
Welcome to Cologne - Welcome to TH Köln!

Online Information Retrieval Evaluation using the STELLA Framework

Timo Breuer, Zeljko Carevic, Leyla Jael Castro,
Daniel Hienert, and Philipp Schaer

STELLA Community Workshop @ TH Köln: 20 June 2022; Cologne, Germany

Technology
Arts Sciences
TH Köln

gesis
Leibniz-Institut
für Sozialwissenschaften

ZB MED
Wissen für
Mensch & Umwelt

Schedule for today…

13:00 Welcome and Introduction
13:15 Overview on STELLA I and LiLAS Lab
14:00 Presentation of use case TIB
14:20 Presentation of use case DIPF
14:40 Presentation of use case ZBW
15:00 Coffee break
15:30 2 Breakout sessions (e.g. technological / methodological / evaluation strategy)
16:15 Round-up and discussion, plans for STELLA II
17:00 Summary and farewell
17:30 end of workshop / evening

If you like: Dinner together with the JCDL Doctoral Consortium at Mainzer Hof (5 minutes walk, arrive between 18-19h).

Did you know...?

- After WW2 Deutsche Lufthansa was re-founded in Cologne and its headquarter was here at Claudiusstraße since 1955
- This building previously was part of University of Cologne.
- Until 1970 this room (Rotunde) was the meeting room of the managing directors.
- Rotunde was designed to resemble the design of an air traffic control tower.

Image source: Werner Müller, http://www.luftfahrtarchiv-koeln.de/

Why are we here?

Common interest in Online Evaluation of IR/RecSys in Academic Systems

- **Outcome I: Learn more about the STELLA project and the framework and about results so far**

  → Overview talk on STELLA by Philipp

- **Outcome II: Learn more about your ideas and usecases**

  → Usecase presentations/pitches by you

- **Outcome III: Make plans for STELLA II** (You all submitted a LOI for STELLA II - Thank you!)

  → Small workshops in the afternoon to work on ideas, usecases and research questions

Common interests...

More realistic academic retrieval and recommendation **evaluation**

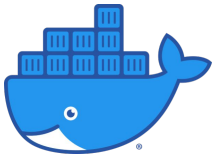
Partner web sites from the **social sciences** and **life science**

- I - Living Labs

- II - Academic Search

Evaluate within academic **live search systems** with **real users**

More than paper retrieval: **Data-set retrieval**, **cross-language retrieval**, **cross-content recommendations**...



STELLA evaluation infrastructure based on **Docker**

Why we care about Living Labs!

Offline ('TREC-style') evaluation has a long history but it's limitations, flaws, and critiques...

- T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel (2009): Improvements that don't add up: Ad-hoc retrieval results since 1998.
- J. Lin (2018): The neural hype and comparisons against weak baselines.
- M. F. Dacrema, P. Cremonesi, and D. Jannach (2019): Are we really making much progress? A worrying analysis of recent neural recommendation approaches.
- ...

What do we mean by Living Lab evaluation?

We don't create artificial evaluation situations but observe how regular users interact with real systems while using it!

- We observe implicit behaviour (like clicks, bookmarks, likes, etc.)
- We infer differences in user behaviour from experimental differences in the search systems (using A/B testing, interleaving, …)
- We run statistical tests to confirm the differences

Why we care about Academic Search!

The problems we face in Academic Search are not simple ac-hoc search issues

- Many systems still rely on bibliographical data with specialized vocabularies, thesauri, complex classification systems
- This introduces classic retrieval problems like wording issues due to domain or field differences
- Scientific communication (citations) are rarely exploited due to a lack of citation-based pools
- Scientific retrieval has to deal with more than papers! Research datasets, surveys, …

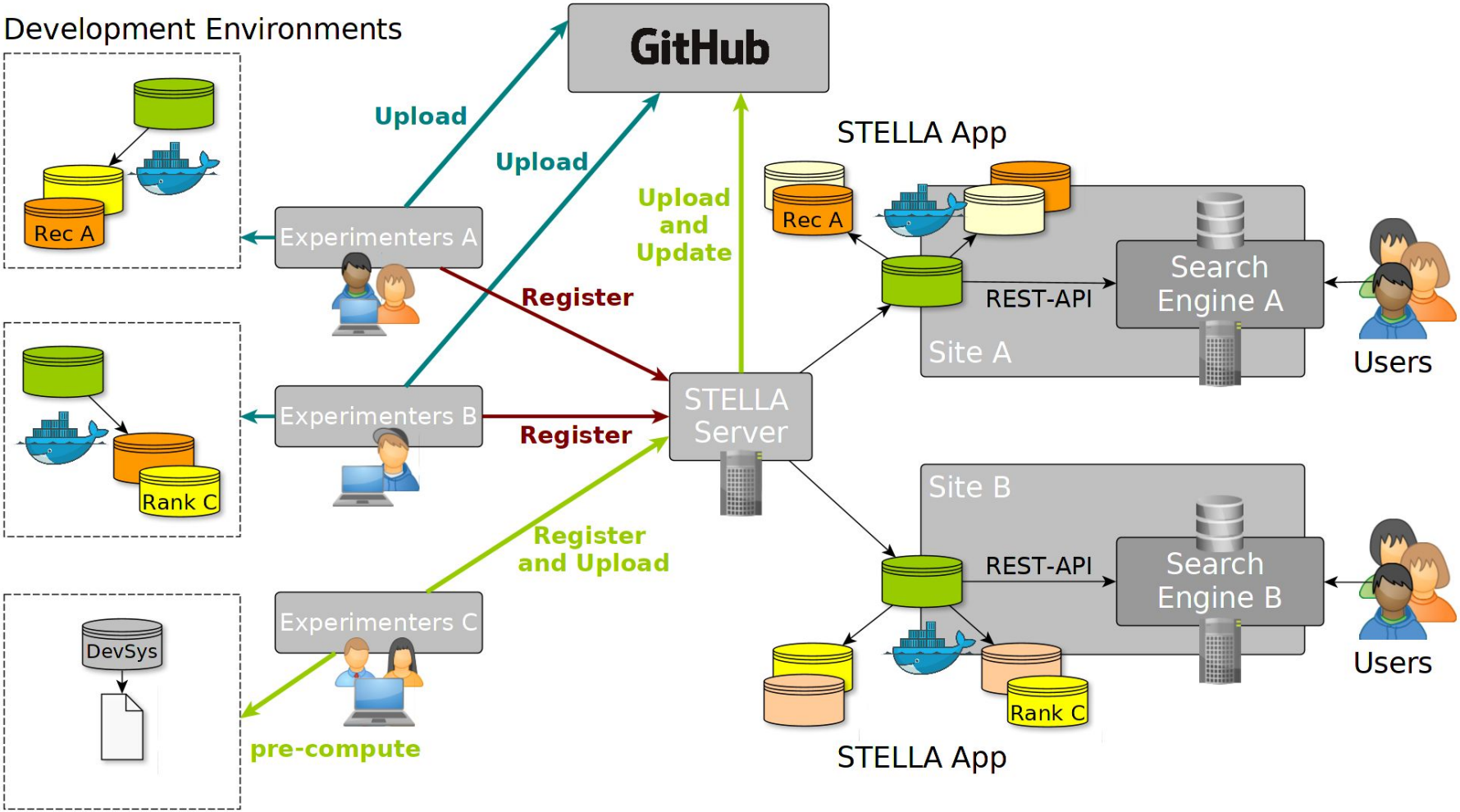
Scientific retrieval has developed and improved relatively little since it's peak (see TREC-COVID).

