Jan Heinrich Reimer Sebastian Schmidt Maik Fröbe Lukas Gienapp Harrisen Scells Benno Stein Matthias Hagen Martin Potthast

The Archive Query Log: Mining Millions of Search Result Pages of Hundreds of Search Engines from 25 Years of Web Archives

The Archive Query Log (AQL)

- Large log of queries and archived search result pages (SERPs)
- Mined from the Internet Archive's Wayback Machine
- 356 million queries, 137 million SERPs, 1 billion results
- 550 search providers across 25 years

Private and public query logs

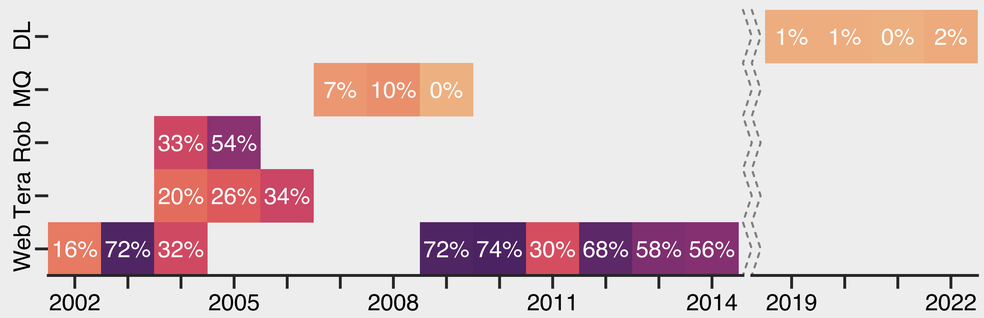
- Review of 492 publications on query logs
- 41 private and 14 public query logs
- Public logs are smaller, more focussed, and less diverse
- Most logs contain session and click data (the AQL does not)

The AQL-22 at a glance

Search provider	URLs	Queries	unique	SERPs	Results
Google	89.4 M	72.7 M	20.0 M	28.0 M	223.1 M
YouTube	41.8 M	41.4 M	11.3 M	15.9 M	339.2 M
Baidu	78.5 M	69.6 M	2.9 M	26.8 M	107.6 M
QQ	0.5 M	0.5 M	0.1 M	0.2 M	2.1 M
Facebook	3.1 M	0.2 M	0.0 M	0.1 M	0.7 M
Yahoo!	8.8 M	2.8 M	1.2 M	1.1 M	9.2 M
Amazon	66.8 M	0.8 M	0.3 M	0.3 M	7.8 M
Wikipedia	68.5 M	1.7 M	0.6 M	0.7 M	7.0 M
JD.com	4.4 M	3.9 M	0.4 M	1.5 M	16.0 M
360	1.5 M	1.1 M	0.1 M	0.4 M	3.5 M
⋮ 540 others	646.8 M	161.8 M	27.8 M	62.4 M	693.9 M
∑ 550	**1010.2 M**	**356.5 M**	**64.5 M**	**137.3 M**	**1410.0 M**

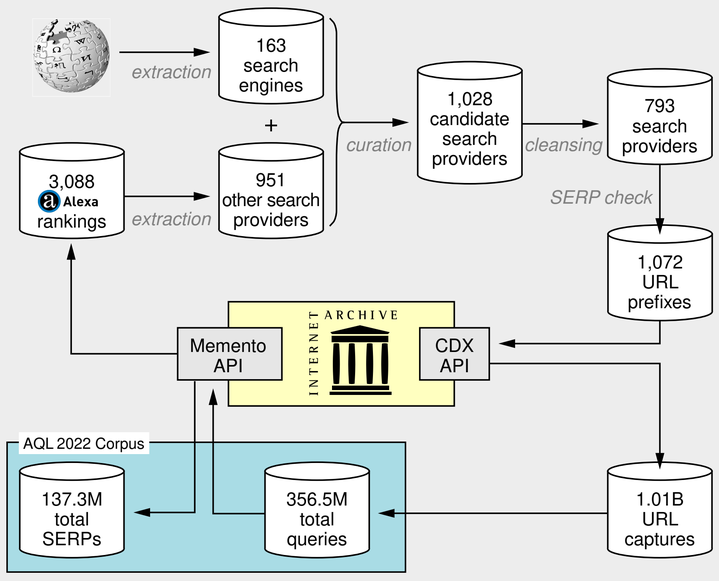
Use cases

- Transparent insights into search industry
- Comparisons of search engines over time
- Training data for (neural) retrieval models



Query overlap with TREC topics (exact match).

Mining the Archive Query Log



1. List popular search providers

- 163 search engines
  (from Wikipedia's "List of search engines")
- 951 popular websites with a search bar
  (fused Alexa rankings from 2010–2022)

2. Collect archived URLs

- Collect provider domains
  (e.g., google.com; manual and from public lists)
- Identify URL prefixes of SERPs
  (e.g., `/search?q=`; manually annotated)
- Fetch 1.1B captures from Internet Archive
  (via CDX API, filter domains and URL prefixes)

3. Parse queries from URLs

`https://google.com/search?q=covid+19+usa+map&start=10&ei=...`
URL prefix | query | offset

`https://chefkoch.de/rs/s0/backen%20dinkelmehl/Rezepte.html`
URL prefix | page | query

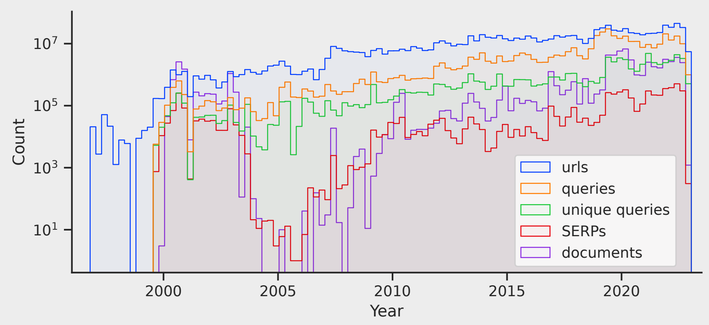
- Gather parser parameters manually
- Parse query, page, offset

4. Parse SERP HTML

- Sample SERPs, annotate expected results
- Apply existing parsers
- Compare parsed result with annotations
- Adapt / extend parsers

→ 70 SERP parsers, 444 approval tests

Analysis



Time coverage of data types in the AQL-22, per quarter.

Query characteristics

- 104 different languages
- Top languages: Chinese, English
- Most queries: 5–20 characters
  (but also longer queries, e.g., pasted text)
- 1.3 % contain obscene terms
- 81 % duplicated
  (different time, SERP page offset, or user)

SERP characteristics

- Top languages: English, Russian
  (Chinese not among the top languages → bias)
- Popular websites often among top results

Top	W				IMDb	other	self
5	2.9 %	0.8 %	0.6 %	0.4 %	0.3 %	25.1 %	69.6 %
10	2.2 %	0.7 %	0.5 %	0.3 %	0.3 %	25.4 %	70.4 %

Most frequent domains in top-5 or top-10 search results.

Conclusions and access

- Largest, most diverse query log ever made publicly available
- Enables researchers to tackle new and existing challenges
  (e.g., new retrieval models, query suggestion / prediction, diachronic analyses)
- Privacy-sensitive dataset → sandboxed public access via TIRA.io

Resources

`github.com/webis-de/archive-query-log`

`tira.io/task/archive-query-log`

`doi.org/10.1145/3539618.3591890`

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU, `doi.org/10.3030/101070014`)