Understanding Search Behavior in Wikipedia.

Bruno Scarone

August 15, 2022

Contents

1 Introduction

The objective of this report is to detail the work performed and results obtained as part of the research initiative titled "Understanding Search Behavior of Users" [1], conducted per the request of the Search Platform team at Wikimedia Foundation (WMF) [2]. The high-level goal of the initiative is to provide a better understanding of how and why WMF's internal search is being utilized by the different types of users of the platform.

1.1 Research questions

The report primarily targets the following questions:

1. Understand differences in search behavior on the web vs mobile user interface (including browsers).
2. Understand country/regional differences, especially in emerging countries and specific languages.

These were selected from a prioritized list of research questions which are of interest for the Search Platform team. Other elements relevant to the study were taken from the following Phabricator[1] task [3], namely:

4. Top queries and top keywords: Evaluation of common or relevant query patterns.
5. Top returned documents (articles) and top clicked through documents.

1.2 Collaborators

In this section we enumerate collaborators and stakeholders that contributed to the present research initiative:

- Leila Zia (lzia@wikimedia.org), Head of Research, Wikimedia Foundation.
- Ricardo Baeza-Yates (r.baeza-yates@northeastern.edu), External collaborator, Northeastern University.
- Erik Bernhardson (ebernhardson@wikimedia.org), Software Engineer, Search Platform Team, Wikimedia Foundation.
- Fabian Kaelin (fkaelin@wikimedia.org), Senior Research Engineer, Search Platform Team, Wikimedia Foundation.
- Martin Gerlach (mgerlach@wikimedia.org), Research Scientist, Wikimedia Foundation.
- Mike Pham (mpham@wikimedia.org), Senior Product Manager, Wikimedia Foundation.
- Guillaume Lederrey (glederrey@wikimedia.org), Engineering Manager, Search Platform Team, Wikimedia Foundation.

[1] Phabricator is a collaboration platform open to all Wikimedia contributors, mostly used for managing work in software projects.

1.3 Structure of this report

The remaining of this report is structured as follows: Section 2 introduces the main data sources used, together with data retention constraints that apply to them, as well as the selected time ranges and methodology for computing the results.

Section 3 contains the general results of the analysis when no splitting over the output data is performed. Later, in Section 4, the same data is presented (when possible) grouped and analyzed by the employed access method. Section 5 aims at performing the analysis of the search behavior of users primarily based on the project's language being accessed, as well as on the countries from which the requests come from.

Finally, Section 6 provides recommendations based on the work performed and details possible future lines of work that are considered of interest.

Name	Details	Max. data retention
Web request logs (wmf.webrequest table)	[4]	90 days
Search logs (event.mediawiki_cirrussearch_request table; abbreviated emcr)	[5]	90 days
discovery.query_clicks_daily table (abbreviated dqcd)	-	90 days
wmf.mediawiki_wikitext_current table	[6]	-
event.searchsatisfaction table (abbreviated ess)	[7]	-
Pageview hourly (wmf.pageview_hourly table)	[8]	-

Table 1: Data sources used for this report.

2 Data and Methods

In this section the main data sources used, together with data retention constraints that apply to it are presented, as well as the selected time ranges and methodology for computing the results.

2.1 Data sources

The different data sources utilized in the analysis are listed in Table 1, together with their description page as well as the data retention policy that applies to them. Additional remarks obtained in consultation with the Engineering division of the Search Platform team are listed in Table 2.

2.2 Time ranges considered

The quantities analyzed in Section 3 and 4 have been computed for 2 one-week time ranges: 2-8/May/2022 and 4-10/Jul/2022. Both time ranges span the first complete week (from Monday to Sunday) of a month and the selection was done as an initial validation check of the results. At times in Section 3, when results are considered to be similar for both time ranges (2-8/May/2022 and 4-10/July/2022) only one set of results is shown, based on the fact that both are included in Section 4.

Section 5.1 uses again two time ranges for validation purposes at the monthly level (Feb/2021 and Jul/2022) and Section 5.2 is computed for 4-10/Jul/2022, since it depends on the number of words per project and the user hits, quantified and validated in Section 5.1.

2.3 Data retention policy

The data retention policy that currently applies to the data sources considered in this report is explained next. The policy has been validated in consultation with the Engineering division of the Search Platform team.

As shown in Table 1, the following tables have **90 days** data retention policy:

- Web request logs

Data source	Remarks
Search logs (emcr)	The table contains 1 tuple for every Mediawiki execution (Web requests, but also jobs) that performs a request against the elasticsearch instance. This is the primary dataset that is considered to be reasonably complete, no sampling is applied and it gives detailed information about the incoming request, what is returned by the elasticsearch instance and what is returned to the user.
event.searchsatisfaction	Contains the desktop web frontend data collection.
discovery.query_clicks_daily	Several remarks for this data set are listed below: ▪ Only contains searches that have clickthroughs. The hourly table, which the daily table is built from, has the unclicked searches as well, but also includes more bots and is unsessionized. ▪ The dataset no longer has the full webrequest log, it only has page loads that came from Special:Search directly (i.e., full text searches). Additionally, it only retains searches that have a click. ▪ The dataset contains, without particular verification in several years, the full set of clickthroughs on Special:Search for mobile and desktop access. It is known by the team that 80% of search sessions never make it to Special:Search, they interact entirely through autocomplete, that takes them directly to a page. The timestamp in the attribute clicks comes from the Hive to_unix_timestamp function, thus it should be in seconds.

Table 2: Data sources' additional remarks.

- Search logs (`event.mediawiki_cirrussearch_request` table)
- `discovery.query_clicks_daily` table

This policy generically applies for tables that may contain *Personal Identifiable Information* (PII) from users, which is the case of the web request and search logs. It is noted by the team that the `dqcd` table may be able to retain data for longer periods of time upon minor adjustment. The `wmf.pageview_hourly` table does not have any data retention constraint.

2.4 Infrastructure and software

All code used to generate the results included in this report was run on Wikimedia Foundation's production cluster [9]. In particular, the databases were queried using `Apache PySaprk` [10] version 2.4.4.

3 General Search behavior

This section contains the general results of the analysis, when no segmentation is performed over the results. Later in Section 4, the output data is presented (when possible) grouped and analyzed by the employed access method [2]. At times in this section, when results are considered to be similar for both time ranges (2-8/May/2022 and 4-10/July/2022) only one set of results is shown, based on the fact that both time ranges are included in Section 4.

3.1 Number of sessions

We start by computing the total number of sessions for the following data source and time ranges:

Data sources and time ranges considered

- Data source: `emcr INNER JOIN dqcd` on

```
emcr.search_id = dqcd.request_set_token
```

- Time ranges: 2-8/May/2022 and 4-10/July/2022.

The total counts are displayed below:

- Total number of sessions on 2-8/May/2022: $4,087,262$.
- Total number of sessions on 4-10/July/2022: $3,605,793$.

Data analysis

Both time ranges have a total number of sessions with the same order of magnitude with their difference being $481,469$.

3.2 Number of clicks per session

In this section, we calculated the distribution of sessions according to their total number of clicks. We considered the following data sources and time ranges.

Data sources and time ranges considered

- Data sources: `emcr INNER JOIN dqcd` on

```
emcr.search_id = dqcd.request_set_token
```

- Time ranges: 2-8/May/2022 and 4-10/July/2022.

[2] The data used in Sections 3.5 and 3.8.4 cannot be grouped by access method and is therefore not included in Section 4.

size	Total - 2-8/May/2022	Total - 4-10/July/2022
1	3254109	2879725
2	538015	470076
3	149607	129059
4	59681	51379
5	29134	25414
6	16699	14299
7	10294	9086
8	6895	5965
9	4735	4152
10	3474	3109

Table 3: Number of sessions with a total number of clicks in $\{1, \ldots, 10\}$ for on 2-8/May/2022 and 4-10/July/2022.

size clicks	count
405	1
364	1
320	1
307	1
291	1

Table 4: Top 5 values for the number of clicks within a session on 4-10/Jul/2022.

We recall, as explained in Section 2.1, that the dqcd table does not contain sessions without clicks on the results page. The number of sessions with a total number of clicks in $\{1, \ldots, 10\}$, for the 2 time ranges are shown in Table 3.

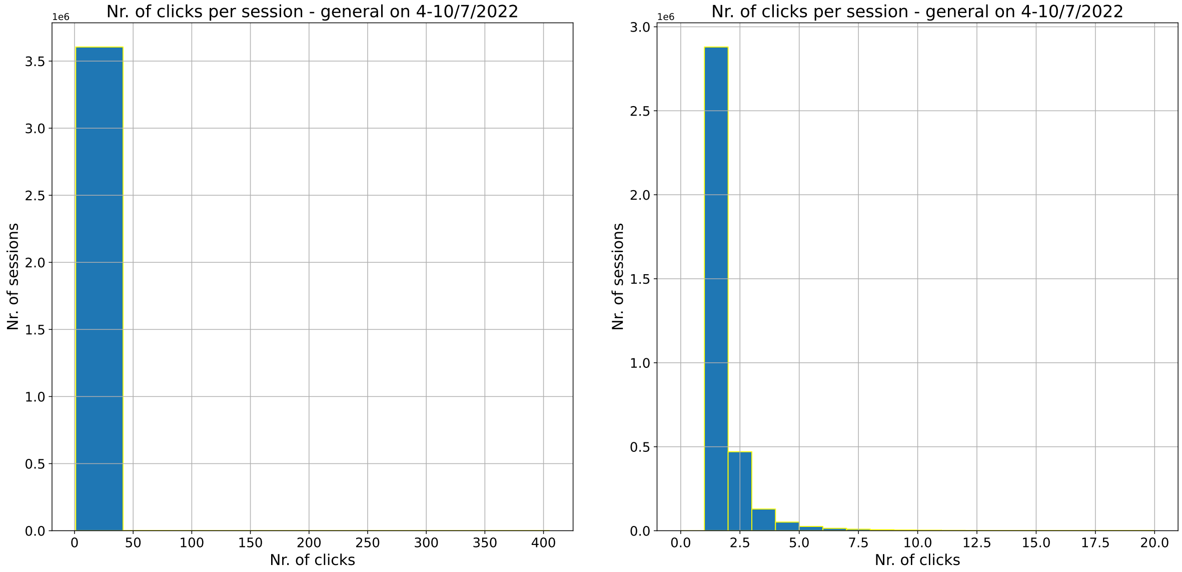
The full set of clicks per session using 10 bins (matplotlib's default value) and a zoomed-in version restricted to the interval $[0s, 20s]$ using one-second bins (bins $= (0s, 1s, \ldots, 20s)$) are plotted in the histograms displayed in Figure 1. The highest values of the distributions for 4-10/July/2022 are shown in Table 4.

Data analysis

- For both time ranges, it can be seen that the majority of sessions have a number of clicks whose size lies within $[1, 5]$.
- We can only compute total time spent for a minority of the sessions ($\sim 20\%$)
- Further question: Do people find what they want quickly or leave the platform unsatisfied?

3.3 Dwell time

Based on the clicks collected per session discussed in Section 3.2 we computed the associated *dwell times*. Dwell time on Web pages, defined as the length of time a user spends on a given



(a) Histogram for the complete data set using 10 bins.

(b) Histogram for the sessions with number of clicks in $[0, 20]$, using one-second bins.

Figure 1: Number of clicks per sessions on 4-10/Jul/2022.

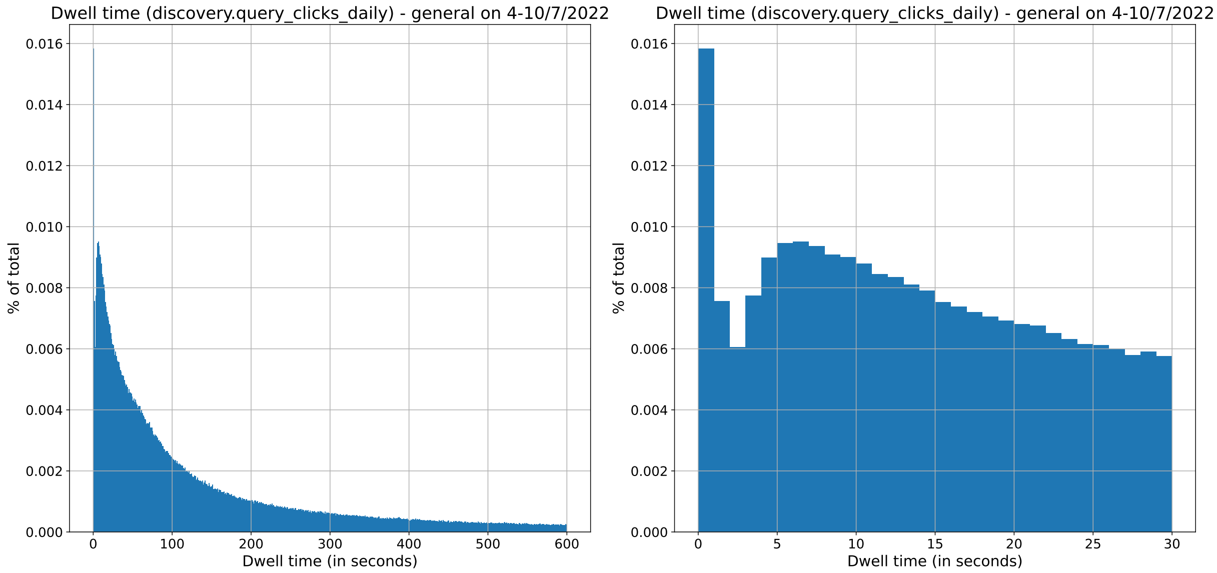
document, is considered a significant indicator of document relevance besides clickthrough [11], which is why the metric is included in this report.

The set of dwell times whose duration lies in the range $[0s, 600s]$ and the subset $[0s, 30s]$ using one-second bins (bins $= (0s, 1s, \ldots, 600s)$ and bins $= (0s, 1s, \ldots, 30s)$) respectively, are plotted in the histograms displayed in Figure 2. The total number of computed dwell times for the entire data set is $1,462,964$.

Data Analysis

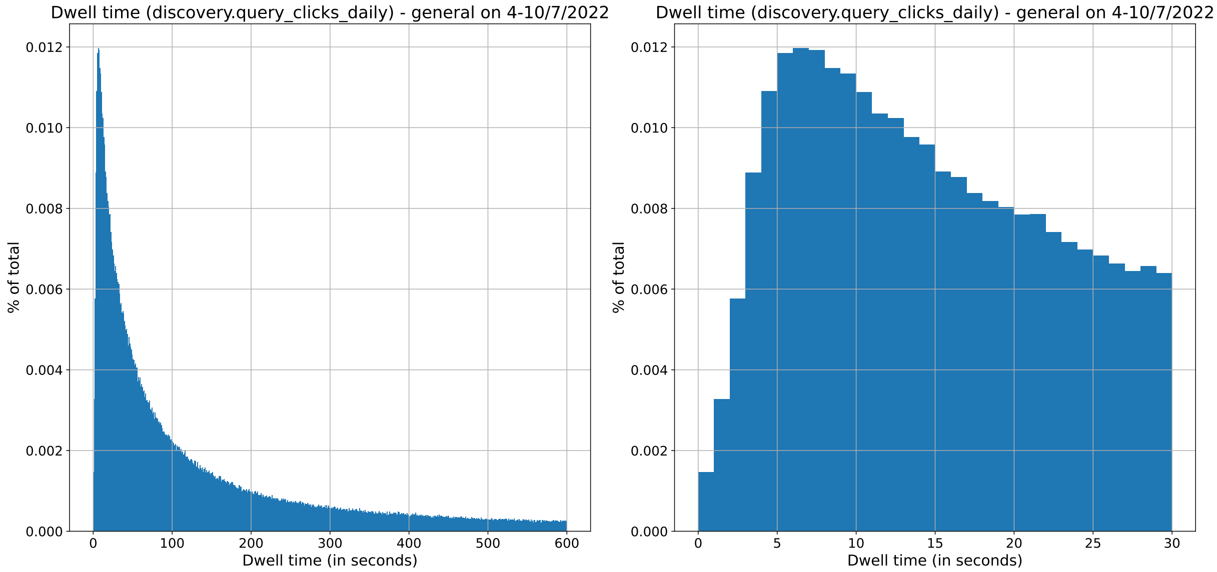
- 2 modes can be observed in the histograms, the first one is hypnotized to correspond to "accidental clicks".
- All histograms exhibit a "long-tail", meaning that although dwell time durations are concentrated in $[0s, 300s]$, several occur far from this range, i.e. far from the "head" or the central part of the distribution.

By observing the data, it was noted that multiple consecutive clicks to the same page (indicated by the "page id" attribute) and from the same referrer were being made among several sessions. In order to validate this idea, consecutive requests to the same page id and from the same referrer were filtered from the sessions. The results are shown in Figure 3. As one can see in the plots, the first mode vanishes, confirming the previously formulated hypothesis. The total number of dwell times after the filtering in the complete data set is $809,404$



(a) Subrange $[0s, 600s]$ using one second bins. (b) Subrange $[0s, 30s]$ using one second bins.

Figure 2: Histograms for dwell times for 4-10/July/2022.



(a) Subrange $[0s, 600s]$ using one second bins. (b) Subrange $[0s, 30s]$ using one second bins.

Figure 3: **Filtered** histograms for dwell times for 4-10/July/2022.

Time range	% of clicks with dwell time
2-8/May/2022	$\sim 29\%$
4-10/July/2022	$\sim 29\%$

Table 5: Percentage of clicks that "have a dwell time".

How many clicks "have a dwell time"? Now, we count the percentage of clicks to which a dwell time can be associated. We say that a click c in a session *has a dwell time* if there exists a consecutive click c' within the same session. In this case, as introduced before, the dwell time is defined as $c'.\texttt{timestamp} - c.\texttt{timestamp}$.

For answering the question, we will first count the number of clicks that **do not** have a dwell time. We know that all clicks within single-click sessions do not have a dwell time and also that for every other session, the last click of it does not have it as well. That is, for every session with a non-empty set of clicks there is one click without a dwell time. Thus,

$$\texttt{clicks\_with\_no\_dt} = \text{nr. of sessions} \; - \; \text{nr. of sessions with } 0 \text{ clicks}$$

We proceed to compute the desired quantities for the 2 time ranges using this formula. We recall that, as explained before, there are no sessions with 0 clicks.

For the time range **2-8/May/2022** we have:

- Total number of sessions is $4,087,262$.
- Total number of clicks is $5,751,214$.

Thus, the percentage of clicks with no dwell time is $\sim 71\%$, which implies that $\sim 29\%$ of clicks have a dwell time.

For the time range **4-10/July/2022** we have:

- Total number of sessions is $3,605,793$.
- Total number of clicks is $5,068,757$.

Thus, the percentage of clicks with no dwell time is $\sim 71\%$, which implies that $\sim 29\%$ of clicks have a dwell time. The results are summarized in Table 5.

Data Analysis

For both time ranges, a dwell time can be computed for a minority of the clicks ($\sim 29\%$).

3.4 Average page length per dwell time bin

An additional element of interest is to evaluate the relation between the duration of dwell times and the size of the pages where these times were measured. We recall that given 2 consecutive clicks c and c' on search results within a user session, the corresponding dwell time is defined as $\texttt{dw}(c, c') = c'.\texttt{timestamp} - c.\texttt{timestamp}$. In this case, we further define the associated page length as follows

$$\texttt{page\_length}(\texttt{dw}(c, c')) = \texttt{size}(\text{page pointed by } c)$$

Implementation	Total
Without page lengths	1,462,964
With page lengths	1,441,344

Table 6: Validating the total number of dwell times obtained.

were the size of the page is measured in Bytes. In particular, in our implementation we considered the page indicated by c.`wiki_db` and c.`pageid` and obtained its size[3] by querying the `wmf.mediawiki_wikitext_current` table[4] (given by the attribute `revision_text_bytes`). As an implementation note, we observe that in order to obtain a unique value for the size, one needs to select a particular *revision* (identified by the `revision_id` attribute) and *snapshot* (identified by the `snapshot` attribute). The time range as well as the complete set of data sources used are shown next.

Data sources and time ranges considered

- Data sources:
  - To obtain clicks: `emcr INNER JOIN dqcd` on

    `emcr.search_id = dqcd.request_set_token`

  - To obtain page lengths: `wmf.mediawiki_wikitext_current`
- Time ranges: 4-10/July/2022.

After computing page lengths, we compared the total number of dwell times obtained against the ones shown previously in Section 4.4. The results are shown in Table 6, where one can see that the order of magnitudes coincide and there is a difference of $21,620$ tuples. As future work, it would be interesting to inspect the reasons behind this difference. In the next section, the differences per access method are also provided, which can be incorporated into the analysis.

The results obtained are shown in Figure 4 (subrange $[0s, 600s]$) and Figure 5 (subrange $[0s, 100s]$). The mean value plotted on the graph corresponds to the mean of among all pages considered (i.e., not restricted to the interval plotted on the charts).

[3] In this report we use the terms "size" and "length" of a page interchangeably.

[4] Table schema available at this webpage.

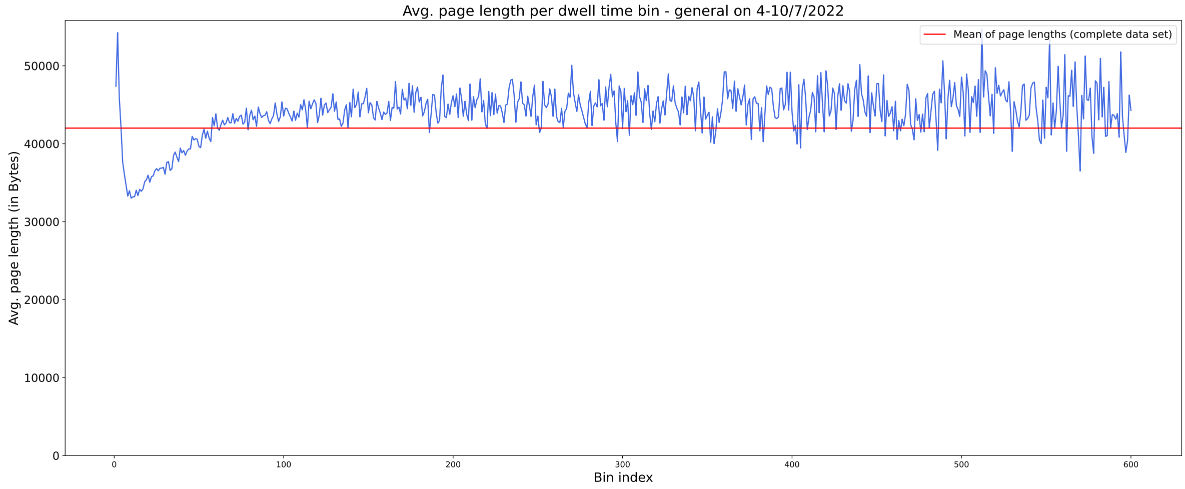


Figure 4: Average page length per dwell time bin (subrange $[0s, 600s]$) on 4-10/Jul/2022.

