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

Jan Heinrich Reimer
Friedrich-Schiller-Universität Jena

Sebastian Schmidt
Leipzig University

Maik Fröbe
Friedrich-Schiller-Universität Jena

Lukas Gienapp
Leipzig University

Harrisen Scells
Leipzig University

Benno Stein
Bauhaus-Universität Weimar

Matthias Hagen
Friedrich-Schiller-Universität Jena

Martin Potthast
Leipzig University and ScaDS.AI

ABSTRACT

The Archive Query Log (AQL) is a previously unused, comprehensive query log collected at the Internet Archive over the last 25 years. Its first version includes 356 million queries, 166 million search result pages, and 1.7 billion search results across 550 search providers. Although many query logs have been studied in the literature, the search providers that own them generally do not publish their logs to protect user privacy and vital business data. Of the few query logs publicly available, none combines size, scope, and diversity. The AQL is the first to do so, enabling research on new retrieval models and (diachronic) search engine analyses. Provided in a privacy-preserving manner, it promotes open research as well as more transparency and accountability in the search industry.

KEYWORDS

query log, search engine result page, information retrieval history

ACM Reference Format:

1 INTRODUCTION

Search engine query logs are a rich resource for many information retrieval applications [4], such as user behavior and user experience analysis, or query suggestions and query reformulations. When a query log also includes users’ clicks and dwell time on search results, this is a valuable source of implicit relevance feedback about their information needs. Modern search engines use this feedback to train retrieval models for re-ranking [94, 118]. However, query logs are also highly sensitive data that affect a number of stakeholders [11, 73]: First and foremost are user privacy concerns.

Over time, if a user frequently uses a search engine, their query log can be enough to personally identify them and reveal a lot about their state of mind and health. To some extent, this also applies to persons or organizations mentioned or implied in queries or search results. Not least, relevance feedback from a query log is an important asset for search providers, commercial or otherwise.

All of the above are good reasons for not publishing query logs. Another, yet questionable, reason for major search providers is that extracted from them, and archived search result pages (SERPs) shown are the respective numbers of archived URLs, the queries (known domains) (total) (total) (unique) (estimate) (estimate)

<table>
<thead>
<tr>
<th>Search provider</th>
<th>URLs (total)</th>
<th>Queries (total)</th>
<th>Queries (unique)</th>
<th>SERPs (estimate)</th>
<th>Results (estimate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>89.4 M</td>
<td>72.7 M</td>
<td>20.0 M</td>
<td>34.0 M</td>
<td>270.9 M</td>
</tr>
<tr>
<td>YouTube</td>
<td>41.8 M</td>
<td>41.4 M</td>
<td>11.3 M</td>
<td>19.3 M</td>
<td>411.8 M</td>
</tr>
<tr>
<td>Baidu</td>
<td>78.5 M</td>
<td>69.6 M</td>
<td>2.9 M</td>
<td>32.5 M</td>
<td>130.7 M</td>
</tr>
<tr>
<td>QQ</td>
<td>0.5 M</td>
<td>0.5 M</td>
<td>0.1 M</td>
<td>0.2 M</td>
<td>2.6 M</td>
</tr>
<tr>
<td>Facebook</td>
<td>3.1 M</td>
<td>0.2 M</td>
<td>0.0 M</td>
<td>0.1 M</td>
<td>0.8 M</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>8.8 M</td>
<td>2.8 M</td>
<td>1.2 M</td>
<td>1.3 M</td>
<td>11.2 M</td>
</tr>
<tr>
<td>Amazon</td>
<td>66.6 M</td>
<td>0.8 M</td>
<td>0.3 M</td>
<td>0.4 M</td>
<td>9.5 M</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>68.5 M</td>
<td>1.7 M</td>
<td>0.6 M</td>
<td>0.8 M</td>
<td>8.5 M</td>
</tr>
<tr>
<td>JD.com</td>
<td>4.4 M</td>
<td>3.9 M</td>
<td>0.4 M</td>
<td>1.8 M</td>
<td>19.4 M</td>
</tr>
<tr>
<td>360</td>
<td>1.5 M</td>
<td>1.1 M</td>
<td>0.1 M</td>
<td>0.5 M</td>
<td>4.2 M</td>
</tr>
<tr>
<td>540 others</td>
<td>646.6 M</td>
<td>161.8 M</td>
<td>27.7 M</td>
<td>75.4 M</td>
<td>839.5 M</td>
</tr>
<tr>
<td>Σ</td>
<td>550</td>
<td>1,010.0 M</td>
<td>356.5 M</td>
<td>166.4 M</td>
<td>1,709.0 M</td>
</tr>
</tbody>
</table>
Table 2: Overview of large-scale query logs used in previous work. Private logs are grouped by source. Each source is referenced by the paper using the largest sample. Number of usages is given under N. Timespan indicates crawled duration; year indicates date of last included query. Fields with ★ are either not available or not specified. Languages are estimated.

