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

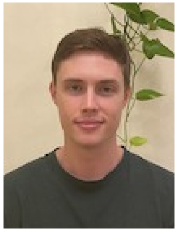
June 1, 2023

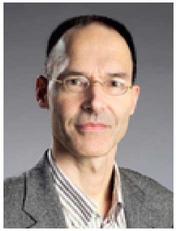
J. Heinrich Reimer[1]

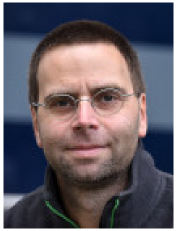
Sebastian Schmidt[2]

Maik Fröbe[1]

Lukas Gienapp[2,3]


Harrisen Scells[2]


Benno Stein[4]


Matthias Hagen[1]

Martin Potthast[2,3]

[1]Friedrich-Schiller-Universität Jena [2]Leipzig University [3]ScaDS.AI [4]Bauhaus-Universität Weimar

`https://webis.de`

46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)

Archive Query Log

The AQL-22 Corpus

INTERNET ARCHIVE →

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
Σ	**550**	**1010.2 M**	**356.5 M**	**64.5 M**	**137.3 M**	**1410.0 M**

- ❑ first large log of SERPs
- ❑ mined from the Internet Archive's Wayback Machine
- ❑ from 550 search engines across 25 years

Archive Query Log

Query Logs

Valuable resources for search engine development and research:

queries	search requests by users
results	retrieved results in a ranked order
user data	identifiers, sessions, clicks

Used for:

- query suggestions and reformulations [e.g., Cui et al. WWW’02]
- analyses of user behavior and experience [e.g., Jansen, Spink. IC’03]
- feedback on retrieval models [e.g., Joachims et al. SIGIR’05]

Archive Query Log

Query Logs

Valuable resources for search engine development and research:

❑ queries	search requests by users
❑ results	retrieved results in a ranked order
❑ user data	identifiers, sessions, clicks

Used for:

- query suggestions and reformulations [e.g., Cui et al. WWW’02]
- analyses of user behavior and experience [e.g., Jansen, Spink. IC’03]
- feedback on retrieval models [e.g., Joachims et al. SIGIR’05]

Archive Query Log

Existing Query Logs

Large query logs are often not publicly available.

Source	Queries	Unique	Results	Task	Lang.	Span	Last Query
AltaVista	575.2 M	153.6 M	–	Web	en	1m	1998
Infoseek	19.9 M	–	–	Web	zh	1m	1998
Microsoft AdCenter	27.9 M	27.9 M	–	Ads	en	2m	2007
Baidu	363.0 M	10.4 M	13.1 M	Web	zh	–	2012
Startpagina	10.0 M	–	–	Web	nl	1m	2014
parsijoo.ir	27.0 M	–	–	Web	fa	2y	2017
CiteSeerX	78.1 M	14.8 M	–	Edu	en	4y	2021
Archive Query Log	356.5 M	64.5 M	1410.0 M	Multi	Multi	25y	2022

(from our focused literature review of 492 publications using query logs)

Incentives not to publish logs:

- user privacy concerns
- commercial value of data on user search behavior
- confidentiality of ranking models
- → Archive Query Log: on par with private logs

Archive Query Log

Existing Query Logs

Public query logs are often limited in size, scope, or diversity.

Source	**Queries**	Unique	**Results**	**Task**	**Lang.**	**Span**	**Last Query**
AOL	36.4 M	10.2 M	19.4 M	Web	en	3m	2006
MSN	14.9 M	–	–	Web	en	1m	2006
Sogou	18.4 M	4.6 M	14.1 M	Web	zh	1m	2009
Yandex	10.1 M	–	49.0 M	Web	ru	–	2011
Bing Images	11.7 M	–	–	Img.	en	–	2013
ORCAS	18.8 M	–	18.8 M	Web	en	–	2020
AOLIA	11.3 M	–	1.5 M	Web	en	–	2022
Archive Query Log	356.5 M	64.5 M	1410.0 M	Multi	Multi	25y	2022

