing pre-computed recommendations. They used Apache Solr BM25 ranking function and applied query expansion and data enrichment by adding the metadata translations and re-ranking the retrieved result using user feedback and KNN. To

---

<sup>13</sup><https://pypi.org/project/google-trans-new/>

**Table 3**

Number of Sessions, impressions, clicks and click through rate (CTR).

Evaluation round	Site	Sessions	Impressions	Clicks	CTR
Round 1	LIVIVO	2852	4658	2452	0.5264
Round 1	GESIS	4568	8390	152	0.0181
Round 2	LIVIVO	12962	25830	11562	0.4476
Round 2	GESIS	6576	12068	250	0.0207

generate the primary recommendations for a publication, they used publication fields as a query to search the indexed dataset.

### 5.3. Team GESIS Research

In addition to the baseline system, team GESIS Research contributed a fully dockerized system in both rounds [18]. `gesis_rec_pyterrier` implements a naive content-based recommendation without any advanced knowledge about user preferences and usage metrics. It uses the metadata available in both entity types, i.e., title, abstract, and topics. They employed the classical tfidf-based weighting model from the PyTerrier framework to obtain first-hand experience with the online evaluation. The indexing and query have been made of the combination of words in title, abstract, and research data topics and publications. They decided to submit the same experimental system for both rounds to gain more user feedback for their unique system. Even though only tfidf-based recommendations are implemented at the current state, it offers a good starting point for further experimentation with PyTerrier and the declarative manner of defining retrieval pipelines.

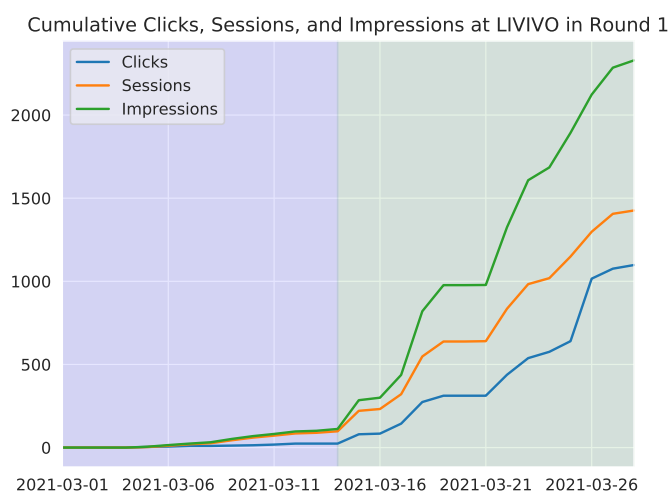
## 6. Results

Our experimental evaluations are twofold. First, we evaluate overall statistics of both rounds and sites. Second, we evaluate the performance of all participating systems based on the click data logged during the active periods. As mentioned before, the first round ran during four weeks from March 1st, 2021 to March 28th, 2021 and the second round for five weeks from April 17th, 2021 until May 24th, 2021 at LIVIVO and for six weeks from April 12th, 2021 until May 24th, 2021 at GESIS. To foster transparency and reproducibility of the evaluations, we release the corresponding evaluation scripts in an open-source GitHub repository<sup>14</sup>.

### 6.1. Overall evaluations of both rounds and sites

Table 3 provides an overview of the traffic logged in both rounds. In sum, substantially more sessions, impressions, and clicks were logged in the second round not only due a longer period but also because more systems contributed as Type B submissions. In the first round, systems deployed at LIVIVO were mostly contributed as Type A submissions, meaning their

<sup>14</sup><https://github.com/stella-project/stella-evaluations>



**Figure 6:** Cumulative sum of logged session data at LIVIVO before (blue) and after (green) the first fully dockerized system went online in the first round.

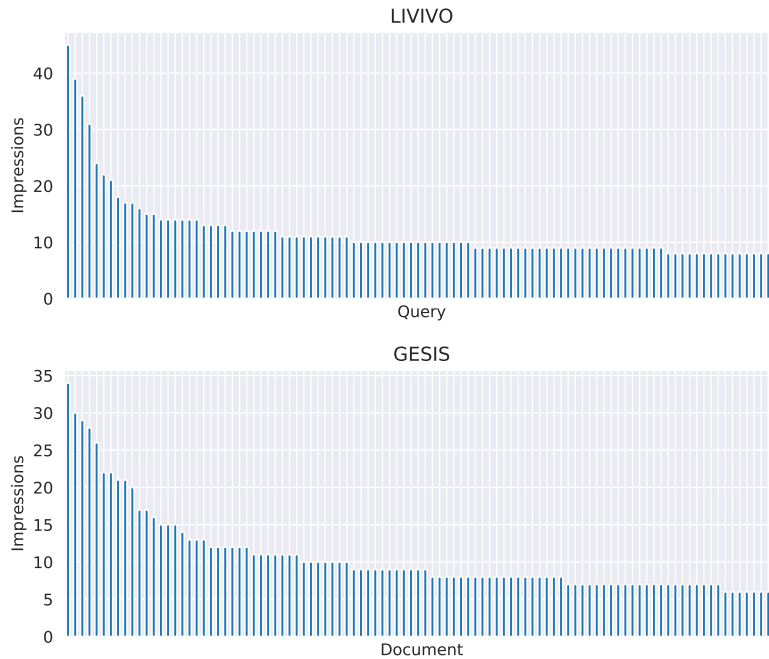
responses were restricted to pre-selected head queries. LIVIVO started the second round with full systems which delivered results for arbitrary queries and thus more session data was logged. GESIS started both rounds with the majority of systems contributed as type B submissions. In comparison to LIVIVO, more sessions and impressions were logged in the first round, but less recommendations were clicked. Similarly, there are less clicks in the second round in comparison to LIVIVO, which is also reflected by the Click-Through Rate (CTR) that is determined by the ratio between Clicks and Impressions. As mentioned before, GESIS introduced the recommendations of research datasets as a new service, and, presumably, users were not aware of this new feature.

Figure 15 shows the distributions of sessions and impressions over the entire time span from the beginning of the first round until the end of the second round. Note that, after the end of the first round, we did not log any interactions until the beginning of the second round. In comparison, the sessions and impressions are more uniformly distributed at GESIS. This can be explained by the deployment of type B systems from the early beginning of the first round and systems could provide recommendations for the entire volume of the publications.

During the first two weeks of the first round, the amount of logged data at LIVIVO is comparatively low due to systems with pre-computed results for pre-selected head queries. After that, the first type B systems was deployed and increasingly more user traffic could be redirected to our infrastructure. Figure 6 illustrates these effects. The cumulative sums of logged sessions, impressions, and clicks rapidly increased after the first Type B system got online in mid-March.