2 BACKGROUND AND RELATED WORK

We take an in-depth look at the use of query logs and search result pages in information retrieval research as well as a brief one at search transparency and accountability and at the Internet Archive.
Despite the fact that query logs are rarely published, researchers in academia have sought alternative means of access, usually by collaborating with search providers of many kinds. Weighted by the number of publications per log, research with very large query logs was conducted in industry and at major search providers (the AOL log being the exception). The AOL log [96] and its recent extension AOLIA [82] are the largest query logs ever made publicly available. The AltaVista log and the Baidu log are the two largest private logs. Our Archive Query Log is on par with the latter two. The ratio of unique queries to all queries averages 0.24. With 0.18 our log is slightly lower due to its multilingual nature. In addition to queries, organic search engine query logs may include information about users, clicks, sessions, and results, while our log only includes queries and results (SERPs and the result documents themselves).

Given the main tasks for which query logs are used, the AQL can be used to study many—though not all—of them at a scale not easily attainable for academic researchers: Query understanding involves analyzing user information behavior. Subtasks include determining the user’s search intent [61] and examining user populations [63], particularly with respect to geographic [105] and temporal [64] dimensions. In addition, much research has focused exclusively on how people search for health information, from both consumer [95] and expert perspectives [102]. Query suggestion involves exploiting query logs to recommend alternative queries to the user. Subtasks include clustering [17], query similarity [25], and query expansion using relevance feedback [44, 45, 59]. The AQL can support both tasks in general, in particular as pre-training data. However, model transfer to a specific application domain will be required.

Session analysis examines how users reformulate their queries across one or more sessions [5], a key subtask being session detection [52, 56]. User modeling involves analyzing logs to build models of user interaction. Subtasks include examining the distributions of query lengths and query terms [65, 103, 110, 111], relevance feedback mechanisms [104], and what users consider relevant [62]. The AQL does not support these tasks; it lacks session or user data.

Learning to rank is about exploiting query logs to derive effective ranking models. Subtasks include developing click models [70, 71] and models that incorporate implicit feedback such as dwell time on pages [3, 114]. The AQL supports this task despite the lack of click data. Craswell et al.’s [38] rationale for the design of the MS MARCO benchmark corroborates this claim. Here, only passages (judged for relevance) from the top-ranked documents returned by Bing for a query are included as ground truth for training, which has proven to be sufficient to yield effective retrieval models. The same is true for the AQL, where the ranked results of third-party retrieval models encode the domain expertise and the implicit relevance feedback from query logs that the respective search providers incorporated into their development.

2.2 Search Engine Result Pages

Search engine result pages (SERPs) are how search engines present results to users in response to a query. SERPs for web search typically consist of a list of links to web pages ranked by their relevance to the user’s query, along with additional information such as snippets, images, and other features designed to help users meet their information needs. SERPs have been studied in information retrieval research for many years to understand how users interact with them, how they can be improved, and how they can present information more effectively to better meet user needs. The AQL contains the SERP for the majority of its queries.

One area of SERP research has focused on understanding how users interact with search results. Researchers have used techniques such as eye-tracking [10, 68, 69] and brain monitoring [88] to study how users perceive SERPs. These studies have led to a deeper understanding of how to improve the presentation of results and the ranking algorithms used by search engines. Another area of SERP research is the study of their design and layout [90, 91]. These longitudinal studies show how SERPs evolve in response to new technologies. The AQL provides millions of archived SERPs which include all necessary assets for showing them in a browser, so that they can be used for user studies and offline experiments.

2.3 Transparency and Accountability

In November 2022, the Digital Market Act [1] and the Digital Services Act [2] came into force in the European Union. The former applies primarily to so-called “gatekeepers” in digital markets, such as Google for the search market, the latter to all digital services that act as so-called (information) intermediaries. Both laws contain provisions that, among other things, require search providers to increase data privacy, transparency, and accountability, with the goal of ensuring fair and open digital markets. In particular, legislators are allowed to exercise regulatory and market investigation powers, which may include looking into the algorithms used. The AQL complements these measures and also gives civilian initiatives the means to conduct independent investigations of search providers. Previous studies on search accountability raise the question of how to inform users about a search engine’s retrieval algorithms to raise awareness of how they work [37, 76, 77] and to ensure unbiased results [51, 78]. While we cannot consider all previous work in this context, a recent overview was provided at the FACTS-IR workshop [92, 93] on fairness, accountability, confidentiality, transparency, and safety in information retrieval. In terms of both algorithm transparency and search engine accountability, archived search result pages are perhaps one of the best representations of a search engine’s behavior, and archiving them on a large scale allows for corresponding post-hoc analyses.