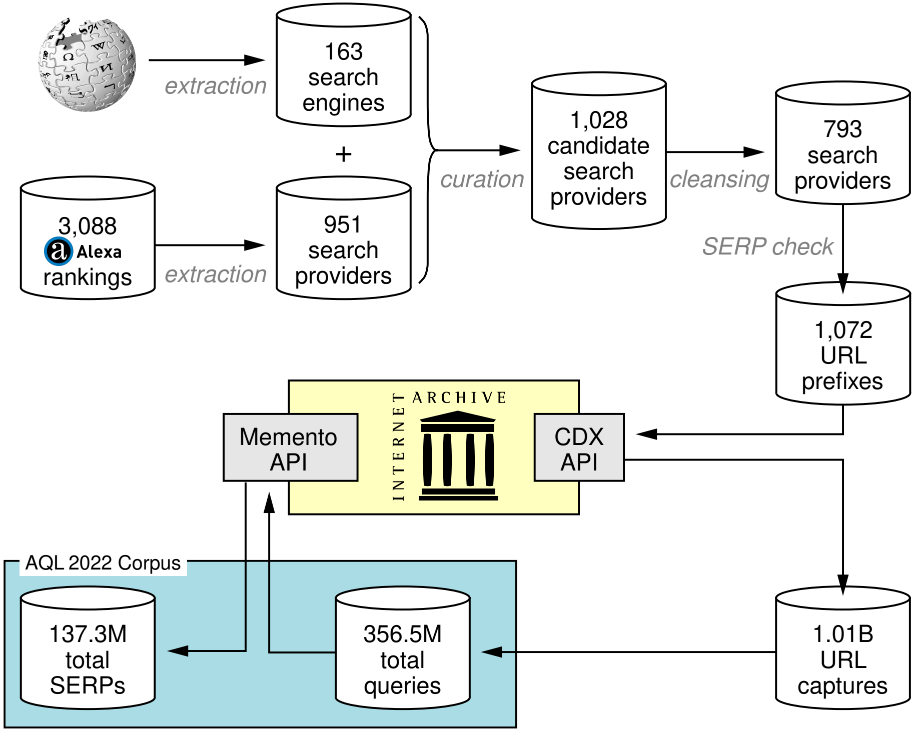
(from our focused literature review of 492 publications using query logs)

Limitations compared to private logs:

- ❑ much smaller
- ❑ single task and language
- ❑ smaller crawling time span
- → Archive Query Log: first to overcome these limitations

Archive Query Log

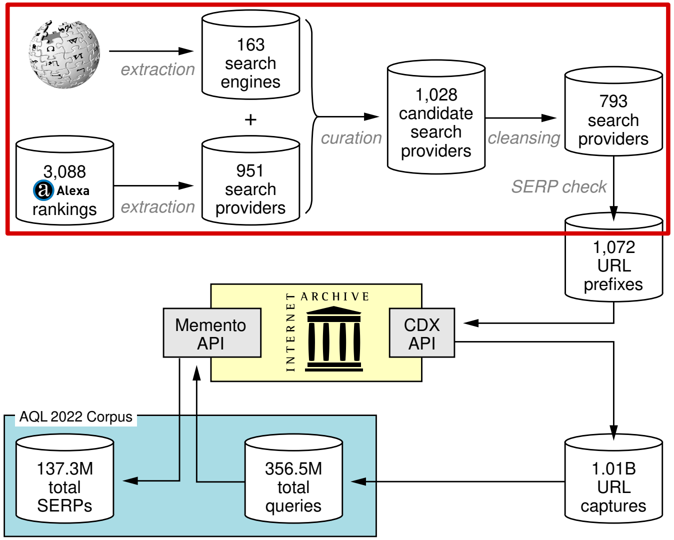
Mining the AQL



1. list popular search providers
2. collect archived URLs
3. parse queries from URLs
4. parse SERP HTML