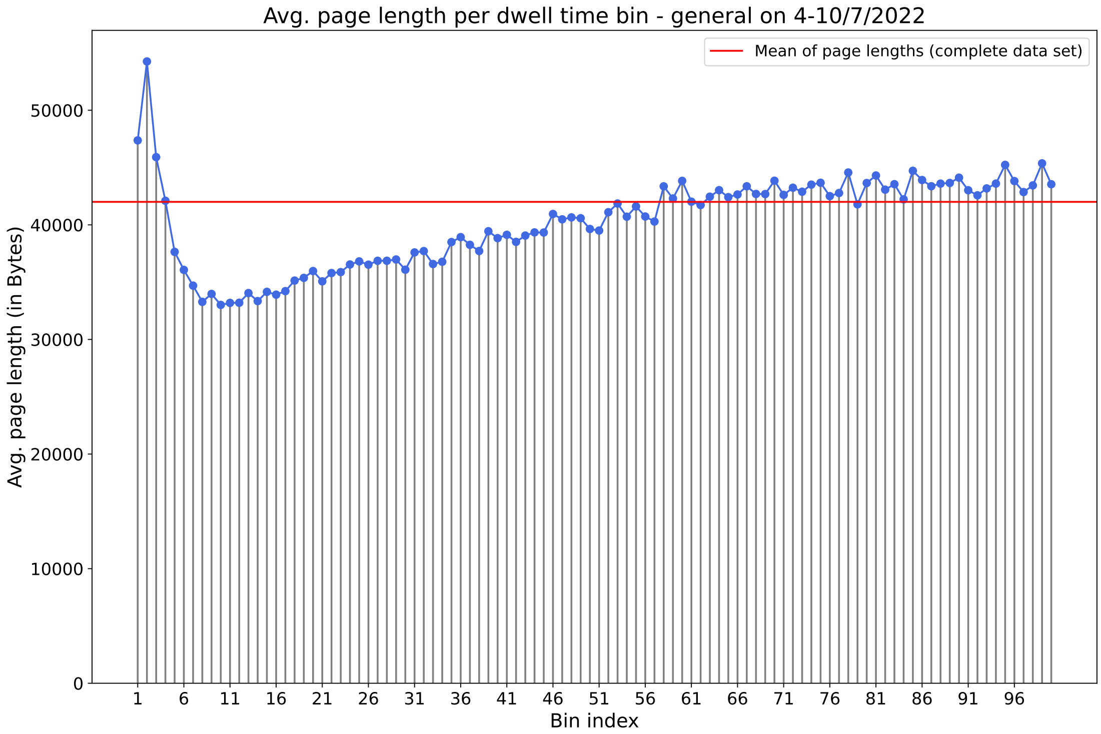


Figure 5: Average page length per dwell time bin (subrange $[0s, 100s]$) on 4-10/Jul/2022.

Data Analysis

- The page length for dwell times smaller than 10 minutes (600 seconds) oscillates and is bounded most of the times between $32,000B = 32KB$ and $50,000B = 50KB$.
- The curves peak at bin 2 and then this peak is followed by a segment, $[4, 57]$, where page length is lower than average.
- After segment $[4, 57]$, lengths in the chart lie predominantly above the mean.

3.5 Approximated dwell time based on checkin time

The schema description page [7] from `event.searchsatisfaction` (abbreviated `ess`) describes the content of the table as: *Tracks the dwell time and bounce rate of a user on pages linked from a search engine result page.*

In particular, the table contains an attribute named `checkin` whose description is (also taken from [7]): *A numeric value representing the number of seconds a user has spent on a page. The pings are at 10s, 20s, ..., 50s, 60s, 90s, 120s, 150s, 180s, 210s, 240s, 300s, 360s, 420s (7 minutes).*

Based on this attribute, we can approximate the dwell time using the last `checkin` value per session (each session is identified with the field `searchSessionId` [7]).

Data sources and time range considered

- Data source: `event.searchsatisfaction.checkin`
- Data quality: `event.searchsatisfaction.checkin` only has records for the year 2022.
- Time range: 2022.

The results for English Wikipedia are shown in Figure 6.

In order to compare these values against the dwell times computed based on `dqcd`, a possible preprocessing strategy is the following:

- Interpolate intermediate bins, so that all have the same length (e.g., based on 60 and 90 generate 60,70,80,90).

3.6 Average ranking position clicked on

We continue our study by computing the average ranking position clicked on the results page by the users. We compute, for each session, the average ranking position clicked on by the user within the session and then average all results to obtain the desired output. We believe this metric to be a valuable measure, therefore we include the following recommendation before presenting the results:

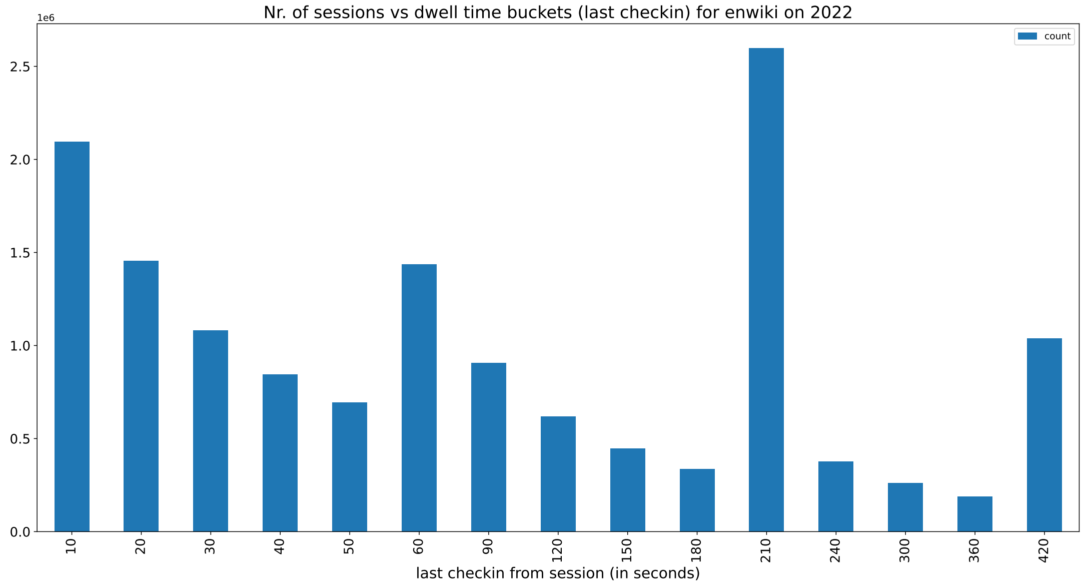


Figure 6: Number of sessions per "last checkin" buckets for English Wikipedia on 2022.

Time range	Avg. ranking pos.
2-8/May/2022	$1.90(2,395,428/4,087,262) + 2.32(1,691,834/4,087,262) = 2.07$
4-10/July/2022	$1.92(2,088,492/3,605,793) + 2.24(1,517,301/3,605,793) = 2.05$

Table 7: Average ranking position.

Recommendation for dashboard's design

Monitor *average ranking position clicked on* periodically as a measure of search quality.

The results for the 2 time ranges are shown in Table 7[5]. We note that the closer the average is to 1, the better the search quality of the results.

Data Analysis

Both results have a high level of similarity, which is reasonable, since the time ranges are not far apart in time (i.e., presumably search results are returned using the same logic in both and the type of users are also similar).

3.7 Number of words per query

Now, we compute the average number of words per query for the following data source and time ranges:

[5] These results were reconstructed based on Table 22 and Tables 13 and 14.

Time range	Avg. nr. of words per query
2-8/May/2022	$2.51(2,395,428/4,087,262) + 2.61(1,691,834/4,087,262) = 2.55$
4-10/July/2022	$2.49(2,088,492/3,605,793) + 2.54(1,517,301/3,605,793) = 2.51$

Table 8: Average nr. of words per query.

Data sources and time ranges considered

- Data source: `emcr INNER JOIN dqcd on`

  `emcr.search_id = dqcd.request_set_token`

- Time ranges: 2-8/May/2022 and 4-10/July/2022.

To implement the measure, we consider for each session, the number of words of the query[6] issued by the user and then average all results to obtain the desired output. Word counting was implemented using the following UDF (user defined function):

```
@F.udf(returnType='int')
def nr_words(query):
    return len(re.findall(r'\w+', str(query)))
```

We observe that the current regular expression being used works correctly for languages where words are separated by characters different from Unicode word characters[7]. In particular, as we will see later, there are Asian languages for which this does not hold. Future work related to this task is included in Section 6.

The results for the 2 time ranges, are shown in Table 8.

Data Analysis

Both time ranges exhibit queries that are similar in size.

3.8 Top k queries

In this subsection, we compute the top k queries for 3 different data sources. We present results for $k = 15$ and the following data sources and time ranges:

- `emcr-dqcd`[8] table - Time ranges: 4-10/Jul/2022.
- Web request logs (`wmf.webrequest` table) - Time ranges: 4-10/Jul/2022.
- Search logs (`emcr` table) - Time ranges: Apr/2022 and 4-10/Jul/2022.

[6] We consider the query to be `LOWER(query)` where query is `emcr.elasticsearch_requests[0].query`.

[7] According to the definition of `\w`, taken from https://docs.python.org/3/library/re.html.

[8] We write `emcr-dqcd` as a shorthand to `emcr INNER JOIN dqcd on emcr.search_id = dqcd.request_set_token`.

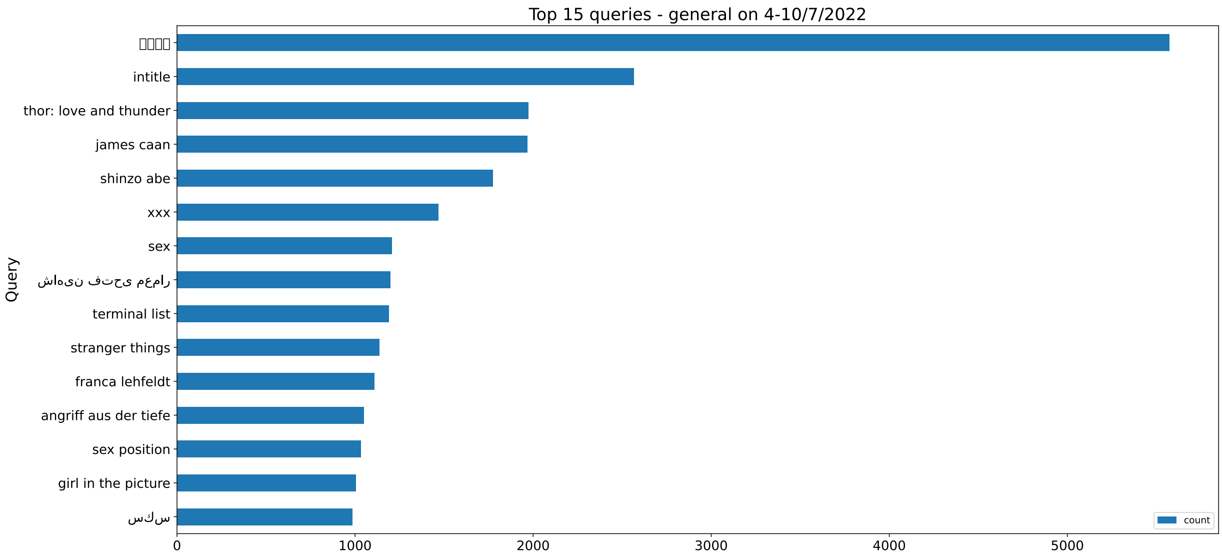


Figure 7: Top 15 queries (dqcd-emcr) for 4-10/July/2022.

The first one of these only contains full text searches. Thus, we have classified and filtered non full text queries from the web request logs to validate the results obtained from the first source. The third data source (search logs) is used to obtain the results from a source that considers all type of searches (i.e., autocomplete and full text).

3.8.1 Full text searches - Initial data source

In this subsection, we compute the top 15 queries for the emcr−dqcd table, which as mentioned in Section 2.1 contains only full text searches.

Data sources and time ranges considered

- Data source: emcr INNER JOIN dqcd on

  emcr.search_id = dqcd.request_set_token

- Time ranges: 4-10/July/2022.

The top 15 queries are shown in Figure 7. The same information presented in tabular form is displayed in Table 9. In the tabular results, when a non Latin-script alphabet was used for the queries, the translated version has been included. For these cases, both language recognition and translation has been done using Google Translate [12].

3.8.2 Classifying queries from web request logs

We start this section by describing the criteria used to distinguish query types in web request logs. This is necessary, since the dqcd table only contains full text searches and thus, for any comparison between the dqcd and web request logs to be meaningful, we need to only consider this type of queries from the second source. First, the complete set of criteria are detailed,

Query	Count
<"Shinzo Abe" in Japanese>	5571
intitle	2566
thor: love and thunder	1973
james caan	1968
shinzo abe	1772
xxx	1466
sex	1205
<"Shahin Fathi Memar" in Persian>	1199
terminal list	1190
stranger things	1135
franca lehfeldt	1107
angriff aus der tiefe	1049
sex position	1033
girl in the picture	1003
<"sex" in Arabic>	985

Table 9: Top 15 queries (`dqcd-emcr`) for 4-10/July/2022.

mainly for documentation purposes. Later, we conclude with a simplified classification rule derived from consultation with the Engineering division of the Search Platform team, which is believed to be correct most of the times.

There are 2 different query types: autocomplete and full text searches. Among the first type (autocomplete), there are 2 additional subtypes:

- From Special Page: Search results.
- From Wikipedia home page (i.e., https://www.wikipedia.org/).

We detail the general set of criteria below.

Autocomplete searches/requests (from Special Page: Search results)

- If the request includes the parameter `action=opensearch`, it is an autocomplete requests for the default mediawiki UI.
- Any request that includes the parameter `action=opensearch` must be sent to `/w/api.php`.
- While having `action=opensearch` is primarily autocomplete from the websites, it also includes anyone who sets up their browser to autocomplete from Wiki. This is believed to be a small group of advanced users.
- Other ways to issue an autocomplete request include
  - Requests to `/w/api.php` that include `generator=prefixsearch`. Difficulty: texttgenerator=prefixsearch may be used for a wide variety of activities such as looking up pages to add links in the wikitext editor.

  - Currently the default UI of MediaWiki is being updated for some languages (e.g. https://fr.wikipedia.org/). This might use different criteria, but were considered to be out of scope for this analysis.

Autocomplete search/requests (from Wikipedia home page)

- These request utilize `generator=prefixsearch` which will generate webrequest logs along with events in `event.mediawiki_cirrussearch_request`. These requests are uncached and are always registered into the backend cirrus event logging. The same difficulty applies here as in the previous case, where `generator=prefixsearch` was mentioned.

Full text search/requests

- For these cases, `uri_path` can be `/w/index.php`, but it could also be `/wiki/Special:Search` or any of the (hundreds of) translations of *Special:Search* (like *Especial:Buscar*, *Spezial:Suche*, etc.).
- While not strictly true, it's probably true in the vast majority of cases that requests to `/wiki/*` or `/w/index.php` that contain a `search=<query>` parameter are full text search requests.

Based on the last point the following heuristic is derived:

Heuristic to identify full text search from web request logs

It is assumed that in the vast majority of cases, requests to `/wiki/*` or `/w/index.php` that contain a `search=<query>` parameter are full text search requests.

The heuristic is implemented by verifying the following condition:

(`uri_path` *contains* "/wiki/" **OR** `uri_path` *contains* "/w/index.php") **AND** `uri_query` *contains* "search="

If the condition is fulfilled, the query is extracted using the following regular expression: `[\w\%+\.\-]*`.

3.8.3 Full text searches - Validation data source

In this section, after classifying the queries and filtering out non full text searches using the criteria explained in Section 3.8.2, we compute the top 15 queries to validate the ones previously obtained from the `emcr-dqcd` table.

Data sources and time ranges considered

- Data source: `wmf.webrequest` (i.e., web request logs)
- Time ranges: 4-10/July/2022.

Query	Count
""	2405661
cleopatra	463878
nasdaq	217241
netscout	216591
xxx	182449
<"Shinzo Abe" in Japanese>	166609
smartbear	161316
stackdriver	120983
+1	104876
cu	95769
android	77663
ubuntu	76122
linux++c	73967
rte	66579
thor: love and thunder	54822

Table 10: Top 15 queries (web request logs) for 4-10/July/2022.

The results are shown in Table 10 for 4-10/July/2022.

Data Analysis

Based on both set of results we have:

- The order of magnitudes differ among the 2 data sources.
- There are terms that occur in both rankings, for example:
    - for 4-10/Jul/2022 the term "xxx" was searched $182,449$ times in web request logs but only $1,466$ in `dqcd-emcr`.

3.8.4 All query types - Top k queries from search logs

In this subsection, we present the top 15 queries based on the `emcr` table, for the following 2 time ranges: April/2022 and 4-10/Jul/2022.

Data sources and time range considered

- Data sources: Search logs (`emcr` table).
- Time ranges: Apr/2022 and 4-10/Jul/2022.

No filters are applied and the only preprocessing consists of lower-casing the query strings. This dataset contains both autocomplete and full text searches. Due to the impossibility of recovering the host of the URI from this table (explained in Section 4.1), the queries cannot be segmented by access method. The results are presented in Figure 8 (Apr/2022) and 9

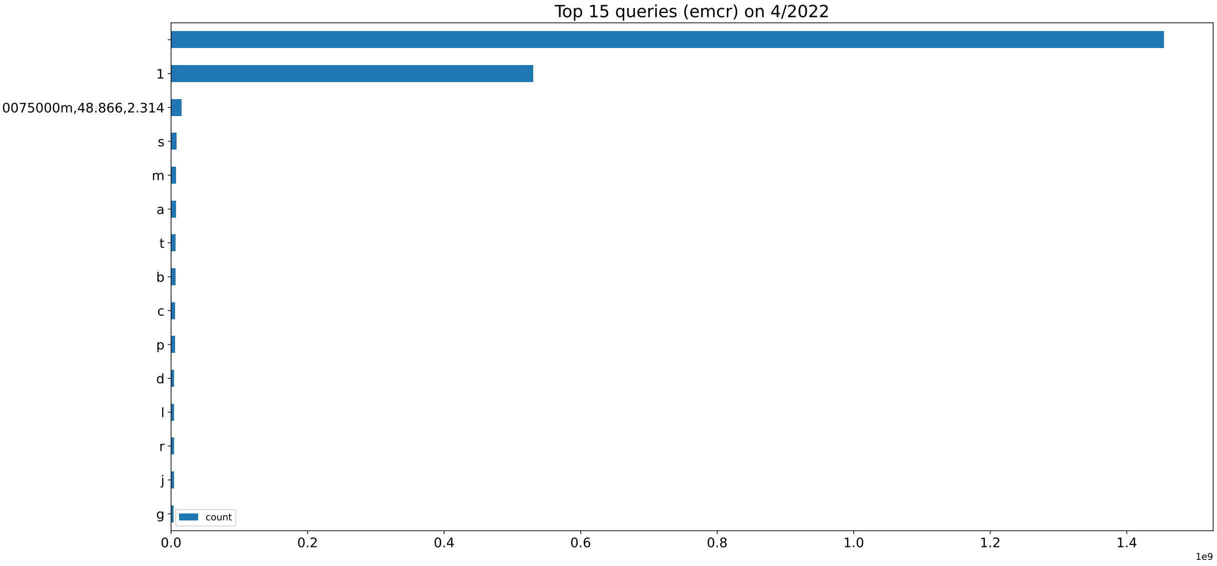


Figure 8: Top 15 queries (emcr) on Apr/2022.

(4-10/Jul/2022), as well as in Table 11 and 12 respectively.

Data Analysis

For both time ranges we have:

- The most issued query is the empty query, followed by single character queries (except the top 3 query, which is assumed to be bot-generated and coincides for both rankings). Intuitively this makes sense, given that among autocomplete searches which are assumed to be the most frequent ones, shorter queries are more likely to be issued by users.

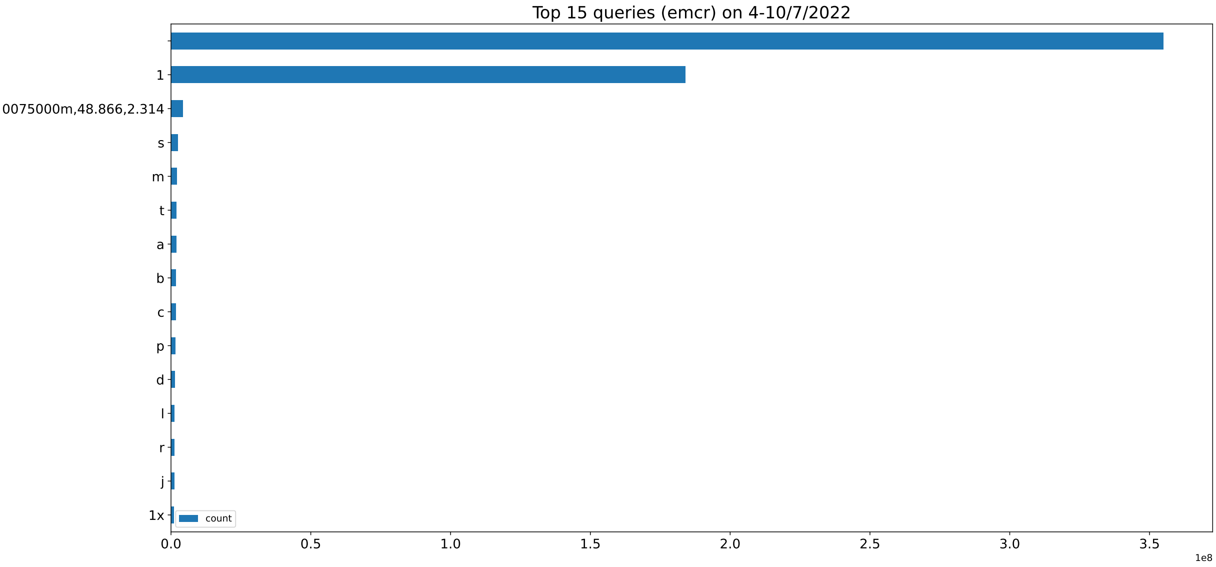


Figure 9: Top 15 queries (emcr) on 4-10/Jul/2022.

Query	Count
""	1453865228
1	529797110
nearcoord:40075000m,48.866,2.314	14953846
s	7968378
m	7002753
a	6907839
t	6200890
b	6016031
c	5818312
p	5236276
d	4342521
l	4230603
r	4096963
j	3906636
g	3625637

Table 11: Top 15 queries (emcr) on Apr/2022.