2.4 Internet Archive

The Internet Archive is a nonprofit digital library that has grown to become the largest and most comprehensive digital library in the world since its inception in 1996. In addition to providing access to extensive archives of books, audio recordings, videos, images, and software, the Internet Archive’s best-known service is probably the Wayback Machine, which provides a digital archive of the web. At the time of writing, it contains 788 billion web pages. We believe that the AQL accumulated both due to accidental crawling by their crawlers and intentional archiving by their users, since any user can request archiving of any publicly accessible URL. AOLIA [82] extends the original AOL log by providing links to archived versions of its search results, originally specified as URLs only.

https://web.archive.org/
3 MINING THE ARCHIVE QUERY LOG

Besides general-purpose search engines, many other websites such as social media platforms offer a search function (a query field) for users. The answer to a query is often encoded as URL linking to a SERP, which is displayed to the user. Like other URLs, these “SERP URLs” can be linked to by web pages and are thus included in automated web crawls. The Internet Archive, as the world’s largest digital library of archived web pages, is likely to include many SERPs, a fact which can be exploited for large-scale query log mining. This section describes a multi-step process to mine a query log from the Internet Archive’s Wayback Machine, which eventually becomes the Archive Query Log (see Figure 1).

First, a list of popular search providers including general-purpose search engines and all kinds of media platforms is compiled (see upper part of Figure 1; Section 3.1). Second, for each provider the top-level domains, subdomains, and URL patterns under which SERPs are likely to be found are semi-automatically generated and the URL captures found in the Internet Archive (using the CDX API of the Wayback Machine) are aggregated (Section 3.2). Third, the queries are extracted from the URLs using provider-specific parsers (Section 3.3). Fourth, the HTML content of archived SERPs is downloaded and its search results snippets are extracted (Section 3.4). Both queries and snippets form the AQL 2022 Corpus (Section 3.5); the corpus will be made accessible for shared tasks and experiments via the TIRA platform [97] as discussed in Section 4.

3.1 Search Provider Collection

Our search provider collection shall contain both (1) websites that primarily act as search engines, and (2) highly relevant websites that have been identified by their Alexa Rank. 5

Regarding (1), we exploit a dedicated list of search engines on Wikipedia which we extend manually. 6 Regarding (2), we take the 3,088 archived snapshots of the Alexa top-1M ranking between June 2010 and November 2022 7 and apply reciprocal rank fusion [36] considering the 1,000 highest ranked domains of each snapshot. The resulting list of 13,647 domains is narrowed down to 951 search providers by checking whether a search bar is present on the website’s landing page of the respective provider. For this purpose we load the landing page directly or from an Internet Archive snapshot from 2022, render the page if JavaScript content is found, and check for HTML forms or <div> elements containing the pattern “search” in its attributes.

The merged list of 1,028 unique candidate search providers is used to identify relevant URL patterns as well as suitable approaches for query extraction (see Section 3.2). Further manual curation steps weed out providers because they have been identified as spam, do not encode the search query in their URL, or offer only an autocomplete search that links directly to a page. Also, search providers are merged because more than one of their second-level domains appears in the merged Alexa list (e.g., so.com is merged into 360.com). After these analyses and curation steps, the list still includes 793 search providers.

3.2 Provider Domains and URLs

Since many search providers are available under multiple top-level domains and/or subdomains, 8 we expand the list of the 793 provider domains manually as well as with publicly available lists 9.

However, on high-traffic domains, only a small fraction of all archived URLs is likely to link SERPs and thus is relevant for our purposes. To identify common prefixes of URLs that contain queries, we submit multiple test queries using the search provider’s query field and examine the URL the request is redirected to. For discontinued or otherwise inaccessible websites, we resort to the most recent functional snapshot of the search provider’s homepage in the Internet Archive. The final list of 1,072 URL prefixes is used to filter the list of available captures with the help of the Internet Archive’s CDX API. This API allows to request a list of URLs by the crawling date they were archived in the Wayback Machine for a certain domain or URL prefix. Via the CDX API, a list of all available captures for each of the 1,072 URL prefixes is retrieved and filtered for successful captures with HTML content (i.e., HTTP status code 200). This process further narrows the search provider list with archived SERPs to 550.