Archive Query Log

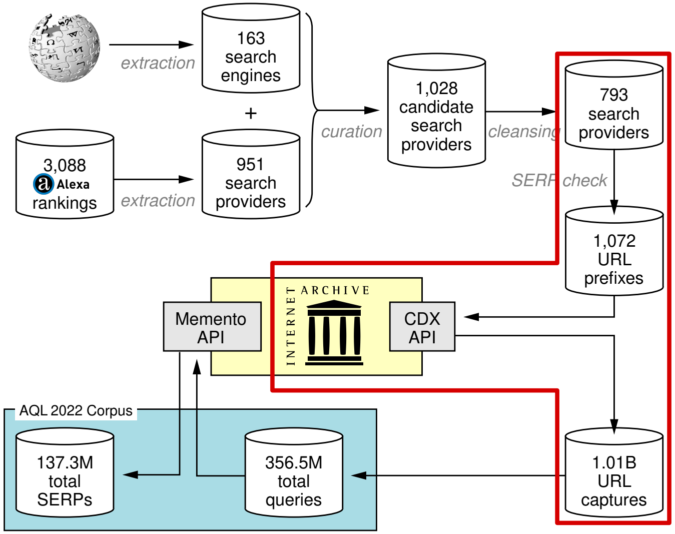
1. Search Provider Collection



- ❑ 163 search engines
  - – from Wikipedia's "List of search engines" [en.wikipedia.org/wiki/List_of_search_engines]
- ❑ 951 popular websites with a search bar
  - – rank-fuse 3088 Alexa rankings from 2010 to 2022 [Cormack et al., SIGIR'09]
  - – filter for websites with a search bar
- ❑ manually remove duplicates and spam

Archive Query Log

2. Provider Domains and URLs

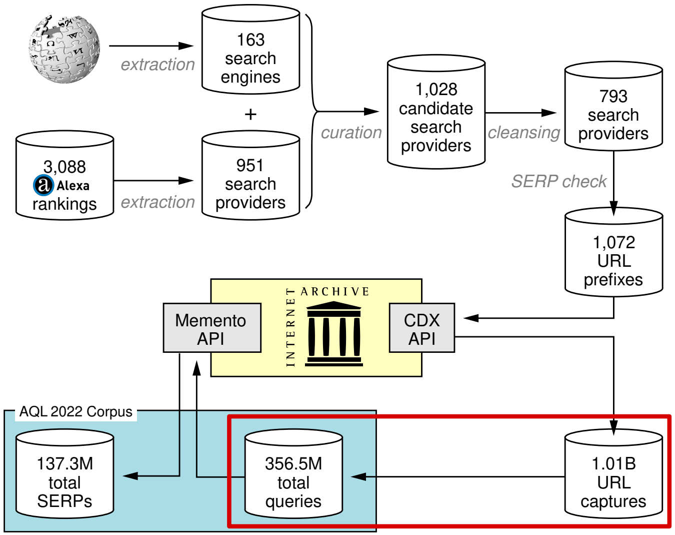


- ❑ find other (sub)domains of search providers
- ❑ manually check for search functionality (query in URL)
- ❑ determine common prefixes
- ❑ fetch available captures from the Internet Archive's CDX API

  [github.com/internetarchive/wayback/tree/master/wayback-cdx-server]

Archive Query Log

3. Query Extraction



- queries encoded in URL parts [rfc-editor.org/rfc/rfc2396]
- find parser parameters manually
- parse query, page, offset

