wikiviews—A Stata interface for the Wikipedia API

Ulrich Kohler [1]

[1] University of Potsdam
Faculty of Economic and Social Sciences

2021 German Stata Conference
June 25th 2021
Hosted by the University of Potsdam

Aim of presentation

- Introducing Stata command `wikiviews`.
- `wikiviews` is a Stata frontend to the Wikimedia REST API; see `https://www.mediawiki.org/wiki/Wikimedia_REST_API`.
- The Wikimedia REST API offers access to Wikimedia's content and metadata in machine-readable formats
- `wikiviews` is designed to create Stata data from the responses of the API.
- Presentation describes the command with some entertaining examples of possible usages.

Technical Background

- The API is accessed through the following URL:

  `https://wikimedia.org/api/rest_v1/metrics/pageviews/per-article`

- Specific requests are created by defining an endpoint using the following elements:
  - Language, e.g. en, de, es, ...
  - Project, e.g. wikipedia, mediawiki, ...
  - Access, i.e. desktop, mobile-app, mobile-web, or all
  - Agent, i.e. user, spider, automated, or all
  - Page name, i.e. the name of a project page (e.g. “Günter_Netzer”)
  - Granularity, i.e. daily or monthly
  - Start, i.e. the start date
  - End, i.e. the end date

API in a browser

Copy/Pasting the endpoint URL
`https://wikimedia.org/api/rest_v1/metrics/pageviews/per-article/en.wikipedia/all-access/all-agents/Günter_Netzer/daily/2021062300/2021062400`
into the address field of a browser, returns the string


```
{"items":[{"project":"en.wikipedia","article"
:"Günter_Netzer","granularity":"daily","timestamp"
:"2021062300","access":"all-access","agent"
:"all-agents","views":266}]}
```


`wikiviews` creates endpoint(s) from the user input, sends them to the API and load the API's returns into Stata.

Extention of the API's history

- The API works for page accesses since July 1st 2015.
- For statistics earlier than July 1st 2015, `wikiviews` accesses `wikipediatrends`, a data base provided by (Meissner, 2020).
- Similar funtionality, but a bit less flexible, and only Wikipedia.
- See `https://petermeissner.de/blog/2019/10/09/wikipediatrend-v2.1.4/`

Syntax of wikiviews

`wikiviews per-article` *`anything`* `[, clear`
`access(`*`string`*`) agent(`*`string`*`) dumpdir(`*`string`*`)`
`end(`*`string`*`) granularity(`*`string`*`) language(`*`string`*`)`
`project(`*`string`*`) start(`*`string`*`) ]`

anything is the page name, and `dumpdir()` a directory to which the API's responses are being stored. The other options refer to the elements of the endpoint URL.

`wikiviews regex` *`string`* `[, clear dumpdir(`*`string`*`) ]`

string is search term in regular expression syntax and `dumpdir()` a directory to which the API's responses are being stored.

Most simple case

The most simple case refers to the page views of the given page in the English Wikipedia for the most recent full month from all agents and access types:

```
. wikiviews per-article "Günter_Netzer", clear
. list
```

1.

date 01may2021	count 3869	language en	project wikipedia	access all-access
agent all-agents	article Günter_Netzer	granul~y monthly	start 2021050100	end 2021060100

Accessing more than one page

Several pages can be accessed by enlisting page names

```
. wikiviews per-article "Günter Netzer" "Franz Beckenbauer", clear
. list article count
```

	article	count
1.	Günter Netzer	3869
2.	Franz Beckenbauer	44849

or by means of a string variable containing page names:

```
. clear
. input str20 players
                players
  1. "Günter Netzer"
  2. "Franz Beckenbauer"
  3. end
. wikiviews per-article players, clear
. list article count
```

	article	count
1.	Günter Netzer	3869
2.	Franz Beckenbauer	44849

Time series

Options `start()`, `end()` and `granularity()` can be used to fine tune the observation window:

`start(<year><month><day>)` defines the begin of the observation period. It defaults to yesterday morning or the start of the previous month, depending on `granularity()`. Start dates before December 9th 2007 are set to `20071209`.