Altogether, 1.1B URL captures along with their crawling date are collected, an average of 1.8M URLs per provider. Most of the captures originate from search engines (226M URLs, 22 %) with Google being the largest contributor of archived URLs (89M URLs, 9 %).

3.3 Query Extraction

To extract the query from a SERP URL, the URL is parsed into its components 10 and the query encoding is identified as one of three possible categories: (1) URL parameter, (2) path segment, or (3) fragment identifier. Examples of the first two patterns are illustrated in Figure 2.

---

5https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server
6The Alexa Rank was a ranking system that reflected the global popularity of websites based on estimated visits; it was shut down end of 2022.
7See https://en.wikipedia.org/wiki/List_of_search_engines
8https://web.archive.org/web/*/s3.amazonaws.com/alexa-static/top-1m.csv.zip
9See https://github.com/JamieFarrelly/Popular-Site-Subdomains
This is supported by the fact that search engines are the main contributors (70M queries, 20%). On average, 648,092 queries are extracted (162M queries, 46%) such as Google (73M queries, 20%) and Baidu (70M queries, 20%).

The resulting set of parsers is applied to all available captures of search result pages. For each search provider, we manually examined the captured URLs in a similar way as for URL prefixes (see Section 3.2) and derived suited URL parser types and parameters. Regular expressions are optionally used to limit parsing to specific URLs, and to further refine the parsers (e.g., removing prefixes such as page in search/page). The resulting set of parsers is applied to all available captures of search result pages, ordered by preference, so that the first parser that returns a non-empty query, page, or offset is used.

Altogether a total of 356.5M URLs containing queries are collected. Again, the majority of queries stems from search engines (162M queries, 46%) such as Google (73M queries, 20%) and Baidu (70M queries, 20%). On average, 648,092 queries are extracted per search provider. This unfiltered set of queries contains large amounts of duplicates (288M, 81%) for which we identify three reasons: (1) the same query is captured at different times, (2) the query is captured at approximately the same time but with different result page offsets, and (3) the same query is captured as issued from different users (e.g., if a user identifier is included in the URL). This is supported by the fact that search engines are the main contributors of duplicates (131M); however, government sites have the highest share (91%).

We create a set of unique queries for each search provider by selecting a representative query URL (the capture with the shortest query string) from each group with the same parsed query. If a group of duplicates has multiple captures with the same query parameter and URL length, the representative URL is chosen by lexicographic order. We refrain from using the capture’s timestamp as a tie-breaker to not favor older or newer captures. The deduplication results in 64.5M unique queries across the full list of 550 search providers. There are 117,353 unique queries per each provider on average. Again, the search engines make up the majority of deduplicated queries (31M, 45%).

### 3.4 SERP Acquisition and Parsing

Previous query logs rarely contain results for the logged queries (see Section 2). The AQL however, since it is obtained from web page captures in the Internet Archive, naturally contains the full ranking of results for the majority of queries. By downloading and parsing the archived SERPs, one has access to the full ranking of results for each search query (including the processed query itself, as it appears in the query field of the SERP). Parsing the search results including result titles, referenced URLs, snippets, and the query from a SERP facilitates not only the comparison of different search provider’s ranking functions but also the evaluation of their query understanding and reformulation techniques.

We download the SERP HTML content for the unique search queries identified in the previous step and save it in WARC format. For the 20 most popular search providers SERPs are downloaded for all unique URLs; for the remaining providers the download is limited to a maximum of 25,000 due to resource constraints. Connection timeouts and other errors during download are handled by repeating the download up to 10 times, after which we consider the archived SERP snapshot to be unavailable. Altogether, a total of 166.4M SERPs are collected most of which originate from search engines (ca. 40 %) and media sharing platforms (ca. 20 %). The downloads are ongoing, and we plan on scaling them up (see Section 5) to compile the full set of estimated SERPs available (see Section 4).

From the downloaded SERPs the search result ranking is extracted and processed using a configurable parser pipeline based on

---

2. As split according to RFC 2396.
FastWARC [19], Beautiful Soup,14 and Approval Tests.15 In detail, the CSS path or selector16 to the result list items is specified, as well as the path from each individual result item to its title element, the referenced URL anchor, and the snippet text. Similarly, the processed query is parsed based on the CSS path to the query field. Figure 3 shows an archived Bing SERP and highlights how CSS paths are used to select relevant HTML tags. Each search provider can have multiple parser configurations, ordered by preference, that, for example, account for a changed HTML structure after redesigns of a search provider’s SERPs.