Query	Count
""	354838676
1	183866985
nearcoord:40075000m,48.866,2.314	4184178
s	2442179
m	1970845
t	1957627
a	1946391
b	1767523
c	1698966
p	1465146
d	1264804
l	1228905
r	1197580
j	1146450
1x	1068828

Table 12: Top 15 queries (emcr) on 4-10/Jul/2022.

4 Search behavior based on client type

This section details the results obtained as part of the analysis when aggregating the results by client (i.e. platform) type.

We start by introducing the method used to categorize the different client types. The client type characterization used is the one given in the webrequest logs' description page [4] by the access_method field, described next[9]:

Client type used to access the site, which can be one of the following: desktop, mobile web or mobile app. The categorization is made using the following criteria:

- Mobile app requests are identified by the user agent including WikipediaApp or Wikipedia/5.0.
- Mobile web requests are identified by the hostname containing a subdomain of m, zero, wap or mobile.
- Any other request is classified as desktop.

This is implemented in the getAccessMethod function in refinery-source [13] code repository. Mobile app access method was excluded from the analysis, since its order of magnitude was significantly smaller than the one from the remaining 2 access methods. When one of the data sources of interest did not natively include the access_method field, it was reconstructed using the getAccessMethod function previously mentioned. We describe how this was implemented for the data sources used in the analysis next.

[9] Also taken from [4].

4.1 Access method computation

The access_method field was reconstructed using the following data sources:

- search logs, i.e. event.mediawiki_cirrussearch_request (emcr)
- discovery.query_clicks_daily (dqcd)

The implementation used for this purpose is detailed next. First, in order to use getAccessMethod function mentioned before, one needs to import it. This can be done by executing the following commands:

```
spark.sql("ADD JAR /srv/deployment/analytics/\
    refinery/artifacts/refinery-hive-shaded.jar").show()
spark.sql("CREATE TEMPORARY FUNCTION get_access_method as\
    'org.wikimedia.analytics.refinery.hive.GetAccessMethodUDF'")
```

Next, looking at the signature of the function we need 2 parameters: the user agent of the request and the host of the URI. By joining the dqcd table with emcr on emcr.search_id = dqcd.request_set_token one can get the user agent from

http.request_headers["user-agent"].

Regarding the host of the URI, since it is not possible to recover it from emcr, it was extracted from the URI given by clicks[0].referer using the following regular expression

:[www.]?([a-zA-Z0-9.]+

We continue with the computed measures and the corresponding analysis, as was done in the previous section.

4.2 Number of sessions by access method

The total number of sessions registered in the given time ranges, grouped by access_method are given in Table 13 and 14.

Data sources and time ranges considered

- Data source: emcr INNER JOIN dqcd on

  emcr.search_id = dqcd.request_set_token

- Time ranges: 2-8/May/2022 and 4-10/July/2022.

Data analysis

- For both time ranges, there is a majority of users that access the desktop version of the sites.
- Further question: Is this associated with the type of search being considered (full text) or the type of content exposed on Wikipedia?

desktop	mobile web	Total
2,395,428 ($\sim 59\%$)	1,691,834 ($\sim 41\%$)	4,087,262

Table 13: Total number of sessions grouped by `access_method` on 2-8/May/2022.

desktop	mobile web	Total
2,088,492 ($\sim 58\%$)	1,517,301 ($\sim 42\%$)	3,605,793

Table 14: Total number of sessions grouped by `access_method` on 4-10/July/2022.

4.3 Number of clicks per session

As explained in Section 2.1, we recall that the `dqcd` table does not contain sessions without clicks on the results page. The number of sessions with a set of clicks whose size lies the interval $\{1, \dots, 10\}$ and the 2 time ranges, grouped by `access_method` are shown in Table 15 and 16.

Data sources and time ranges considered

- Data source: `emcr INNER JOIN dqcd` on

  `emcr.search_id = dqcd.request_set_token`

- Time ranges: 2-8/May/2022 and 4-10/July/2022.

Based on this data, we have

Data analysis

- The majority of sessions with 1 click are desktop sessions.
- The difference evens out when considering the number of clicks in $\{2, \dots, 10\}$.

size	desktop (count)	mobile web (count)	desktop (%)	mobile web (%)
1	1983754	1270355	60.96	39.04
2	267639	270376	49.75	50.25
3	72658	76949	48.57	51.43
4	29086	30595	48.74	51.26
5	14158	14976	48.6	51.4
6	8377	8322	50.16	49.84
7	5272	5022	51.21	48.79
8	3547	3348	51.44	48.56
9	2271	2464	47.96	52.04
10	1650	1824	47.5	52.5

Table 15: Number of clicks per session grouped by `access_method` on 2-8/May/2022.

size	desktop (count)	mobile web (count)	desktop (%)	mobile web (%)
1	1735380	1144345	60.26	39.74
2	229996	240080	48.93	51.07
3	61938	67121	47.99	52.01
4	24643	26736	47.96	52.04
5	12075	13339	47.51	52.49
6	6986	7313	48.86	51.14
7	4349	4737	47.86	52.14
8	2850	3115	47.78	52.22
9	1996	2156	48.07	51.93
10	1478	1631	47.54	52.46

Table 16: Number of clicks per session grouped by access_method for 4-10/July/2022.

size clicks	count
405	1
291	1
276	1
274	1
268	1

(a) Desktop access method

size clicks	count
364	1
320	1
307	1
184	1
179	1

(b) Mobile web access method

Table 17: Top 5 values for the number of clicks within a session for the 2 access methods being considered.

The highest values of the distributions for 4-10/July/2022 are shown in Table 17a (desktop) and 17b (mobile web).

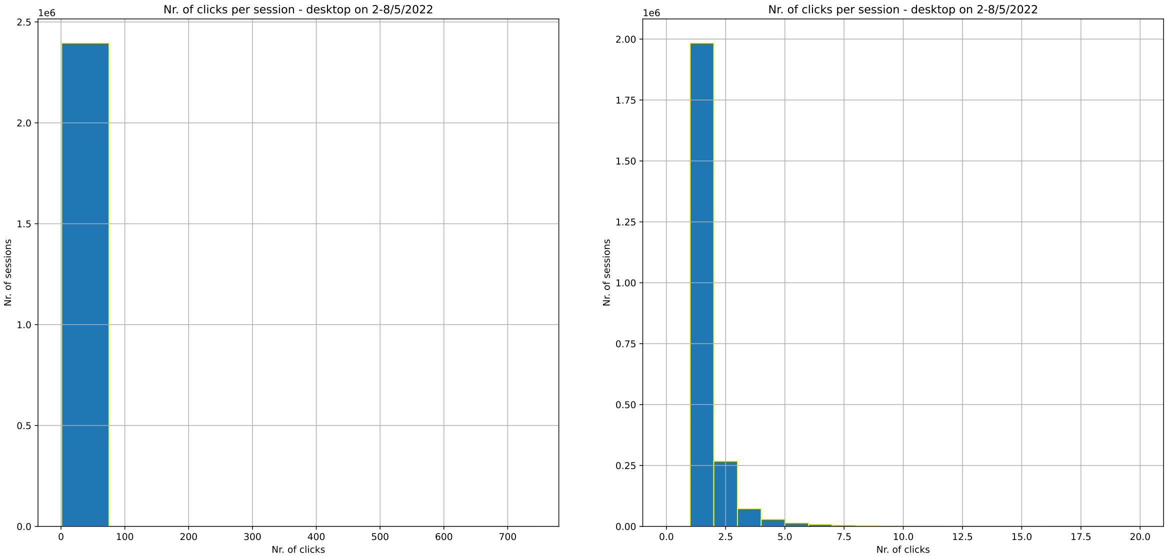
The full set of clicks per session using 10 bins (matplotlib's default value) and a zoomed-in version restricted to the interval $[0s, 20s]$ using one-second bins (bins $= (0s, 1s, \ldots, 20s)$) are plotted in the histograms displayed in the following figures:

- Time range: 2-8/May/2022 and access_method $=$ desktop, Figure 10a and 10b.
- Time range: 2-8/May/2022 and access_method $=$ mobile web, Figure 11a and 11b.
- Time range: 4-10/July/2022 and access_method $=$ desktop, Figure 12a and 12b.
- Time range: 4-10/July/2022 and access_method $=$ mobile web, Figure 13a and 12b.

Data analysis

In all cases, it can be seen that the majority of sessions have a number of clicks whose size lies within $[1, 5]$.

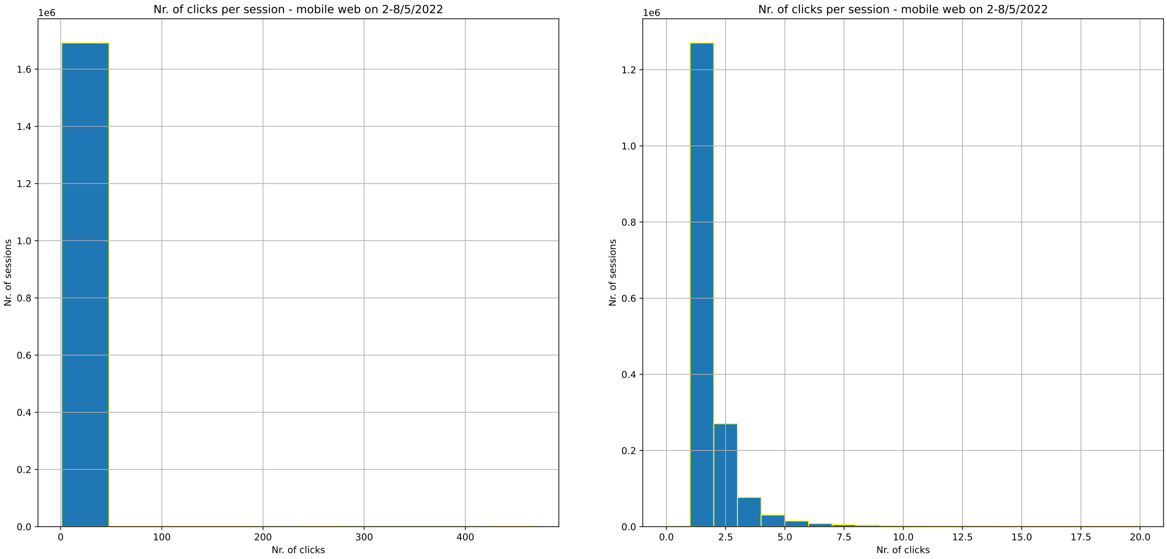
The data sets for 4-10/July/2022 have also been plotted using a logarithmic scale, displayed in Figure 14 (desktop) and 15 (mobile web).



(a) Histogram for the complete data set using 10 bins.

(b) Histogram for the sessions with number of clicks in $[0, 20]$, using one-second bins.

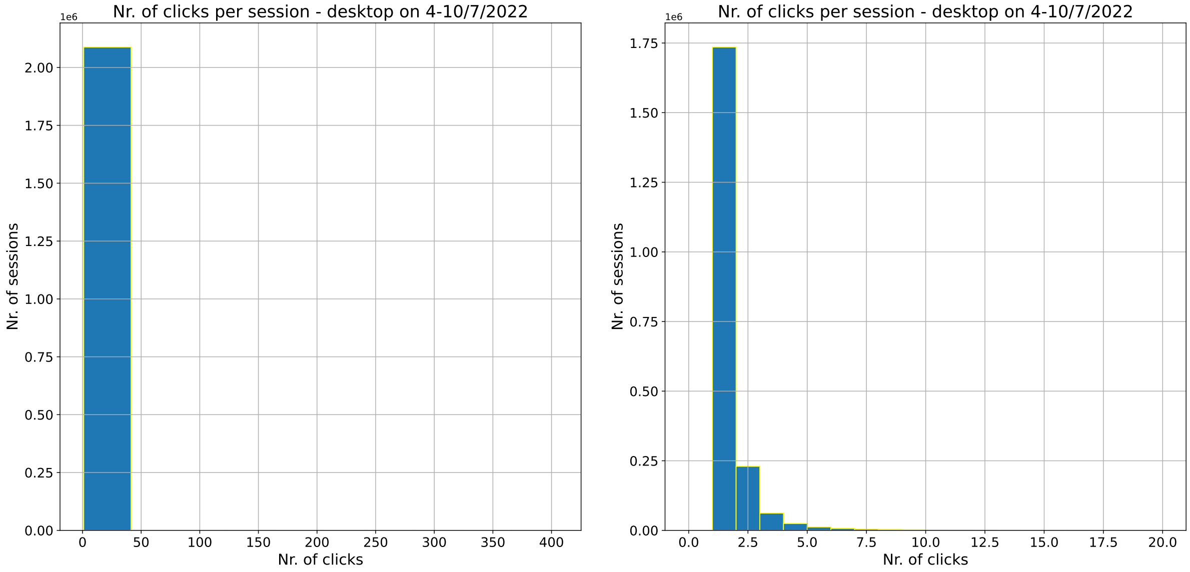
Figure 10: Number of clicks per desktop sessions for 2-8/May/2022.



(a) Histogram for the complete data set using 10 bins.

(b) Histogram for the sessions with number of clicks in $[0, 20]$, using one-second bins.

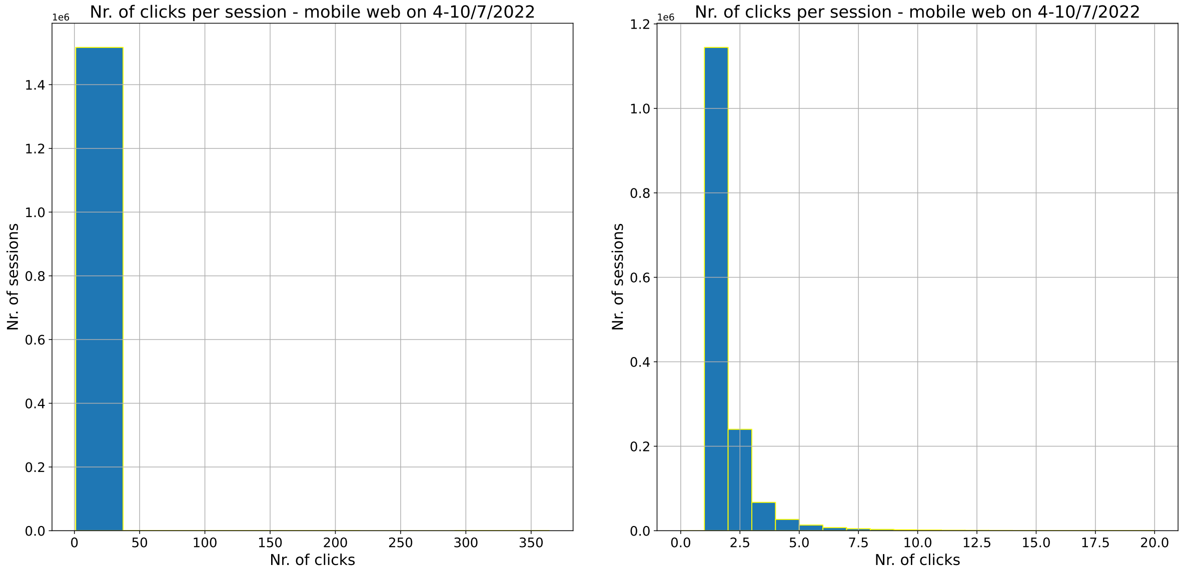
Figure 11: Number of clicks per mobile web sessions for 2-8/May/2022.



(a) Histogram for the complete data set using 10 bins.

(b) Histogram for the sessions with number of clicks in $[0, 20]$, using one-second bins.

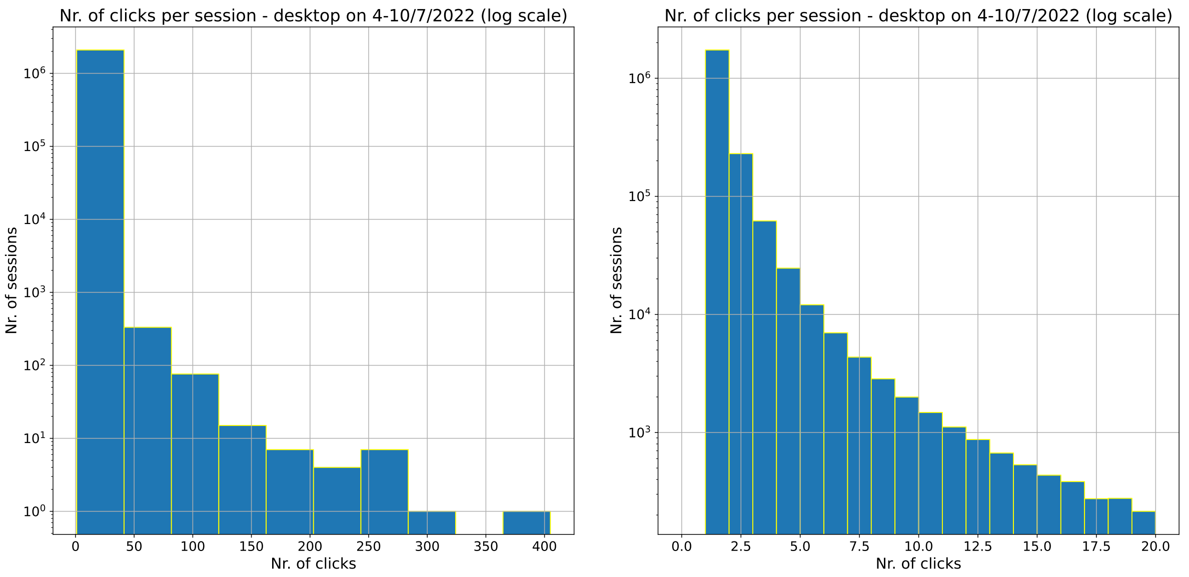
Figure 12: Number of clicks per desktop sessions for 4-10/July/2022.



(a) Histogram for the complete data set using 10 bins.

(b) Histogram for the sessions with number of clicks in $[0, 20]$, using one-second bins.

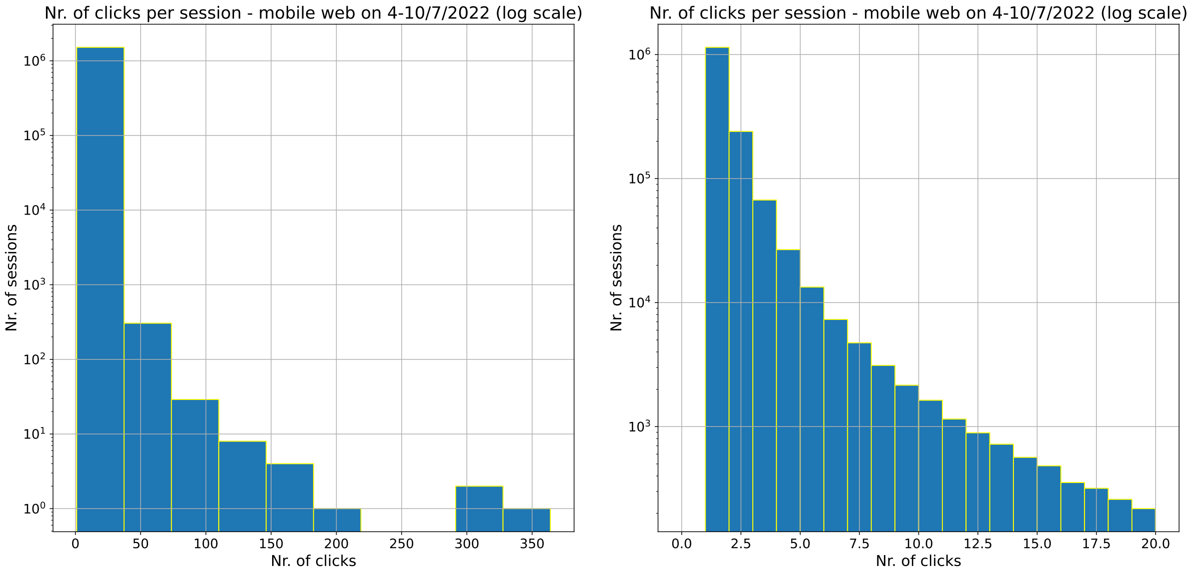
Figure 13: Number of clicks per mobile web sessions for 4-10/July/2022.



(a) Histogram for the complete data set using 10 bins.

(b) Histogram for the sessions with number of clicks in $[0, 20]$, using one-second bins.

Figure 14: Number of clicks per desktop sessions for 4-10/July/2022 (log. scale).



(a) Histogram for the complete data set using 10 bins.

(b) Histogram for the sessions with number of clicks in $[0, 20]$, using one-second bins.

Figure 15: Number of clicks per mobile web sessions for 4-10/July/2022 (log. scale).

desktop	mobile web	Total
829,494 ($\sim 50\%$)	834,458 ($\sim 50\%$)	1,663,952

(a) Time range 2-8/May/2022

desktop	mobile web	Total
719,425 ($\sim 49\%$)	743,539 ($\sim 51\%$)	1,462,964

(b) Time range 4-10/July/2022

Table 18: Total number of dwell times grouped by `access_method`.

4.4 Dwell time

As was done in Section 3.3, this time based on the clicks collected per session shown in Section 4.3, we computed the associated *dwell times*.

The percentage of the dwell times considered by access method is displayed in Table 18a (2-8/May/2022) and 18b (4-10/July/2022).

The set of dwell times whose duration lies in the range $[0s, 600s]$ and the subset $[0s, 30s]$ using one-second bins (bins $= (0s, 1s, \ldots, 600s)$ and bins $= (0s, 1s, \ldots, 30s)$) respectively, are plotted in the histograms displayed in the following figures:

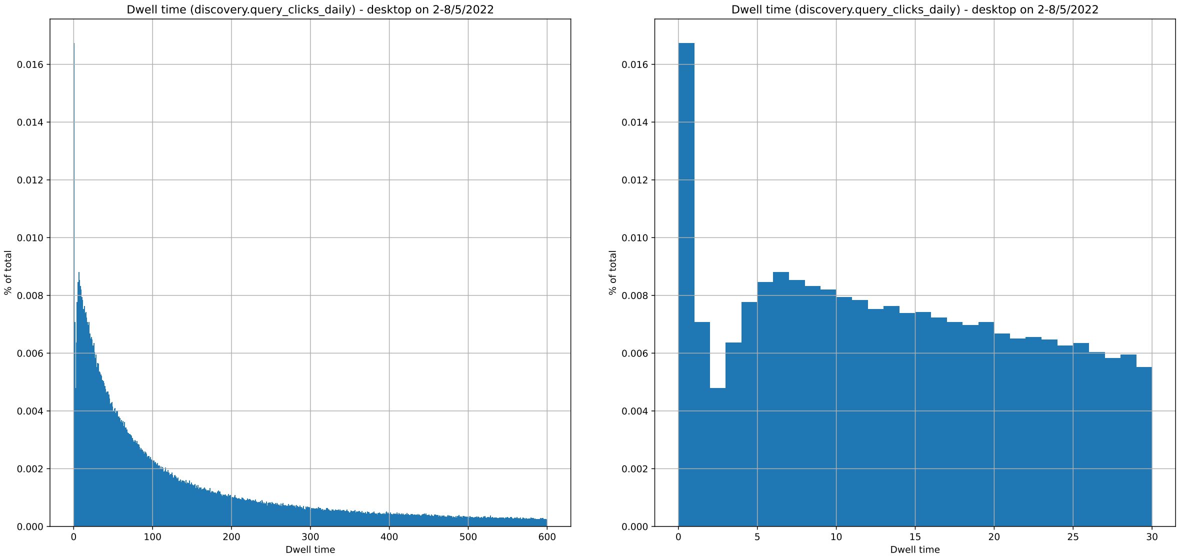
- Time range: 2-8/May/2022 and `access_method` = desktop, Figure 16a and 16b.
- Time range: 2-8/May/2022 and `access_method` = mobile web, Figure 17a and 17b.
- Time range: 4-10/July/2022 and `access_method` = desktop, Figure 18a and 18b.
- Time range: 4-10/July/2022 and `access_method` = mobile web, Figure 19a and 19b.