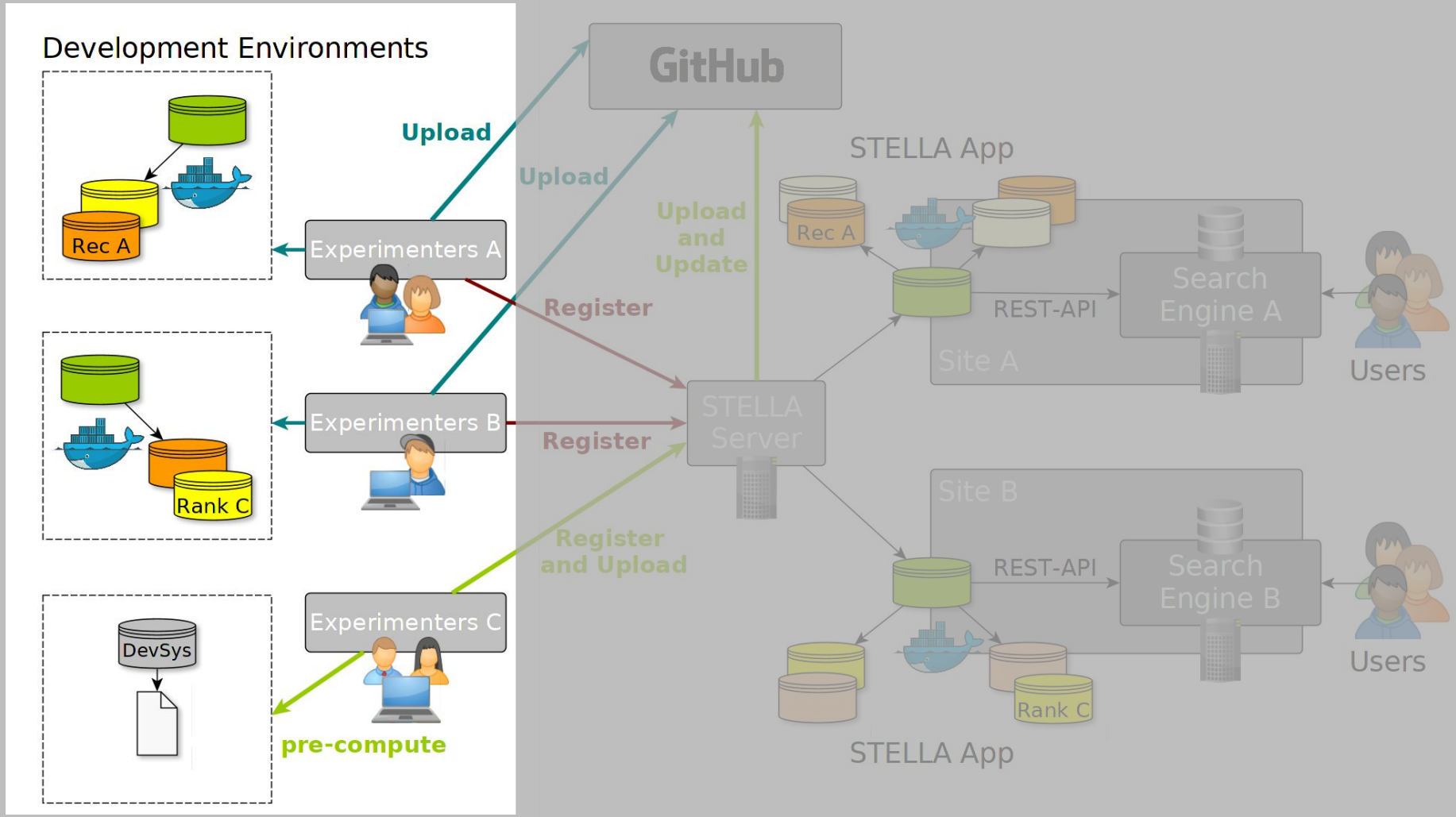
Interdisciplinary endeavours like BIR, BIRNDL or SDP show what's possible...

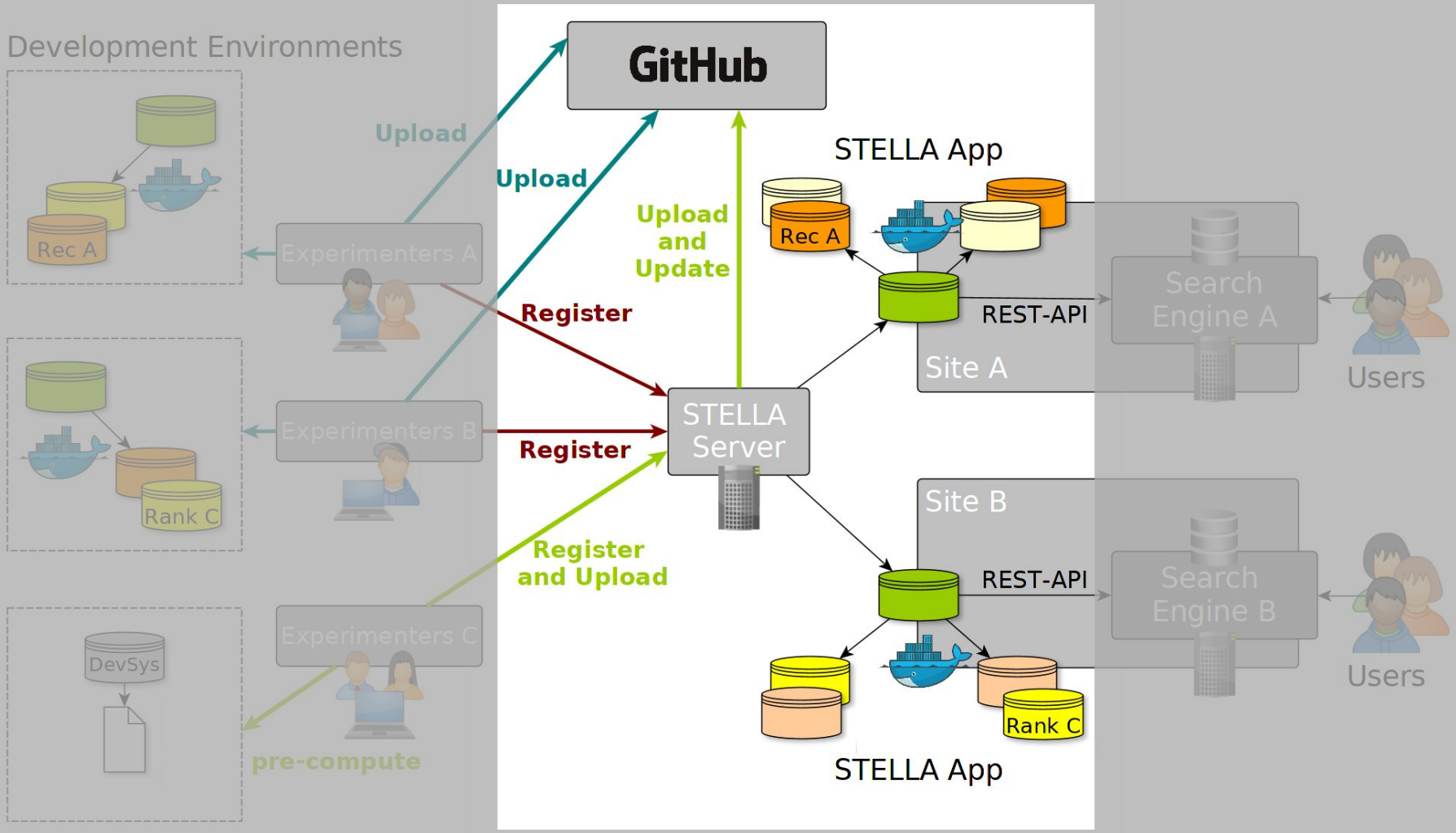
Living Lab evaluations with STELLA

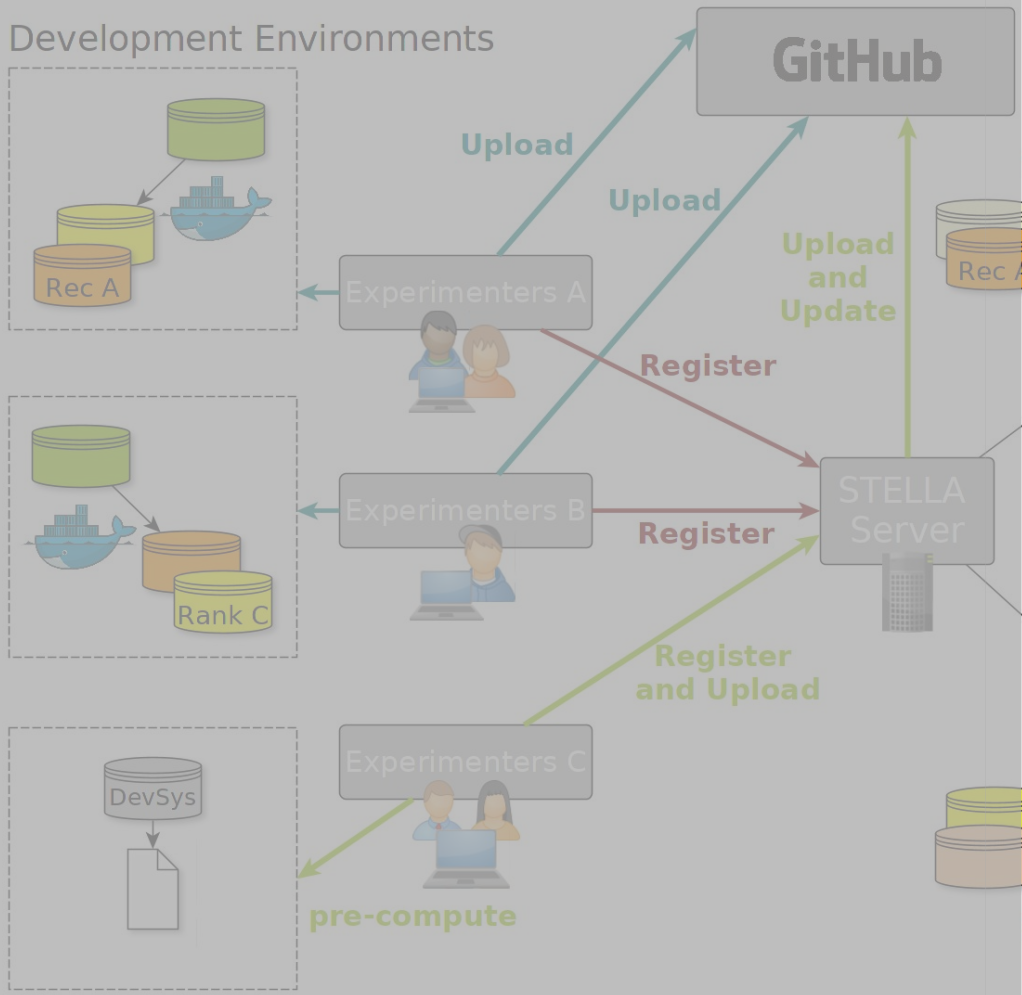
Cranfield	Living lab
System-oriented experiments with document collection, queries, relevance judgments	**User-oriented IR experiments** with log-based evaluations of experiments with **real users**
Static **explicit (topical) relevance labels**	Users are in their natural task environments deliver **implicit relevance feedback**