We derive parser configurations (CSS paths) for the 50 most popular search providers by generating Approval Tests according to following workflow for a provider: (1) Randomly sample 10 SERPs from the downloaded SERPs. (2) For each SERP manually annotate the expected ranking and query. (3) Apply the existing parser configurations to the sampled SERPs. (4) Compare the parsed results to the annotations. (5) If the annotations do not match, inspect the HTML page in a browser, adapt or extend missing patterns, and add them to the provider’s parser configurations. New configurations are added iteratively until all sampled SERPs are correctly parsed. Altogether, 70 parser configurations for SERPs and 57 parsers for processed queries of the 50 most popular providers are derived. With additional manual tests for the 10 most popular providers (see Table 1), the parsers pass a test suite of 444 Approval Tests. The code base is available open source.17

3.5 The Archive Query Log 2022 (AQL-22)

We merge the filtered URL captures, queries, and SERPs into a single corpus to be used in subsequent analyses (see Section 4). This corpus, the Archive Query Log 2022, consists of two artifacts: (1) a set of queries and (2) a set of ranked documents (search result snippets). Both artifacts are stored in a GZIP-compressed, newline-delimited JSON format.18 To create the query set, each captured URL is assigned a unique identifier based on the full URL string and timestamp of the capture.19 The captured URL is associated with its parsed query, the location of the stored copy of the SERP, and the processed query and search results parsed from that SERP. In addition, we include the URL to the SERP’s archived snapshot on the Wayback Machine and tag the query language based on the parsed query text using cld3.20

The set of result documents is created by concatenating all ranked search results (i.e., rank, snippet text with title, and document URL to the referenced web page) from all parsed SERPs. Each document is assigned a unique identifier based on the document URL, the timestamp of its origin query, and the rank of the snippet on the SERP. We also associate each document with the attributes of the corresponding query in the query set. Two additional fields are the URL to the nearest available snapshot of the SERP on the Wayback Machine and the snippet language as tagged using cld3 based on the snippet’s title and text.

Figure 4: Distribution of query lengths for 5 search providers contributing the highest amount of queries. The remaining search providers are grouped as “others”.

4 ANALYSIS

In order to provide a better understanding of the corpus, we conduct some analyses on the query and SERP characteristics and highlight potential use cases. Detailed analyses of the AQL will be the subject of future work. At the time of writing, we could download and parse all URLs and queries. However, due to computational constraints, only a subset of all available SERPs could be parsed yet (cf. Section 4.4) and are used for our analyses.

4.1 Query Characteristics

A central feature of the AQL is its diversity. In addition to the variety of search providers, it also features a total of 104 different languages21 with Cantonese and English being the most frequently used query languages (see Table 3). The query length in the AQL follows a skewed distribution, with most queries containing between 5 and 20 characters. Figure 4 provides a visualization for the 5 search providers with the most queries. We inspect samples of queries with 5, 10, 100, and 1000 characters. Very short queries are often Mandarin keywords (e.g., 长袖衬衫男, “men’s long sleeve shirt”) or single English words (e.g., comic, font) or hashtags (e.g., #чемпионат, "championship"). Most longer queries extensively use search operators like site: and order:, include literature references, or include long multi-line text like stack traces from errors in programming.

Second, we evaluate whether obscene or unwanted terms comprise a large share of the AQL. We use lists of obscene words for 27 languages22 and expand the list of English terms with new expressions found in the downloaded queries. We check each query from the two most dominant languages, Cantonese and English, for their lists of obscene terms. Overall, only 1.30 % of all queries contain obscene terms. The highest share of these obscene queries

---

14https://pypi.org/project/beautifulsoup4/
15https://pypi.org/project/approvaltests/
16See https://facelessuser.github.io/soupsieve/ for a list of supported selectors. The CSS path to an HTML element can be inferred using a web browser’s developer tools.
17AQL code: https://github.com/webis-de/archive-query-log
18https://jsonlines.org/
19Name-based SHA-1 UUID according to RFC 4122: https://rfc-editor.org/rfc/rfc4122
20Google’s Compact Language Detector; https://github.com/google/cld3
21Out of 107 detectable with cld3.
Table 3: The Archive Query Log 2022 (AQL-22) in detail. Categories manually annotated. Top-3 languages tagged by cld3. Ticks in the timelines indicate days of archival from Jan 1999 to Dec 2022. Number of SERPs and results estimated (c.f. Section 4.4).