The logged impressions follow a power-law distribution for both rankings and recommendations as shown in Figure 7. Most of the impressions can be attributed to a few top-k queries (rankings) or documents (recommendations). Table 4 and 5 show the top ten queries and documents for that rankings and recommendations were made. Query strings were normalized by





**Figure 7:** Impressions vs. Query/Document

lower-casing and removing special characters. As it can be seen from Table 4 the COVID-19 pandemic has a clear influence on the query distributions: the most frequent and the fifth most frequent query are “covid19” and “covid”, respectively. Three of the ten most frequent queries are definitely German queries (“demenz”, “pflege”, “schlagenfall”). Others are either domain-specific or can also be interpreted as English queries. In Table 6 we report statistics about the queries logged during both rounds at LIVIVO. In both rounds, interaction data was logged for 11,822 unique queries with an average length of 2.9840 terms and each session had 1.9340 queries on average. Nine out of the ten most frequent target items of the recommendations at GESIS are publications with German titles as shown in Table 5.

Likewise the total number of clicks over queries and documents is extremely thin-tailed (cf. Figure 11 and 12). A large amount of the clicks at LIVIVO were made for the query “polyvinyl and nasal and packing” and LIVIVO’s internal server logs indicate a crawling process here. All other queries received 23 or less clicks. As mentioned before, less clicks were made at GESIS. Three clicks were made at maximum on recommendations for the most frequently clicked documents.

Similar power-law distributions can be observed for the total number of clicks over documents (rankings) and datasets (recommendations) in Figure 13 and 14, respectively. A few documents and datasets receive most of the clicks. Details about the corresponding items can be found in Table 13 and 14.

Another important aspect to be considered as part of the system evaluations is the position bias inherent in the logged data. Click decisions are biased towards the top ranks of the

**Table 4**

Top ten queries at LIVIVO during our logging periods of round 1 and 2. Query strings were lower-cased and special characters removed.

Rank	Query string	Impressions
1	covid19	45
2	demenz	39
3	guillian barre syndrome	36
4	polyvinyl and nasal and packing	31
5	covid	24
6	pflege	22
7	cancer	21
8	parkinson	18
9	depression	17
10	schlaganfall	17

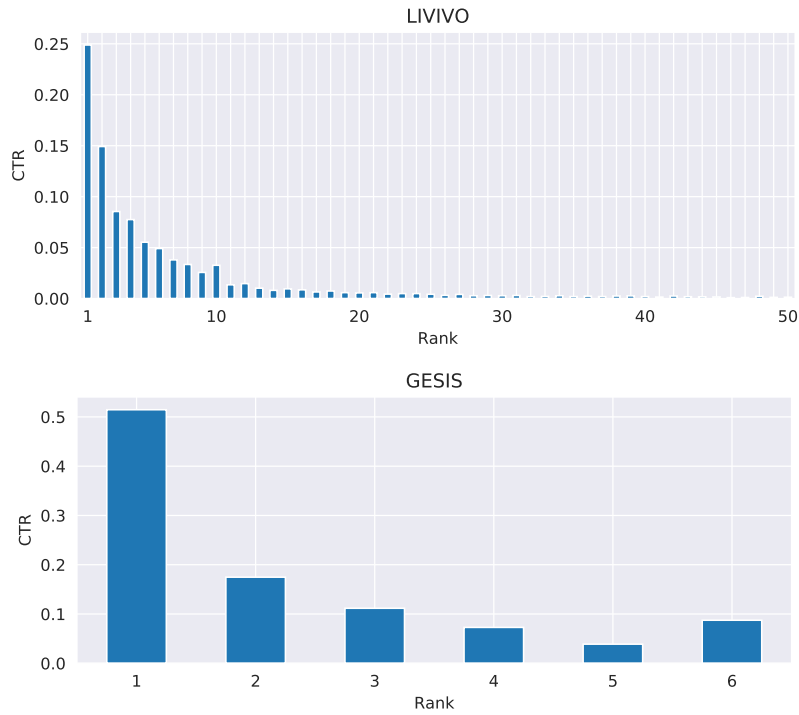
**Table 5**

Top ten documents for which recommendations were made at GESIS during our logging periods of round 1 and 2

Rank	Document title	Impressions
1	Die Nichtwähler : Politische Normalität oder wachsende Distanz zu den Parteien?	34
2	Doing Gender: Soziale Praktiken der Geschlechterunterscheidung	30
3	ZUMA-Informationssystem. Elektronisches Handbuch sozialwissenschaftlicher Erhebungsinstrumente	29
4	Situiertes Wissen : die Wissenschaftsfrage im Feminismus und das Privileg einer partialen Perspektive	28
5	Party identification, ideological preference, and the left-right dimension among western mass publics	26
6	Die soziale Konstruktion von Geschlecht : Erkenntnisperspektiven und gesellschaftstheoretische Fragen	22
7	Konsensfiktionen in Kleingruppen: dargestellt am Beispiel von jungen Ehen	22
8	SWLS Satisfaction with Life Scale	21
9	Entwicklung einer Skala zur Messung von Arbeitszufriedenheit (SAZ)	21
10	Gesundheitliche Ungleichheit / Health Inequalities	20

result lists as shown in Figure 8. For both use cases, the rankings and recommendations were displayed to users as vertical lists. Note that, GESIS restricted the recommendations to the first six recommended datasets and no pagination over the following recommended items was possible. LIVIVO shows ten results per page to its users, and as it can be seen from the logged data, users rarely click results beyond the fifth page.

In addition to “simple” clicks on ranked items, we logged specific SERP elements that were clicked at LIVIVO. Table 5 already provided an overview on which elements were logged and Figure 9 shows the CTR of these elements also follows a power-law distribution. The number



**Figure 8:** Click-through Rate (CTR) vs. Rank

**Table 6**

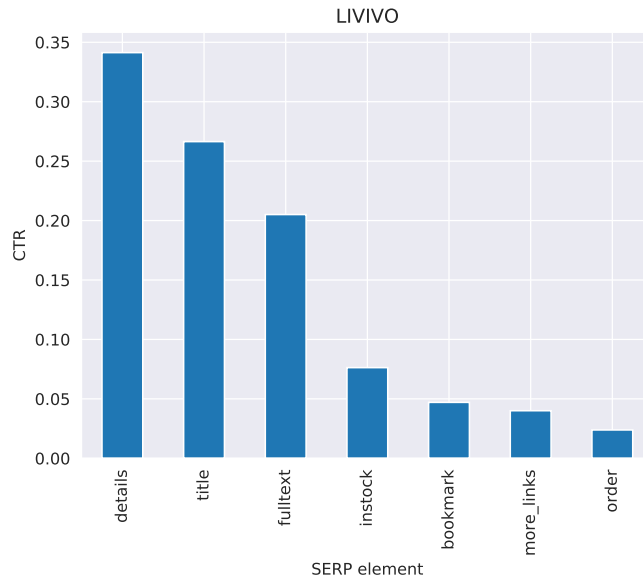
Statistics of the queries at LIVIVO

Number of Unique Queries	11822
Average Query Length [Terms]	2.9840
Average Number of Queries per Session	1.9340
Average Number of Clicks per Query	0.4547

of clicks is the highest for the *Details* button and it is followed by the *Title* and *Fulltext* click options. In comparison, the other four logged elements receive substantially less clicks.

