MZES Social Science Data Lab

legislatoR

Political, sociodemographic, and Wikipedia-related data on political elites

Sascha Göbel
University of Konstanz

Simon Munzert
Hertie School of Governance

May 6, 2019

Motivation

Motivation

Why a data package on political elites?

- **Continued demand** for individual-level data by researchers, students, analysts, and journalists
- **Status quo:** recurrent data collection with the same purpose is **inefficient and restricting**
- **Existing data** structures **limited** in scope, hidden **behind paywalls**, or **difficult to access**

Motivation

Why a data package on political elites?

- **Continued demand** for individual-level data by researchers, students, analysts, and journalists
- **Status quo:** recurrent data collection with the same purpose is **inefficient and restricting**
- **Existing data** structures **limited** in scope, hidden **behind paywalls**, or **difficult to access**

Why use & provide Wikipedia data?

- Contains archives full of **politicians' biographies**
- Widely employed and primary **web information source**
- Often deemed **neutral and trustworthy**

Motivation

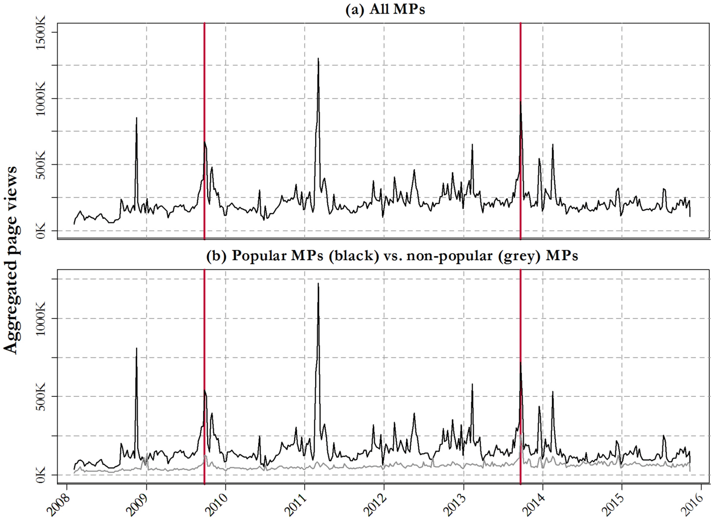


Figure: Page views of German MPs' Wikipedia entries.

Motivation

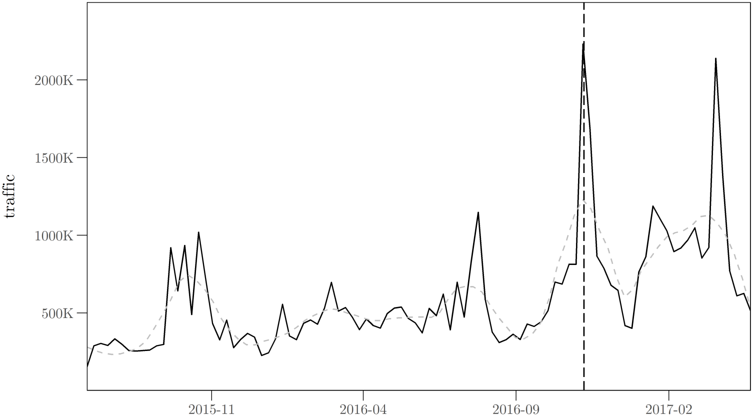


Figure: Page views of US Representatives' Wikipedia entries.

Motivation

legislatoR

- Resource **efficient**: one stop shop for broad and deep data
- Easily **accessible** from R with simple function calls
- Facilitates data **integration** and **replication** efforts
- It’s **free**!

Content and Structure

Content and data structure

Content

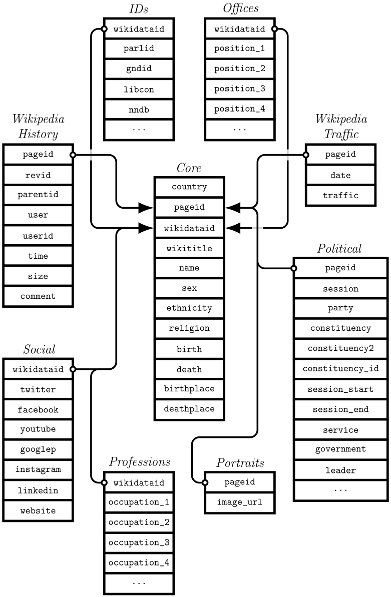
- Currently **32,533 elected politicians**, all sessions of parliamentary chambers of Austria, Canada, the Czech Republic, France, Germany, Ireland, Scotland, UK, and the US (House and Senate)

Content and data structure

Content

- Currently **32,533 elected politicians**, all sessions of parliamentary chambers of Austria, Canada, the Czech Republic, France, Germany, Ireland, Scotland, UK, and the US (House and Senate)
- **nine datasets** for each legislature:
  1. *Core* (basic sociodemographic data)
  2. *Political* (basic political data)
  3. *History* (full revision records of Wikipedia biographies)
  4. *Traffic* (daily user traffic on Wikipedia biographies)
  5. *Social* (social media handles and personal website URLs)
  6. *Portrait* (URLs to Wikipedia portraits, facial recognition estimates)
  7. *Office* (public offices)
  8. *Profession* (professions)
  9. *IDs* (identifiers linking to another file, database, or website)

Content and data structure



IDs

wikidataid
parlid
gndid
libcon
nndb
...

Offices

wikidataid
position_1
position_2
position_3
position_4
...

Wikipedia History

pageid
revid
parentid
user
userid
time
size
comment

Core

country
pageid
wikidataid
wikititle
name
sex
ethnicity
religion
birth
death
birthplace
deathplace

Wikipedia Traffic

pageid
date
traffic

Political

pageid
session
party
constituency
constituency2
constituency_id
session_start
session_end
service
government
leader
...

Social

wikidataid
twitter
facebook
youtube
googlep
instagram
linkedin
website

Professions

wikidataid
occupation_1
occupation_2
occupation_3
occupation_4
...

Portraits

pageid
image_url

Content and data structure

Structure

- **relational** Individual-level data – all datasets joinable with *Core* dataset via two keys (Wikipedia page or Wikidata ID)

Content and data structure

Structure

- **relational** Individual-level data – all datasets joinable with *Core* dataset via two keys (Wikipedia page or Wikidata ID)

Sources

- **Automated** data extraction (XPath, Web APIs)
- Face++ Cognitive Services API, Wikimedia Commons, Wikidata API, Wikipedia, Wikipedia API

Content and data structure

Structure

- **relational** Individual-level data – all datasets joinable with *Core* dataset via two keys (Wikipedia page or Wikidata ID)

Sources

- **Automated** data extraction (XPath, Web APIs)
- Face++ Cognitive Services API, Wikimedia Commons, Wikidata API, Wikipedia, Wikipedia API

Issues

- Dependence on data coverage of sources, collaborative effort (but that's a feature, not a bug!)
- In part substantial amount of missings

Usage

Usage

Installation

```
> # install from GitHub
> devtools::install_github("saschagobel/legislatoR")
> # load and attach
> library(legislatoR)
```

Usage

Installation

```
> # install from GitHub
> devtools::install_github("saschagobel/legislatoR")
> # load and attach
> library(legislatoR)
```

For up-to-date information, check out
https://github.com/saschagobel/legislatoR