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

**The AQL-22 at a glance**

<table>
<thead>
<tr>
<th>Search provider</th>
<th>URLs</th>
<th>Queries</th>
<th>unique</th>
<th>SERPs</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>89.4 M</td>
<td>72.7 M</td>
<td>20.0 M</td>
<td>28.0 M</td>
<td>223.1 M</td>
</tr>
<tr>
<td>YouTube</td>
<td>41.8 M</td>
<td>41.4 M</td>
<td>11.3 M</td>
<td>15.9 M</td>
<td>339.2 M</td>
</tr>
<tr>
<td>Baidu</td>
<td>78.5 M</td>
<td>69.6 M</td>
<td>2.9 M</td>
<td>26.8 M</td>
<td>107.6 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>2.1 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>0.7 M</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>8.8 M</td>
<td>2.8 M</td>
<td>1.2 M</td>
<td>1.1 M</td>
<td>9.2 M</td>
</tr>
<tr>
<td>Amazon</td>
<td>66.8 M</td>
<td>0.8 M</td>
<td>0.3 M</td>
<td>0.3 M</td>
<td>7.8 M</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>68.5 M</td>
<td>1.7 M</td>
<td>0.8 M</td>
<td>0.7 M</td>
<td>7.0 M</td>
</tr>
<tr>
<td>JD.com</td>
<td>4.4 M</td>
<td>3.9 M</td>
<td>0.4 M</td>
<td>1.5 M</td>
<td>16.0 M</td>
</tr>
<tr>
<td>360</td>
<td>1.5 M</td>
<td>1.1 M</td>
<td>0.1 M</td>
<td>0.4 M</td>
<td>3.5 M</td>
</tr>
<tr>
<td>540 others</td>
<td>646.8 M</td>
<td>161.8 M</td>
<td>27.8 M</td>
<td>62.4 M</td>
<td>693.9 M</td>
</tr>
<tr>
<td>Total</td>
<td>550</td>
<td>1010.2 M</td>
<td>356.5 M</td>
<td>64.5 M</td>
<td>137.3 M</td>
</tr>
</tbody>
</table>

**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)

**Use cases**

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

**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**
   - 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**

**Query characteristics**

- 104 different languages
- Top languages: Chinese, English
- Most queries: 5–20 characters
- 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
- Popular websites often among top results

<table>
<thead>
<tr>
<th>Top</th>
<th>Language</th>
<th>URL Prefixes</th>
<th>Other</th>
<th>Self</th>
</tr>
</thead>
<tbody>
<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>
| 2022| 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](https://github.com/webis-de/archive-query-log)
- [tira.io/task/archive-query-log](https://tira.io/task/archive-query-log)
- [doi.org/10.1145/3539618.3591898](https://doi.org/10.1145/3539618.3591898)

---

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/10.13038/10.1007/101070014)