## 6.2. System evaluations

An overview of all systems participating in our experiments is provided in Table 7. In the first round, three type A systems (*lemuren\_elk*, *tekmas*, *save\_fami*) were submitted and deployed at LIVIVO. They were also deployed in the second round, but did not receive any updates between the two rounds. Since there were no type B submissions in the first round for LIVIVO, we deployed the type B system *livivo\_rank\_pyserini* after two weeks in mid-March. It provided results for the entire volume of publications and rankings were based on the BM25 method. It was implemented with Pyserini [19] and the corresponding default



**Figure 9:** Click distribution on SERP elements at LIVIVO

settings<sup>15</sup>. In contrast to the other systems, it was online for the last two weeks of the first round only. In the second round, it was online in the first days until the other type B systems were ready to be deployed since we wanted to distribute the user traffic among the participants’ systems only. In the second round, two type B systems `lemuren_elastic_only` and `lemuren_elastic_preprocessing` were contributed. Both systems build up on Elasticsearch, whereas they differ by the pre-processing as outlined before. At GESIS, `gesis_rec_pyterrier`, submitted as type B system, was online in both rounds. In the first round, the only type A submission was `gesis_rec_precom` that was substituted in the second round by `tekma_n`. Both baseline systems at LIVIVO (`livivo_base`) and GESIS (`gesis_rec_pyserini`) were integrated as type B systems, remained unmodified, and could deliver results for every request.

Table 8 compares the experimental systems’ outcomes and the corresponding logged interactions and session data during the first round. Regarding the *Outcome* measure, none of the experimental systems was able to outperform the baseline systems. Note that the reported *Outcomes* of the baseline systems result from comparisons against all experimental systems. The systems with pre-computed rankings (type A submissions) received a total number of 32 clicks over a period of four weeks at LIVIVO. Since interaction data was sparse in the first round, we only received enough data for `livivo_rank_pyserini` to conduct significance tests. The reported p-value results from a Wilcoxon signed-rank test and shows a significant difference between the experimental and baseline system.

Table 9 shows the results of the second round. `tekma_n` was contributed as type A submission, but results were pre-computed for the entire volume of publications at GESIS. It replaced `gesis_rec_precom` and achieved a higher CTR compared to the other recommender systems.

<sup>15</sup>[https://github.com/stella-project/livivo\\_rank\\_pyserini](https://github.com/stella-project/livivo_rank_pyserini)

**Table 7**  
System overview

System name	Task	Type	Experimental	Round 1	Round 2
lemuren_elk	1	A	●	●	●
tekmas	1	A	●	●	●
save_fami	1	A	●	●	●
livivo_rank_pyserini	1	B	●	◐	◐
lemuren_elastic_only	1	B	●	○	●
lemuren_elastic_preprocessing	1	B	●	○	●
livivo_base	1	B	○	●	●
tekma_n	2	A	●	○	●
gegis_rec_precom	2	A	●	●	○
gegis_rec_pyterrier	2	B	●	●	●
gegis_rec_pyserini	2	B	○	●	●

**Table 8**  
Outcomes of Round 1. Dagger symbols (†) indicate baseline systems. Significant differences are denoted by an asterisk symbol (\*).

System	Win	Loss	Tie	Outcome	Sessions	Impressions	Clicks	CTR
gegis_rec_pyserini†	36	36	1	0.50	2284	4195	37	0.0088
gegis_rec_pyterrier	26	28	1	0.48	1968	3675	28	0.0076
gegis_rec_precom	10	8	0	0.56	316	520	11	0.0212
livivo_base†	332	234	67	0.59	1426	2329	677	0.2907
livivo_rank_pyserini	215	302	64	0.42*	1260	2135	517	0.2422
lemuren_elk	4	8	1	0.33	45	55	10	0.1818
tekmas	6	10	1	0.38	64	77	8	0.1039
save_fami	9	12	1	0.43	57	62	14	0.2258

Likewise, it achieves an *Outcome* of 0.62, which might be an indicator that it outperforms the baseline recommendations given by `gegis_rec_pyserini`. Unfortunately, we are not able to conduct any meaningful significance tests due to the sparsity of click data. At LIVIVO, the systems with pre-computed rankings (type A submissions) received a comparable amount of clicks similar to the first round. In sum, all three systems received a total number of 35 clicks over a period of five weeks. Even though, click data is sparse and interpretations have to be made carefully, the relative ranking order of these three systems is preserved in the second round (e.g. in terms of the *Outcome*, total number of clicks, or CTR).

In the second round, no experimental system could outperform the baseline system at LIVIVO. Both experimental type B systems `lemuren_elastic_only` and `lemuren_elastic_preprocessing` achieve significantly lower *Outcome* scores as the baseline. However, the second system has substantially lower *Outcome* and CTR scores. Both systems share a fair amount

of the same methodological approach and only differ by the processing of the input text. In this case, the system performance does not seem to benefit from this specific pre-processing step, when interpreting clicks as positive relevance signals. The third type B system at LIVIVO `livivo_rank_pyserini` did not participate the entire second round, since we took it offline as soon as the other type B systems were available. Despite having participated in comparatively less experiments than in the first round (1260 sessions vs. 243 sessions), the system achieves in both rounds comparable results in terms of *Outcome* and CTR scores. This circumstance raises the question for how long systems have to be online to deliver reliable performance estimates. Figure 10 provides an overview of how the *Outcome* score evolves over aggregated sessions for different systems and rounds. As the figures show, after a certain number of sessions, the outcome tends to stabilize. In our future work, we want to investigate how much sessions (or online time) is required to deliver meaningful estimates of system performance in terms of the *Outcome* and other measures derived from interleaving experiments.

Previous studies showed that a system is more likely to win if its documents are ranked at higher positions [20]. As part of our experimental evaluations, we can confirm this circumstance. We also determined the Spearman correlation between an interleaving outcome (1: win, -1: loss, 0: tie) and the highest ranked position of a document contributed by an experimental system. At both sites, we see a weak but significant correlation (LIVIVO:  $\rho = -0.0883$ ,  $p = 1.3535e - 09$ ; GESIS:  $\rho = -0.3480$ ,  $p = 4.7422e - 07$ ).