Data Analysis

The same characteristics as the ones observed in Section 3.3 apply for these cases, namely:

- 2 modes can be observed in the histograms, the first one is hypnotized to correspond to "accidental clicks".
- All histograms exhibit a "long-tail", meaning that although dwell time durations are concentrated in $[0s, 300s]$, several occur far from this range, i.e. far from the "head" or the central part of the distribution.

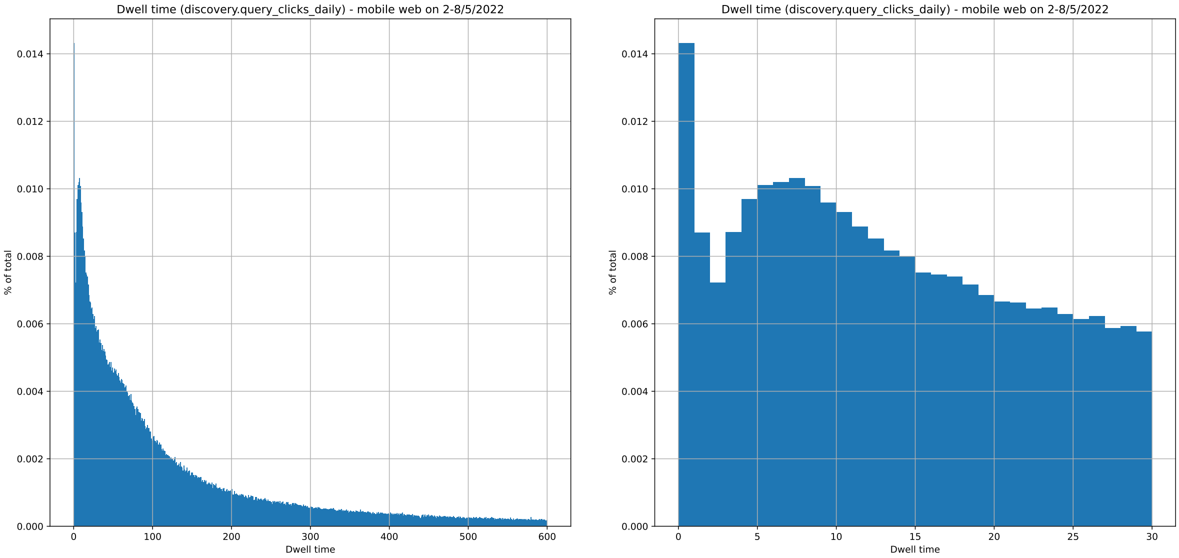
As with the plots shown in Section 3.3, it was noted that multiple consecutive clicks to the same page (page id) and from the same referrer were being made among several sessions. The same filtering was applied, yielding the results shown in Figures 20 and 21. Again, as one can see in the plots, the first mode vanishes, confirming that the previously formulated hypothesis is in fact independent from the access method used.



(a) Subrange $[0s, 600s]$ using one second bins.

(b) Subrange $[0s, 30s]$ using one second bins.

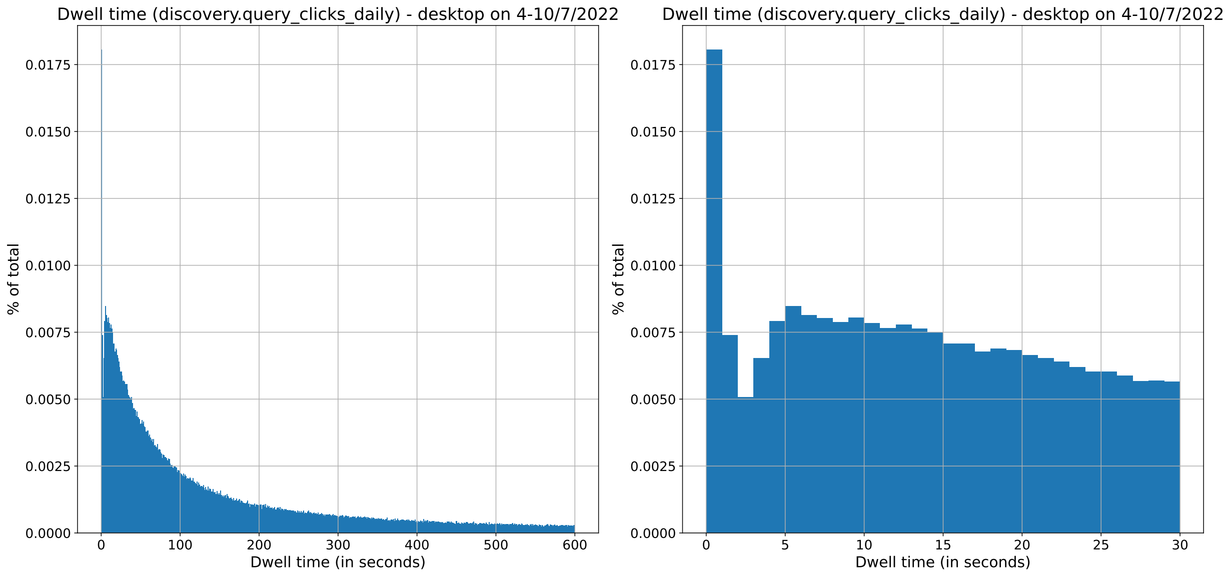
Figure 16: Histograms for dwell times - desktop sessions for 2-8/May/2022.



(a) Subrange $[0s, 600s]$ using one second bins.

(b) Subrange $[0s, 30s]$ using one second bins.

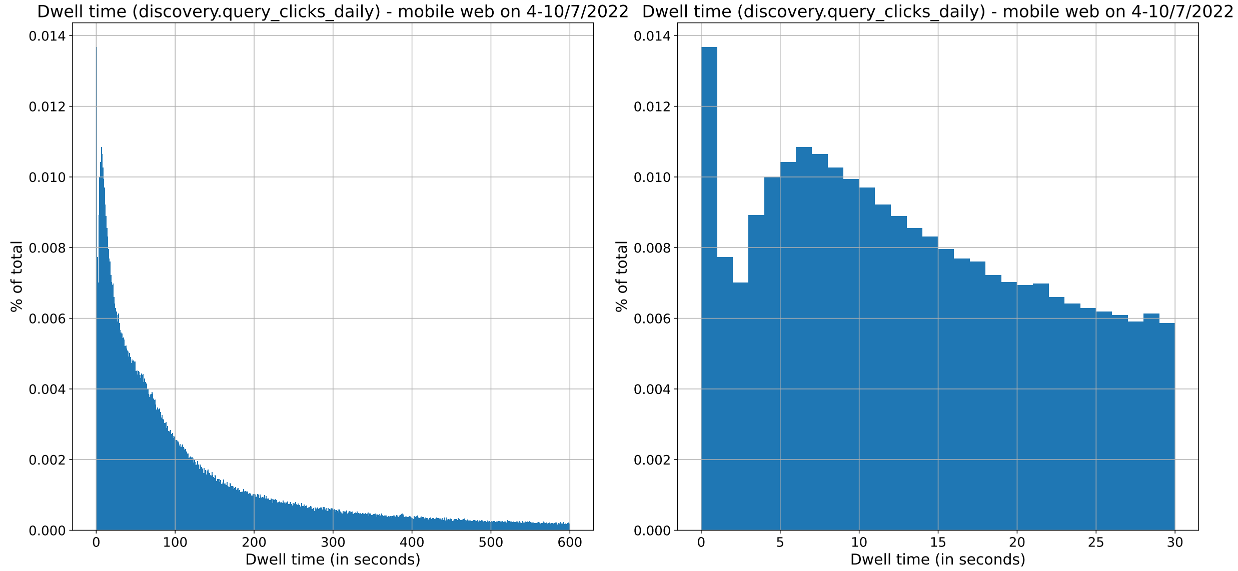
Figure 17: Histograms for dwell times - mobile web sessions for 2-8/May/2022.



(a) Subrange $[0s, 600s]$ using one second bins.

(b) Subrange $[0s, 30s]$ using one second bins.

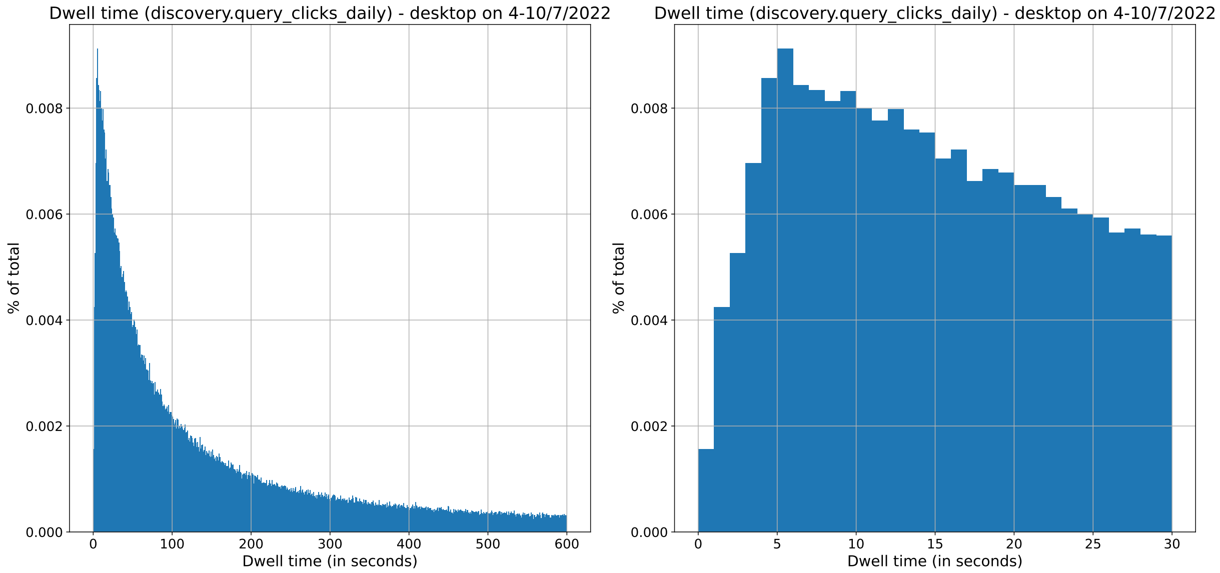
Figure 18: Histograms for dwell times - desktop sessions for 4-10/July/2022.



(a) Subrange $[0s, 600s]$ using one second bins.

(b) Subrange $[0s, 600s]$ using one second bins.

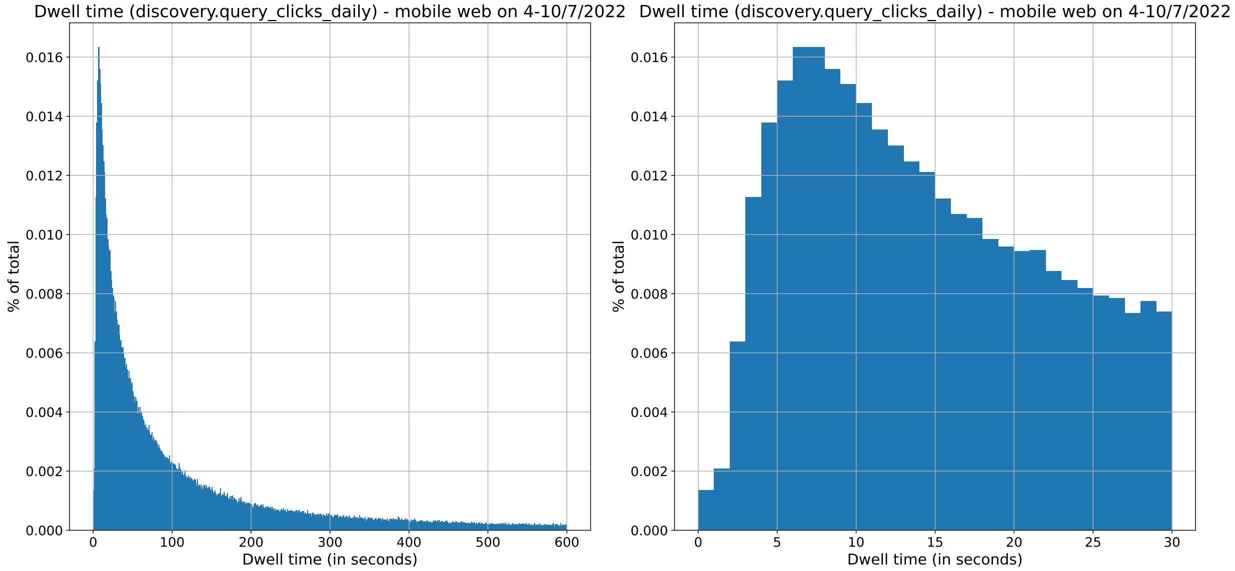
Figure 19: Histograms for dwell times - mobile web sessions for 4-10/July/2022.



(a) Subrange $[0s, 600s]$ using one second bins.

(b) Subrange $[0s, 30s]$ using one second bins.

Figure 20: **Filtered** histograms for dwell times - desktop sessions for 4-10/July/2022.



(a) Subrange $[0s, 600s]$ using one second bins.

(b) Subrange $[0s, 600s]$ using one second bins.

Figure 21: **Filtered** histograms for dwell times - mobile web sessions for 4-10/July/2022.

How many clicks "have a dwell time"? As done in Section 3.3, we now count the percentage of clicks to which a dwell time can be associated. Analogously as before, we will make use of

$$\texttt{clicks\_with\_no\_dt} = \text{nr. of sessions} - \text{nr. of sessions with } 0 \text{ clicks}$$

We proceed to compute the desired quantities for the 2 time ranges and the 2 access methods using this formula. We recall that, as noted before, there are no sessions with 0 clicks.

For the time range **2-8/May/2022** and **desktop** access method we have:

- Total number of "desktop" sessions is $2,395,428$.
- Total number of "desktop" clicks is $3,224,922$.

Thus, the percentage of "desktop" clicks with no dwell time is $\sim 74\%$, which implies that $\sim 26\%$ of "desktop" clicks have a dwell time.

For the time range **2-8/May/2022** and **mobile web** access method we have:

- Total number of "mobile web" sessions is $1,691,834$.
- Total number of "mobile web" clicks is $2,526,292$.

Thus, the percentage of "mobile web" clicks with no dwell time is $\sim 67\%$, which implies that $\sim 33\%$ of "mobile web" clicks have a dwell time.

For the time range **4-10/July/2022** and **desktop** access method we have:

- Total number of "desktop" sessions is $2,088,492$.
- Total number of "desktop" clicks is $2,807,917$.

Thus, the percentage of "desktop" clicks with no dwell time is $\sim 74\%$, which implies that $\sim 26\%$ of "desktop" clicks have a dwell time.

For the time range **4-10/July/2022** and **mobile web** access method we have:

- Total number of "mobile web" sessions is $1,517,301$.
- Total number of "mobile web" clicks is $2,260,840$.

Thus, the percentage of "mobile web" clicks with no dwell time is $\sim 67\%$, which implies that $\sim 33\%$ of "mobile web" clicks have a dwell time.

The results are summarized in Table 19.

Data Analysis

As before, for both time ranges, a dwell time can be computed for a minority of the clicks (between $26 - 30\%$). When segmenting by access method, more "mobile web" clicks have a dwell time compared to "desktop" clicks (6% difference).

Time range	Access method	% of clicks with dwell time
2-8/May/2022	desktop	$\sim 26\%$
2-8/May/2022	mobile web	$\sim 33\%$
4-10/July/2022	desktop	$\sim 26\%$
4-10/July/2022	mobile web	$\sim 33\%$

Table 19: Percentage of clicks that "have a dwell time" by access method.

Implementation	Desktop	Mobile Web	Total
Without page lengths	719,425	743,539	1,462,964
With page lengths	703,648	737,696	1,441,344

Table 20: Validating the total number of dwell times obtained.

4.5 Average page length per dwell time bin

In this section we group the results presented in Section 3.4 by access method. We recall that given 2 consecutive clicks c and c' on search results within a user session, the associated dwell time is defined as $\texttt{dw}(c, c') = c'.\texttt{timestamp} - c.\texttt{timestamp}$. In this case, we define the associated page length as follows

$$\texttt{page\_length}(\texttt{dw}(c, c')) = \texttt{size}(\text{page pointed by } c)$$

were the size of the page is measured in Bytes. We consider the same data sources and time range as in Section 3.4:

Data sources and time ranges considered

- Data sources:
  - To obtain clicks: `emcr INNER JOIN dqcd on`

    `emcr.search_id = dqcd.request_set_token`

  - To obtain page length: `wmf.mediawiki_wikitext_current`
- Time ranges: 4-10/July/2022.

Following what was done in Section 3.4, after computing page lengths, we compared the total number of dwell times obtained against the ones shown previously in Section 4.4. The results are displayed in Table 20. This time the counts per access method are also provided.

The results obtained for desktop access method are shown in Figure 22 (subrange $[0s, 600s]$), Figure 23 (subrange $[0s, 100s]$) and Figure 24 (subrange $[0s, 30s]$), while the case of mobile web access for the same subranges on Figures 25, 26 and 27 respectively. We recall that the mean value plotted on both graphs corresponds to the mean of the total of considered pages for the given access method (i.e., not restricted to the interval plotted on the charts).

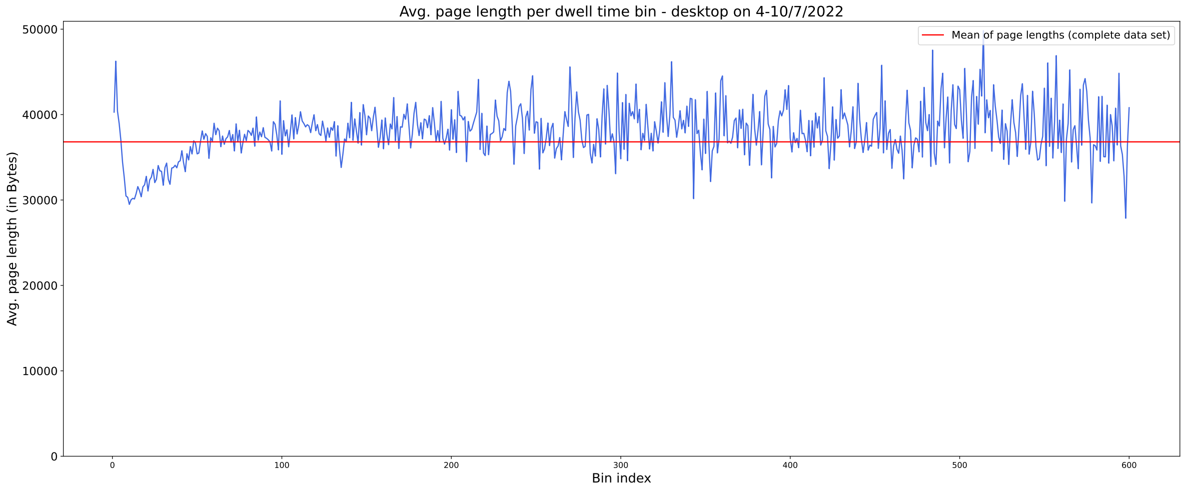


Figure 22: Average page length per dwell time bin (subrange $[0s, 600s]$) for desktop access method on 4-10/Jul/2022.

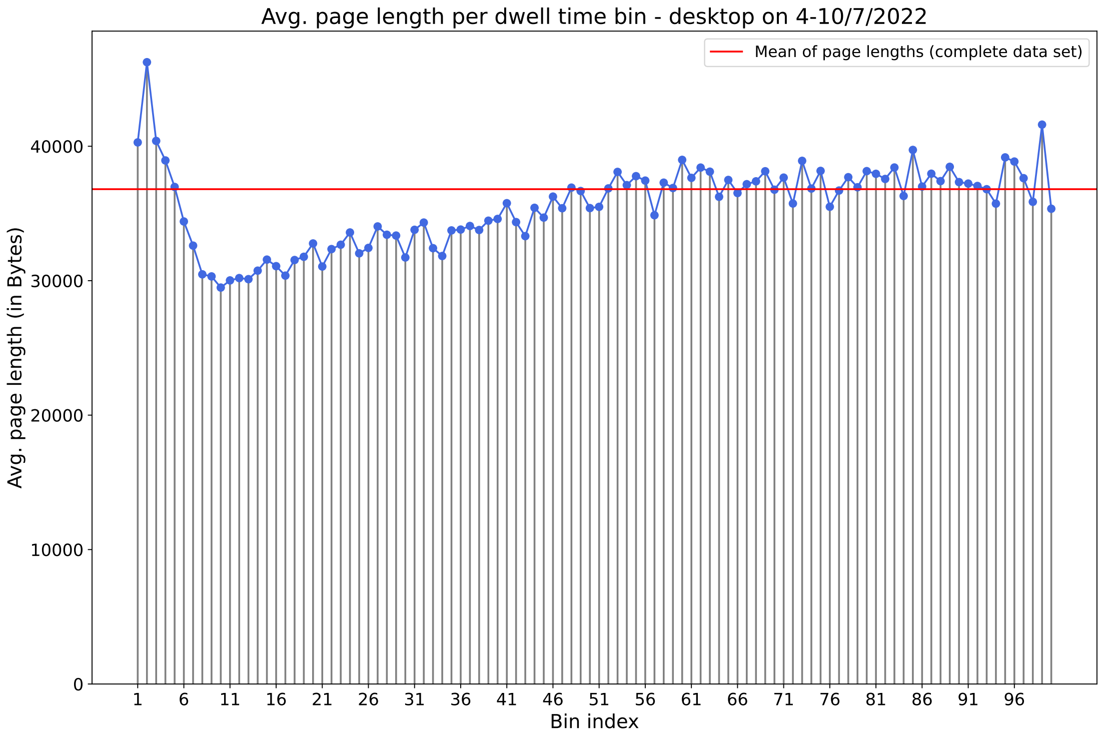


Figure 23: Average page length per dwell time bin (subrange $[0s, 100s]$) for desktop access method on 4-10/Jul/2022.