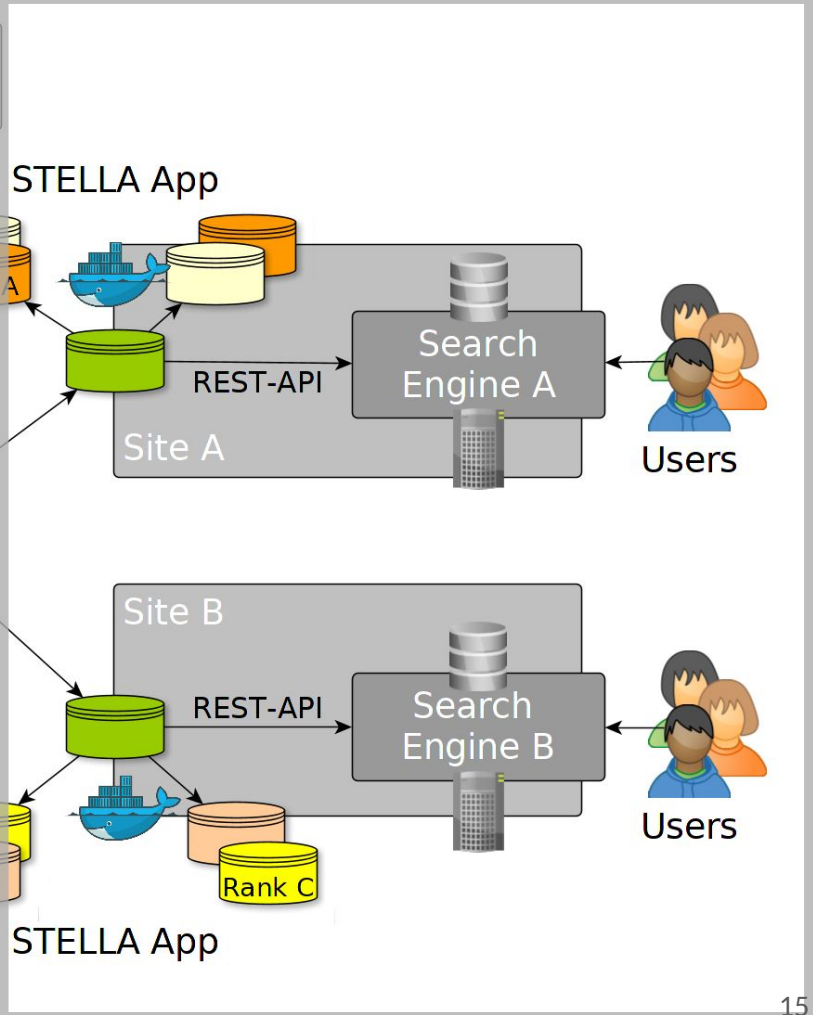
The STELLA Infrastructure


Development Environments
GitHub
Upload
Upload
Upload and Update
Register
Register
Register and Upload
pre-compute
Experimenters A
Experimenters B
Experimenters C
Rec A
Rank C
DevSys
STELLA Server
STELLA App
Rec A
REST-API
Search Engine A
Site A
Users
Site B
REST-API
Search Engine B
Rank C
Users
STELLA App

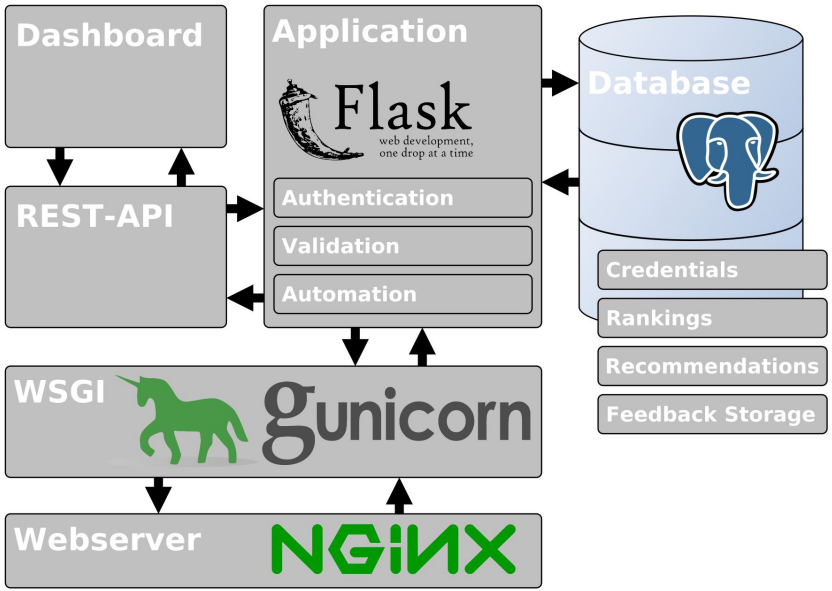

Development Environments
Rec A
Rank C
DevSys
Experimenters A
Experimenters B
Experimenters C
Upload
Upload
Register
Register
Upload and Update
Register and Upload
pre-compute
GitHub
STELLA Server
STELLA App
Rec A
REST-API
Search Engine A
Site A
Users
Site B
REST-API
Search Engine B
Rank C
Users
STELLA App


Development Environments
Rec A
Rank C
DevSys
Experimenters A
Experimenters B
Experimenters C
Upload
Upload
Upload and Update
Register
Register
Register and Upload
pre-compute
GitHub
STELLA Server
STELLA App
Rec A
REST-API
Site A
Search Engine A
Users
Site B
REST-API
Rank C
STELLA App
Search Engine B
Users


Development Environments
GitHub
Upload
Upload
Upload and Update
Register
Register
Register and Upload
pre-compute
Rec A
Rank C
DevSys
Experimenters A
Experimenters B
Experimenters C
STELLA Server

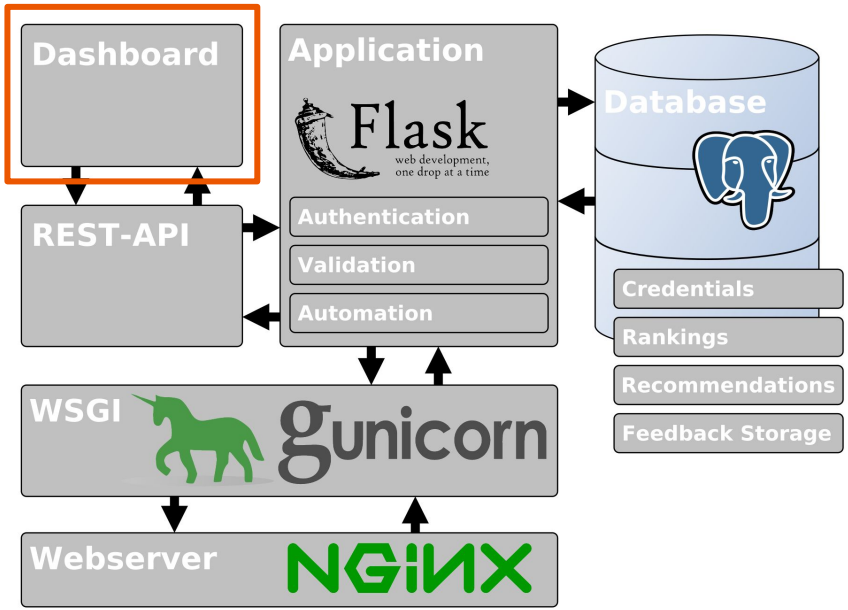

STELLA App
Rec A
REST-API
Search Engine A
Site A
Users
Site B
REST-API
Search Engine B
Users
Rank C
STELLA App