One shortcoming of the previous measures derived from interleaving experiments is the simplified interpretation of click interactions. As outlined in Section 4, by weighting clicks differently, it is possible to account for the meaning of the corresponding SERP elements. Table 10 shows the total number of clicks on SERP elements for each systems and the *Normalized Reward* (nReward) resulting from the weighting scheme given in Figure 5. We compare the total number of clicks of those (interleaving) experiments in which the experimental and baseline systems delivered results. As it can be seen, comparing systems by clicks on different SERP elements, provides a more diverse analysis. For instance, some of the systems achieve higher numbers of clicks (and CTRs) for some SERP elements in direct comparison to the baseline systems. `livivo_rank_pyserini`, `lemuren_elastic_only` got more clicks on the *Bookmark* element than the baseline system, while all systems achieve lower numbers of total clicks.

None of the systems could outperform the baseline system in terms of the nReward measure, but in comparison to the *Outcome* scores, there is a more balanced ratio between the nReward scores that also accounts for the meaning of specific clicks. Likewise, it accounts for clicks even if the experimental system did not “win” in the interleaving experiment. In Table 10 we compare the total number of clicks over multiple sessions. While the Win, Loss, Tie, and Outcome only measure if there have been more clicks in a single experiment, the nReward also considers those clicks that were made in experiments in which the experimental system did not necessarily win.

## 7. Conclusions

The Living Labs for Academic Search (LiLAS) lab re-introduced the living lab paradigm with a focus on tasks in the domain of academic search. The lab offered the possibility to participate

**Table 9**

Outcomes of Round 2. Dagger symbols (†) indicate baseline systems. Significant differences are denoted by an asterisk symbol (\*).

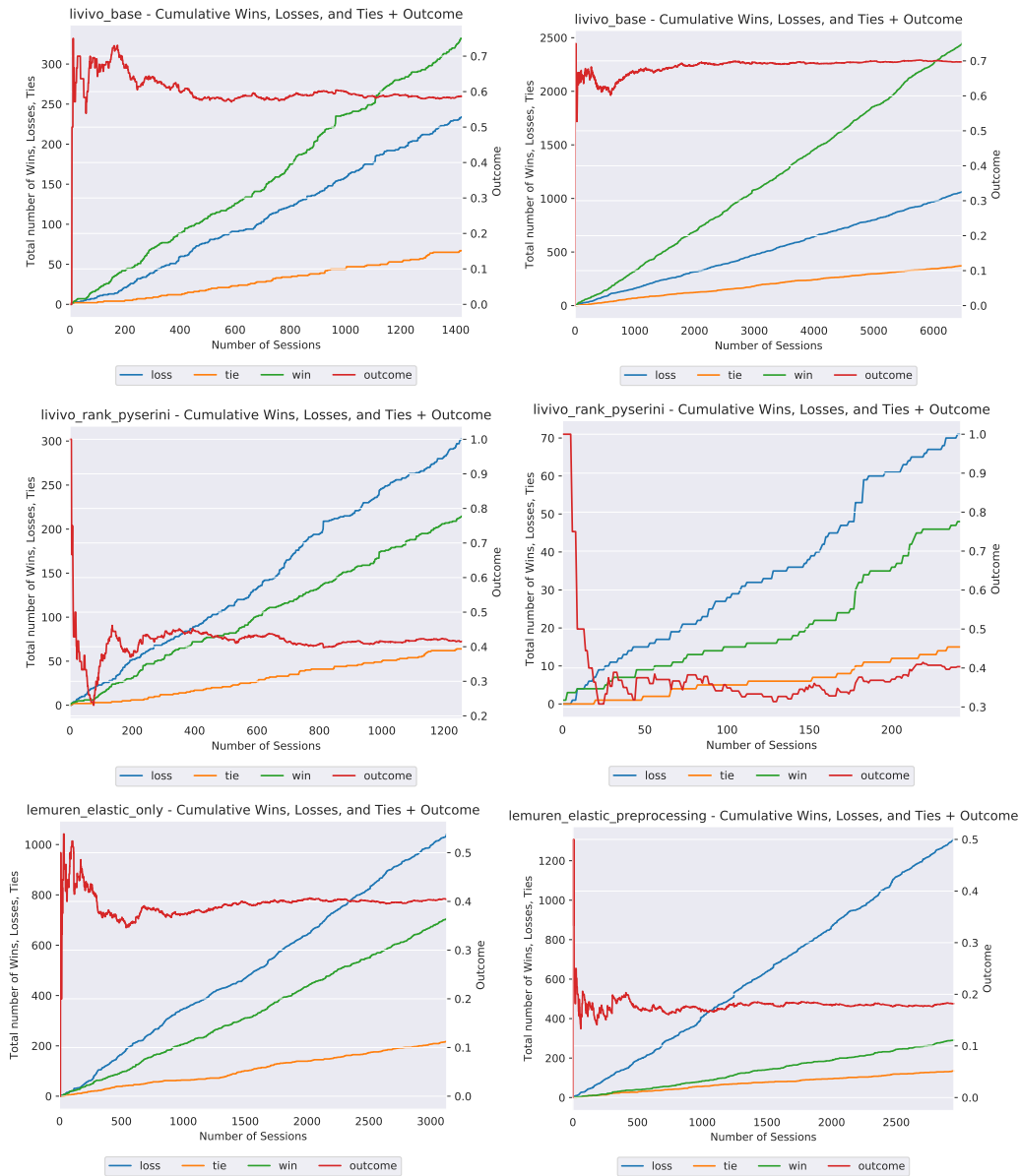
System	Win	Loss	Tie	Outcome	Sessions	Impressions	Clicks	CTR
gesis_rec_pyserini†	51	68	2	0.43	3288	6034	53	0.0088
gesis_rec_pyterrier	26	25	1	0.51	1529	2937	27	0.0092
tekma_n	42	26	1	0.62	1759	3097	45	0.0145
livivo_base†	2447	1063	372	0.70	6481	12915	3791	0.2935
livivo_rank_pyserini	48	71	15	0.40	243	434	112	0.2581
lemuren_elastic_only	707	1042	218	0.40*	3131	6274	1273	0.2029
lemuren_elastic_preprocessing	291	1308	135	0.18*	2948	6026	570	0.0946
lemuren_elk	6	13	0	0.32	61	69	10	0.1449
tekma_s	4	7	1	0.36	36	42	5	0.1190
save_fami	7	6	3	0.54	62	70	20	0.2857

in two different tasks, which were either dedicated to ad-hoc search in the Life Sciences or research data recommendations in the Social Sciences. Participants were provided with datasets and access to the underlying search portals for experimentation. For both tasks, participants could contribute their experimental systems either by pre-computed outputs for selected queries (or target items) or as fully-fledged dockerized systems. In total, we evaluated nine experimental systems out of which seven were contributed by three participating groups. In sum, two groups contributed experiments that cover pre-computed rankings and fully dockerized systems at LIVIVO and pre-computed recommendations at GESIS. The GESIS research team contributed another completely dockerized recommendation system. Our experimental setup is based on interleaving experiments that combine experimental results with those from the corresponding baseline systems at LIVIVO and GESIS. In accordance with the living lab paradigm, our evaluations are based on user interactions, i.e. in the form of click feedback.

