Accessing Social Science Data with R

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

Last updated: December 13, 2020 cengel @ stanford

The Stanford Libraries (SL) offer a number of datasets for social science research. This is a document detailing if and how those can be accessed when using R. I am also adding packages tp access social science data beyond Stanford. It is work in progress. (Download as pdf.)

Data from SL

- 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: Loads US Census boundary and attribute data as `tidyverse` and `sf`-ready data frames
    - totalcensus: Download summary files and extract data at block, block group, and tract level.
    - censusapi: A wrapper for the US Census Bureau APIs that returns data frames of Census data and metadata.
    - idbr: US Census Bureau's International Data Base
  - References, slighty outdated: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 and sign 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
- IPUMS
  - R package `ipumsr` to import census, survey and geographic data provided by IPUMS into R. Great vignette, also for Value labels, Current Population Survey (CPS), Geographic Data, and use of NHGIS.
- 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)
- PolicyMap
  - Stanford access via https://stanford.policymap.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)
- Washington Post
  - Full text from 1968, though with varying coverage.
  - To request access see the searchworks link.
- World Bank
  - Currently 60 databases
  - Has API
  - R packages:
    * `WDI` (GitHub) - (Tutorial)
    * `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`
- `citecorp`: Client for the Open Citations Corpus
- `tradestatistics` R package to use the Trade Statistics API
- `rdryad` is a package to interface with the Dryad data repository API.
- `rplos`: Interface to the Search API for PLoS (Public Library of Science) Journals
- `rbace`: 'Bielefeld' Academic Search Engine ('BASE') Client
- 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.)
- `pdfetch`: Economic and financial time series from public sources, including the St Louis Fed's FRED system, Yahoo Finance, the US Bureau of Labor Statistics, the US Energy Information Administration, the World Bank, Eurostat, the European Central Bank, the Bank of England, the UK's Office of National Statistics, Deutsche Bundesbank, and INSEE.
- `fredr`: An R client for the Federal Reserve Economic Data (FRED) API.
- `eechidna`: 2013 and 2016 Australian Federal Election (House of Representatives) and the 2011 Australian Census
- `rtimes`: Interface to Congress, Campaign Finance, Article Search, and Geographic APIs from the New York Times and ProPublica. Covers only a subset of the APIs.
- `crminer` and `rcrossref`: Text mining client for Crossref. Includes functions for getting getting links to full text of articles, fetching full text articles from those links or Digital Object Identifiers (DOIs), and text extraction from PDFs. `rcrossref` is for metadata.
- `rdpla`: Client for the Digital Public Library of America (DPLA), using its REST API
- `internetarchive`: Search the Internet Archive, retrieve metadata, and download files.
- `patentsview`: An R Client for PatentsView with functions to simplify the PatentsView API query language and parse the data that comes back.
- `pleiades`: Interface to the Pleiades Archeological Database
- `USAboundaries`: Historical boundaries of the United States. Map the United States (or the colonies that became the United States) on any date from 1629 to 2000. Contains both county and state/territory level polygons.
- `rdatacite` Client for the web service methods provided by DataCite
- `roadoi`: Find Free Versions of Scholarly Publications via Unpaywall
- `data.world`: High-level tools for working with data.world data sets.

- `jstor`: Import journal data from DfR (JSTOR)
- `fulltext`: A single interface to full text sources 'scholarly' data, including 'Biomed Central', Public Library of Science, 'Pubmed Central', 'eLife', and more. (Manual)
- `qualtRics`: Convenience functions to pull survey results straight into R using the Qualtrics API. (**package archived 2018-02-05**)
- `essurvey`: Download data from the European Social Survey.
- `RefManageR`: Import and work with bibliographic references. Stores with `BibTeX` and `BibLaTeX` references, interfaces with `NCBI Entrez`, `CrossRef`, and `Zotero`, extracts references from locally stored PDF and generates bibliographies for `RMarkdown`.
- `nomisr`: UK official statistics from the 'Nomis' database.Includes Census, Labour Force Survey, DWP benefit statistics and other economic and demographic data from the Office for National Statistics.
- `fingertipsR`: Public health indicators in England
- `Quandl`: Access to financial, economic, and alternative datasets from Quandl. (Documentation)
- `quantmod`: Quantitative Financial Modelling & Trading Framework. (Documentation)
- `tidyquant`: A wrapper to various 'xts', 'zoo', 'quantmod', 'TTR' and 'PerformanceAnalytics' package functions that returns the objects in the tidy 'tibble' format.
- `rdhs`: Management and analysis of Demographic and Health Survey (DHS) data.
- `refnet`: Read, organize, geocode, analyze, and visualize Clarivate Web of Knowledge/Web of Science, format reference data files for scientometric, social network, and Science of Science analyses. Not on CRAN.
- geospatial data:
  - `bikedata` data from public hire bicycle systems,including London, New York, Chicago, Washington DC, Boston, Los Angeles, and Philadelphia.
  - `FedData`: Download geospatial Data from federated data sources, including the The National Elevation Dataset digital elevation models, the Global Historical Climatology Network, the National Land Cover Database, and more.
  - `getlandsat`: Get Landsat 8 Data from Amazon Public Data Sets
  - `geonames`: Interface to the "Geonames" Spatial Query Web Service
  - `MODIStsp`: automates the creation of time series of rasters derived from MODIS Land Products data
  - `stats19` Open Road Traffic Casualty Data from Great Britain
- `rdataretriever`: Provides an R interface to the Data Retriever via the Data Retriever's command line interface. The Data Retriever automates the tasks of finding, downloading, and cleaning public datasets, and then stores them in a local database.
- data wrangling
  - `naniar`: explore missing values
  - `visdat`: visualise a dataframe (http://visdat.njtierney.com)

(A number of packages are from https://ropensci.org/. You may want to check there for new ones.)

- misc
  - `archivr` Automated preservation of urls in Web Archives (More)