STELLA Server



STELLA Server

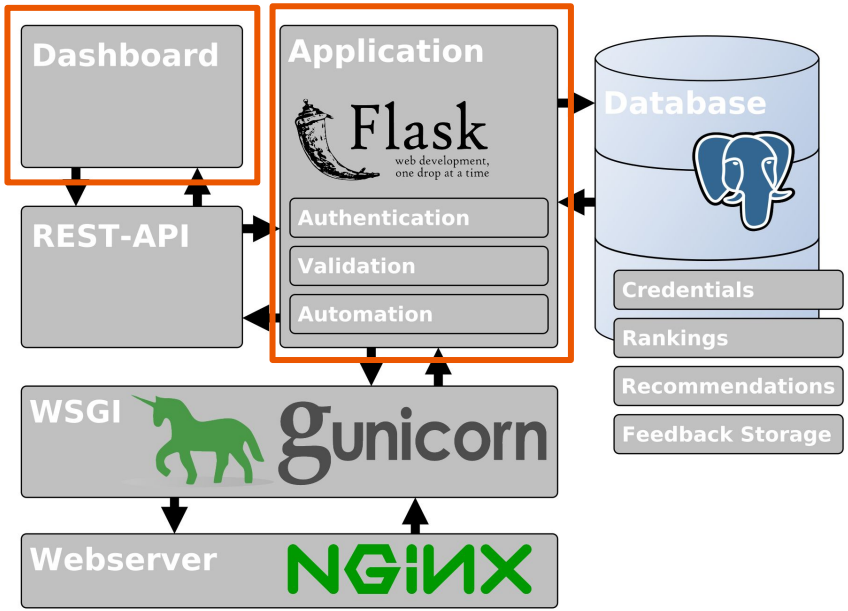
Dashboard service for participants and administrators



STELLA Server

User administration
Automatic updates of the STELLA app

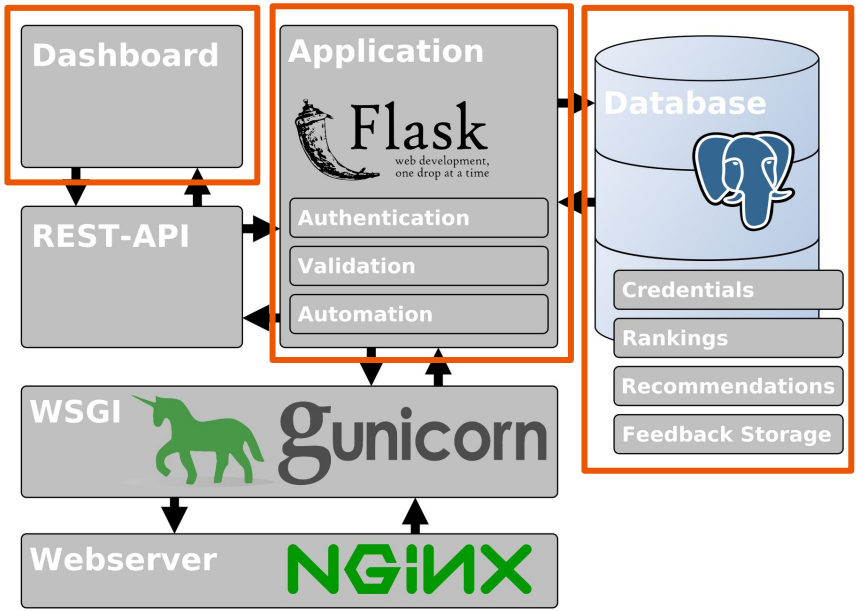
Dashboard service for participants and administrators



STELLA Server

User administration
Automatic updates of the STELLA app

Dashboard service for participants and administrators



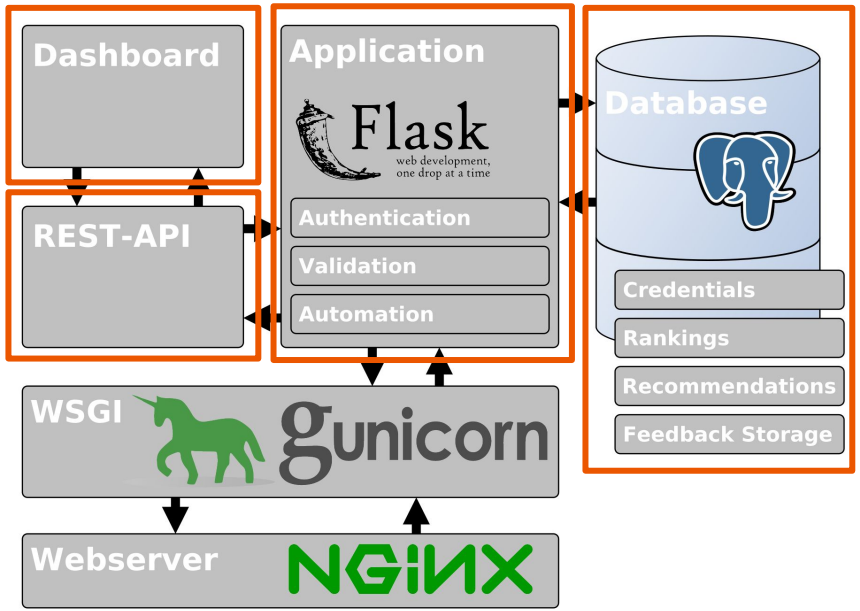
Data storage (click feedback) for analysis, training, etc.

STELLA Server

User administration
Automatic updates of the STELLA app

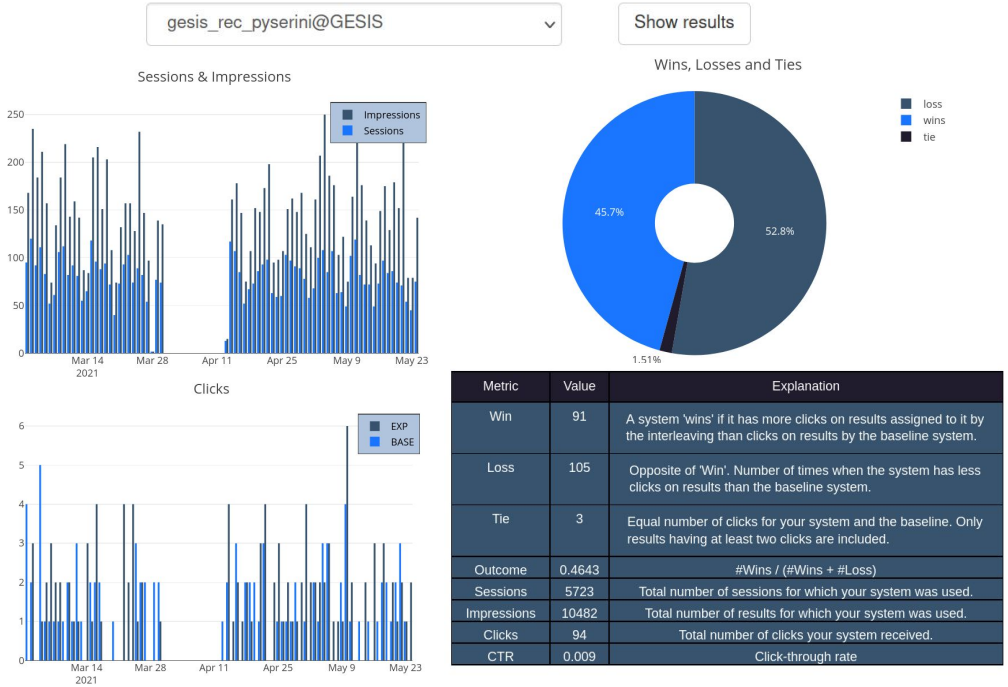
Dashboard service for participants and administrators

REST-API for the STELLA app



Data storage (click feedback) for analysis, training, etc.

Dashboard - Participants



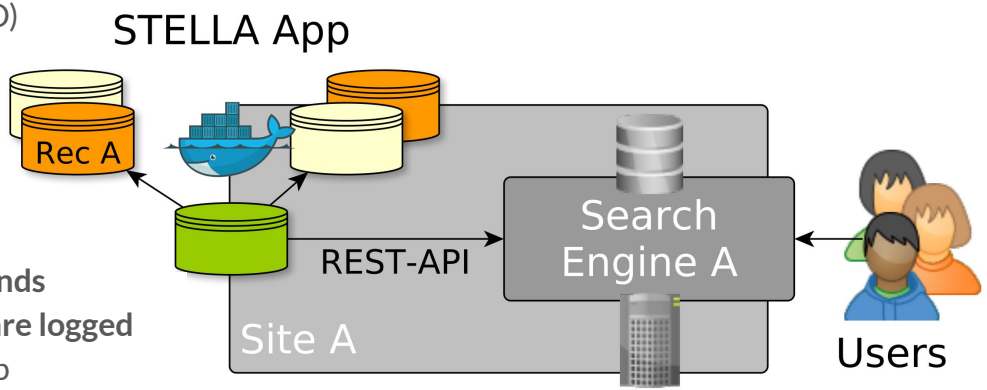
Metric	Value	Explanation
Win	91	A system 'wins' if it has more clicks on results assigned to it by the interleaving than clicks on results by the baseline system.
Loss	105	Opposite of 'Win'. Number of times when the system has less clicks on results than the baseline system.
Tie	3	Equal number of clicks for your system and the baseline. Only results having at least two clicks are included.
Outcome	0.4643	#Wins / (#Wins + #Loss)
Sessions	5723	Total number of sessions for which your system was used.
Impressions	10482	Total number of results for which your system was used.
Clicks	94	Total number of clicks your system received.
CTR	0.009	Click-through rate