A key component of the underlying infrastructure is the integration of experimental ranking and recommendation systems as micro-services that are implemented with the help of Docker. The LiLAS lab was the first test-bed to use this evaluation service and it exemplified some of the benefits resulting from the new infrastructure design. First of all, completely dockerized systems can overcome the restrictions of results limited to filtered lists of top-k queries or target items. Significantly more data and click interactions can be logged if the experimental systems can deliver results on-the-fly for arbitrary requests of rankings and recommendations. As a consequence, this allows much more data aggregation in a shorter period of time and provides a solid basis for statistical significance tests.

Furthermore, the deployment effort for site providers and organizers is considerably reduced. Once the systems are properly described with the corresponding Dockerfile, they can be rebuild on purpose, exactly as the participants and developers intended them to be. Likewise, the entire infrastructure service can be migrated with minimal costs due to Docker. However, we





**Figure 10:** Dockerized systems that were deployed at LIVIVO in Round 1 and 2 (livivo\_base and livivo\_rank\_pyserini) and Round 2 (lemuren\_elastic\_only and lemuren\_elastic\_preprocessing).

hypothesize that one reason for the low participation might be the technical overhead for those who were not already familiar with Docker. On the other hand, the development efforts pay off. If the systems are properly adapted to the required interface and the source code is available in a public repository, the (IR) research community can rely on these artifacts that make the experiments transparent and reproducible.

**Table 10**

Experimental systems of round 2 and the corresponding number of clicks on SERP elements, total number of clicks, and the *Reward* score.

	Bookmark	Details	Fulltext	In Stock	More Links	Order	Title	Total Clicks	nReward
livivo_rank_pyserini	182	341	176	55	62	28	263	1107	0.4367
livivo_base	180	443	228	154	57	29	329	1420	0.5633
lemuren_elastic_only	63	832	481	107	105	54	638	2280	0.4045
livivo_base	56	1066	646	295	129	85	858	3135	0.5955
lemuren_elastic_preprocessing	23	355	257	23	28	21	285	992	0.2143
livivo_base	69	1190	762	301	119	82	934	3457	0.7857
lemuren_elk	1	13	16	0	2	0	10	42	0.4242
livivo_base	1	24	7	14	1	0	20	67	0.5758
tekmas	2	11	2	2	1	0	6	24	0.3430
livivo_base	0	13	6	7	0	1	9	36	0.6570
save_fami	11	21	9	3	1	1	16	62	0.5496
livivo_base	8	13	7	5	2	1	6	42	0.4504
All experimental systems	282	1573	941	190	199	104	1218	4507	0.3485
livivo_base	314	2749	1656	776	308	198	2156	8157	0.6515

Thus, we address the reproducibility of these living lab experiments mostly from a technological point of view, in the sense that we can repeat the experiments in the future with reduced efforts, since the participating systems are openly available and should be reconstructible with the help of the corresponding Dockerfiles. Future work should investigate how feasible it is to rely on the Dockerfiles for the long-term preservation. Since experimental systems are rebuilt each time with the help of the Dockerfile, updates of the underlying dependencies might be a threat to the reproducibility. An intuitive solution would be the integration of pre-built Docker images that may allow a longer reproducibility. Apart from the underlying technological aspects, the reproducibility of the actual experimental results has to be investigated. Our experimental setup would allow to answer questions with regard to the reproducibility of the experimental results over time and also across different domains (e.g. Life vs. Social Sciences).

Most of the evaluation measures are made for interleaving experiments that also depend on the results of the baseline system and not solely on those of an experimental system. We have not investigated yet, if the experimental results follow a transitive relation: if the experimental system A outperforms the baseline system B, denoted as  $A \succ B$ , and the baseline system B outperforms another experimental system C ( $B \succ C$ ), can we conclude that system A would also outperform system C ( $A \succ C$ )? As the evaluations showed, click results are heavily biased towards the first ranks and likewise they are context-dependent, i.e. they depend on the entire result list and single click decisions have to be interpreted in relation

to neighboring and previously seen results and further evaluations in these directions would require counterfactual reasoning. Nonetheless, in the second round it was illustrated how our infrastructure service can be used for incremental developments and component-wise analysis of experimental systems. The two experimental systems `lemuren_elastic_only` and `lemuren_elastic_preprocessing` follow a similar approach and only differ by the pre-processing component that has been shown not to be of any benefit.

In addition to established outcome measures of interleaving experiments (Win, Loss, Tie, Outcome), we also account for the meaning of clicks on different SERP elements. In this context, we implement the Reward measure that is the weighted sum of clicks on different elements corresponding to a specific result. Even though most of the experimental systems could not outperform the baseline systems in terms of the overall scores, we see some clear differences between the system performance, which allow us to assess a system's merits more thoroughly, when the evaluations are based on different SERP elements.

Overall, we consider our lab as a successful advancement to previous living lab experiments. We were able to exemplify the benefits of fully dockerized systems delivering results for arbitrary results on-the-fly. Furthermore, we could confirm several previous findings, for instance the power laws underlying the click distributions. Additionally, we were able to conduct more diverse comparison by differentiating between clicks on different SERP elements and accounting for their meaning. Unfortunately, we could not attract many participants, leaving some aspects not tested, e.g. how many systems/experiments can be run simultaneously considering the limitations of the infrastructure design, hardware requirements, server load and user traffic. Likewise, no experimental ranking system could outperform the baseline system. In the future, it might be helpful to provide participants with open and more transparent baseline systems they can build upon. Some of the pre-computed experimental ranking and recommendations seem to deliver promising results; however, the evaluations need to be interpreted with care due to the sparsity of the available click data. As a way out, we favor continuous evaluations freed from the time limits of rounds, in order to re-frame the introduced living lab service as an ongoing evaluation challenge. The corresponding source code can be retrieved from a public GitHub project<sup>16</sup> and we plan to release the aggregated session data as a curated research dataset.

## Acknowledgments

This work was supported by DFG (project no. 407518790).