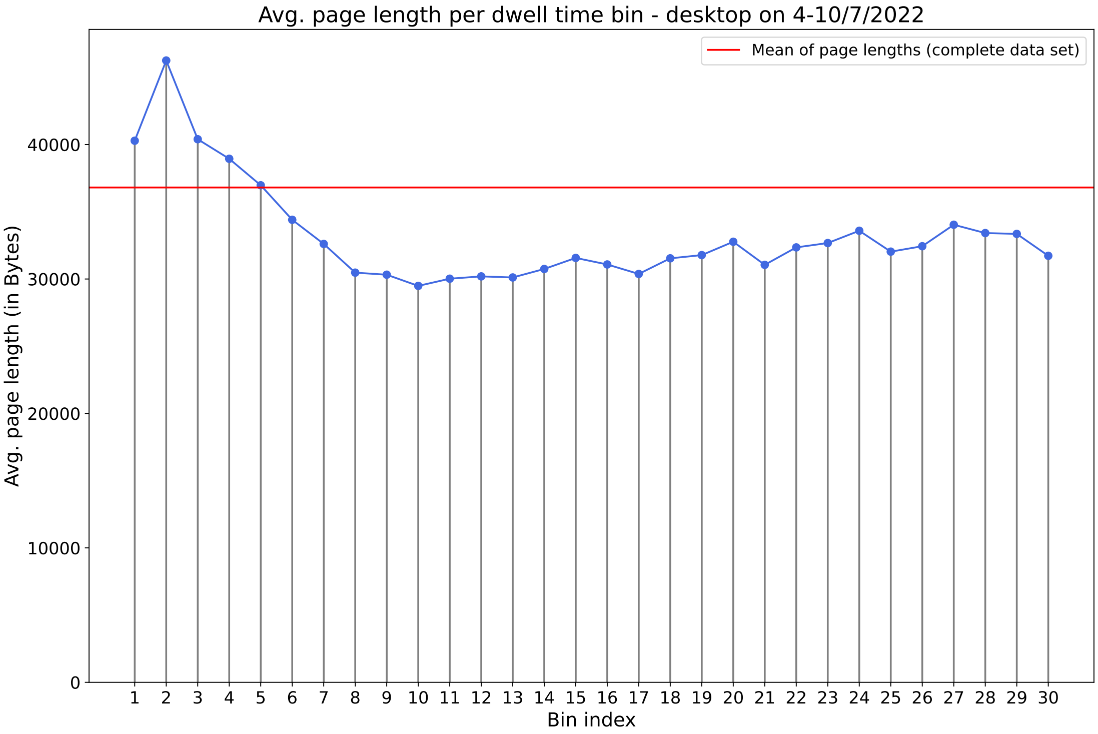


Figure 24: Average page length per dwell time bin (subrange $[0s, 30s]$) for desktop access method on 4-10/Jul/2022.

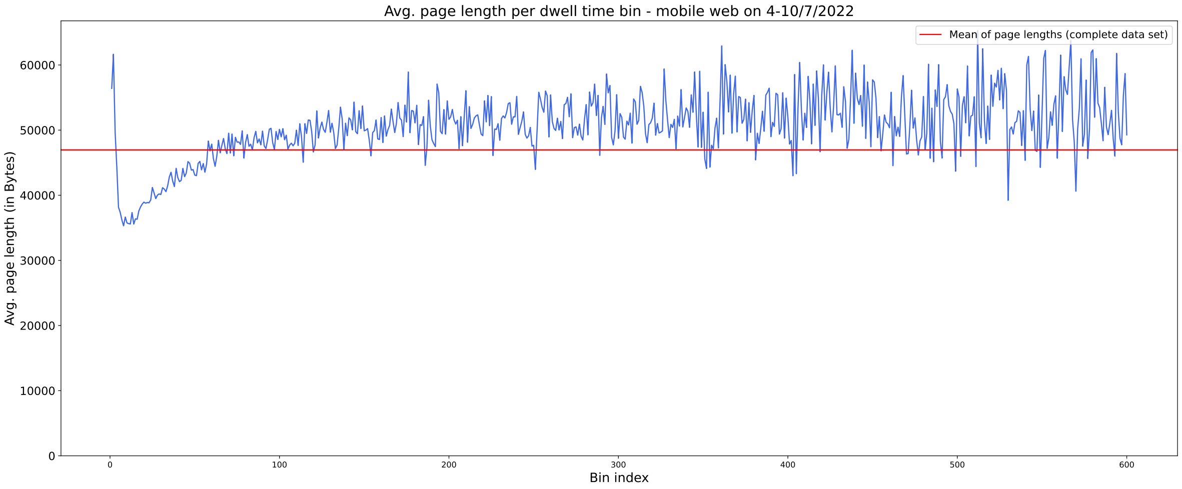


Figure 25: Average page length per dwell time bin (subrange $[0s, 600s]$) for mobile web access method on 4-10/Jul/2022.

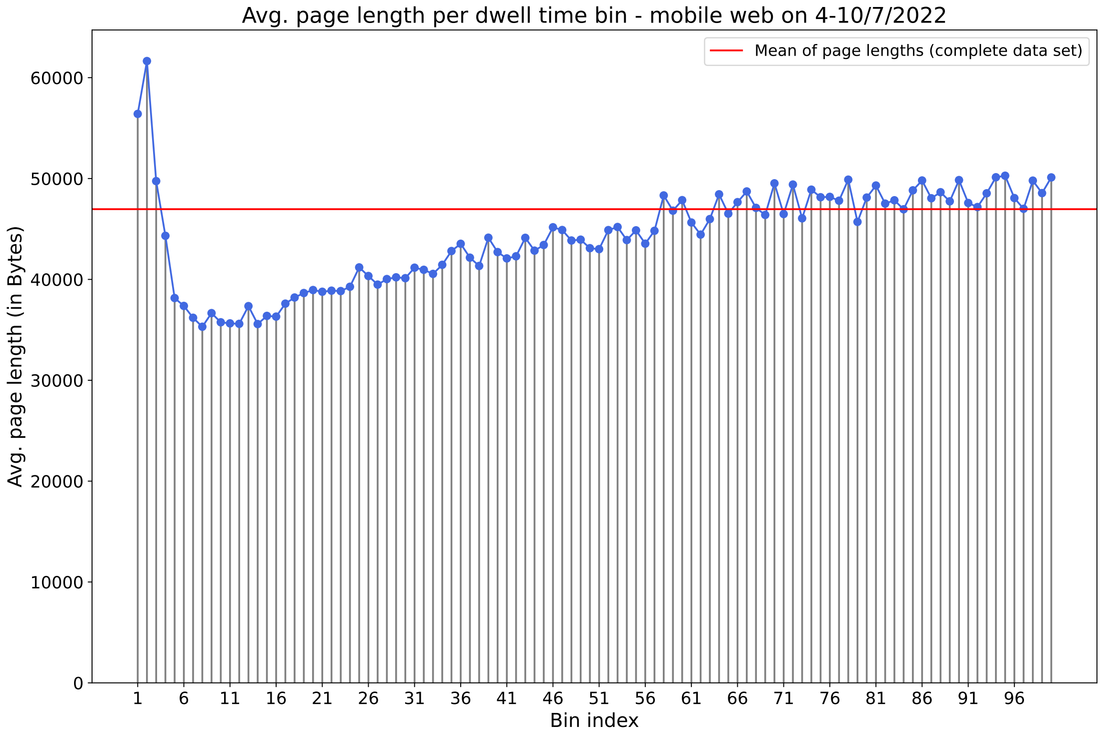


Figure 26: Average page length per dwell time bin (subrange $[0s, 100s]$) for mobile web access method on 4-10/Jul/2022.

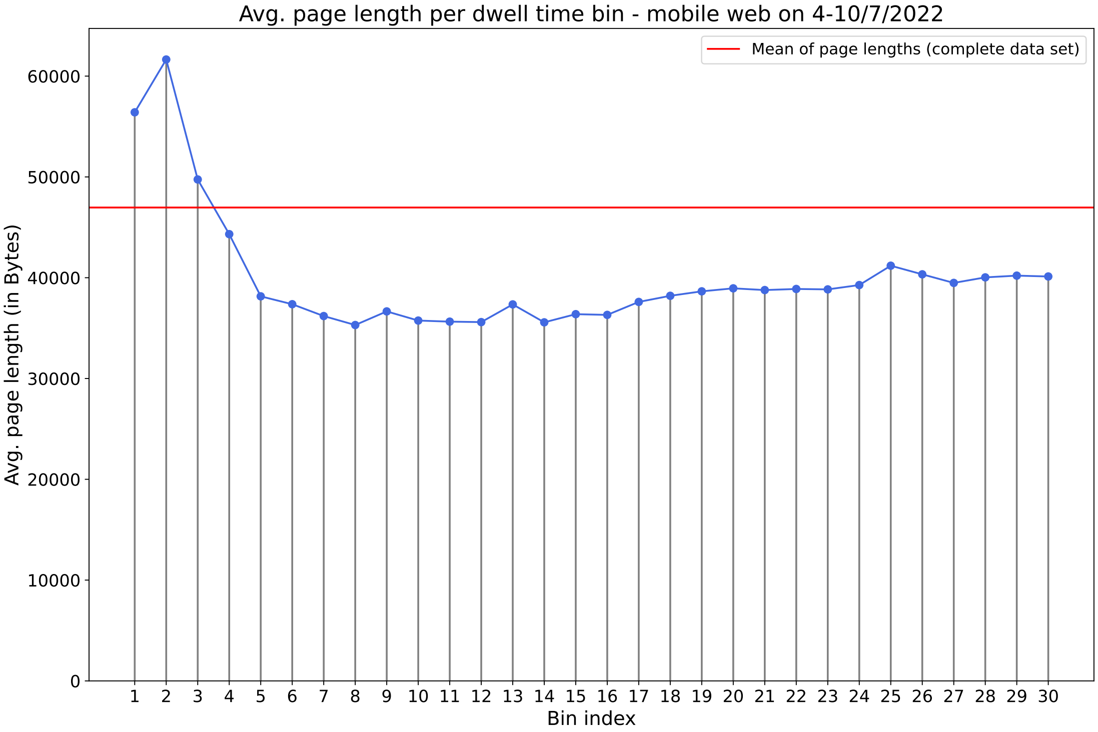


Figure 27: Average page length per dwell time bin (subrange $[0s, 30s]$) for mobile web access method on 4-10/Jul/2022.

Time range	Access method	Avg. ranking pos.
2-8/May/2022	desktop	1.90
2-8/May/2022	mobile web	2.32
4-10/July/2022	desktop	1.92
4-10/July/2022	mobile web	2.24

Table 21: Average ranking position by access method.

Data Analysis

The results for both access methods are similar in nature:

- For both access methods, the page length for dwell times smaller than 10 minutes (600 seconds) oscillates and is bounded between $30,000B = 30KB$ and $50,000B = 50KB$ for the desktop case and $35,000B = 35KB$ and $50,000B = 50KB$ for the mobile web case.
- For both access methods, the curves peak at bin 2 and then this peak is followed by a segment ($[5, 48]$ for desktop and $[5, 57]$ for mobile web) where page length is lower than average.
- For the mobile web case, after segment $[5, 57]$, lengths lie predominantly above the mean.

4.6 Average ranking position clicked on

In this section, we compute for each session (this time filtered by access method), the average ranking position clicked on by the user within the session and then average all results to obtain the desired output. The results for the 2 time ranges, grouped by access method are shown in Table 21. We recall that the closer the average is to 1, the better the search quality of the results.

Data Analysis

According to this measure, search quality is better for desktop access method.

4.7 Number of words per query

Now, we compute the average number of words per query, this time grouped by access method, usisng the same implementation previously discussed in Section 3.7.

Time range	Access method	Avg. nr. of words per query
2-8/May/2022	desktop	2.51
2-8/May/2022	mobile web	2.61
4-10/July/2022	desktop	2.49
4-10/July/2022	mobile web	2.54

Table 22: Average nr. of words per query by access method.

Data sources and time ranges considered

- Data source: `emcr INNER JOIN dqcd` on

  `emcr.search_id = dqcd.request_set_token`

- Time ranges: 2-8/May/2022 and 4-10/July/2022.

The results for the 2 time ranges, grouped by access method are shown in Table 22.

Data Analysis

Queries are longer on average for mobile platforms.

4.8 Top k queries

In this subsection, we compute the top k queries grouped by access method and present plots for $k = 15$. We consider the same initial and validation data sources as the ones introduced in Section 3.8.

Data sources and time ranges considered

- Initial data source: `emcr INNER JOIN dqcd` on

  `emcr.search_id = dqcd.request_set_token`

- Validation data source: `wmf.webrequest` (i.e., web request logs)
- Time ranges: 2-8/May/2022 and 4-10/July/2022.

We recall that due to the impossibility of recovering the host of the URI from this table (mentioned in Section 4.1), the queries from it cannot be segmented by access method and are therefore not included in this section of the report.

4.8.1 Full text searches - Initial data source

We start with the results for 2-8/May/2022. The top 15 queries are shown in Figure 28 (desktop) and 29 (mobile web). The same information presented in tabular form is displayed in Table 23 and 24 respectively. As in Section 3.8.1, when a non Latin-script alphabet was

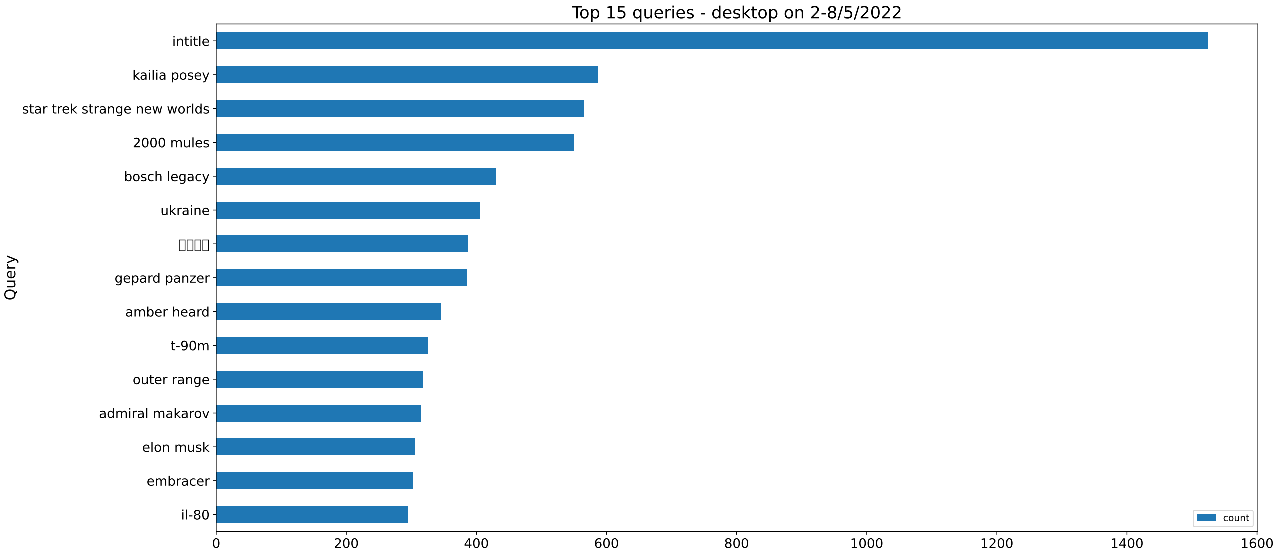


Figure 28: Top 15 queries (`dqcd-emcr`) for desktop access method for 2-8/May/2022.

used for a query appearing in the results, the translated version has been included. For these cases, both language recognition and translation has been done using Google Translate [12].

The top 15 queries for the time range 4-10/July/2022 are shown in Figure 30 (desktop) and 31 (mobile web). The same information presented in tabular form is displayed in Table 25 and 26 respectively.

Data Analysis

- For both time ranges, there are terms that occur in both rankings.
- Mobile web queries refer predominantly to adult content.

4.8.2 Validation against web request logs

After classifying the queries and filtering out non full text searches using the criteria explained in Section 3.8.2, we obtained the segmented results shown in Table 27 (desktop) and 28 (mobile web) for 2-8/May/2022 and Table 29 (desktop) and 30 (mobile web) for 4-10/July/2022.

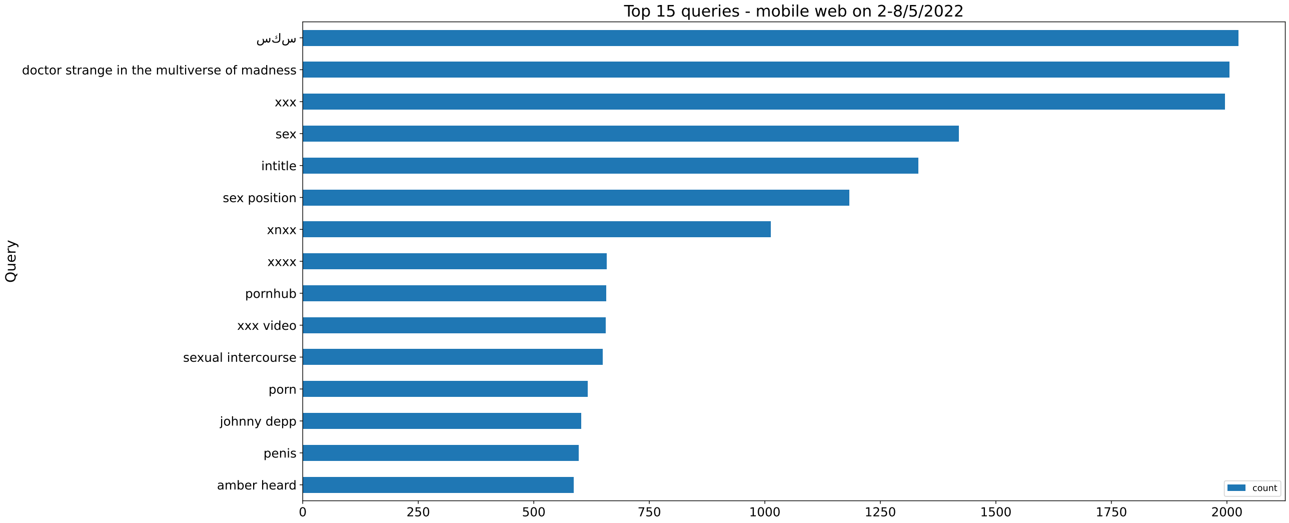


Figure 29: Top 15 queries (dqcd-emcr) for mobile web access method for 2-8/May/2022.

Query	Count
intitle	1525
kailia posey	586
star trek strange new worlds	565
2000 mules	550
bosch legacy	430
ukraine	406
<"Shanghai Epidemic" in Chinese>	387
gepard panzer	385
amber heard	346
t-90m	325
outer range	317
admiral makarov	314
elon musk	305
embracer	302
il-80	295

Table 23: Top 15 queries (dqcd-emcr) for desktop access method for 2-8/May/2022.

Query	Count
<"sex" in Arabic>	2025
doctor strange in the multiverse of madness	2005
xxx	1996
sex	1420
intitle	1332
sex position	1183
xnxx	1013
xxxx	658
pornhub	657
xxx video	655
sexual intercourse	649
porn	617
johnny depp	602
penis	597
amber heard	586

Table 24: Top 15 queries (dqcd-emcr) for mobile web access method for 2-8/May/2022.

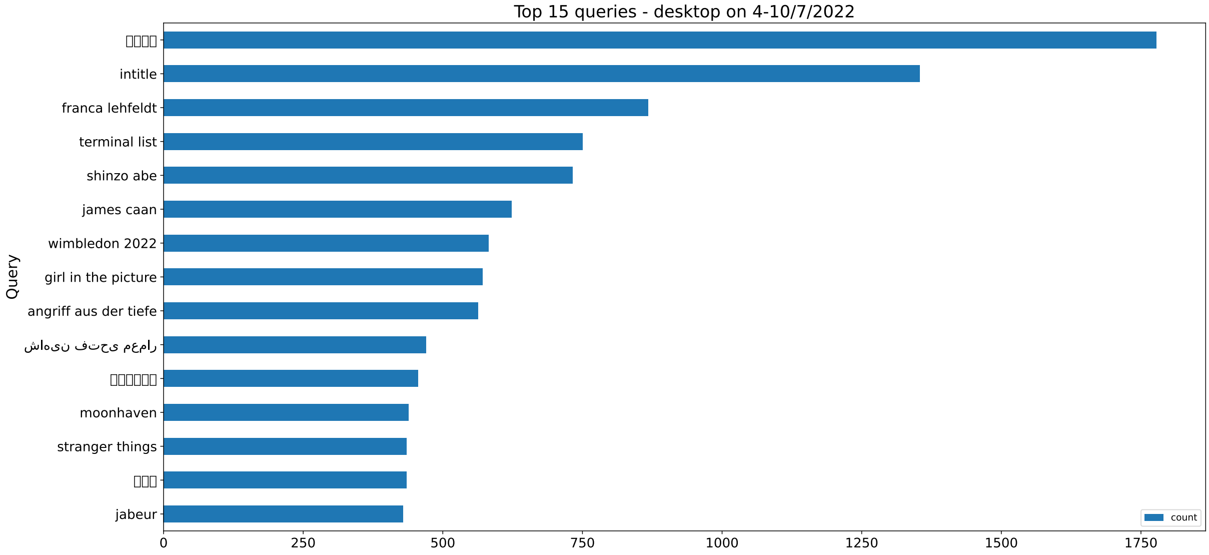


Figure 30: Top 15 queries (dqcd-emcr) for desktop access method for 4-10/July/2022.