Dashboard - Administration

Status	Name	Submission date	Site	Task	Type	Repository	Activate	Deactivate	Delete	Feedback Data
running	gesis rec pyserini	2019-06-10	GESIS	Recommendation	Docker Container					
running	gesis rec pyterrier	2019-06-10	GESIS	Recommendation	Docker Container					
running	livivo base	2019-06-10	LIVIVO	Ranking	Docker Container					
submitted	livivo rank pyterrier	2019-06-10	LIVIVO	Ranking	Docker Container					
running	livivo rank pyserini	2019-06-10	LIVIVO	Ranking	Docker Container					
submitted	gesis rec precom	2019-06-10	GESIS	Recommendation	Pre-computed Run					
running	tekma n	2021-04-12	GESIS	Recommendation	Pre-computed Run					
running	lemuren elastic only	2021-04-15	LIVIVO	Ranking	Docker Container					

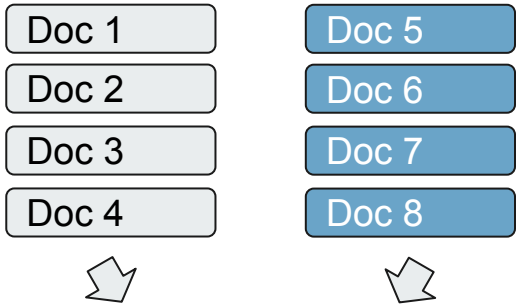
STELLA App

- **Broker** between Sites (GESIS, LIVIVO) and STELLA infrastructure
- Every site will **deploy one instance** of the STELLA app
- **Multi-container application** with all experimental systems
- Sites run STELLA apps in their **backends**
- **User interactions** (implicit, explicit) **are logged** and will be written to the STELLA app
- STELLA app temporarily **stores feedback and sends it to the STELLA server**

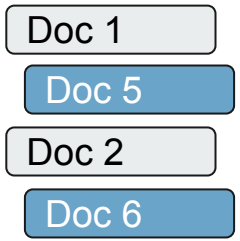


Team Draft Interleaving (TDI) [Radlinski et al., CIKM, 2008]

Ranking A and B



Interleaved ranking



Algorithm 2 Team-Draft Interleaving

Input: Rankings $A = (a_1, a_2, \dots)$ and $B = (b_1, b_2, \dots)$
Init: $I \leftarrow (); TeamA \leftarrow \emptyset; TeamB \leftarrow \emptyset;$
while $(\exists i : A[i] \notin I) \wedge (\exists j : B[j] \notin I)$ **do**
  if $(|TeamA| < |TeamB|) \vee$
    $((|TeamA| = |TeamB|) \wedge (RandBit() = 1))$ **then**
    $k \leftarrow \min_i \{i : A[i] \notin I\}$ ..... *top result in A not yet in I*
    $I \leftarrow I + A[k];$ ..... *append it to I*
    $TeamA \leftarrow TeamA \cup \{A[k]\}$ ..... *clicks credited to A*
  else
    $k \leftarrow \min_i \{i : B[i] \notin I\}$ ..... *top result in B not yet in I*
    $I \leftarrow I + B[k]$ ..... *append it to I*
    $TeamB \leftarrow TeamB \cup \{B[k]\}$ ..... *clicks credited to B*
  end if
end while
Output: Interleaved ranking I, $TeamA$, $TeamB$

[Radlinski et al., CIKM, 2008]

REST-API

GET /stella/api/v1/ranking?query=<string:query>

Example
GET /stella/api/v1/ranking?query=vaccine&page=0&rpp=10

Output


```
{'body': {'1': {'docid': 'M27622217', 'type': 'BASE'},
          '2': {'docid': 'M27251231', 'type': 'EXP'},
          '3': {'docid': 'M27692969', 'type': 'BASE'},
          '4': {'docid': 'M26350569', 'type': 'EXP'},
          '5': {'docid': 'M26715777', 'type': 'EXP'},
          '6': {'docid': 'M26650940', 'type': 'BASE'},
          '7': {'docid': 'M27098271', 'type': 'EXP'},
          '8': {'docid': 'M28381438', 'type': 'BASE'},
          '9': {'docid': 'M27763523', 'type': 'EXP'},
          '10': {'docid': 'M27157745', 'type': 'BASE'},
 'header': {'container': {'base': 'rank_elastic_base', 'exp': 'rank_elastic'},
            'page': 0,
            'q': 'vaccine',
            'rid': 3,
            'rpp': 20,
            'sid': 1}}
```


POST /stella/api/v1/ranking/<int:rid>/feedback

Example
POST /stella/api/v1/ranking/3/feedback

Output


```
{'clicks': {'1': {'clicked': False,
                  'date': None,
                  'docid': 'M26923455',
                  'type': 'EXP'},
            '2': {'clicked': False,
                  'date': None,
                  'docid': 'M25600519',
                  'type': 'EXP'},
            '3': {'clicked': True,
                  'date': '2020-07-29 16:06:51',
                  'docid': 'M27515393',
                  'type': 'EXP'}},
 'end': '2020-07-29 16:12:53',
 'interleave': True,
 'start': '2020-07-29 16:06:51'}
```

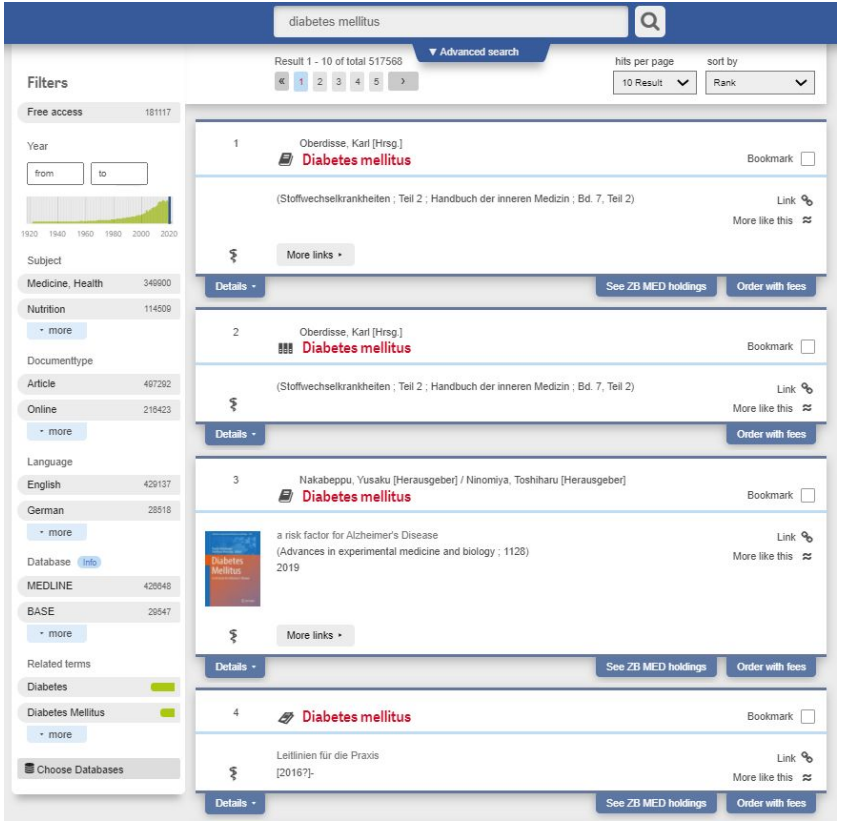
STELLA in the Wild
LiLAS @ CLEF

LiLAS - Living Labs for Academic Search

- **CLEF Lab in 2020 and 2021**
  - https://clef-lilas.github.io/
- **Task 1: Ad-hoc Search Ranking**
  - Given a query, find the most relevant documents
- **Task 2: Research Data Recommendation**
  - Given a *seed* publication, recommend research datasets

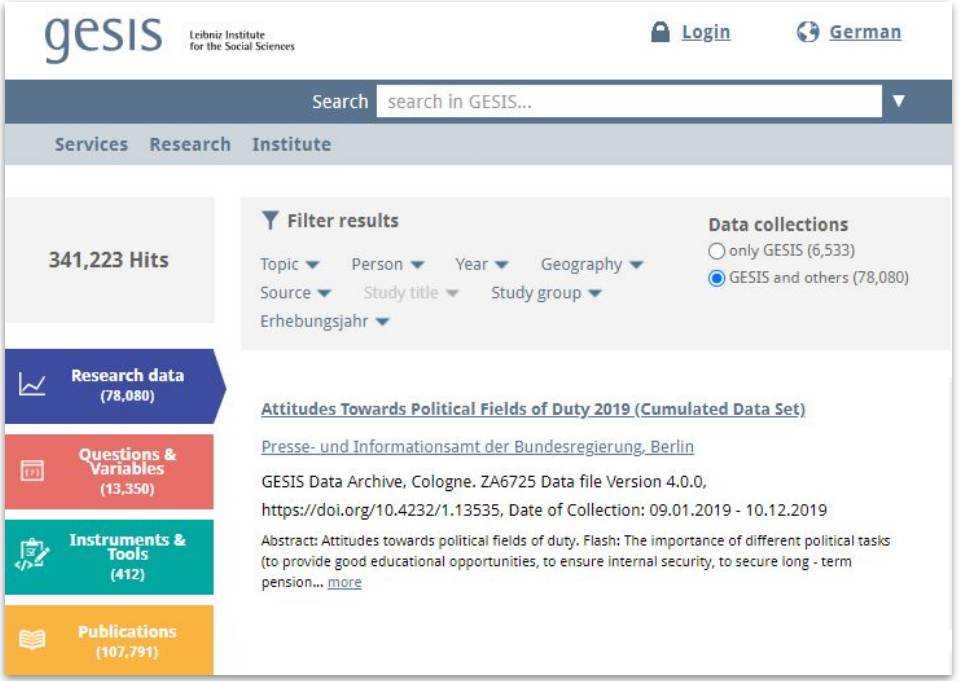
Task 1: Biomedical Ad-hoc Search in LIVIVO

- 68 million publication metadata sets
- 31 different source databases
- 200,000 searches per month
- English, German, Spanish, French
- Data set (for training and precompute task): 1000 head queries, 100 candidate documents per query, and 35 million metadata sets



Task 2: Data Set Recommendation with GESIS Search

- 84,000 research data sets
- 114,000 publications
- metadata and full texts available (different languages)



Login German Contact FAQ

Search search in GESIS...