## References

- [1] A. Lommatzsch, B. Kille, F. Hopfgartner, L. Ramming, Newsreel multimedia at mediaeval 2018: News recommendation with image and text content, in: Working Notes Proceedings of the MediaEval 2018 Workshop, CEUR-WS, 2018.
- [2] A. Schuth, K. Balog, L. Kelly, Overview of the living labs for information retrieval evaluation (LL4IR) CLEF lab 2015, in: J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. J. F. Jones,

---

<sup>16</sup><https://github.com/stella-project>

- E. SanJuan, L. Cappellato, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015*, Toulouse, France, September 8-11, 2015, Proceedings, volume 9283 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 484–496. doi:10.1007/978-3-319-24027-5\_47.
- [3] K. Balog, A. Schuth, P. Dekker, P. Schaer, N. Tavakolpoursaleh, P.-Y. Chuang, Overview of the trec 2016 open search track, in: *Proceedings of the Twenty-Fifth Text REtrieval Conference (TREC 2016)*. NIST, 2016.
- [4] D. J. de Solla Price, *Little Science, Big Science*, Columbia University Press, New York, 1963.
- [5] J. Schaible, T. Breuer, N. Tavakolpoursaleh, B. Müller, B. Wolff, P. Schaer, Evaluation infrastructures for academic shared tasks, *Datenbank-Spektrum* 20 (2020) 29–36. doi:10.1007/s13222-020-00335-x.
- [6] F. Hopfgartner, K. Balog, A. Lommatzsch, L. Kelly, B. Kille, A. Schuth, M. Larson, Continuous Evaluation of Large-Scale Information Access Systems: A Case for Living Labs, in: N. Ferro, C. Peters (Eds.), *Information Retrieval Evaluation in a Changing World*, volume 41, Springer International Publishing, Cham, 2019, pp. 511–543. doi:10.1007/978-3-030-22948-1\_21, series Title: The Information Retrieval Series.
- [7] E. M. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, L. L. Wang, TREC-COVID: constructing a pandemic information retrieval test collection, *CoRR abs/2005.04474* (2020). arXiv:2005.04474.
- [8] W. Yang, K. Lu, P. Yang, J. Lin, 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'19*, ACM Press, Paris, France, 2019, pp. 1129–1132. doi:10.1145/3331184.3331340.
- [9] T. G. Armstrong, A. Moffat, W. Webber, J. Zobel, 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 '09*, ACM, Hong Kong, China, 2009, pp. 601–610. doi:10.1145/1645953.1646031.
- [10] Z. Carevic, P. Schaer, 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, April 13, 2014, volume 1143 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2014, pp. 37–44.
- [11] P. Schaer, T. Breuer, L. J. Castro, B. Wolff, J. Schaible, N. Tavakolpoursaleh, Overview of lilas 2021 - living labs for academic search, in: K. S. Candan, B. Ionescu, L. Goeriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Twelfth International Conference of the CLEF Association (CLEF 2021)*, volume 12880 of *Lecture Notes in Computer Science*, 2021.
- [12] B. Müller, C. Poley, J. Pössel, A. Hagelstein, T. Gübitz, LIVIVO – the Vertical Search Engine for Life Sciences, *Datenbank-Spektrum* 17 (2017) 29–34. URL: <https://doi.org/10.1007/s13222-016-0245-2>. doi:10.1007/s13222-016-0245-2.
- [13] D. Hienert, D. Kern, K. Boland, B. Zapilko, P. Mutschke, 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, June 2-6, 2019, 2019, pp. 148–157. doi:10.1109/JCDL.2019.00030.

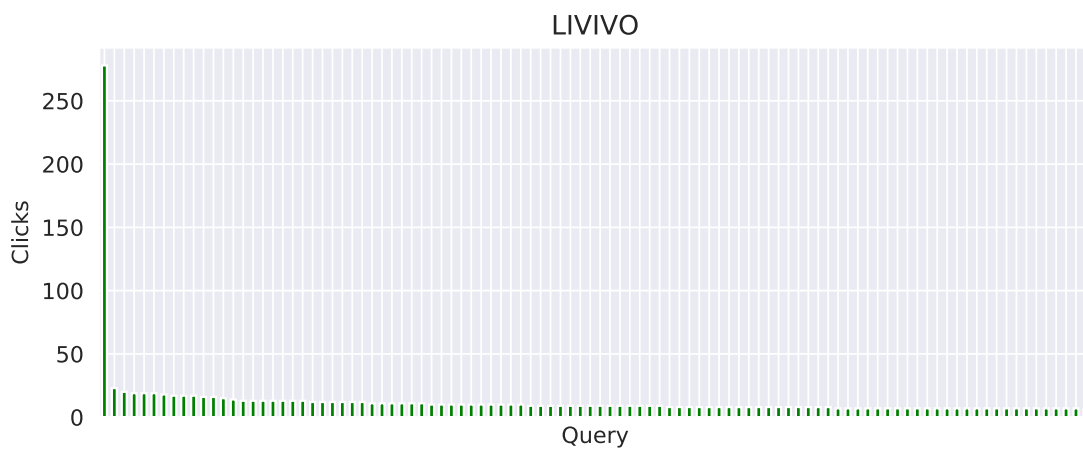
- [14] F. Radlinski, M. Kurup, T. Joachims, How does clickthrough data reflect retrieval quality?, in: J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K. Choi, A. Chowdhury (Eds.), Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008, ACM, 2008, pp. 43–52. doi:10.1145/1458082.1458092.
- [15] K. Gingstad, Ø. Jekteberg, K. Balog, Arxivdigest: A living lab for personalized scientific literature recommendation, in: M. d’Aquin, S. Dietze, C. Hauff, E. Curry, P. Cudré-Mauroux (Eds.), CIKM ’20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, ACM, 2020, pp. 3393–3396. doi:10.1145/3340531.3417417.
- [16] A. H. M. Tran, A. Kruff, J. Thos, C. Kraß, M. Reiners, F. Ax, S. Brech, S. Gharib, V. Pawlas, Ad-hoc retrieval of scientific documents on the livivo search portal, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2021.
- [17] J. Keller, L. P. M. Munz, Tekma at clef-2021: Bm-25 based rankings for scientific publication retrieval and data set recommendation, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2021.
- [18] N. Tavakolpoursaleh, S. Schaible, Pyterrier-based research data recommendations for scientific articles in the social sciences, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2021.
- [19] J. Lin, X. Ma, S. Lin, J. Yang, R. Pradeep, R. Nogueira, Pyserini: An easy-to-use python toolkit to support replicable IR research with sparse and dense representations, CoRR abs/2102.10073 (2021). arXiv:2102.10073.
- [20] R. Jagerman, K. Balog, M. de Rijke, Opensearch: Lessons learned from an online evaluation campaign, J. Data and Information Quality 10 (2018) 13:1–13:15.

## A. Appendix

**Table 11**

Top ten queries clicked at LIVIVO during both rounds. Query strings were normalized and special characters removed.