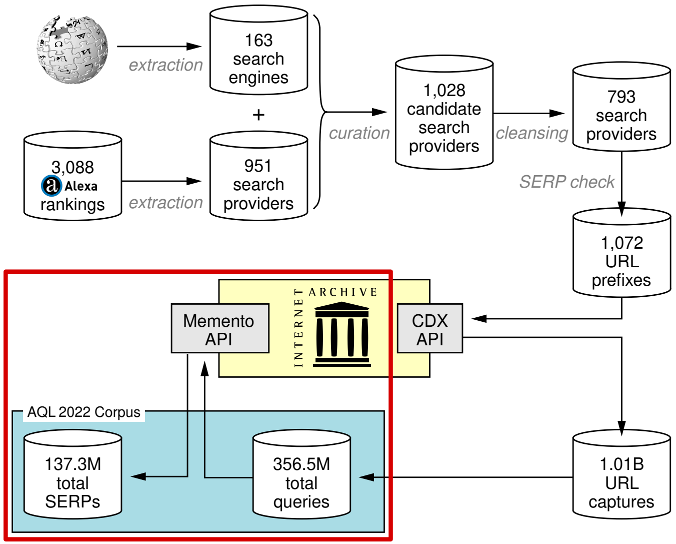
Examples:

$$\underbrace{\texttt{google.com/search}}_{\text{URL prefix}}\texttt{?}\underbrace{\texttt{q=covid+19+usa+map}}_{\text{query}}\texttt{\&}\underbrace{\texttt{start=10}}_{\text{offset}}$$

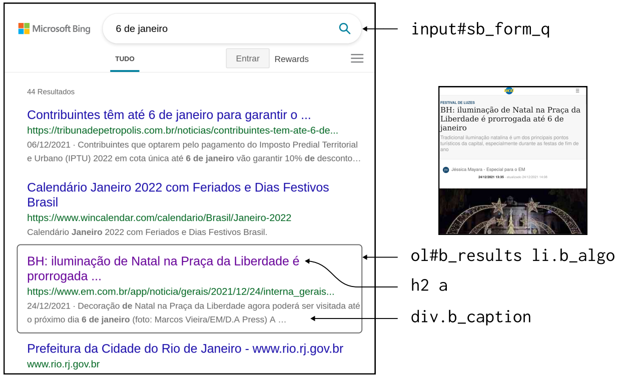
$$\underbrace{\texttt{chefkoch.de/rs/}}_{\text{URL prefix}}\underbrace{\texttt{s0}}_{\text{page}}\texttt{/}\underbrace{\texttt{backen\%20dinkelmehl}}_{\text{query}}\texttt{/Rezepte.html}$$

Archive Query Log

4. SERP Acquisition and Parsing



- ❑ sample SERPs, annotate expected results
- ❑ apply existing parsers
- ❑ compare parsed result with annotations
- ❑ adapt/extend parsers



6 de janeiro

44 Resultados

Contribuintes têm até 6 de janeiro para garantir o ...
https://tribunadepetropolis.com.br/noticias/contribuintes-tem-ate-6-de...
06/12/2021 · Contribuintes que optarem pelo pagamento do Imposto Predial Territorial e Urbano (IPTU) 2022 em cota única até **6 de janeiro** vão garantir 10% **de** desconto…

Calendário Janeiro 2022 com Feriados e Dias Festivos Brasil
https://www.wincalendar.com/calendario/Brasil/Janeiro-2022
Calendário **Janeiro** 2022 com Feriados e Dias Festivos Brasil.

BH: iluminação de Natal na Praça da Liberdade é prorrogada ...
https://www.em.com.br/app/noticia/gerais/2021/12/24/interna_gerais...
24/12/2021 · Decoração **de** Natal na Praça da Liberdade agora poderá ser visitada até o próximo dia **6 de janeiro** (foto: Marcos Vieira/EM/D.A Press) A …