<table>
<thead>
<tr>
<th>Search provider</th>
<th>Category</th>
<th>URLs</th>
<th>Queries</th>
<th>SERPs</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>Search engine</td>
<td>89,364,948</td>
<td>72,673,044</td>
<td>19,953,592</td>
<td>en, th, zh</td>
</tr>
<tr>
<td>YouTube</td>
<td>Media sharing</td>
<td>41,846,525</td>
<td>41,365,166</td>
<td>11,250,179</td>
<td>ru, ko, ja</td>
</tr>
<tr>
<td>Baidu</td>
<td>Search engine</td>
<td>78,506,825</td>
<td>69,619,339</td>
<td>2,900,878</td>
<td>zh, ga, ja</td>
</tr>
<tr>
<td>QQ</td>
<td>Web portal</td>
<td>515,895</td>
<td>513,608</td>
<td>51,228</td>
<td>zh, ja, lb</td>
</tr>
<tr>
<td>Facebook</td>
<td>Social media</td>
<td>3,131,212</td>
<td>159,087</td>
<td>35,492</td>
<td>ca, en, bs</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>Web portal</td>
<td>8,787,707</td>
<td>2,827,103</td>
<td>1,232,589</td>
<td>en, la, de</td>
</tr>
<tr>
<td>Amazon</td>
<td>E-commerce</td>
<td>66,795,164</td>
<td>776,127</td>
<td>315,068</td>
<td>en, ja, zh</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Wiki</td>
<td>68,547,099</td>
<td>1,707,058</td>
<td>621,971</td>
<td>sv, zh, en</td>
</tr>
<tr>
<td>JD.com</td>
<td>E-commerce</td>
<td>4,370,884</td>
<td>3,902,604</td>
<td>370,473</td>
<td>zh, br, ja</td>
</tr>
<tr>
<td>360</td>
<td>Search engine</td>
<td>1,495,365</td>
<td>1,090,152</td>
<td>65,596</td>
<td>zh, ja, mg</td>
</tr>
<tr>
<td>Weibo</td>
<td>Social media</td>
<td>6,245,012</td>
<td>5,324,385</td>
<td>1,868,458</td>
<td>zh, ja, en</td>
</tr>
<tr>
<td>Reddit</td>
<td>Forum</td>
<td>94,162</td>
<td>89,492</td>
<td>36,852</td>
<td>en, la, de</td>
</tr>
<tr>
<td>Vk.com</td>
<td>Social media</td>
<td>643,354</td>
<td>153,642</td>
<td>46,134</td>
<td>ru, sr, ky</td>
</tr>
<tr>
<td>CSDN</td>
<td>Forum</td>
<td>21,863</td>
<td>946</td>
<td>736</td>
<td>zh, en, vi</td>
</tr>
<tr>
<td>Bing</td>
<td>Search engine</td>
<td>11,263,539</td>
<td>6,152,425</td>
<td>2,253,965</td>
<td>en, zh, pt</td>
</tr>
<tr>
<td>Twitter</td>
<td>Social media</td>
<td>55,499,532</td>
<td>48,084,528</td>
<td>3,869,382</td>
<td>ja, en, gl</td>
</tr>
<tr>
<td>Twitch</td>
<td>Streaming</td>
<td>21,931</td>
<td>15,225</td>
<td>11,445</td>
<td>en, de</td>
</tr>
<tr>
<td>eBay</td>
<td>E-commerce</td>
<td>7,927,123</td>
<td>5,507,532</td>
<td>1,379,646</td>
<td>zh, en, la</td>
</tr>
<tr>
<td>Naver</td>
<td>Search engine</td>
<td>1,063,991</td>
<td>756,153</td>
<td>400,490</td>
<td>ja, ko, vi</td>
</tr>
<tr>
<td>AlliExpress</td>
<td>E-commerce</td>
<td>4,620,331</td>
<td>1,861,642</td>
<td>55,849</td>
<td>en, lb, fy</td>
</tr>
<tr>
<td>530 others</td>
<td></td>
<td>559,225,614</td>
<td>93,871,236</td>
<td>17,806,322</td>
<td>en, zh</td>
</tr>
<tr>
<td>∑</td>
<td></td>
<td>1,009,988,486</td>
<td>356,450,494</td>
<td>64,544,345</td>
<td>zh, en, ga</td>
</tr>
</tbody>
</table>

Figure 5: Time coverage of the total amount of different data types collected for the AQL-22, per quarter.

were observed on pornography (19.48 %), torrent (3.73 %), and forum (2.87 %) websites. For non-pornographic search providers, most stem from heroturko.org (16.13 %, e-commerce), reddit t.com (5.05 %, forum), and kat.cr (4.08 %, forum).

Regarding time coverage, Figure 5 shows that the archival of SERPs dropped between 2004 and 2010 for unknown reasons, which might indicate that using more specialized SERP parsers are required, or that results were loaded using Javascript. The queries, however, extend over the whole timespan, with tens of thousands of queries recorded for the early 2000’s as well. Table 3 contains an overview of the 20 most popular services’ time coverage.