Rank	Query string	Clicks
1	polyvinyl and nasal and packing	278
2	nabelschnur	23
3	rtms and doc	20
4	vegan	19
5	fußball subkultur	19
6	anrede	19
7	skoliose and schroth	18
8	clown therapy	17
9	vegan ernährung kind	17
10	mammakarzinom	17

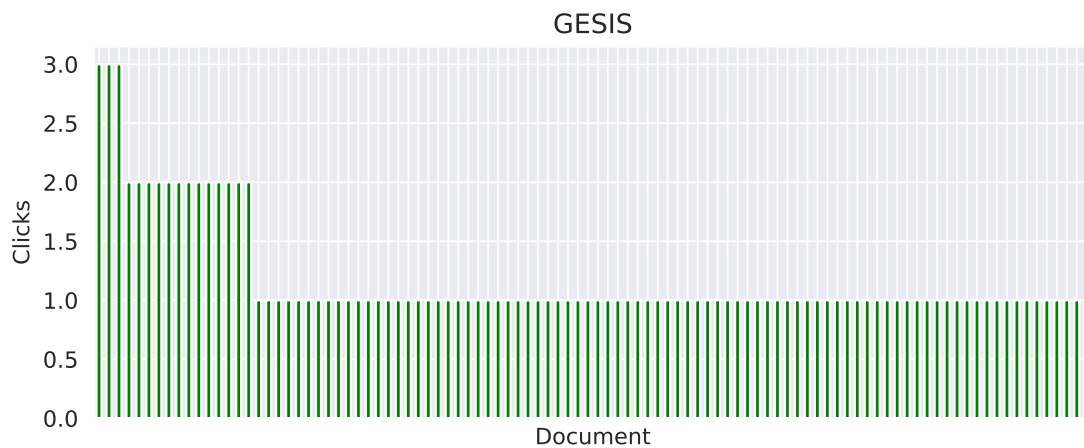


**Figure 11:** Number of clicks versus Query at LIVIVO in both rounds.

**Table 12**

Top ten documents for which recommendations were clicked at GESIS during both rounds.

Rank	Document title	Clicks
1	SWLS Satisfaction with Life Scale	3
2	Nichtwähler der Bundestagswahl 2017	3
3	Die Nichtwähler : Politische Normalität oder wachsende Distanz zu den Parteien?	3
4	Sozialer Zusammenhalt in Deutschland 2017	2
5	Die Abwertung der Anderen : eine europäische Zustandsbeschreibung zu Intoleranz, Vorurteilen und Diskriminierung	2
6	The political participation of disabled people in Europe: rights, accessibility and activism	2
7	Trade union decline and what next: is Germany a special case?	2
8	ALLBUScompact - Kumulation 1980-2014 Variable Report	2
9	Medienkritikfähigkeit messbar machen: Analyse medienbezogener Fähigkeiten bei Eltern von 10- bis 15-Jährigen	2
10	Substanzkonsum in der Allgemeinbevölkerung in Deutschland : Ergebnisse des Epidemiologischen Suchtsurveys 2015	2

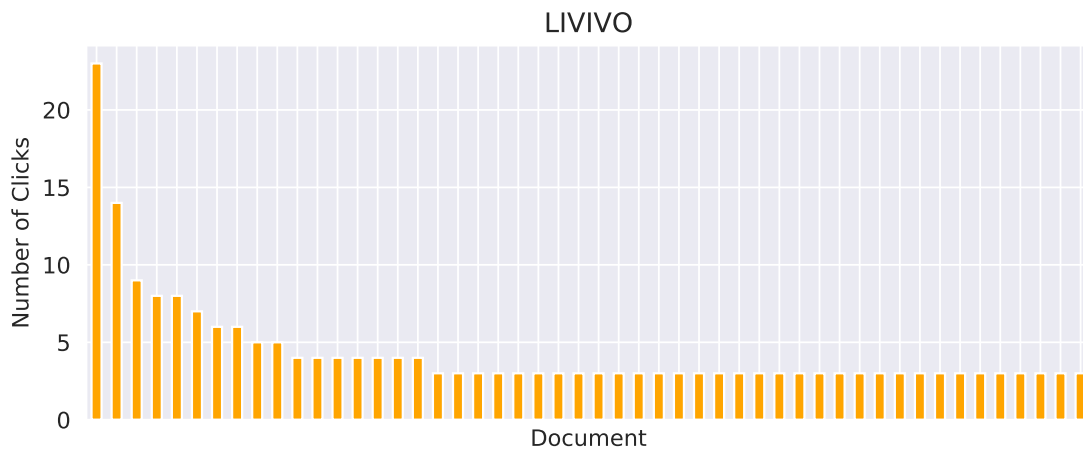
**Figure 12:** Number of clicks versus Document at GESIS in both rounds.



**Table 13**

Titles of the top ten documents clicked at LIVIVO during both rounds.

Rank	Query string	Clicks
1	Taub im Kopf? – Chancen und Risiken in der Entwicklung von hörenden Kindern gehörloser Eltern	23
2	Superheilmittel Vitamin C	14
3	Guolin Qigong	9
4	Taschenlehrbuch Histologie (1058335)	8
5	Palliativversorgung von Kindern, Jugendlichen und jungen Erwachsenen	8
6	Medical clown support is associated with better quality of life of children with food allergy starting oral immunotherapy	7
7	Wikibooks	6
8	Taschenlehrbuch Histologie (B4466071)	6
9	Physiotherapie bei Parkinson-Syndromen	5
10	Histologie	5