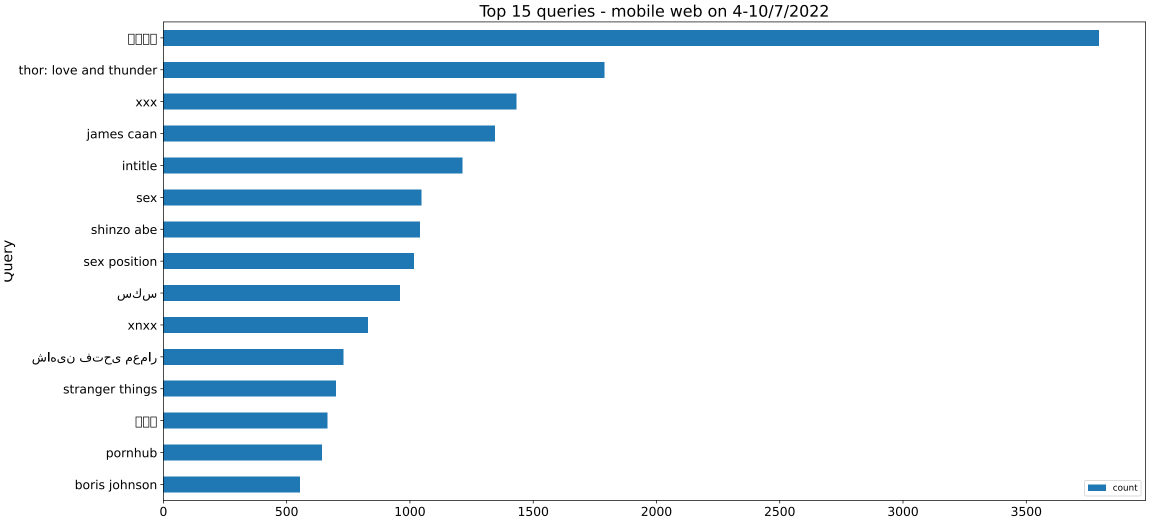


Figure 31: Top 15 queries (dqcd-emcr) for mobile web access method for 4-10/July/2022.

Query	Count
<"Shinzo Abe" in Japanese>	1777
intitle	1354
franca lehfeldt	867
terminal list	750
shinzo abe	732
james caan	623
wimbledon 2022	582
girl in the picture	571
angriff aus der tiefe	563
<"Shahin Fathi Memar" in Persian>	470
<"Shinzo Abe" -different spelling- in Jap.>	455
moonhaven	438
stranger things	435
<"suffrage party" in Japanese>	435
jabeur	429

Table 25: Top 15 queries (dqcd-emcr) for desktop access method for 4-10/July/2022.

Query	Count
<"Shinzo Abe" in Japanese>	3794
thor: love and thunder	1789
xxx	1431
james caan	1345
intitle	1212
sex	1046
shinzo abe	1040
sex position	1015
<"sex" in Arabic>	960
xnxx	829
<"Shahin Fathi Memar" in Persian>	729
stranger things	700
<"Xi Jinping" in Chinese>	665
pornhub	642
boris johnson	553

Table 26: Top 15 queries (`dqcd-emcr`) for mobile web access method for 4-10/July/2022.

Query	Count
""	829362
mishcon.com	630941
mishcon de reya	628667
etsy.com	491306
nasdaq	202324
netscout	201617
smartbear	161423
isaac+newton	159341
johann+goethe	159074
+1	133847
albert++einstein	128730
stackdriver	120860
willard++gibbs	99919
james+maxwell	99690
cu	84038

Table 27: Top 15 queries (web request logs) for desktop access method for 2-8/May/2022.

Query	Count
""	1782564
xxx	128298
cleopatra	58745
w	42511
<"sex" in Arabic>	40290
xnxx	35151
sex	28715
xxxx	25620
special	15817
wikipedia	14886
file	14717
<"Rina Ikoma" in Japanese>	12381
google	11910
porn	11781
wwwxxx	11404

Table 28: Top 15 queries (web request logs) for mobile web access method for 2-8/May/2022.

Query	Count
""	792416
nasdaq	217130
netscout	216591
smartbear	161316
stackdriver	120983
+1	104858
<"Shinzo Abe" in Japanese>	100060
cu	95607
android	77257
ubuntu	75903
linux++c	73967
xxx	72690
rte	66539
thor: love and thunder	54714
"fut+succedee"	52415

Table 29: Top 15 queries (web request logs) for desktop access method for 4-10/July/2022.

Query	Count
""	1612882
cleopatra	461851
xxx	109759
<"Shinzo Abe" in Japanese>	66473
w	37740
jmeter	35285
<"sex" in Arabic>	32347
xnxx	28282
sex	24642
windex.php	23379
special	21459
xxxx	21334
x	15907
<"Xi Jinping" in Chinese>	12451
<"Rina Ikoma" in Japanese>	11778

Table 30: Top 15 queries (web request logs) for mobile web access method for 4-10/July/2022.

Data Analysis

For both time ranges we have:

- The order of magnitudes differ among the 2 data sources.
- There are terms that occur in both rankings, for example:
  - for 2-8/May/2022 (mobile web) the term "xxx" was searched $1,782,564$ times in web request logs but only $2,025$ in `dqcd-emcr`.
  - for 4-10/Jul/2022 (mobile web) the term "xxx" was searched $109,759$ times in web request logs but only $1,431$ in `dqcd-emcr`.
- As before, mobile web queries refer predominantly to adult content.

5 Search behavior based on language and country

The objective of this section is to analyze the search behavior of users primarily based on the language of the prioritized projects as well as on the countries accessing them.

5.1 User hits per country for a given language

In this subsection, we present the distribution of user hits per country (top k countries) to different Wikipedia projects based on their language, also segmenting the hits by platform type (i.e. access method).

Data sources and time ranges considered

- Data source: `wmf.pageview_hourly`.
- Data quality
  - When filtering out bots and automated traffic (i.e., `agent_type` = "user") from the hits to English Wikipedia (i.e., `project` = "en.wikipedia") and considering only valid (i.e., non-null) page ids,`wmf.pageview_hourly` does not have records with `access_method` = "mobile app" for the year 2022.
- Time ranges: Feb/2021 and Jul/2022.

We considered the top 10 most spoken languages, which are given in Table 31. The corresponding figures for the time ranges are also indicated in the table, as well as the top priority languages[10] for the Search Platform team highlighted in **bold**.

When generating the charts, bots and automated traffic were filtered out (i.e., `agent_type` = "user") and only valid (i.e., non-null) page ids were considered. The countries are ranked according to the total number of user hits.

Data analysis - Feb/2021

- Due to its order of magnitude mobile app usage has to be (visually) analyzed separately. In consultation with the Search Platform Team, it has been decided to exclude it from the analysis.
- For almost all languages (except Chinese and Indonesian), the top 3 countries have a majority of mobile accesses.
- China ranks 12th in user hits to Chinese Wikipedia (zk.wikipedia).
- There are languages with different "profiles" (e.g., Arabic vs English): for some, mobile access is significantly prevalent while for others mobile vs desktop is more balanced.

[10] Taken from [14].

Language	Time range	Figure
English	Feb/2021	Figure 32
English	Jul/2022	Figure 42
Chinese	Feb/2021	Figure 33
Chinese	Jul/2022	Figure 43
Hindi	Feb/2021	Figure 34
Hindi	Jul/2022	Figure 44
Spanish	Feb/2021	Figure 35
Spanish	Jul/2022	Figure 45
French	Feb/2021	Figure 36
French	Jul/2022	Figure 46
Arabic	Feb/2021	Figure 37
Arabic	Jul/2022	Figure 47
Bengali	Feb/2021	Figure 38
Bengali	Jul/2022	Figure 48
Russian	Feb/2021	Figure 39
Russian	Jul/2022	Figure 49
Portuguese	Feb/2021	Figure 40
Portuguese	Jul/2022	Figure 50
Indonesian	Feb/2021	Figure 41
Indonesian	Jul/2022	Figure 51

Table 31: Results for top 10 most spoken languages for the 2 time ranges considered.

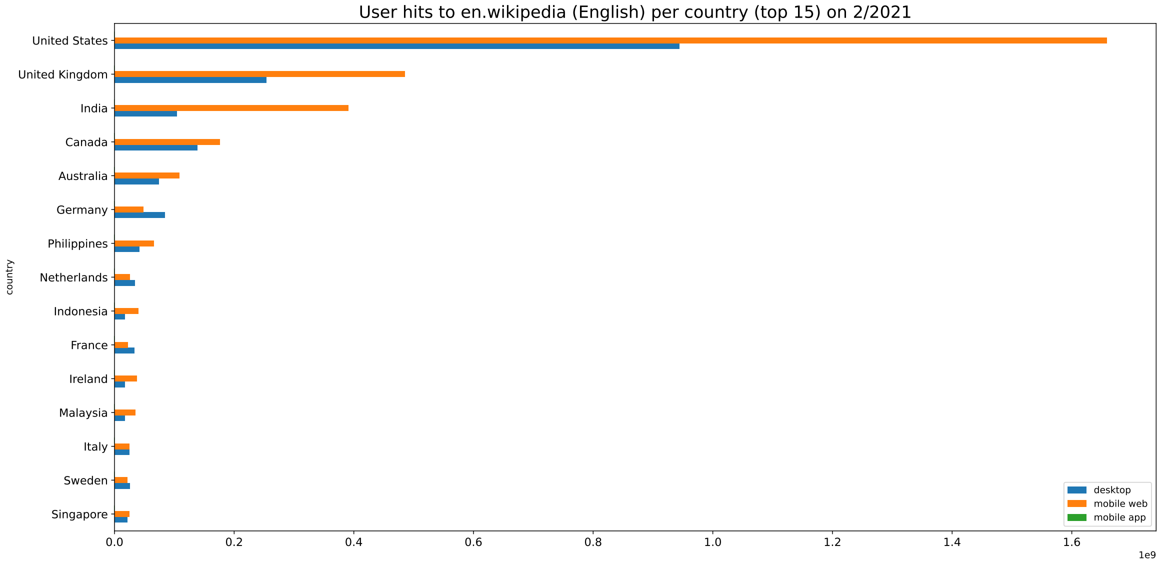


Figure 32: User hits to English Wikipedia per country by access method (top 15) on Feb/2021

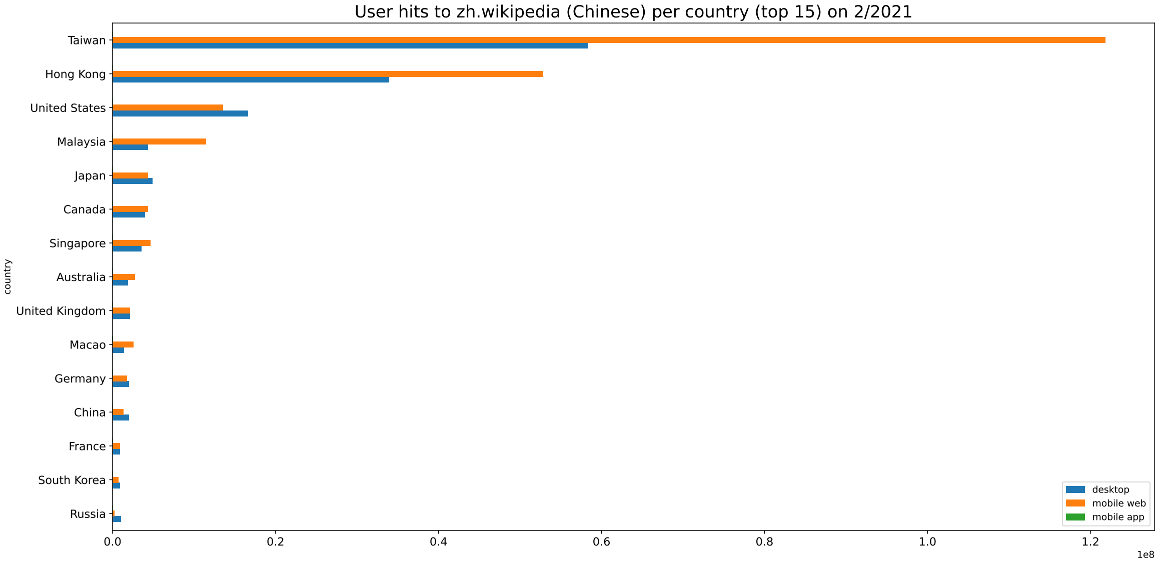


Figure 33: User hits to Chinese Wikipedia per country by access method (top 15) on Feb/2021

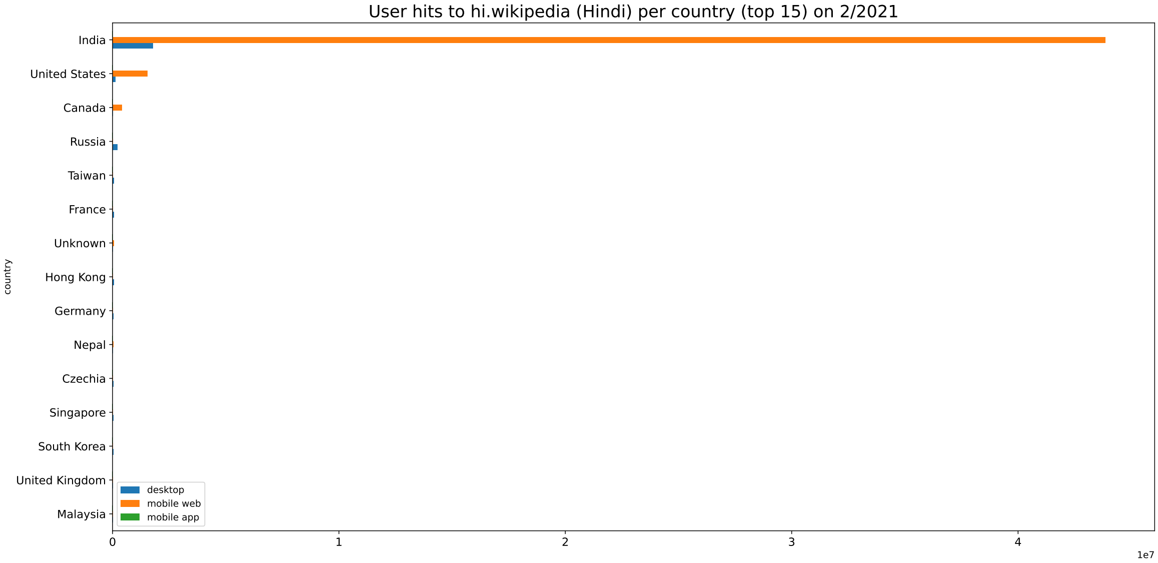


Figure 34: User hits to Hindi Wikipedia per country by access method (top 15) on Feb/2021

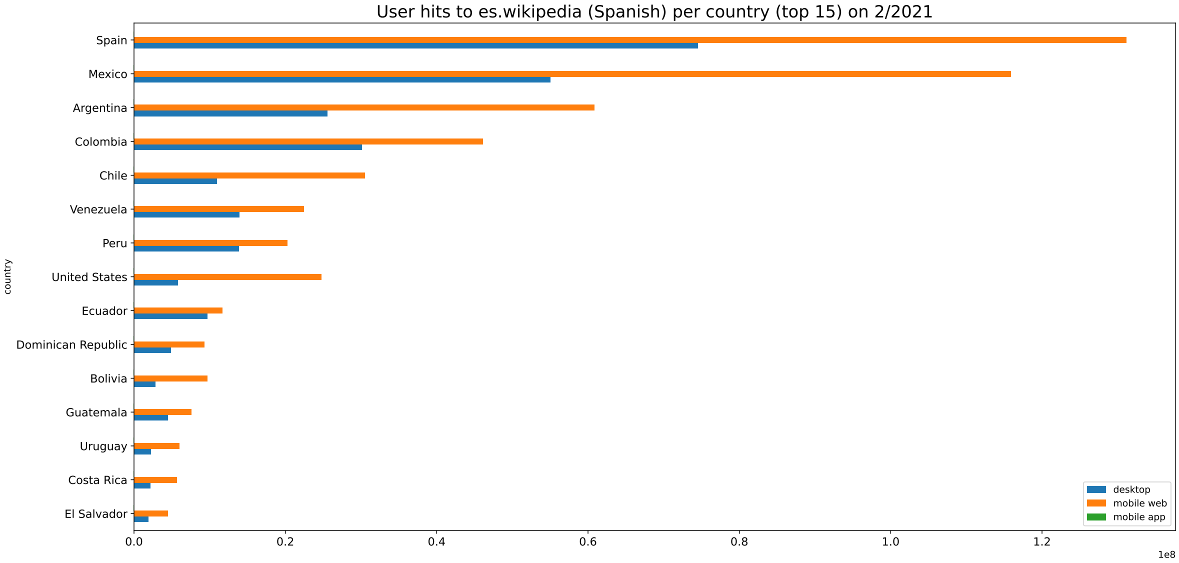


Figure 35: User hits to Spanish Wikipedia per country by access method (top 15) on Feb/2021

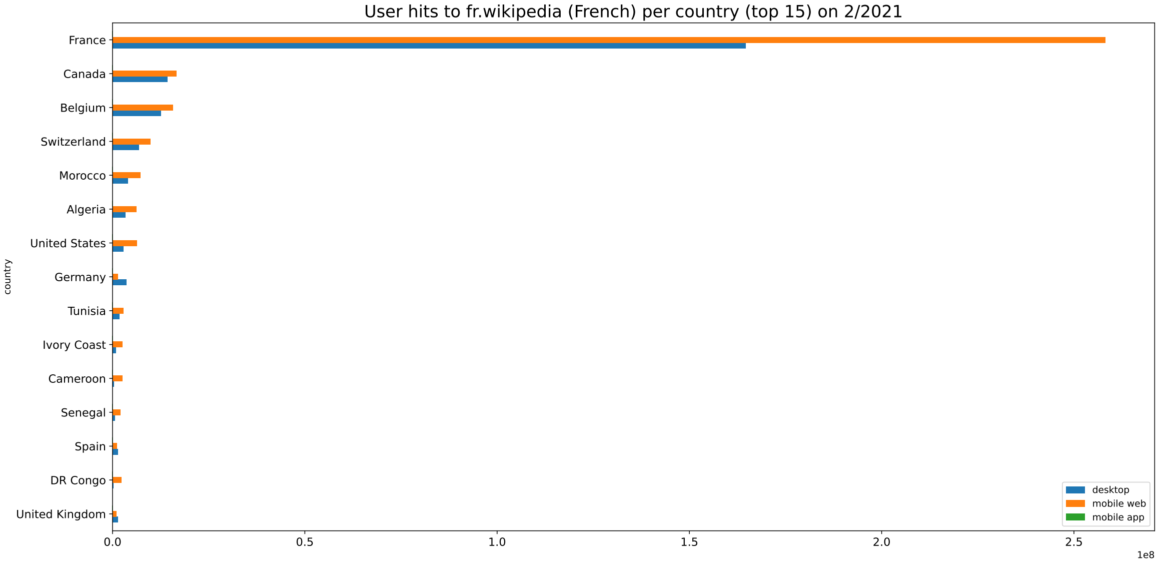


Figure 36: User hits to French Wikipedia per country by access method (top 15) on Feb/2021

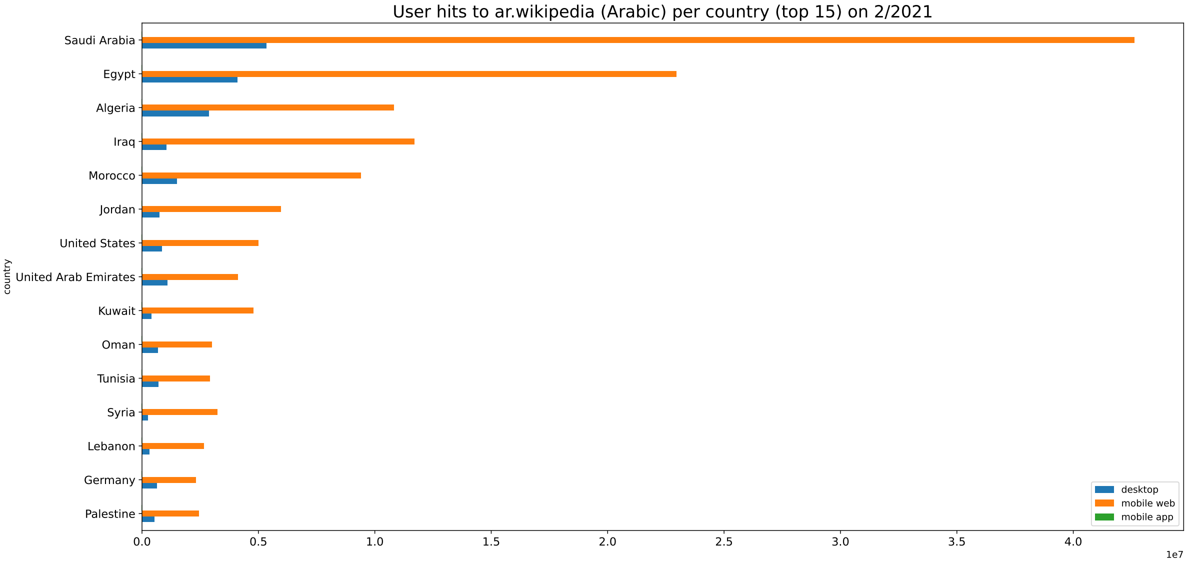


Figure 37: User hits to Arabic Wikipedia per country by access method (top 15) on Feb$/$2021

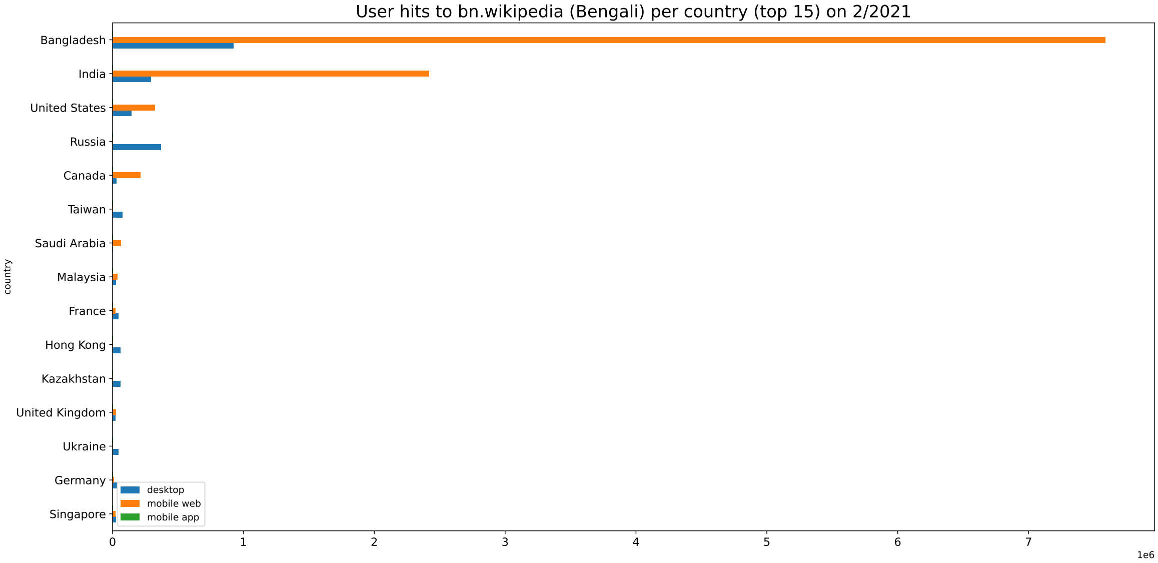


Figure 38: User hits to Bengali Wikipedia per country by access method (top 15) on Feb$/$2021