4.2 SERP Characteristics

In Table 4, we consider the most frequently referenced URLs from search results as an indicator of plausible rankings. Excluding frequent self-references (e.g., to internal redirect pages), by far the most frequently ranked domain is wikipedia.org contributing 2.9 % of all top-5 results and 2.2 % of the top-10. Other popular domains like youtube.com and facebook.com also frequently appear on high ranks.

Table 4: Most frequent document domains in the top-5 or the top-10 search results compared to references to the search provider’s own domain (∩) or 791,646 other domains (−−).

<table>
<thead>
<tr>
<th>Top</th>
<th>101</th>
<th>103</th>
<th>105</th>
<th>107</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count urls</td>
<td>queries</td>
<td>unique queries</td>
<td>SERPs</td>
<td>Results</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>----------------</td>
<td>-------</td>
<td>---------</td>
</tr>
<tr>
<td>5</td>
<td>2.9 %</td>
<td>0.8 %</td>
<td>0.6 %</td>
<td>0.4 %</td>
</tr>
<tr>
<td>10</td>
<td>2.2 %</td>
<td>0.7 %</td>
<td>0.5 %</td>
<td>0.3 %</td>
</tr>
</tbody>
</table>

The AQL opens up a variety of use cases for the IR community. We highlight two promising applications.

First, we evaluate the exact overlap of the queries in the AQL with the collections used in various TREC tracks from 2004 to 2022 [7, 8, 21, 27–34, 39–43, 108, 109]. As shown in Figure 6, the highest overlap exists with the Web tracks, specifically in 2010 (74 %), 2003, and 2009 (both 72 %). The lowest overlap was found with the Deep Learning tracks, ranging between 0–2 %. The high overlap on older Web tracks poses an interesting opportunity for enriching existing benchmarks. While query logs have been used previously in shared community tasks [83], shared tasks often specify only one query for each topic. We propose to sample semantically similar queries...
from the AQL to generate topics with user query variations [16] automatically. On the other hand, the low overlap with the Deep Learning tracks highlights a sampling bias in creating the Deep Learning topics. The topics were sampled from the official eval set of MS MARCO, which includes only natural language questions from a Bing query log [43, 89]. The AQL, on the other hand, contains a much broader range of queries, including queries from other search providers and non-question-like queries. Therefore, we propose using the AQL to create new, “harder” Deep Learning topics that are more representative of other kinds of queries users submit.

Second, we demonstrate how global trends are reflected in the AQL on the example of the Covid-19 pandemic. In Figure 7, we count the occurrences of the terms covid 19, sars cov 2, and corona virus each month since the outbreak in 2019. A peak can be observed during the first global lockdowns in early 2020, but overall, interest in the pandemic has yet to stagnate. The example showcases how the AQL enables unique opportunities for diachronically analysing global trends.

### 4.4 Total Size Estimates

As Section 3 explains, we have only downloaded and parsed a subset of all available SERPs. Based on our results so far, we estimate 85% of all SERP snapshots to be available for download. Assuming an estimated parsing success rate of 55% and 10.6 results per SERP, we expect the total number of parsed SERPs in the AQL-22 to be 166.4M with 1.7B search results. As outlined in Section 5, we continue to download and parse SERPs and look forward to expanding the AQL with the IR community.

## 5 DISCUSSION

Access to query logs has long been an insurmountable barrier to answering critical questions about the search economy at large—if not to ask them in the first place. As a consequence, the media and general public were left with no choice but to trust search engines on questions such as “How accountable are organizations operating search engines in terms of measures of interest, like representation and fairness?”, “How have these accountability measures changed for these organizations over time?”, and “How honest has self-reported accountability of these organizations been?”. As the most extensive public query log to date, the AQL enables detailed analyses of and thus facilitates the public discourse on the search industry. It also furthers the research on information retrieval, whose retrieval models are often presumed to be behind or at least detached from those of industry players (e.g., [13]). Using the AQL, researchers will also be able to answer questions such as “How far is academic information retrieval research behind industry?”; “How much do query logs contribute compared to other ranking signals?”, and not least “What are domain-specific differences?”.

However, with the scale of the AQL, several ethical and legal considerations also arise, particularly around personally identifiable information or illegal content. To address inherent risks, we release the data by imposing a barrier to access that minimizes potential harms while giving the information retrieval community as much freedom to conduct their research as possible. In addition, we also acknowledge the challenges of creating such extensive collections and discuss our plans for opening contributions to the AQL.

