daten.berlin.de Usage Statistics

Figure 1: logo for "daten.berlin.de Usage Statistics" dataset

This dataset contains usage statistics (page impressions and visits) for the Berlin Open Data Portal https://daten.berlin.de. Statistics are collected per month, both for the domain as such, and for all datasets (pages below `/datensaetze`).

Statistics are given in both CSV (split over two files) and JSON (one combined file).

Until 2019-12-31, usage statistics were collected with our internal *BerlinOnline Site Statistics* (BOSS) tool. As of 2020-01-01 we have stopped using BOSS on the Berlin Open Data Portal, and have replaced it with Webtrekk Analytics. Webtrekk has been in use since February 2019. While BOSS and Webtrekk provide the same metrics, the actual results differ. We have written a little bit on how and possible why BOSS and Webtrekk differ with respect to their results.

The historic BOSS data has been moved to `data/historical`, while the current Webtrekk data resides in `data/current`.

Requirements

The code to extract the usage statistics is written in Ruby. It has been tested with Ruby 2.7.1.

The required gems are defined in the Gemfile. In particuler, these are:

- webtrekk_connector
- ruby-keychain
- activesupport

If you have bundler, you can install the required gems as follows:

```
bundle install
```

daten_berlin_de.domain_stats.csv

Download here: daten_berlin_de.domain_stats.csv

Domain-wide statistics. One row per month, columns for page impressions, visits and average time spent on site (in seconds).

```
month,impressions,visits,visit_duration_avg_seconds
2021-04,26514,8937,147.94
```

```
2021-03,25116,8505,153.71
2021-02,20832,7212,152.06
2021-01,31221,11199,162.95
...
```

daten_berlin_de.page_stats.datensaetze.csv

Download here: daten_berlin_de.page_stats.datensaetze.csv

Per-dataset statistics. One row per dataset, two columns per month (page impressions and visits).

```
page,2013-04-01 pi,2013-04-01 pv, ... ,2018-05-01 pi,2018-05-01 pv
liste-der-h%C3%A4ufigen-vornamen-2017,,, ... ,279,246
alkis-berlin-amtliches-liegenschaftskatasterinformationssystem,,, ... ,211,185
...
```

daten_berlin_de.stats.json

Download here: daten_berlin_de.stats.json.tgz (compressed)

The structure of the data is as follows:

- `/source` - From which source system the statistics were generated. One of `[ "Webtrekk", "Boss" ]`.
- `/timestamp` - when these usage statistics were generated
- `/stats/site_uri` - domain of the data portal
- `/stats/earliest` - first month for which domain-wide statistics have been collected
- `/stats/latest` - last month for which domain-wide statistics have been collected
- `/stats/totals` - domain-wide statistics
- `/stats/totals/{MONTH}/impressions` - domain-wide page impressions during `MONTH`
- `/stats/totals/{MONTH}/visits` - domain-wide visits during `MONTH`
- `/stats/pages/datensaetze` - statistics for individual datasets
- `/stats/pages/datensaetze/page_uri` - parent page for all datasets
- `/stats/pages/datensaetze/earliest` - first month for which dataset-specific statistics have been collected
- `/stats/pages/datensaetze/latest` - last month for which dataset-specific statistics have been collected
- `/stats/pages/datensaetze/sub_page_counts/{MONTH}/{DATASET}/impressions` - page impressions recorded for DATASET during MONTH
- `/stats/pages/datensaetze/sub_page_counts/{MONTH}/{DATASET}/visits` - visits recorded for DATASET during MONTH


```
{
  "timestamp": "2018-06-12 14:29:47 +0200",
```

```
  "stats": {
    "site_uri": "daten.berlin.de",
    "earliest": "2011-09-01",
    "latest": "2018-05-01",
    "totals": {
      "2018-05": {
        "impressions": 28563,
        "visits": 10436
      },
      ...
      "2011-09": {
        "impressions": 54454,
        "visits": 23765
      }
    },
    "pages": {
      "datensaetze": {
        "page_uri": "daten.berlin.de%2Fdatensaetze",
        "earliest": "2013-04-01",
        "latest": "2018-05-01",
        "sub_page_counts": {
          "2018-05": {
            "liste-der-h%C3%A4ufigen-vornamen-2017": {
              "impressions": 279,
              "visits": 246
            },
            "alkis-berlin-amtliches-liegenschaftskatasterinformationssystem": {
              "impressions": 211,
              "visits": 185
            },
            ...
          }
        }
      }
    }
  }
}
```


Normalization of Dataset Names

In August 2024, the Open Data Portal received a major update, which resulted in slightly changed URLs for many datasets. There is a mapping from the old to the new datasets: https://github.com/berlinonline/berlin_dataset_name_mapping

Starting in November 2024, the dataset names in the usage statistics have been normalized to show the new dataset names everywhere, to allow comparisons

through time.

In cases where requests to both the old and the new name were made in the same month, both have been combined to a single entry, with the sum of the impressions and visits. This normalization is implemented in the map_dataset_names.py script.

For example, there is the following mapping:

```
old_name,new_name
verlauf-der-berliner-mauer-1989-wms,verlauf-der-berliner-mauer-1989-wms-bc24fb23
```

In the usage data, there is:

```
...
    "2024-10": {
      "verlauf-der-berliner-mauer-1989-wms-bc24fb23": {
        "impressions": 247,
        "visits": 210
      },
      ...
      "verlauf-der-berliner-mauer-1989-wms": {
        "impressions": 1,
        "visits": 1
      },
...
```

These two entries have been combined to:

```
...
    "2024-10": {
      "verlauf-der-berliner-mauer-1989-wms-bc24fb23": {
        "impressions": 248,
        "visits": 211
      },
...
```

License

Dataset URL: https://daten.berlin.de/datensaetze/zugriffsstatistik-daten-berlin-de

This page was generated from the github repository at https://github.com/berlinonline/berlin_dataportal_usage.

Last changed: 2024-12-17