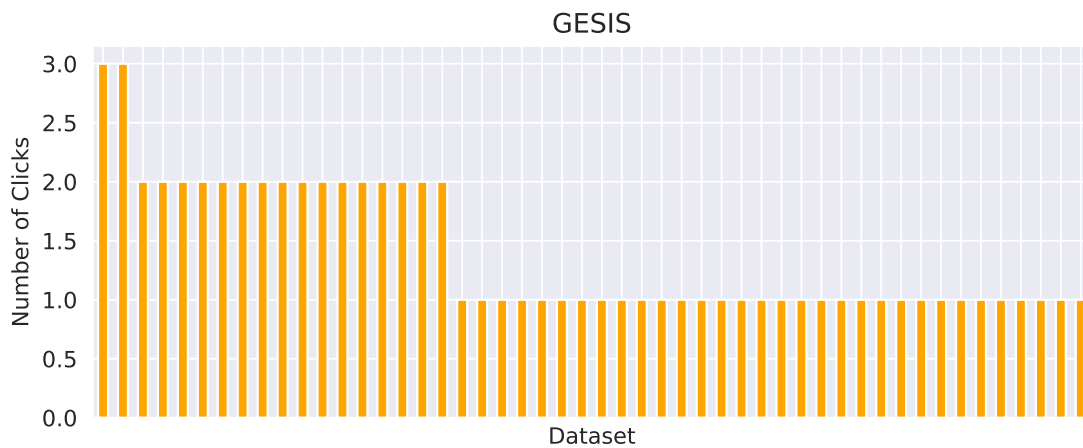


**Figure 13:** Number of clicks versus Document at LIVIVO in both rounds.

**Table 14**

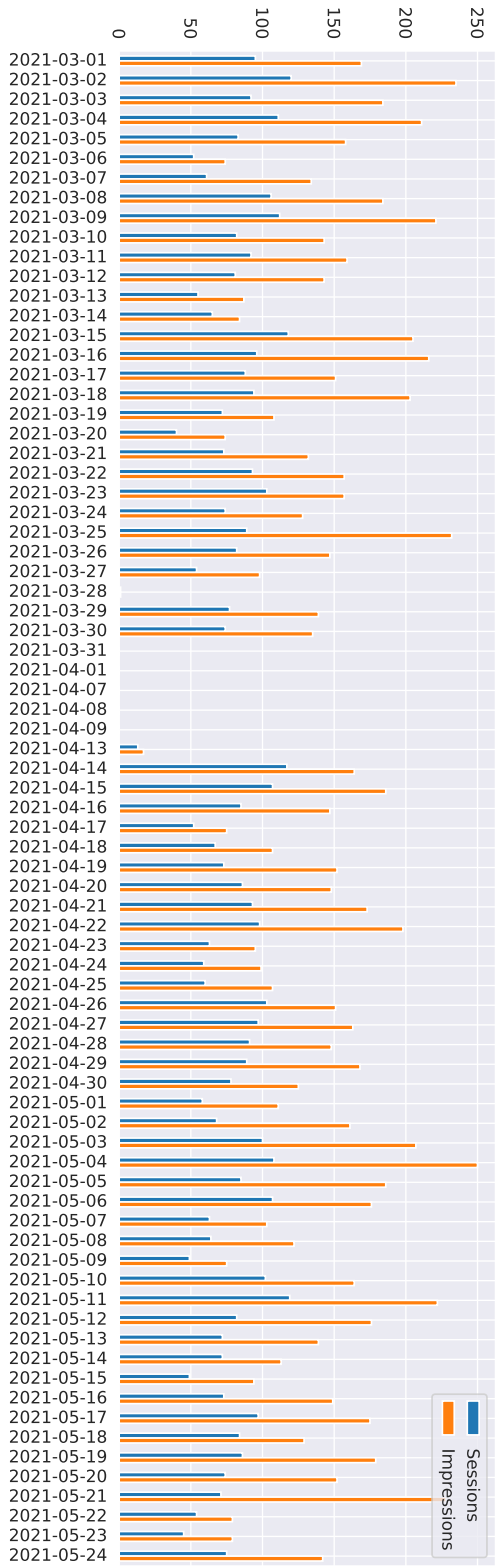
Top ten datasets clicked at GESIS during both rounds.

Rank	Document title	Clicks
1	Nichtwähler in Deutschland 2005 & 2009 Non-voters in Germany 2005 & 2009	3
2	Vertrauen in Staat und Gesellschaft während der Corona-Krise (April 2020)	3
3	EUSI: Datenbank zum Europäischen System Sozialer Indikatoren, 1950-2013	2
4	Landtagswahl in Bayern 2018	2
5	Satisfaction with Life Scale (CAPS-LIFESAT module)	2
6	Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUScompact - Kumulation 1980-2014	2
7	Mannheimer Corona-Studie	2
8	Naturbewusstsein 2015	2
9	Soziales Nachhaltigkeitsbarometer der Energiewende	2
10	Transitions and Old Age Potential: Übergänge und Alternspotenziale (TOP) - 1. und 2. Welle	2

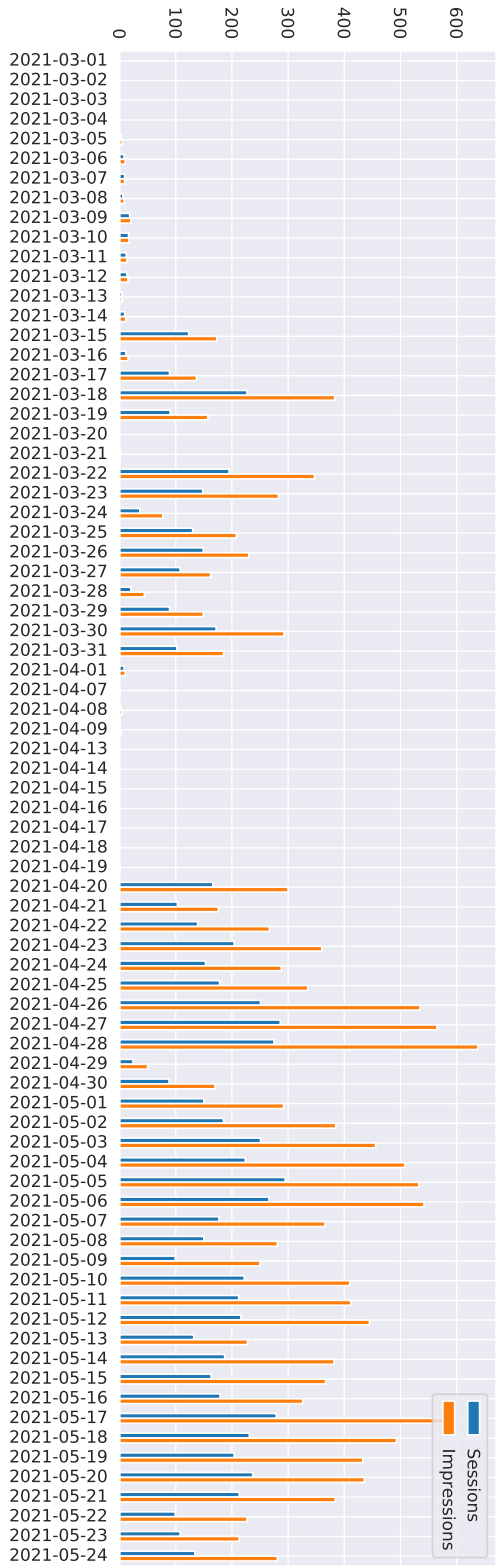


**Figure 14:** Number of clicks versus Dataset at GESIS in both rounds.

Figure 15: Sessions and Impressions at LIVIVO (livivo\_base) and GESIS (gesis\_rec\_pyserini) during both rounds.



Sessions vs. Impressions - gesis\_rec\_pyserini



Sessions vs. Impressions - livivo\_base