Prefeitura da Cidade do Rio de Janeiro - www.rio.rj.gov.br
www.rio.rj.gov.br

FESTIVAL DE LUZES

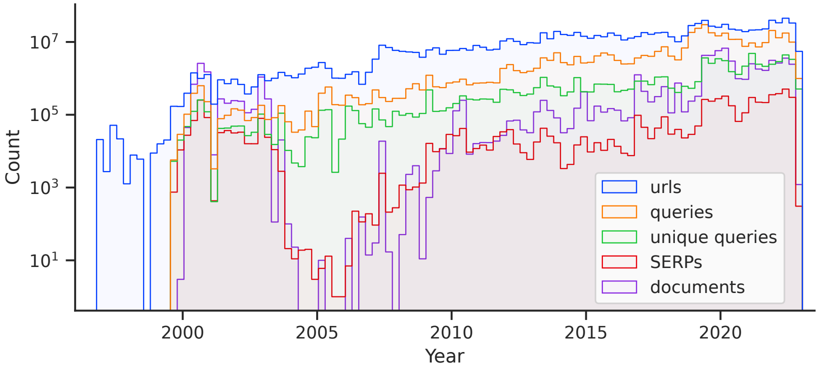
BH: iluminação de Natal na Praça da Liberdade é prorrogada até 6 de janeiro

Tradicional iluminação natalina é um dos principais pontos turísticos da capital, especialmente durante as festas de fim de ano

Jéssica Mayara - Especial para o EM
24/12/2021 13:35 - atualizado 24/12/2021 14:08

Archive Query Log

Analysis



Top	W	▶	f	in	IMDb	•••	↻
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 %

Queries:

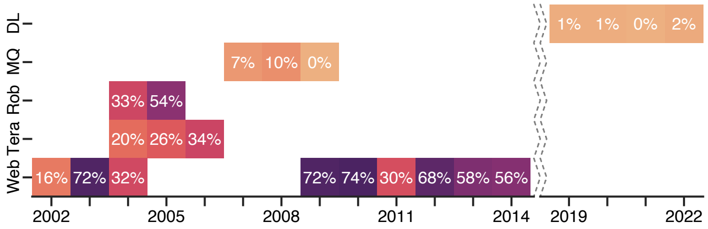
- 104 different languages
- frequent languages: Chinese, English
- most queries 5–20 characters long
- 82% duplicates

SERPs:

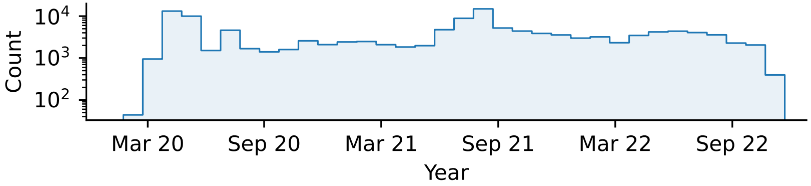
- frequent languages: English, Russian
- popular websites often among top results

Archive Query Log

Use Cases



TREC query overlap



Covid-19-related queries

- transparent insights into search industry
- benchmarks with real user queries
  - → high overlap with some TREC tasks
- diachronic analyses
  - → example with Covid-19-related queries
- training data for neural retrieval models

Archive Query Log

Limitations and Scalability

- ❑ parsers written semi-automatically, cannot interpret dynamic content
  - → explore BERT classification models and wrapper generation
  - → use headless browser to render SERPs
- ❑ slow downloads due to rate limits and network bandwidth
  (93 % of the SERPs still need to be downloaded)
  - → distributed download infrastructure

Access

- ❑ unanswered questions about search economy
  - → AQL facilitates analyses to answer these questions
- ❗ scale opens up ethical and legal questions (e.g., PII, illegal content)
- → privacy-preserving access via TIRA
  - – sandboxed access, results blinded until review
  - – guarantees that no sensitive data is leaked

Archive Query Log

Summary

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
Σ	**550**	**1010.2 M**	**356.5 M**	**64.5 M**	**137.3 M**	**1410.0 M**

Code and Data

- `github.com/webis-de/archive-query-log`
- `tira.io/task/archive-query-log`
- `doi.org/10.1145/3539618.3591890`
- contributions are welcome!

Archive Query Log

Summary

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
Σ 550	**1010.2 M**	**356.5 M**	**64.5 M**	**137.3 M**	**1410.0 M**

Code and Data

- `github.com/webis-de/archive-query-log`
- `tira.io/task/archive-query-log`
- `doi.org/10.1145/3539618.3591890`
- contributions are welcome!

Thank you!