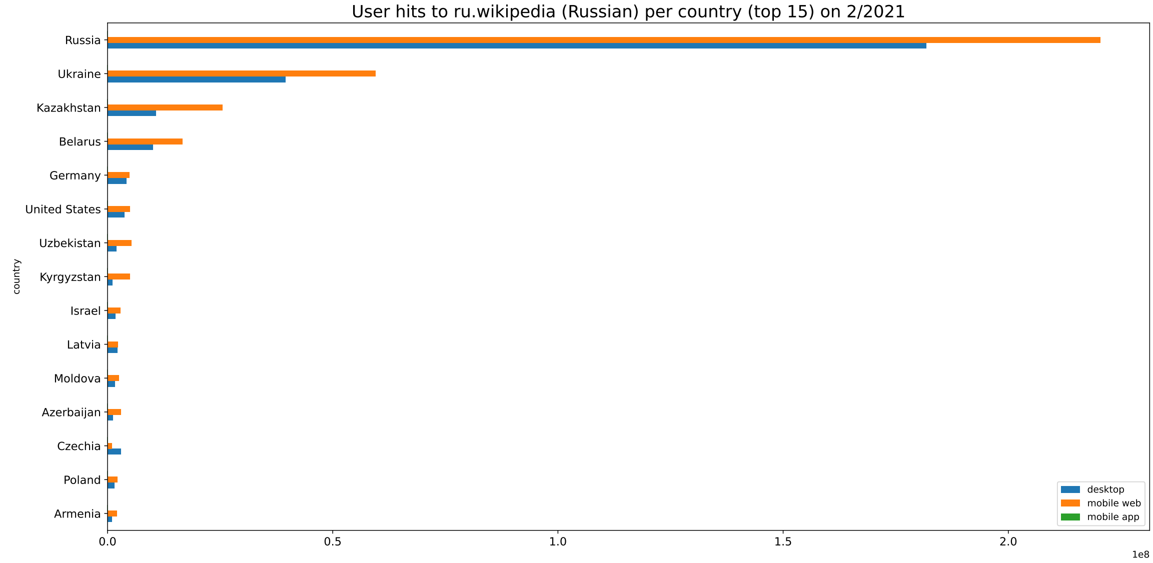


Figure 39: User hits to Russian Wikipedia per country by access method (top 15) on Feb/2021

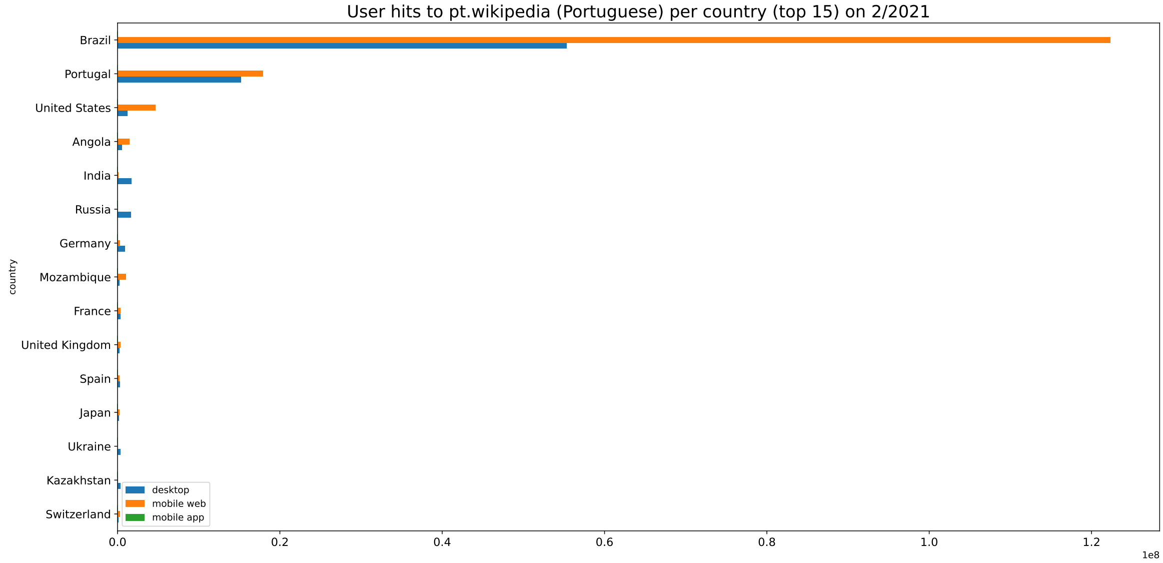


Figure 40: User hits to Portuguese Wikipedia per country by access method (top 15) on Feb/2021

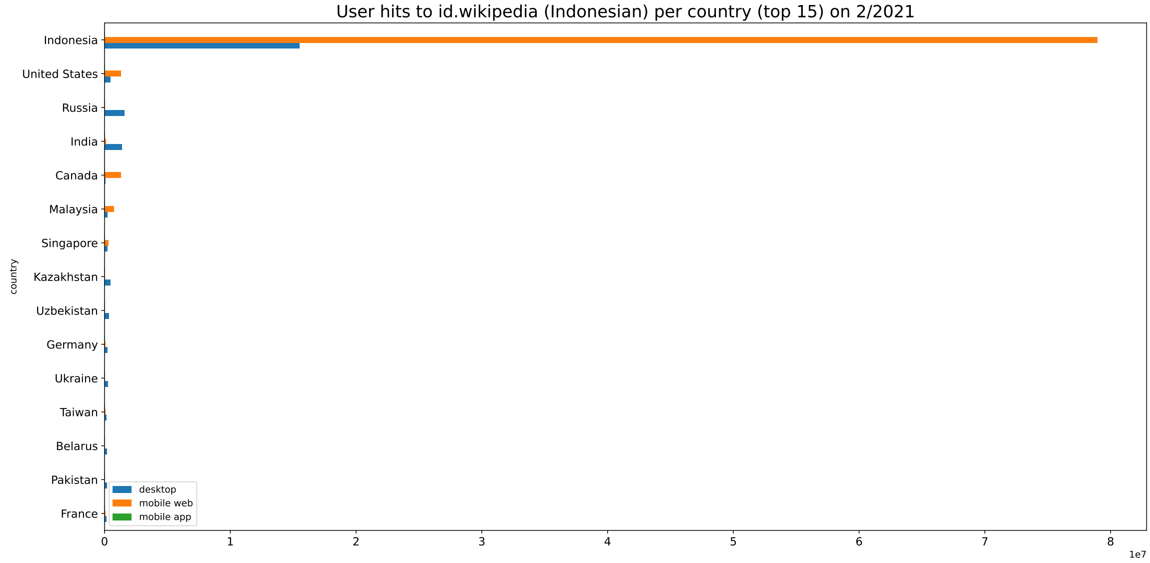


Figure 41: User hits to Indonesian Wikipedia per country by access method (top 15) on Feb/2021

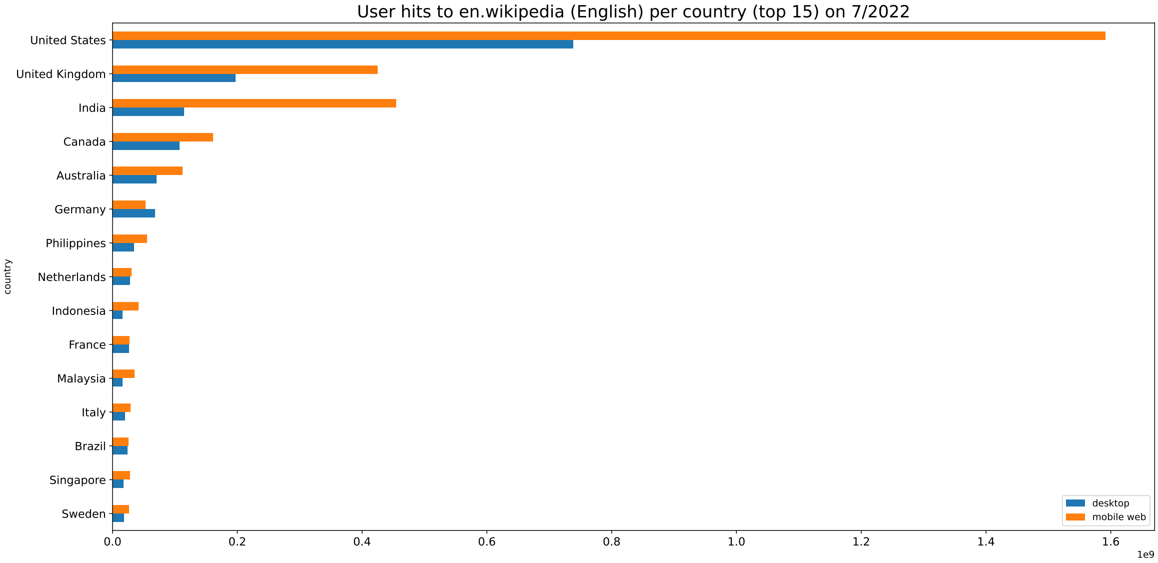


Figure 42: User hits to English Wikipedia per country by access method (top 15) on Jul/2022

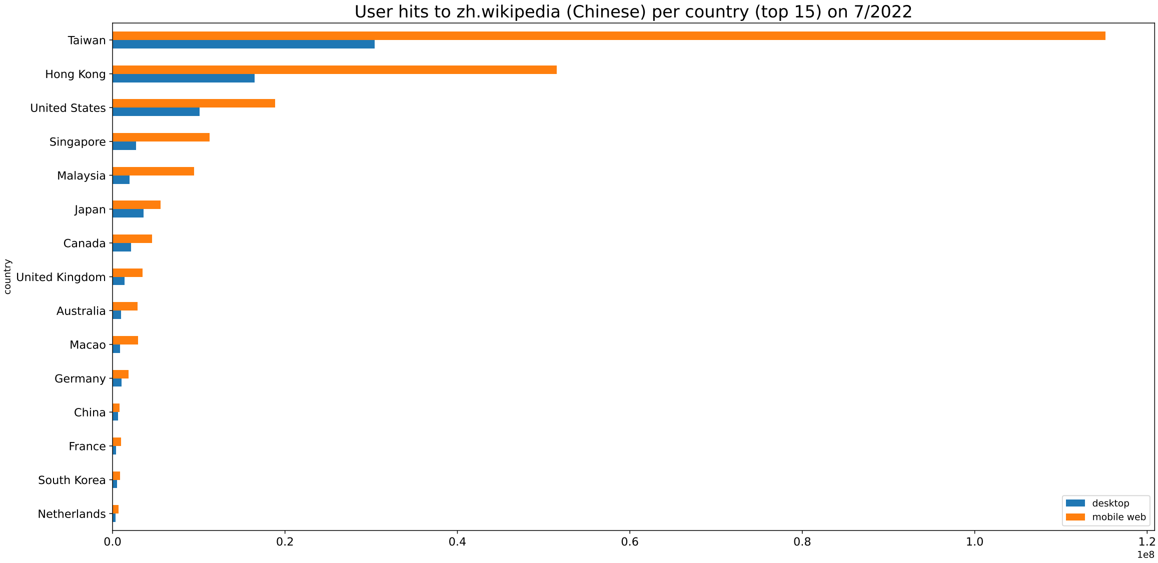


Figure 43: User hits to Chinese Wikipedia per country by access method (top 15) on Jul/2022

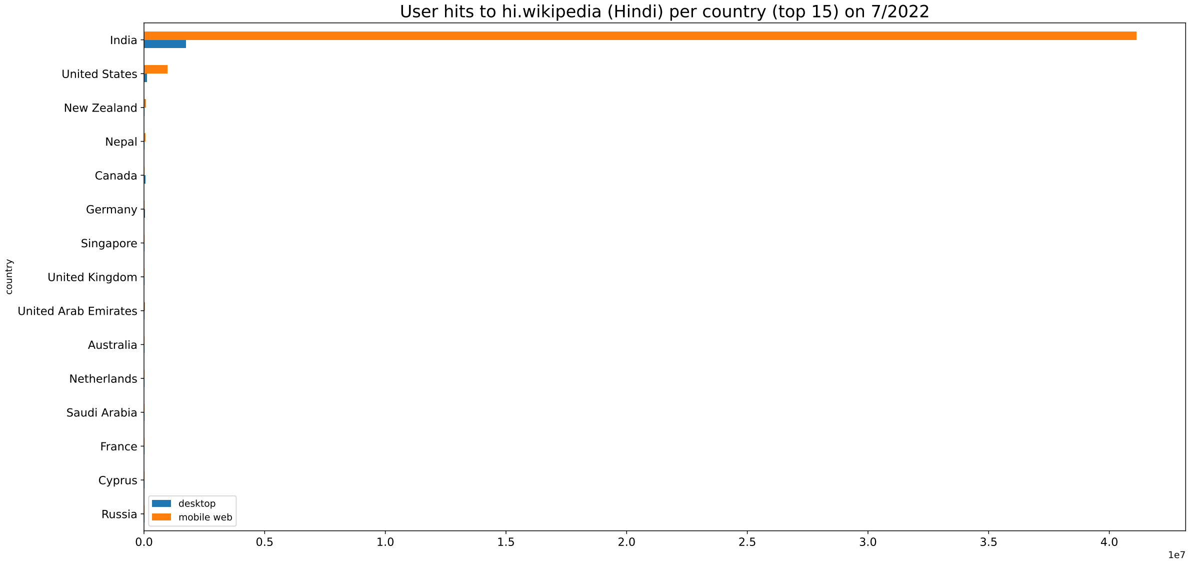


Figure 44: User hits to Hindi Wikipedia per country by access method (top 15) on Jul/2022

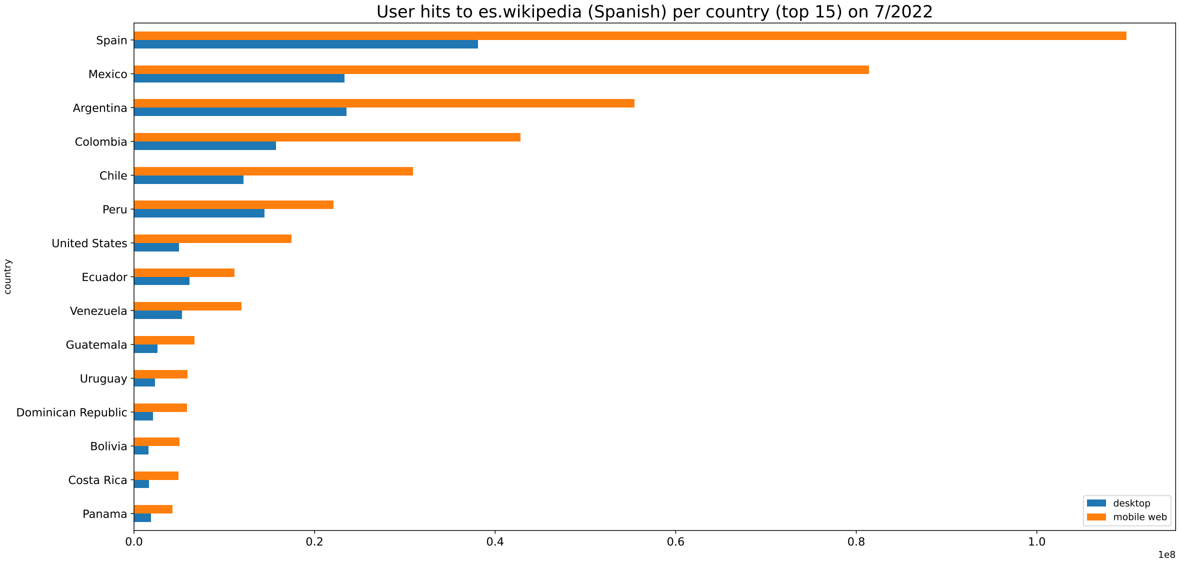


Figure 45: User hits to Spanish Wikipedia per country by access method (top 15) on Jul/2022

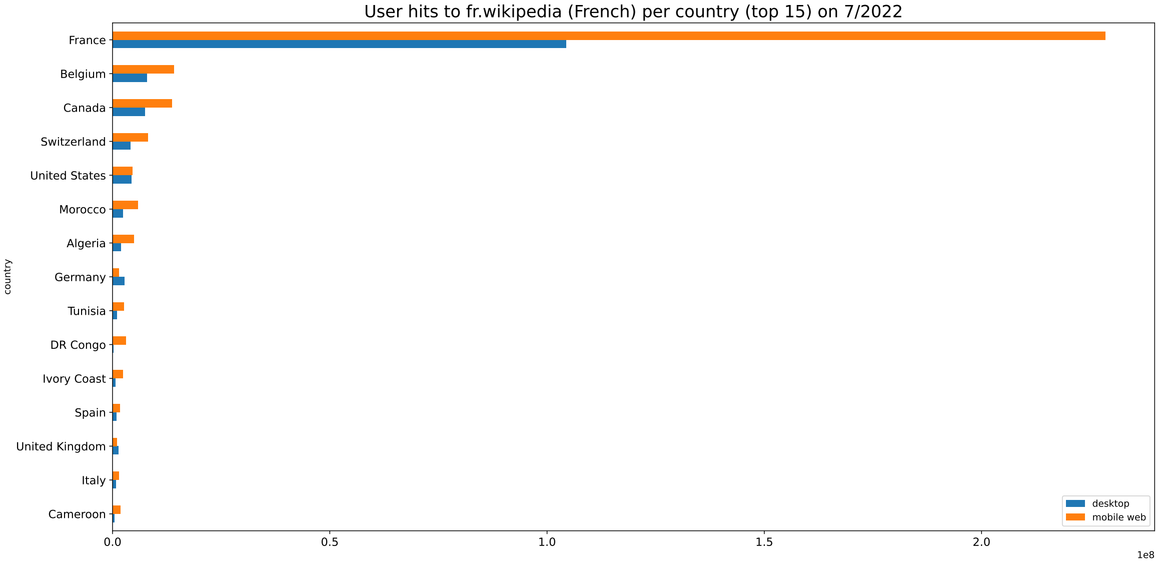


Figure 46: User hits to French Wikipedia per country by access method (top 15) on Jul/2022

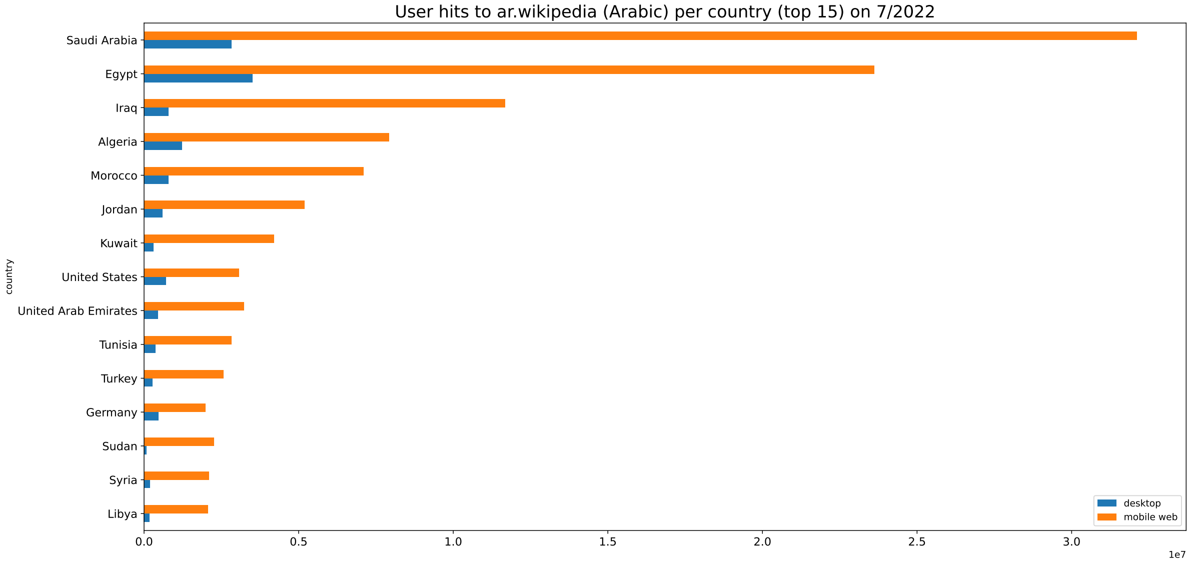


Figure 47: User hits to Arabic Wikipedia per country by access method (top 15) on Jul/2022

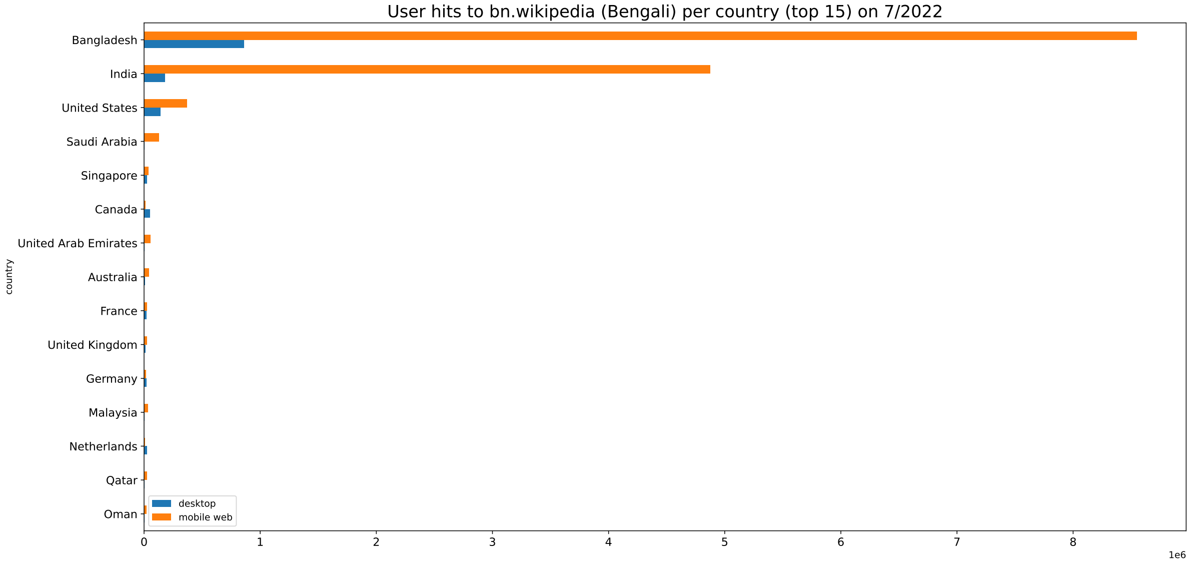


Figure 48: User hits to Bengali Wikipedia per country by access method (top 15) on Jul/2022

