



Open knowledge in R with Wikimedia APIs

Follow along at `git.io/vSi6a`

Mikhail Popov

25 April 2017

Wikimedia Foundation

Wikimedia Foundation is a non-profit that operates free & open projects like **Wikipedia**, **Wiktionary**, and **Wikidata** that anyone can contribute to

No time to talk about me (plus that's always the boring part)¹

A Markdown copy of this deck is at git.io/vSi6a for following along

R packages required to follow along:

```
install.packages(  
  c("pageviews", "WikipediR", "WikidataR",  
    "WikidataQueryServiceR", "magrittr"),  
  repos = c(CRAN = "https://cran.rstudio.com")  
)
```

¹If you're **really** curious just search for **User:MPopov (WMF)** on **Meta-Wiki**

- Running R 3.4.0 on macOS Sierra 10.12.4
- Rendered with **rmarkdown** 1.4 and **knitr** 1.15.1
- The pipe (`%>%`) from **magrittr** is **occasionally** used
- Using the following versions of packages for demos:

Package	Version	Imports
pageviews	0.3.0	jsonlite, httr, curl
WikipediR	1.5.0	httr, jsonlite
WikidataR	1.2.0	httr, jsonlite, WikipediR, utils
WikidataQueryServiceR	0.1.0	httr, dplyr, jsonlite

Wikipedia is a free encyclopedia that anyone can edit

You may have heard of it

It is available in 296 languages

English Wikipedia has over 5.3 million articles

Wikipedia is powered by **MediaWiki**, which includes an **API** that makes it fast and easy to fetch content



WikipediR is a wrapper for MediaWiki API but aimed at Wikimedia's wikis such as Wikipedia. It can be used to retrieve page text, information about users or the history of pages, and elements of the category tree.

```
library(WikipediR)
r_wiki <- page_content(
  language = "en",
  project = "wikipedia",
  page_name = "R (programming language)"
)
r_releases <- r_wiki$parse$text$`*` %>%
  read_html %>%
  xml_find_first("//*[table[@class='wikitable']]") %>%
  rvest::html_table()
```

Release	Date	Description
0.16		This is the last alpha version developed...
0.49	1997-04-23	This is the oldest source release which ...
0.60	1997-12-05	R becomes an official part of the GNU Pr...
0.65.1	1999-10-07	First versions of update.packages and in...
1.0	2000-02-29	Considered by its developers stable enou...
1.4	2001-12-19	S4 methods are introduced and the first ...
2.0	2004-10-04	Introduced lazy loading, which enables f...
2.1	2005-04-18	Support for UTF-8 encoding, and the begi...
2.11	2010-04-22	Support for Windows 64 bit systems....
2.13	2011-04-14	Adding a new compiler function that allo...
2.14	2011-10-31	Added mandatory namespaces for packages....
2.15	2012-03-30	New load balancing functions. Improved s...
3.0	2013-04-03	Support for numeric index values 231 and...

- Use `language` and `project` arguments for Wikimedia's wikis²
- Use `domain` for everything else, such as:
 - Project Gutenberg's wiki:
`domain = "www.gutenberg.org/w/api.php"`
 - Mozilla Foundation's wiki:
`domain = "wiki.mozilla.org/api.php"`
 - Geek Feminism wiki:
`domain = "geekfeminism.wikia.com/api.php"`
 - A Wiki of Ice and Fire:
`domain = "awoiaf.westeros.org/api.php"`
- **Tip:** if using `random_page`, specify `namespaces = 0` to only get articles

²Currently: [Commons](#), [Wikivoyage](#), [Wikiquote](#), [Wikisource](#), [Wikibooks](#), [Wikinews](#), [Wikiversity](#), [Wikispecies](#), [MediaWiki](#), [Meta-Wiki](#), [Wiktionary](#)

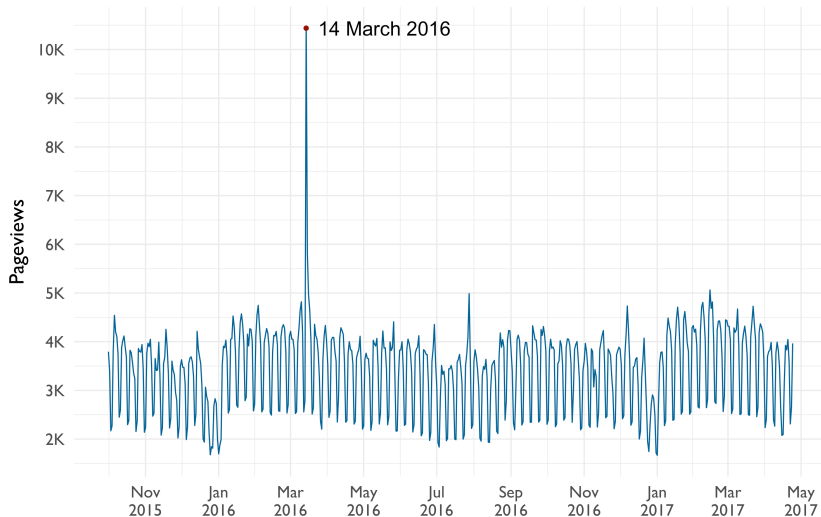
WMF provides an [API for accessing daily and monthly pageviews of any article on any project](#) for counts from 2015 onwards.³ The package `pageviews` allows you to get those counts in R:

```
library(pageviews)
r_pageviews <- article_pageviews(
  project = "en.wikipedia",
  article = "R (programming language)",
  user_type = "user", start = "2015100100",
  end = format(Sys.Date() - 1, "%Y%m%d00")
)
```