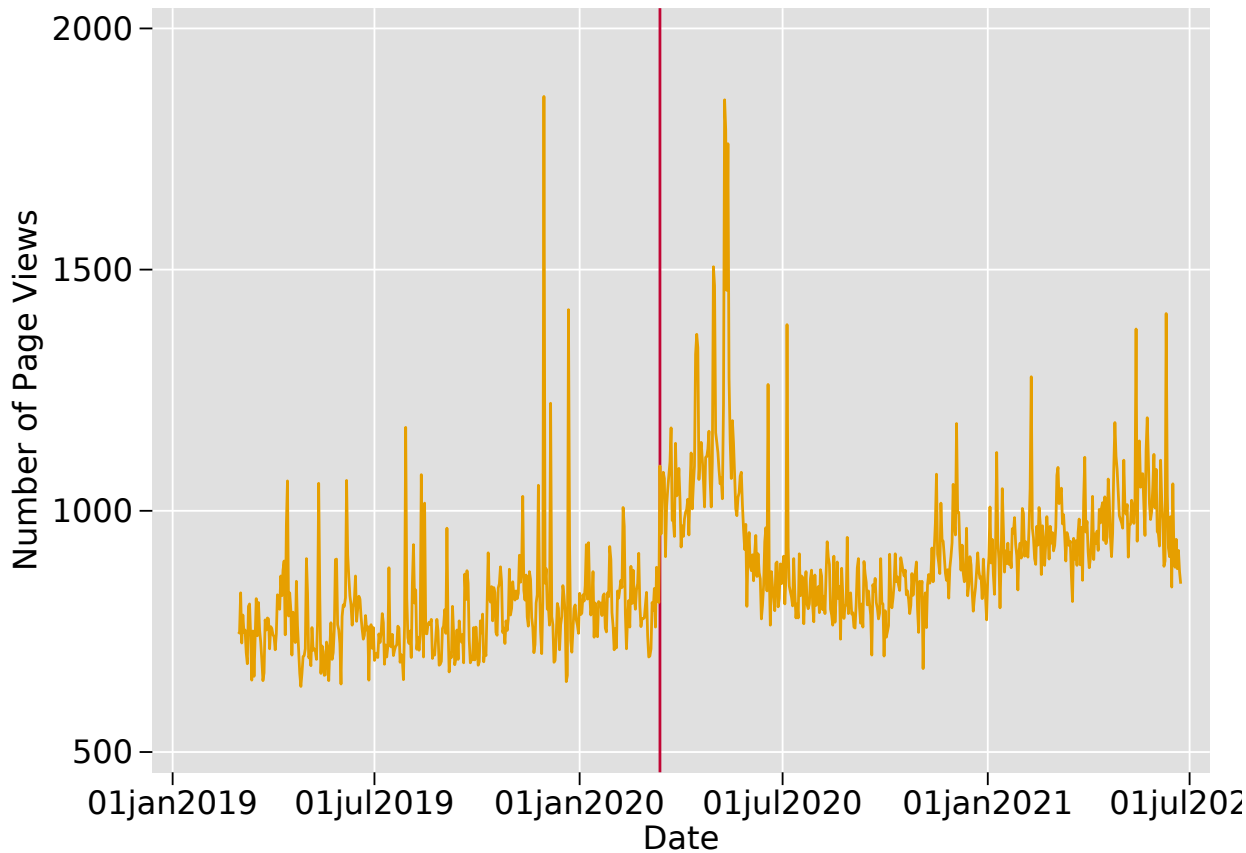
`end(<year><month><day>)` defines the end of the observation period. It defaults to today's morning or the end of the previous month, depending on `granularity()`.

`granularity(`*`daily|monthly`*`)` is either monthly (default) or daily.

The “Bazooka-Effect”

The following example shows the page views of the term “Bazooka” before and after a press conference of the German Minister of Finance, who claimed to take out a “Bazooka” to help German companies in the Corona pandemic.

```
. wikiviews per-article "Bazooka", clear s(20190301) g(daily)
. line count date, xline(`=date("20200313","YMD")´) sort
```