### 5.1 Accessing the AQL

Our goal with releasing the AQL is to do so responsibly and in a privacy-preserving manner. We work towards that goal using the TIRA platform [50, 97] for any analysis on the AQL one wishes to conduct. TIRA has been used since 2012 [53] to facilitate shared tasks with software submissions while ensuring that the submitted software can process the data without giving the participants themselves access. TIRA achieves this through sandboxing, i.e., disconnecting the software from the internet while it is running and thus ensuring that it can not leak data. We added the AQL to TIRA to allow researchers to submit their analysis software as Docker images. The platform is open to the public, and we provide examples and documentation on how to perform analyses on the AQL.23 Specific shared tasks will be developed as well.

TIRA allows running arbitrary software packaged in Docker images on the AQL dataset in a privacy-preserving way, as the analysis results are blinded until reviewed. Specifically, we review the output and the software installed in the Docker image to ensure no sensitive data is leaked. TIRA runs the software in a Kubernetes cluster (1,620 CPU cores, 25.4 TB RAM, 24 GeForce GTX 1080 GPUs) with a timeout of 24 hours, so that almost any evaluation is supported.24 In summary, TIRA provides the ideal means to work with the AQL, ensuring sensitive query log data remains secure and is responsibly used for academic research.

### 5.2 Limitations and Scalability

While creating the AQL, we encountered several technical limitations that guide future optimizations and improvements. First, the various parsers for creating the AQL-22 corpus were written semi-automatically. This approach was error-prone and inefficient, requiring much manual work. When building future versions of the AQL, we plan to train token classification models like BERT [47] to automatically generate query parsers based on a training set derived from our existing parsers. Manually finding the correct

---

23https://tira.io/task/archive-query-log
24We can extend the timeouts and available resources individually if the need arises.
paths for snippets on a SERP is a similarly tedious process that can benefit from wrapper generation [75], which has successfully been applied to web page parsing [23].

Second, dynamic content cannot be interpreted by our existing parsers. The SERPs of DuckDuckGo, for instance, are loaded dynamically using JavaScript and thus cannot be parsed from just the archived HTML snapshot, yet the search results are still archived as a different record. To overcome this limitation, we plan to use a headless browser to render the SERPs and then parse them. A helpful library for this type of content extraction is Scriptor.25

However, the last and most pressing limitation is that all URL captures and SERP contents must first be downloaded from the Internet Archive, which is restricted by both rate limits and network bandwidth. As estimated in Section 4.4, currently 96% of the SERPs still need to be fully downloaded from the Internet Archive and thus could not yet be parsed. Hence, more computational resources are required to use this extensive collection. In this regard, we will reach out to the Internet Archive whose privileged access to their infrastructure will allow for much faster compilation of the data.

5.3 Contributing to the AQL

There is an inherent boundary between search providers and researchers when using query logs. The AQL lowers this boundary by exploiting the web archival process of the Internet Archive. As we have described above, physical limitations such as network speed restrict the rate at which we can further grow the AQL. One way to overcome such limitations is to distribute computations across an open community, an approach that has been successfully employed in mathematics.26 We will therefore open source our code to allow the community to contribute query and SERP parsers.

6 CONCLUSION

The Archive Query Log provides an unparalleled academic resource for information retrieval researchers. It consists of over 356 million queries, over 166 million SERPs, and over 1.7 billion results extracted from the SERPs, all coming from 550 search providers spanning 25 years. The AQL is the largest and most diverse query log ever publicly available. From an academic perspective, the AQL will enable researchers to tackle challenges in information retrieval that were not possible until now, ranging from the development of new retrieval models, the development of query suggestion or query prediction models, to large-scale diachronic analyses of search engines; to name the most salient research avenues. Furthermore, our release plan for accessing the AQL ensures that we minimize the harm to society and will allow researchers to safely research the transparency and accountability of commercial search engines while protecting user privacy.

In this paper, we have documented the initial version of the AQL (i.e., AQL-22). We have plans to release future versions of the AQL that will further expand the collection. First, we will continue to add to the long tail of search providers and continue our efforts to download and extract more data from the Internet Archive. Continuing to grow the types of data provided, the next version of the AQL will also include the content of web pages for each result in a SERP. Not least, we will investigate the training of large re-ranking models based on this data.

Altogether, the AQL is an exceedingly valuable resource for researchers and will enable advances in information retrieval research that were previously insurmountable due to the relatively low scale of query logs. Because of its scope, size, and diversity we consider the AQL a significant contribution to the community, and these dimensions will continue to grow as we build upon and expand future versions of the AQL.

REFERENCES


The Archive Query Log: Mining Millions of Search Result Pages of Hundreds of Search Engines from 25 Years of Web Archives
SIGIR ’23, July 23–27, 2023, Taipei, Taiwan