Services Research Institute

< Back

Contextualizing educational differences in vaccination uptake: A thirty nation survey

Makarovs, Kirils; Achterberg, Peter

Social Science & Medicine, 2017

Database: GESIS Bibliography

Actions
Cite
search in Google Scholar

publication

Recommended Research Data

the following research data is related to the publication

Flash Eurobarometer 287 (Influenza H1N1)

Papacostas, Antonis

GESIS Data Archive, Cologne. ZA5222 Data file Version 1.0.0, https://doi.org/10.4232/1.10224, Date of Collection: 26.11.2009 - 30.11.2009

Abstract: Knowledge about influenza H1N1 (swine flu). Topics: intention to get vaccinated against seasonal influenza; awareness of pandemic H1N1 flu (swine flu); concern of swine flu developing into a serious... more

Materials
Datasets
Questionnaires
Other documents

Actions
Cite

Research data

LiLAS - Confusion matrix

	Task 1	Task 2
Type A	Precomputed rankings ZB MED Wissen für Mensch & Umwelt Doc	Precomputed recommendations gesis Leibniz-Institut für Sozialwissenschaften Doc
Type B	Dockerized ranking system ZB MED Wissen für Mensch & Umwelt docker	Dockerized recommender system gesis Leibniz-Institut für Sozialwissenschaften docker

Evaluation Setup & Metrics

Wins, Losses, and Ties [Schuth et al.; CLEF, 2015]

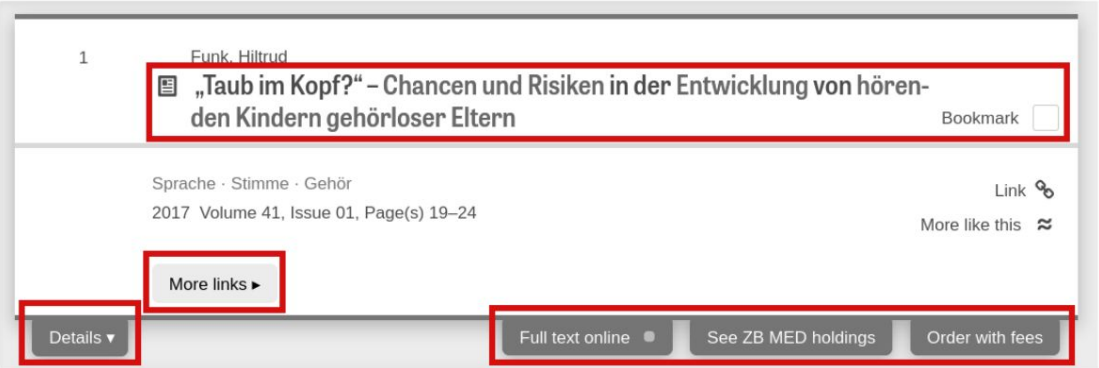
Wins	A system has more clicks on results assigned to it by the interleaving than clicks on results by the baseline system.
Loss	Opposite of 'Win'. Number of times when the system has less clicks on results than the baseline system.
Tie	Equal number of clicks for your system and the baseline. Only results having at least two clicks are included.
Outcome	#Wins / (#Wins + #Loss)

(Normalized) Reward [Gingstad et al., CIKM, 2020]

$$Reward = \sum_{s \epsilon S} w_s c_s$$

$$nReward = \frac{Reward_{\text{exp}}}{Reward_{\text{exp}} + Reward_{\text{base}}}$$

Logged SERP elements at LIVIVO



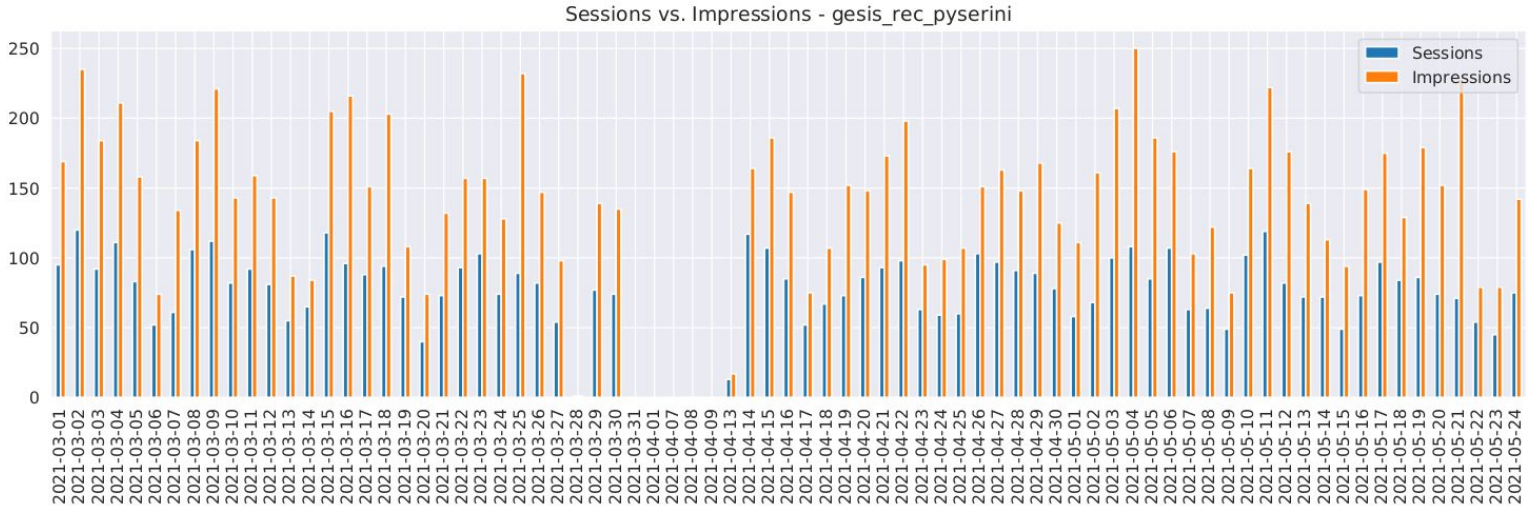
SERP Element	w_s
Bookmark	10
Order	10
Fulltext	8
In Stock	8
More Links	2
Title	1
Details	1

Lab Overview and Experimental Results

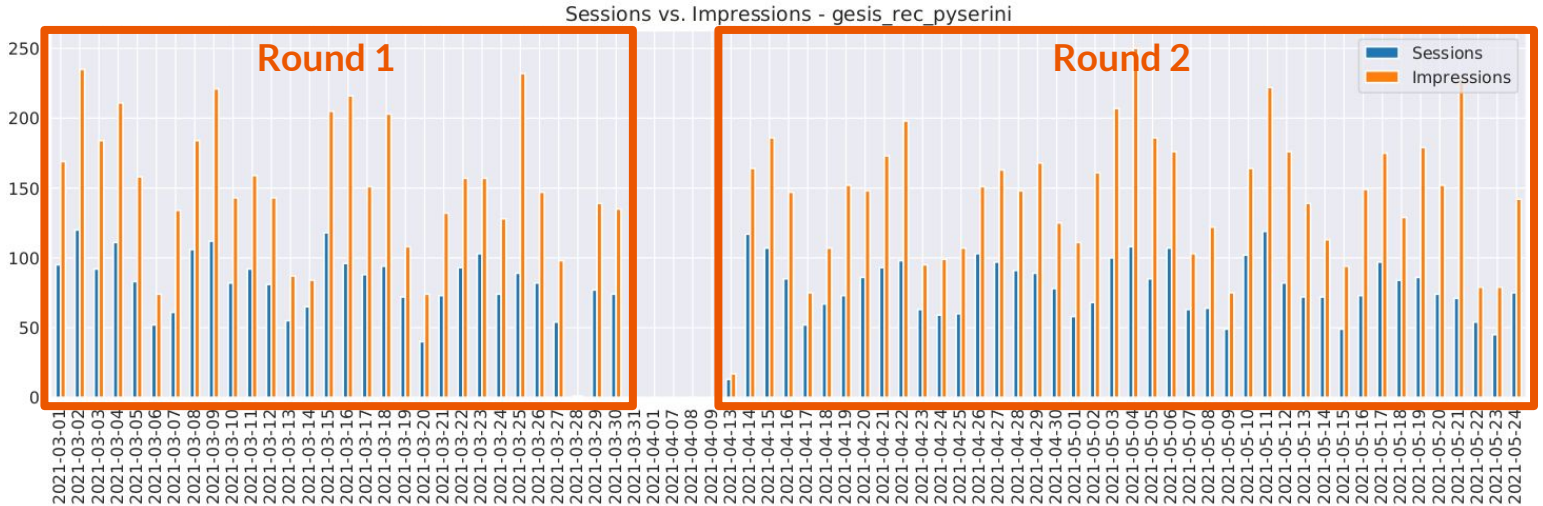
System overview for LiLAS 2021

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	●	○	●
gesis_rec_precom	2	A	●	●	○
gesis_rec_pyterrier	2	B	●	●	●
gesis_rec_pyserini	2	B	○	●	●