International comparision

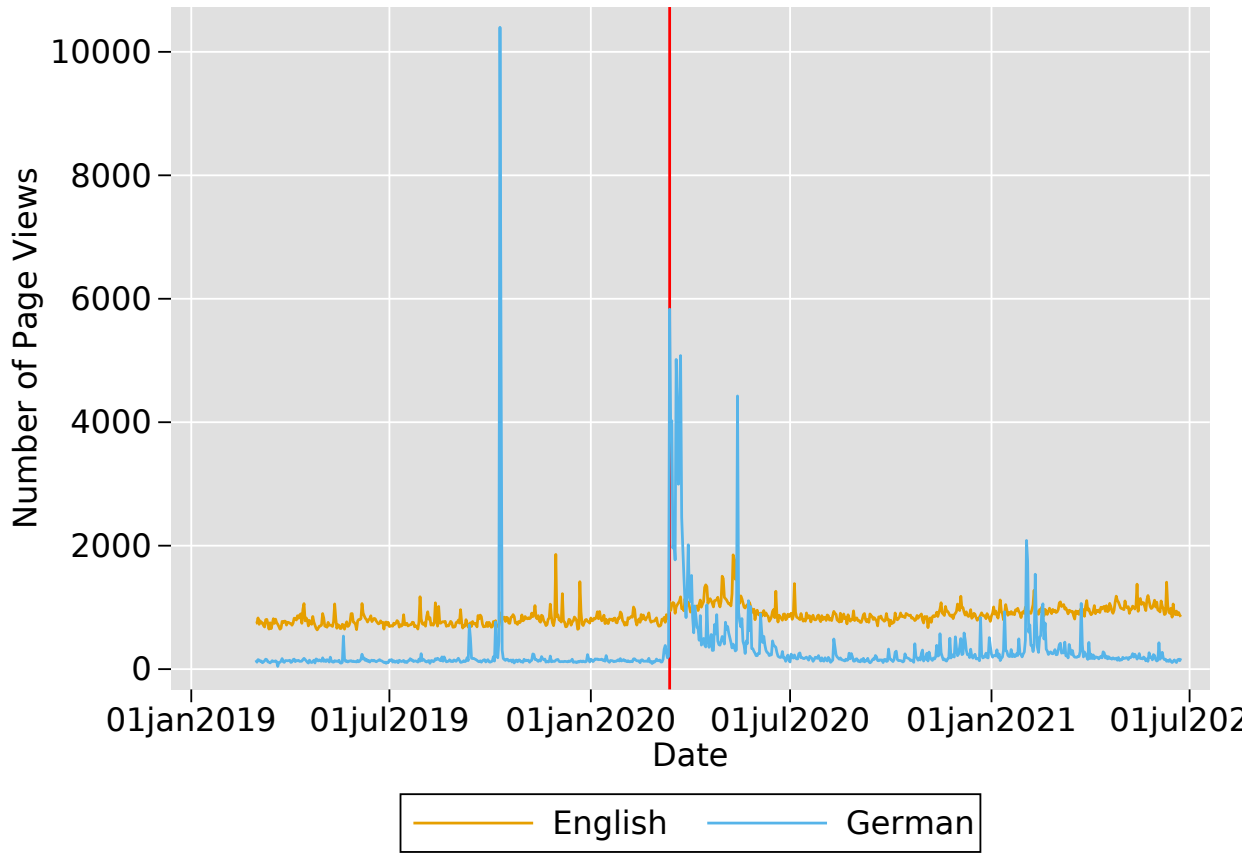
Option `language()` can be used to download page views for Wikipedia projects in other languages than English.

`language(`*`string`*`)` defaults to `en` for English. Other languages can be specified in terms of the Wikipedia language code[1]. The option allows to list several languages.

[1] See `https://meta.wikimedia.org/wiki/Template:List_of_language_names_ordered_by_code`

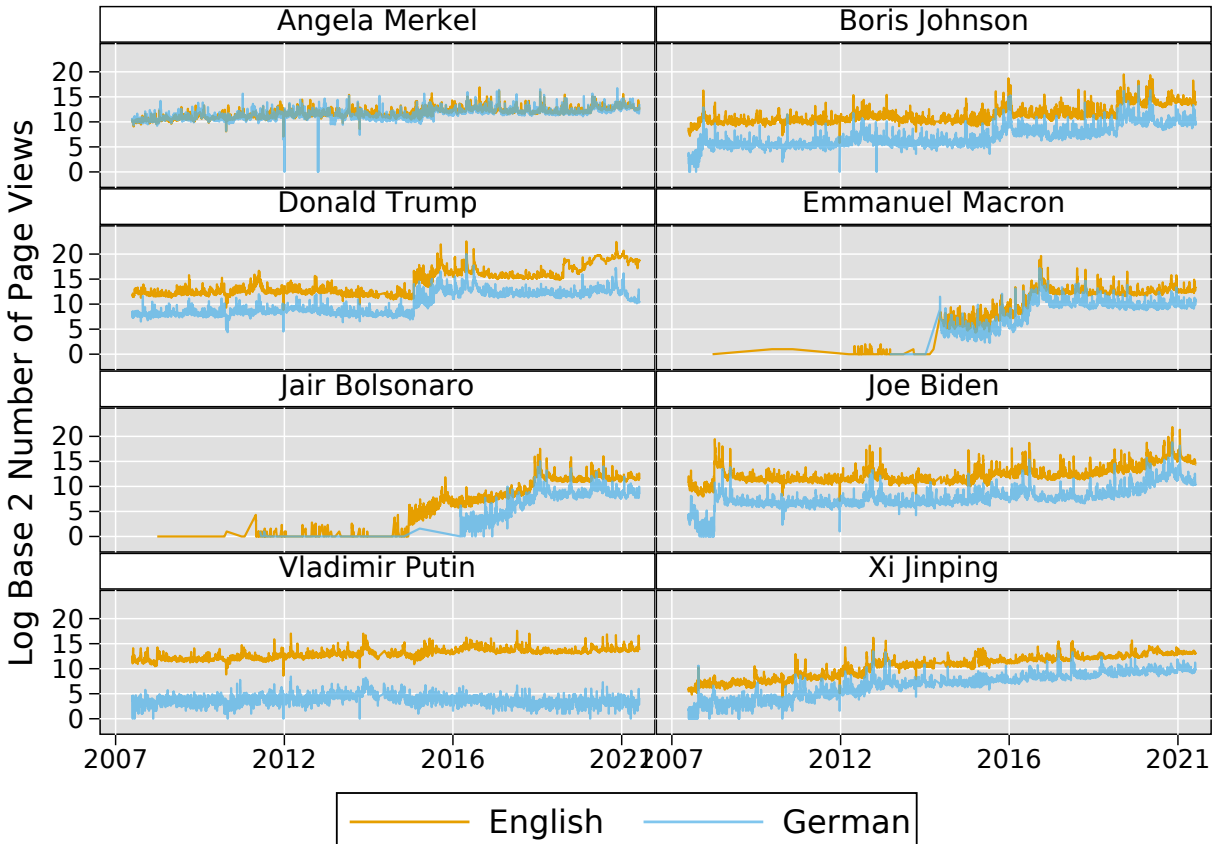
The Bazooka-Effect, reloaded

```
. wikiviews per-article Bazooka, clear l("de" "en")
  s(20190301) g(daily)
. graph tw
  || line count date if language=="en", sort
  || line count date if language=="de", sort
  || , xline(`=date("20200313","YMD")´, lcolor(red))
  legend(order(1 "English" 2 "German"))
```



