Accessing SUL Provided Social Science Data with R

Online at: bit.ly/sul-db-R

Last updated: June 27, 2017 – cengel @ stanford

The Stanford Libraries offer a number of datasets for social science research. This is a document detailing if and how those can be accessed when using R. It is work in progress. (Download as pdf.)

Data providers marked with **(*)** are presenters at the *Social Science Gear Up 2016 event.*

Data from SUL

- Bloomberg **(*)**
  - Restricted to designated machines at GSB
  - Details here: http://libguides.stanford.edu/bloomberg
- Bureau of Economic Analysis
  - has API
  - R package announced for late Oct 2016
- California Attorney General's Open Justice Portal
  - Use `read.csv(url("http:..."))` to load *CSV* from their data portal.
- Census public/RDC **(*)**
  - Public census has APIs
  - RDC restricted to designated, offline machines
  - R packages:
    - acs (some instructions here)
    - UScensus2010
    - tigris
    - tidycensus: Load US Census Boundary and Attribute Data as `tidyverse` and `sf`-Ready Data Frames
  - References, slighty outdated:
    - Getting data from the ACS into R
    - Performing spatial regression modeling in R with ACS data
- CoreLogic **(*)**
  - Stanford only, download *XLS* from Stanford Digital Repository
  - **need to read End User License Agreement before use**
- DataPlanet **(*)**
  - Stanford only via EzProxy and export *CSV*
  - Supposedly has API(?), but request form is not accessible
- Gallup Analytics **(*)**
  - Stanford only, via EzProxy
  - Data download through web interface as *XLSX* only – **broken or disabled(?)**
- ICPSR **(*)**
  - R package `icpsrdata` (GitHub)
- IMF **(*)**
  - Has API
  - R packages:
    - `IMFData` (GitHub)
    - `rWEO` for IMF-WEO data
- MapLight **(*)**
  - Bill Positions has API
  - Some R code from Hadley Wickham (not a package!). Last commit several years back, but still seems to work.

  - Download California Money and Politics Bulk Data Set and Federal Money and Politics Data Set as zipped *CSV*, like this.
- OECD
  - R package `OECD` (GitHub)
- ProQuest Statistical Products
  - Stanford only, via EzProxy
  - Data download as *XLS* (or *PDF*)
- RefUSA
  - Stanford only, via GSB
  - Data download as *CSV*, *TSV*, *Excel* (275 downloads per search)
- Roper iPoll
  - Stanford only, via EzProxy
  - Data download seems tedious. I was only able to download *CSV* for a single question at a time.
- San Mateo County Open Data
  - Use generic Socrata Open Data API
  - R package `RSocrata` (GitHub)
- SimplyMap
  - Stanford extended access via http://simplymap.com/ – need to be either on campus, or use VPN or set up browser proxy. 5 concurrent licenses.
  - Manual data download, no remote access (guest account allows to download *CSV*, Stanford access allows to download *SHP* as well)
- World Bank **(*)**
  - Currently 60 databases
  - Has API
  - R packages:
    * `WDI` (GitHub)
    * `wbstats`
    * `rWBData`
    * `rWBclimate` for the World Bank climate data
- Wharton Research Data Services (WRDS) **(*)**
  - Stanford login via GSB
  - Data access via remote connection to SAS/SHARE server, which allows direct query of WRDS data via standard database queries, using their (remote) R version.
  - Documentation:
    * WRDS Data Directly from Python, R, and MATLAB
    * Using R with WRDS
  - Sparsely documented R package `wrds`

Other R packages

- US Bureau of Labor Statistics (BLS):
  - `rUnemploymentData`
  - `blscrapeR`
- DataCite: `rdatacite`
- Medicare public files: `medicare`
- Social Media for Network Analysis: `SocialMediaLab`
- United States Treasury
  - `Rtreasuryio`: a single, simple function for submitting SQL queries to `treasury.io`
- `enigma`: a client to interact with the Enigma API, including getting the data and metadata for datasets as well as collecting statistics on datasets. (Note that there is another site: Enigma Public "the world's broadest collection of public data" which provides API access as well, not sure how the two are related with regard to this package.)