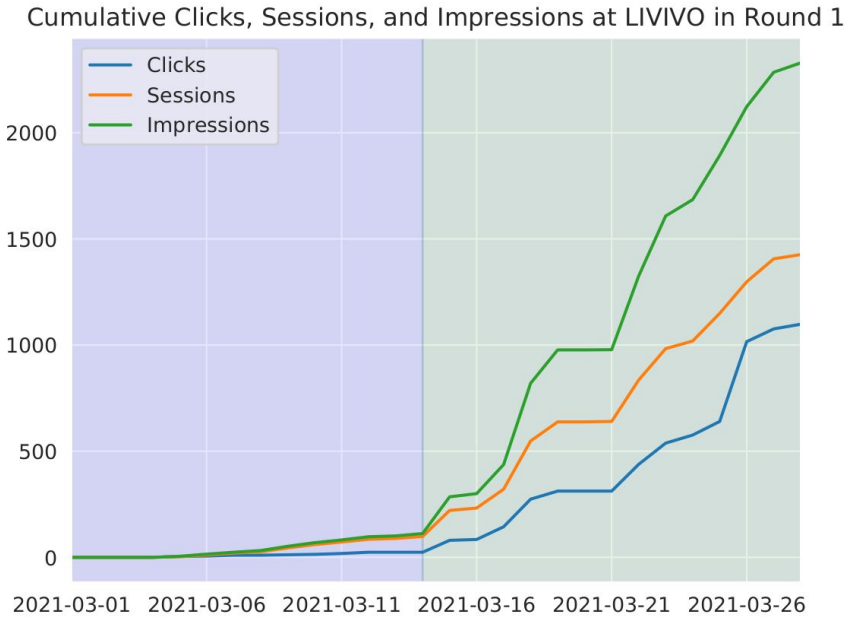
Sessions & Impressions distributions



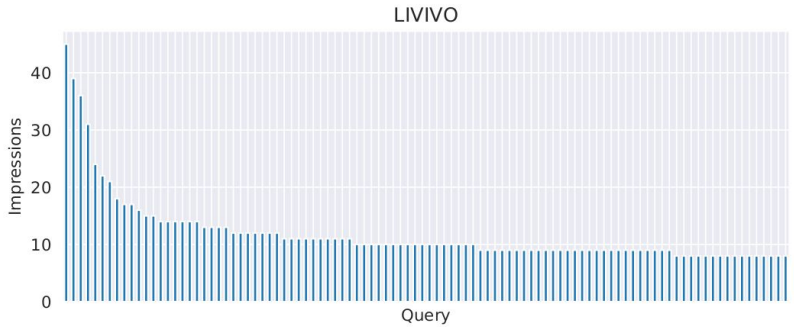
Sessions & Impressions distributions



Precomputed Submissions vs. Dockerized Systems

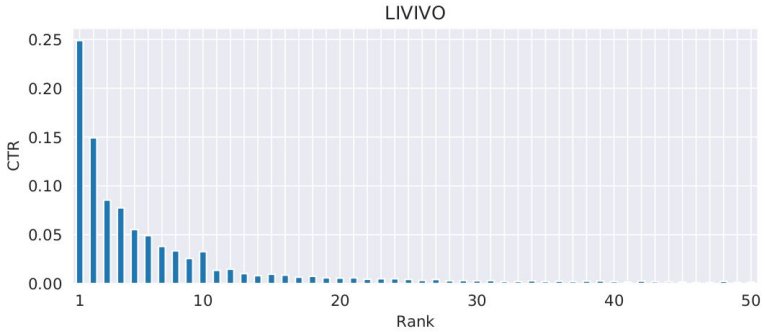


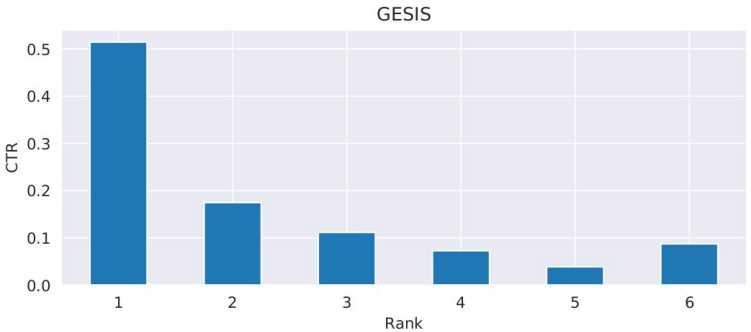
Top 10 head queries at LIVIVO



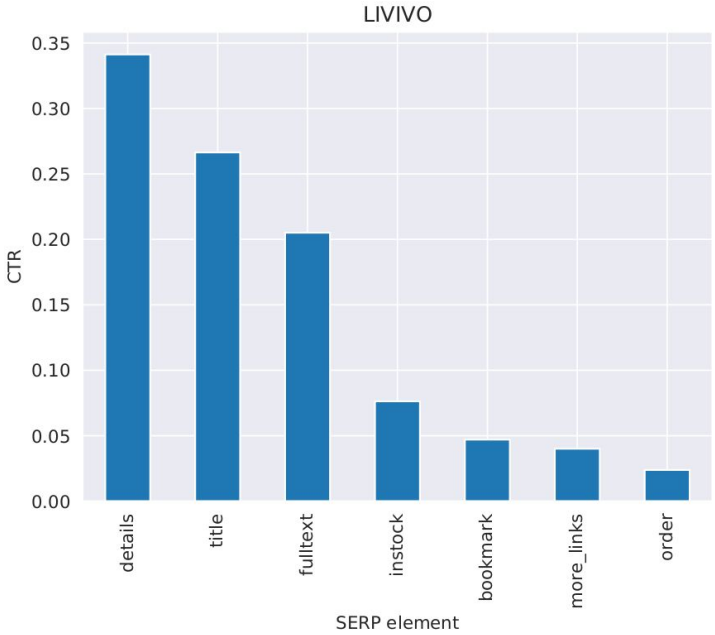
Rank	Query string	Impressions
1	covid19	45
2	demenz	39
3	guillian barre syndrome	36
4	polyvinyl and nasal and packing	31
5	covid	24
6	pflege	22
7	cancer	21
8	parkinson	18
9	depression	17
10	schlaganfall	17

Click-through Rate (CTR) vs. Rank





Click distribution on SERP elements at LIVIVO



Round 1 - Outcome

System	Win	Loss	Tie	Outcome	Sessions	Impressions	Clicks	CTR
gesis_rec_pyserini†	36	36	1	0.50	2284	4195	37	0.0088
gesis_rec_pyterrier	26	28	1	0.48	1968	3675	28	0.0076
gesis_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