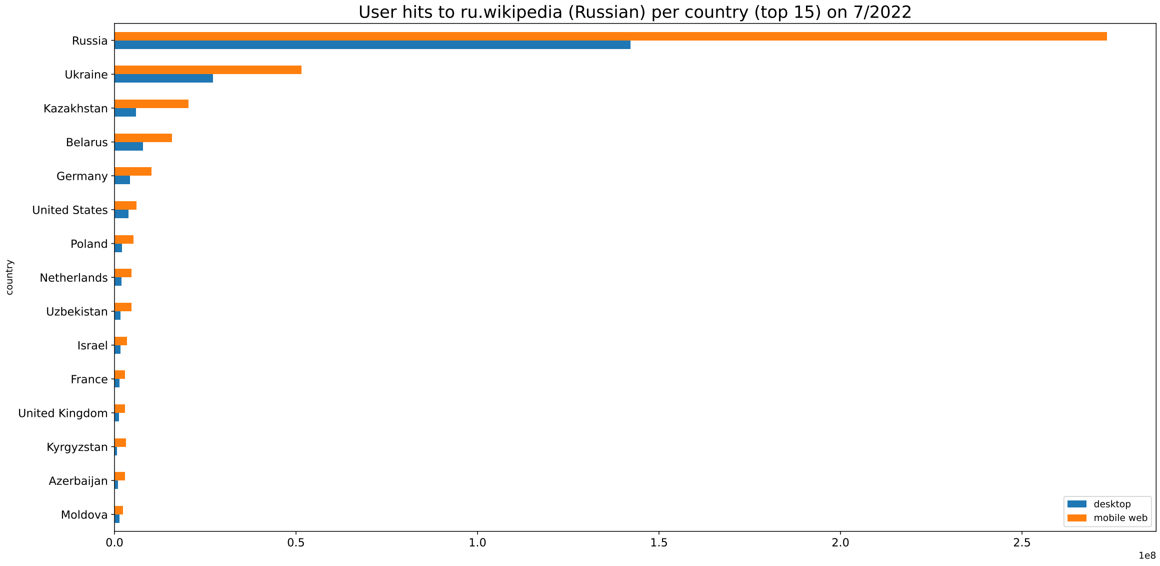


Figure 49: User hits to Russian Wikipedia per country by access method (top 15) on Jul/2022

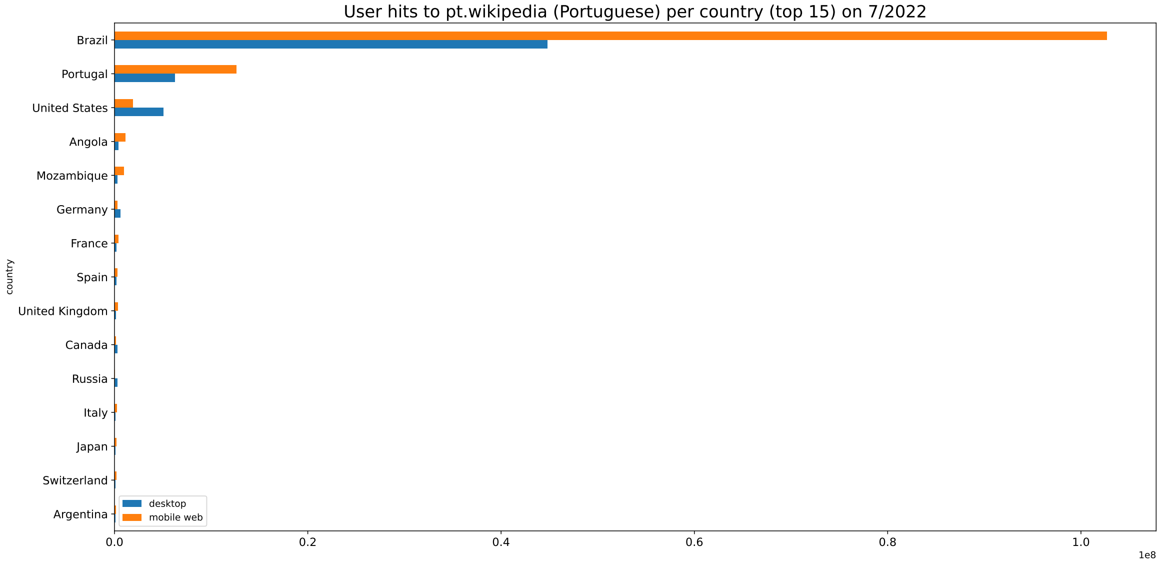


Figure 50: User hits to Portuguese Wikipedia per country by access method (top 15) on Jul/2022

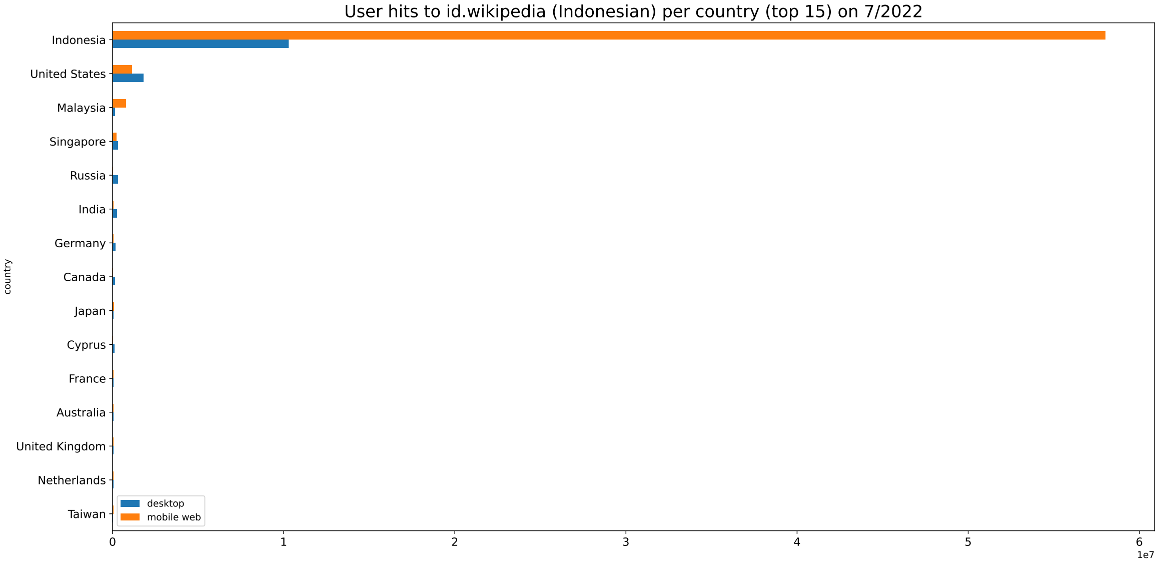


Figure 51: User hits to Indonesian Wikipedia per country by access method (top 15) on Jul/2022

5.2 Guiding content creation

In this subsection, we aim at computing an usage measure for the different Wikipedias. The following simple measure is proposed as an initial approach:

$$\texttt{usage(lan)} = \frac{\texttt{nr\_hits(wiki(len))}}{\texttt{size(wiki(len))}}$$

where `wiki(len)` refers to the Wikipedia associated with language `lan` and `size(wiki(len))` refers to the size measured in number of words[11] in all content pages from the corresponding Wikipedia[12]. The metric measures the number of hits normalized by size, where longer pages need a larger number of hits to reach the same usage level as shorter ones. In a sense, the metric quantifies "content effectiveness", given that when extending or adding new content to a given wiki, it's usage value would only increase if more hits are generated due to the new content.

[11] This measure is considered more appropriate than the number of bytes or the number of characters from the page, since it is independent of the average length of words of the language. For example, Chinese words are shorter on average in terms of characters used than English words. If we decided to use the number of bytes, two documents with the same number of words and hits could have different usage values depending on its language.

[12] Taken from the associated "Statistics" page (e.g., for English https://en.wikipedia.org/wiki/Special:Statistics). The values used in this report were taken on 8/Aug/2022. The "Statistics" special page can be accessed from the home page of the Wikipedia by clicking on the "Special pages" option from the left panel and then "Data and tools "->"Statistics".

Language (lan)	Nr. of words in all content pages from wiki(lan)
English	4,180,601,186
Chinese	606,302,164
Hindi	50,154,715 (appears as "5,01,54,715" on the page.)
Spanish	1,056,914,839
French	1,530,272,385
Arabic	371,156,178
Bengali	59,218,059 (appears as "5,92,18,059" on the page.)
Russian	940,422,067
Portuguese	488,744,118
Indonesian	175,048,431

Table 32: Number of words in all content pages for different Wikipedias.

Language (lan)	usage(lan) - desktop	usage(lan) - mobile web	usage(lan) - total
English	0.107	0.203	0.31
Chinese	0.03	0.09	0.12
Hindi	0.01	0.183	0.192
Spanish	0.038	0.094	0.132
French	0.023	0.044	0.067
Arabic	0.009	0.073	0.082
Bengali	0.005	0.05	0.055
Russian	0.053	0.103	0.157
Portuguese	0.029	0.057	0.086
Indonesian	0.02	0.069	0.089

Table 33: Values for the usage metric for 4-10/Jul/2022 by access method.

Data sources and time ranges considered

- Data source: `wmf.pageview_hourly`.
- Time ranges: 4-10/Jul/2022.

The usage values by access method are shown in Table 33 and in the following charts: Figure 52 (total usage, all access methods), Figure 53 (desktop) and Figure 54 (mobile web). The number of words used for the computation are listed in Table 32.

The position for the different projects across the 4 rankings are shown in Table 34. To measure ranking correlation, we calculated the Kendall's τ coefficients [15] between the ranking according to number of speakers and the other 3, which are shown in Table 35. We recall that given 2 rankings R and R', we have that Kendall's τ coefficients $\tau(R, R') \in [-1, 1]$, where values close to 1 indicate strong positive correlation, values close to -1 indicate strong negative correlation and values close to 0 indicates nonexistent correlation between the rankings.

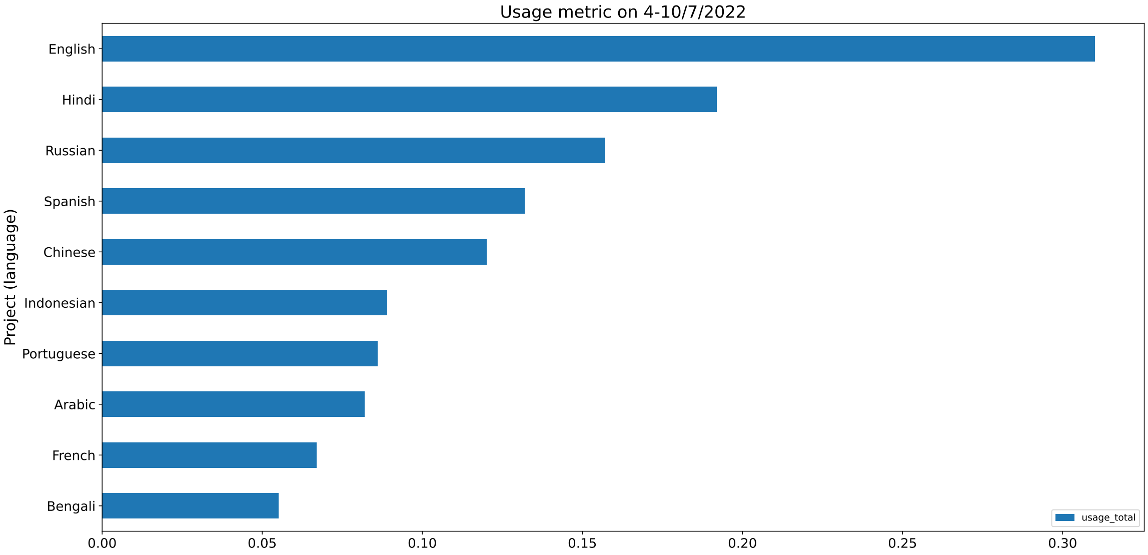


Figure 52: Values for the usage metric (all access methods) for 4-10/Jul/2022.

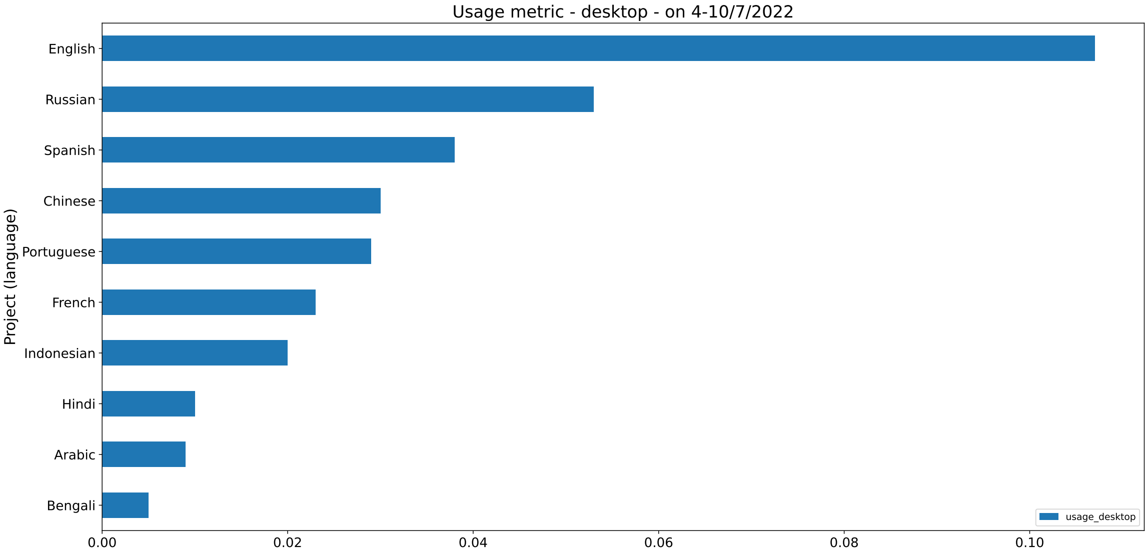


Figure 53: Values for the usage metric (desktop) for 4-10/Jul/2022.

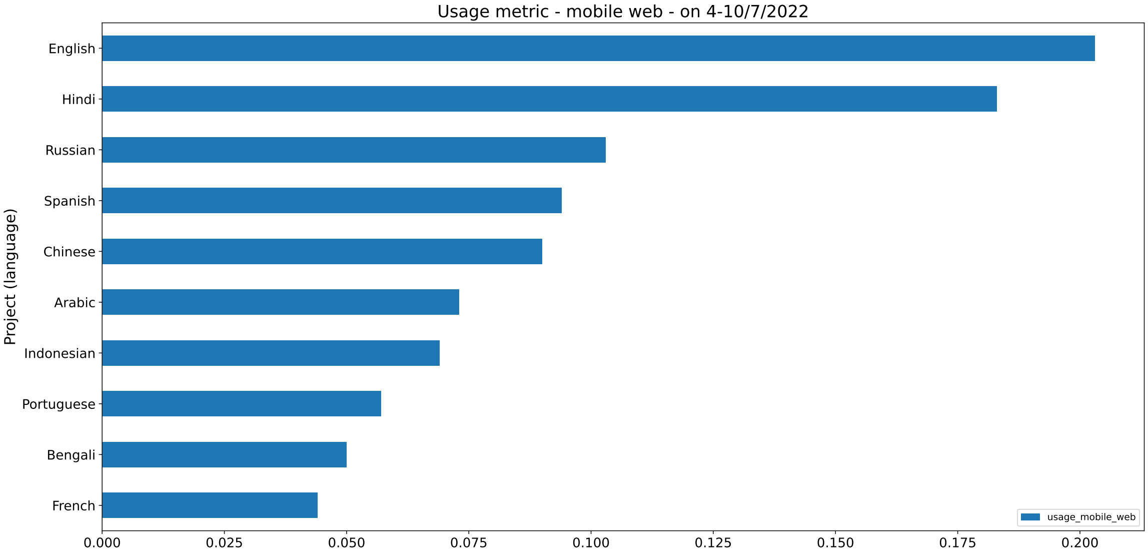


Figure 54: Values for the usage metric (mobile web) for 4-10/Jul/2022.

Lan.	Nr. of speakers	Total usage	Desktop usage	Mobile web usage
English	1	1	1	1
Chinese	2	5	4	5
Hindi	3	2	8	2
Spanish	4	4	3	4
French	5	9	6	10
Arabic	6	8	9	6
Bengali	7	10	10	9
Russian	8	3	2	3
Portuguese	9	7	5	8
Indonesian	10	6	7	7

Table 34: Position for the different projects across the 4 rankings.

R	R'	$\tau(R, R')$
Nr. of speakers	Total usage	-0.2
Nr. of speakers	Desktop usage	-0.24
Nr. of speakers	Mobile web usage	0.29

Table 35: Kendall's τ coefficients between the ranking according to number of speakers and the other 3 (total desktop and mobile web usage).

Data analysis

- English is the first across all categories.
- Bengali is last for the case all access methods and desktop and second-to-last on mobile web access.
- Regarding the Kendall's τ coefficients (Table 35) we have:
  - The ranking according to nr. of speakers and the one according to mobile web usage have a weak positive correlation.
  - The ranking according to nr. of speakers and with the remaining 2 other rankings (total and desktop usage) have a weak negative correlation.

6 Concluding remarks and future work

This section details recommendations based on the work performed and establishes possible future lines of work that are considered of interest.

Data documentation. Regarding the data sources, having accessed to a more detailed documentation of the individual tables would have not only reduced the time of the data source selection process, but also the design of the data processing pipelines used to compute the results. In particular, the documentation of the search logs is scarce at times and the lack of documentation for the `dqcd` table makes it a challenging task to understand the correct semantics of the data. On the global level, having a "data set catalog", where an overview and/or a categorization of the different data sources is provided, would have also contributed in the directions previously mentioned (data set selection and pipeline design).

Next, we continue with future lines of work that resulted from the analysis.

Expanding data set to include autocomplete search We start with future lines of work related to the results presented in Section 3 and 4. In general, and probably the activity believed to be of higher priority is to extend the dataset to incorporate autocomplete searches in the analysis. This is of importance, since as was mentioned in Section 2.1 and 3.8.2, the current data set used (`dqcd`) only contains full text searches, which are in particular a minority ($\sim 20\%$) of all the queries issued by users. A key preprocessing step for several of the metrics included in this report (clicks' analysis, dwell time computation, average word length per session, average ranking position clicked on), which are measures that depend on session information, would be to sessionize and generate click lists for the data set, as is the case of the `dqcd` table. Regarding the 1-click sessions, it would be interesting to study whether in these cases people find what they want quickly or leave the platform unsatisfied. Lastly, another activity considered of less importance, would be to understand the cause behind the tuples difference after adding the page length in Section 3.4.

Top k queries - Understanding differences for initial and validation data sources With respect to top k query analysis, the significant difference between the order of magnitudes obtained considering the `dqcd-emcr` table and the web request logs should be studied further. This also includes better understanding the type of data (i.e., from where it is obtained and which preprocessing pipeline is used to generate it) that is stored in the `dqcd` table, which is itself related to the first point discussed in this section (Data documentation).

Number of words per query As follow-up work related to the number of words per query, it would be interesting to compute the percentage of queries that are issued using languages where the current regular expression does not work correctly, as well as exploring other expressions that account for the languages not currently considered.

Search behavior based on language Considering the results presented in Section 5, possible future work includes: normalizing the charts by countries' population; including scatterplots that, for a given language (i.e., project), includes points of the form <hits for country, country's population> for a subset of countries of interest, as well as barcharts that for a given country, plot, given a subset of languages, the number of hits of that country in that language. Regarding the proposed usage measure, it would be interesting to explore the causes for the Kendall's τ correlation coefficients computed to compare the rankings.

Seasonal differences A further line of future work consists of varying the time ranges considered for the analysis and in particular observe whether the time range influences the results obtained. A natural initial approach for this would be to look for seasonal differences st various levels (e.g., monthly, meteorological seasons, etc.).

References

[1] *Wikipedia Research: Understanding search behavior of users*, https://meta.wikimedia.org/wiki/Research:Understanding_search_behavior_of_users, Accessed: 2022-08-15.

[2] *Wikimedia Foundation: Search Platform team description page*, https://www.mediawiki.org/wiki/Wikimedia_Search_Platform, Accessed: 2022-08-15.

[3] *Wikimedia Foundation Phabricator Task: Create/revive Search Platform team metrics dashboard*, https://phabricator.wikimedia.org/T279105, Accessed: 2022-08-15.

[4] *Web request logs description page*, https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Traffic/Webrequest, Accessed: 2022-08-15.

[5] *Search logs description page*, https://github.com/wikimedia/schemas-event-primary/blob/master/jsonschema/mediawiki/cirrussearch/request/0.0.1.yaml, Accessed: 2022-08-15.

[6] *Mediawiki wikitext current description page*, https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Content/Mediawiki_wikitext_current, Accessed: 2022-08-15.

[7] *Mediawiki wikitext current description page*, https://meta.wikimedia.org/wiki/Schema:SearchSatisfaction, Accessed: 2022-08-15.

[8] *Mediawiki wikitext current description page*, https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Traffic/Pageview_hourly, Accessed: 2022-08-15.

[9] *Production cluster: Analytics clients*, https://wikitech.wikimedia.org/wiki/Analytics/Systems/Clients, Accessed: 2022-08-15.

[10] *Apache Spark: Official website*, https://spark.apache.org/, Accessed: 2022-08-15.

[11] C. Liu, R. W. White, and S. Dumais, "Understanding web browsing behaviors through weibull analysis of dwell time," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '10, Geneva, Switzerland: Association for Computing Machinery, 2010, pp. 379–386. DOI: 10.1145/1835449.1835513. [Online]. Available: https://doi.org/10.1145/1835449.1835513.

[12] *Google Translate: Official website*, https://translate.google.com/, Accessed: 2022-08-15.

[13] *Wikimedia Analytics Refinery: Source code*, https://github.com/wikimedia/analytics-refinery-source/blob/master/refinery-core/src/main/java/org/wikimedia/analytics/refinery/core/Webrequest.java#L277, Accessed: 2022-08-15.

[14] *Wikimedia Foundation Phabricator Task: Get search traffic breakdown for emerging language wikis*, https://phabricator.wikimedia.org/T301902, Accessed: 2022-08-15.

[15] M. G. KENDALL, "A new measure of rank correlation.," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, Jun. 1938. DOI: 10.1093/biomet/30.1-2.81. eprint: https://academic.oup.com/biomet/article-pdf/30/1-2/81/423380/30-1-2-81.pdf. [Online]. Available: https://doi.org/10.1093/biomet/30.1-2.81.

A Contact information

Questions or comments about this report can be sent to the author at scarone.b@northeastern.edu.