A more serious application

```
. use ../somenames, clear
. wikiviews per-article name, l("de" "en") start(20071209) g(daily)
. gen logcount = log(count)/log(2)
. format %tdCCYY date
. graph twoway
  || line logcount date if language == "en", sort
  || line logcount date if language == "de", sort lcolor(%75)
  || , by(article, rows(4) compact note(""))
  legend(order(1 "English" 2 "German")) xtitle("")
  ylab(0(5)20) ytitle(Log Base 2 Number of Page Views)
```



The dumpdir option

- `wikiviews` stores the API's responses into the directory `wikiviews-dump` in the working directory.
- For requests already made, API's responses are taken from the files in `wikiviews-dump`.
- Location of the dumpdir directory can be changed using option `dumdir()`

Note: The dumpdir directory is being stored permanently in order to reduce the number of API requests. Please respect that policy.

Other options

The following options are only used for offical Wikimedia's REST API. They do not affect requests on page views before July 1st 2015.

`access(`*`string`*`)` Whether access is by `desktop`, `mobile-app`, or `mobile-web`. Default is `all-access`.

`agent(`*`string`*`)` Whether agent is a `user`, a `spider`, or `automated`. Default is `all-agents`.

`project(`*`string`*`)` Whether to download page views of `wikipedia.org`, `www.mediawiki.org`, or `commons.wikimedia.org`. Default is `wikipedia.org`.

The regex subcommand

- `wikiviews per-article` only works for known page names.
- `wikiviews regex` can be used to search for page names that match a regular expression.
- `wikiviews regex` relies on Peter Meissners Wikipediatrends API. It does not match names of pages created after July 1st 2015.

Regex usage example

```
. wikiviews regex `".+Beckenbauer.*"´, l(de) clear
. wikiviews per-article pagename2, l(de) start(20080101) clear
. tab article, sum(count)
```

Title of article	Summary of Number of Page Views Mean	Std. dev.	Freq.
Alfons Beckenbauer	333.51087	587.45382	92
Alfons Beckenbauer (Historiker)	42.928571	28.894074	56
Franz Anton Beckenbauer	10.395683	5.4328826	139
Franz Beckenbauer	29595.743	20563.495	167
Franz Beckenbauer (Begriffsklär..	109.78808	78.141036	151
Franz-Beckenbauer-Cup	28.905983	54.949347	117
Franz-Beckenbauer-Stiftung	196.16766	120.95821	167
Stefan Beckenbauer	329.61677	676.07017	167
Stephan Beckenbauer	5790.0658	22274.86	152
Total	4937.8121	14878.205	1,208

Acknowledgements

- I owe the inspiration to start working on `wikiviews` to Lena Hipp and the participants of summer's 2019 seminar on social sciences with big data.
- Graphs in this presentation are designed using Ben Jann's `grstyle` package
- I whish to thank Peter Meissner for the provision of `wikipediatrends` and for quick responses to all may questions.

Bibliograpy I

Meissner, P. 2020. *wikipediatrend: Public Subject Attention via Wikipedia Page View Statistics*. R package version 2.1.6. URL `https://CRAN.R-project.org/package=wikipediatrend`