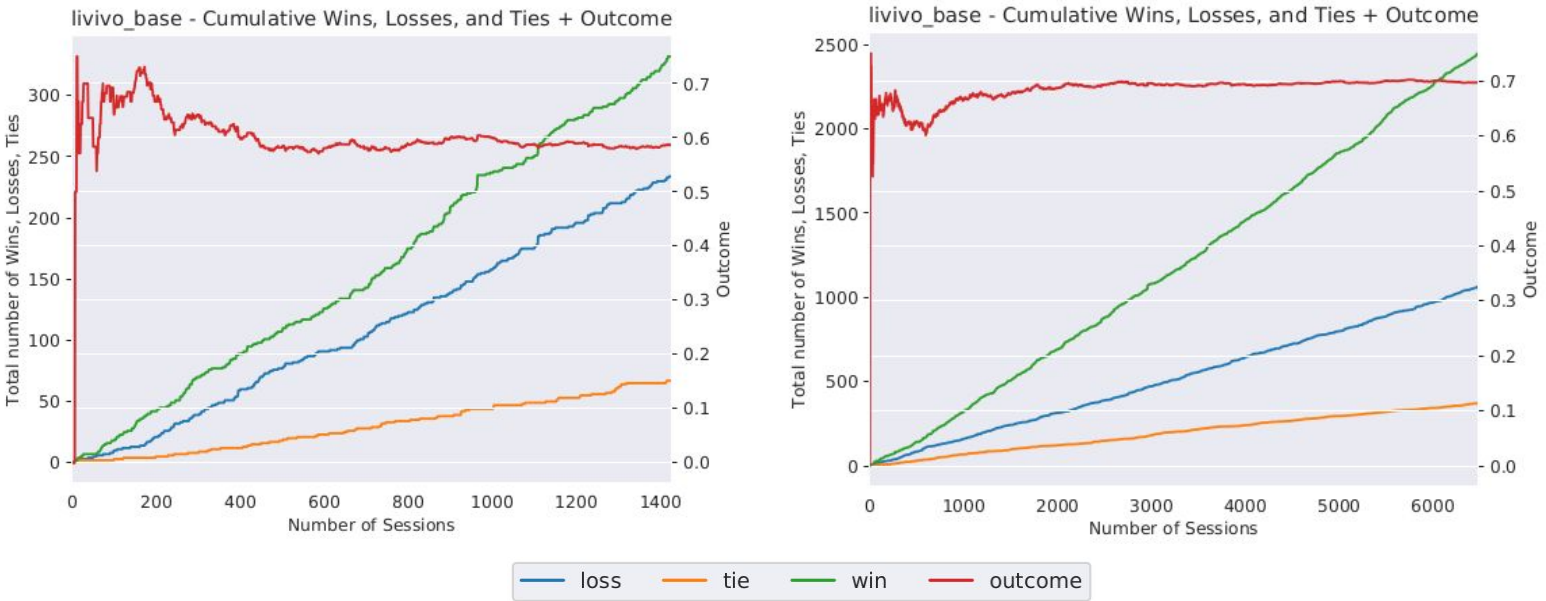
Round 2 - Outcome

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

Round 2 - Outcome vs. Weighted Clicks

	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

Wins, Losses, Ties vs. Number of sessions



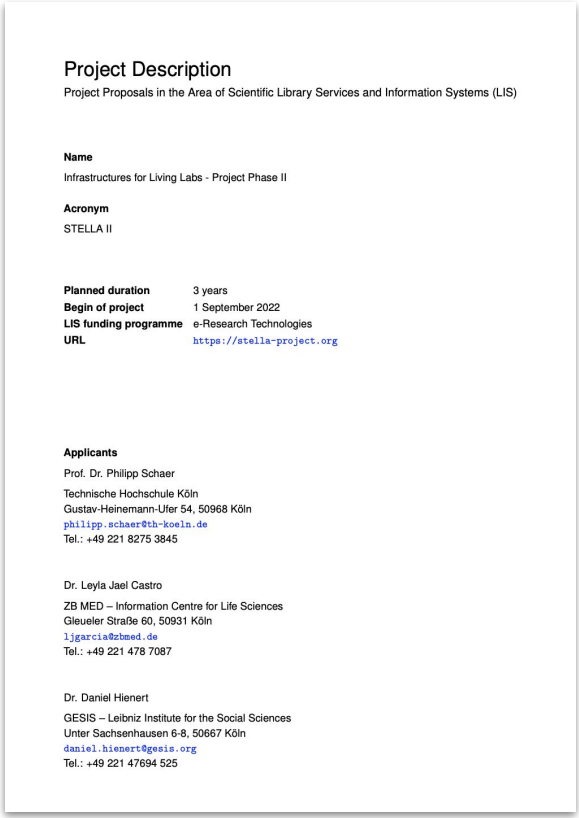
Lessons learned

- **Docker pays off**:
  - technically reproducible systems…
  - … but **additional workload** for participants (and organizers).
  - more feedback data as compared to precomputed results
- **Statistically significant results** with less online time
- **Power-law like distributions** of clicks and queries
  - De-biasing is required for some evaluations
- More in-depth comparison of systems with **weighted clicks and rewards**
- **In the future:** provide participants with **transparent baseline systems**

Outlook to STELLA II

DFG - STELLA II (submitted)

- Improve **reproducibility**: technically and experimentally
- Technological advances: Shared **index** (cf. CIFF), better **logging** functionality (cf. LogUI)
- **Continuous evaluation:** Long-term test collection build-up and simulation studies with STELLA interaction data



Project Description

Project Proposals in the Area of Scientific Library Services and Information Systems (LIS)

Name

Infrastructures for Living Labs - Project Phase II

Acronym

STELLA II

Planned duration	3 years
Begin of project	1 September 2022
LIS funding programme	e-Research Technologies
URL	https://stella-project.org

Applicants

Prof. Dr. Philipp Schaer

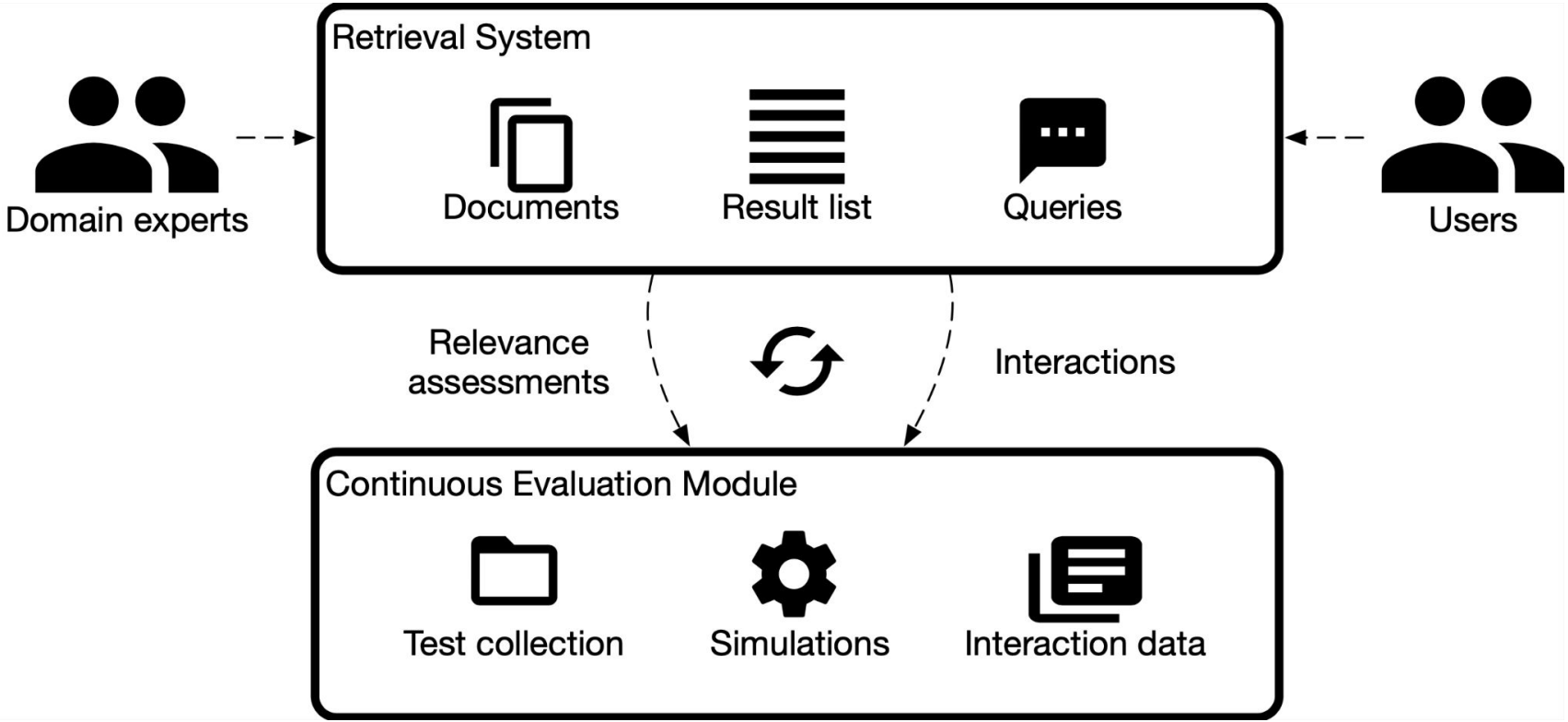
Technische Hochschule Köln
Gustav-Heinemann-Ufer 54, 50968 Köln
philipp.schaer@th-koeln.de
Tel.: +49 221 8275 3845

Dr. Leyla Jael Castro

ZB MED – Information Centre for Life Sciences
Gleueler Straße 60, 50931 Köln
ljgarcia@zbmed.de
Tel.: +49 221 478 7087

Dr. Daniel Hienert

GESIS – Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8, 50667 Köln
daniel.hienert@gesis.org
Tel.: +49 221 47694 525


Retrieval System
Domain experts
Documents
Result list
Queries
Users
Relevance assessments
Interactions
Continuous Evaluation Module
Test collection
Simulations
Interaction data

Thank you for your attention!

https://stella-project.org/

https://github.com/stella-project