³[wikipediatrend](#) package wraps the [stats.grok.se](#) API which has historical Wikipedia pageview data for 2008 up to 2016 from [these pageview count dumps](#)

Daily pageviews of R's entry on English Wikipedia

Desktop and mobile traffic, excluding known bots



- **Wikidata** is a language-agnostic open knowledge base
- Facts are expressed as 3-part statements:
 - Subject (resource)
 - Predicate (property type)
 - Object (property value, can be another resource)
- Examples:
 - “R” (**Q206904**) is an “instance of” (**P31**) a “programming language” (**Q9143**)
 - “RStudio” (**Q4798119**) was “programmed in” (**P277**) “C++” (**Q2407**)
 - “Portland” (**Q6106**) had a “population” (**P1082**) of 583,776 (in 2010)
- Resources and properties have unique numeric identifiers but can have human-friendly labels in any language

```
library(WikidataR)
r_search <- find_item("R")[[8]]
r_search[c("id", "description")] # check the results

## $id
## [1] "Q206904"
##
## $description
## [1] "programming language for statistical computing"
```

```
property <- get_property("P31")  
property$labels$`en`$value # check that we want P31
```

```
## [1] "instance of"
```

```
r_item <- get_item(r_search$id)  
r_item$claims$P31$mainsnak$datavalue$value$id # extract values
```

```
## [1] "Q9143"      "Q341"      "Q20825628" "Q28920142" "Q3839507"  
## [7] "Q1993334"  "Q24529812"
```

This tells us that R is an instance of Q9143, Q341, Q20825628, Q28920142, Q3839507, Q12772052, Q1993334, Q24529812. Great?

- Allows querying Wikidata with **SPARQL**
- Provides a public SPARQL endpoint usable via:
 - Web front-end: query.wikidata.org
 - Web API
(`https://query.wikidata.org/sparql?query=<SPARQL>`)
 - In Python with **SPARQLWrapper**
 - In R with:
 - **SPARQL** package
 - **WikidataQueryServiceR**
- For useful reference links, see
`help("WDQS", package = "WikidataQueryServiceR")`

```
# PREFIXes are optional when using WDQS
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX bd: <http://www.bigdata.com/rdf#>
```

```
SELECT DISTINCT ?instanceOfLabel
WHERE {
    wd:Q206904 wdt:P31 ?instanceOf .
    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "en"
    }
}
```

```
library(WikidataQueryServiceR)
query_wikidata('SELECT DISTINCT ?instanceOfLabel
WHERE {
  wd:Q206904 wdt:P31 ?instanceOf .
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en"
  }
}') %>% head(n = 5L)
```

```
##           instanceOfLabel
## 1      programming language
## 2           free software
## 3 multi-paradigm programming language
## 4           interpreted language
## 5      functional programming language
```

Advanced SPARQL Example

- Prefix **wd**: points to an entity
- Prefix **p**: points not to the object, but to a statement node
- Prefix **ps**: within the statement node retrieves the object (value)
- Prefix **pq**: within the statement node retrieves the qualifier info

```
r_versions_query <- "SELECT DISTINCT
  ?softwareVersion ?publicationDate
WHERE {
  BIND(wd:Q206904 AS ?R)
  ?R p:P348 [
    ps:P348 ?softwareVersion;
    pq:P577 ?publicationDate
  ] .
}"
```

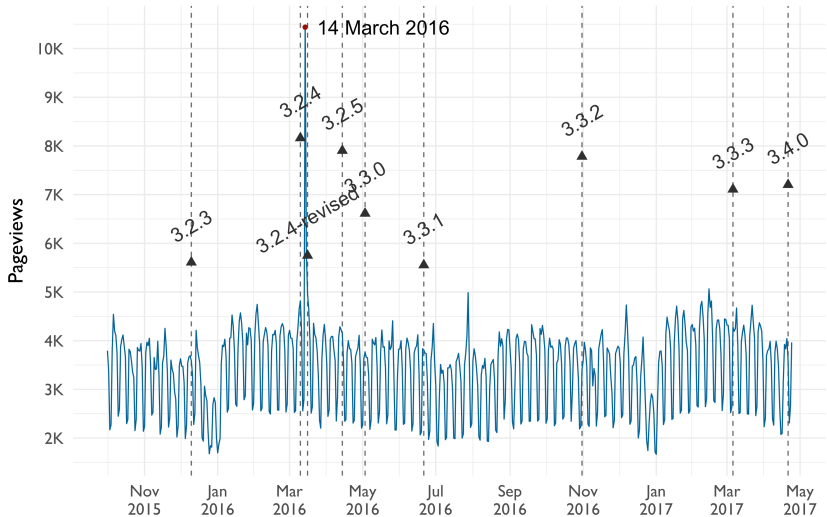


```
r_versions_results <- query_wikidata(r_versions_query)
```

softwareVersion	publicationDate
1.0.0	2000-02-29
2.0.0	2004-10-04
2.15.3	2013-03-01
...	...
3.3.2	2016-10-31
3.3.3	2017-03-06
3.4.0	2017-04-21

R's releases and article pageviews on English Wikipedia

Desktop and mobile traffic, excluding known bots



Source for the whole shebang is up on GitHub: [bearloga/wmf](#), available under [CC BY-SA 4.0](#)

Specifically: [wmf/presentations/talks/Cascadia R Conference 2017/](#)

Contact Info

- Twitter: [bearloga](#)
- WMF-related: mikhail@wikimedia.org
(PGP public key: people.wikimedia.org/~bearloga/public.asc)
- General: mikhail@mpopov.com
(PGP public key on keybase.io/mikhailpopov)