Access to official statistics from R: an overview

Statistics Netherlands

Olav ten Bosch, Edwin de Jonge

uRos2023 12-14 December 2023

Contents

- What is the awesome list?
  - History, concept, status

2018

2019

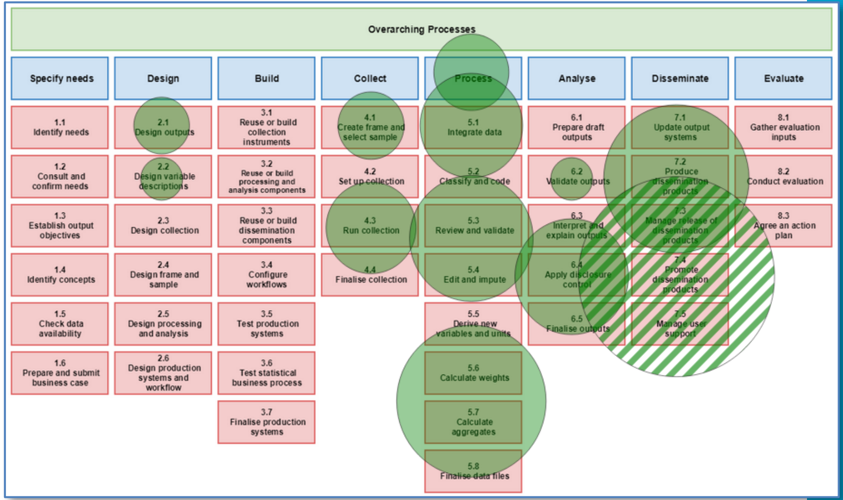
- Category “access to official statistics”
  - Packages, data providers, standards, features
  - What are potential improvements in this software landscape?

- Wrap-up

Awesome list of official statistics software

- When: started at ***UNECE SDE conference*** april 2017 (The Hague)
- Why: to ***collectively remember useful software*** in official statistics
- Who: maintained by ***statistical community***
- What: a ***community approach*** to knowledge management
- How:
  - Using ***awesome concept***
  - A ***public*** list of awesome things
  - Clear and simple ***criteria***
  - ***awesomeofficialstatistics.org***

# awesome packages by GSBPM



Overarching Processes

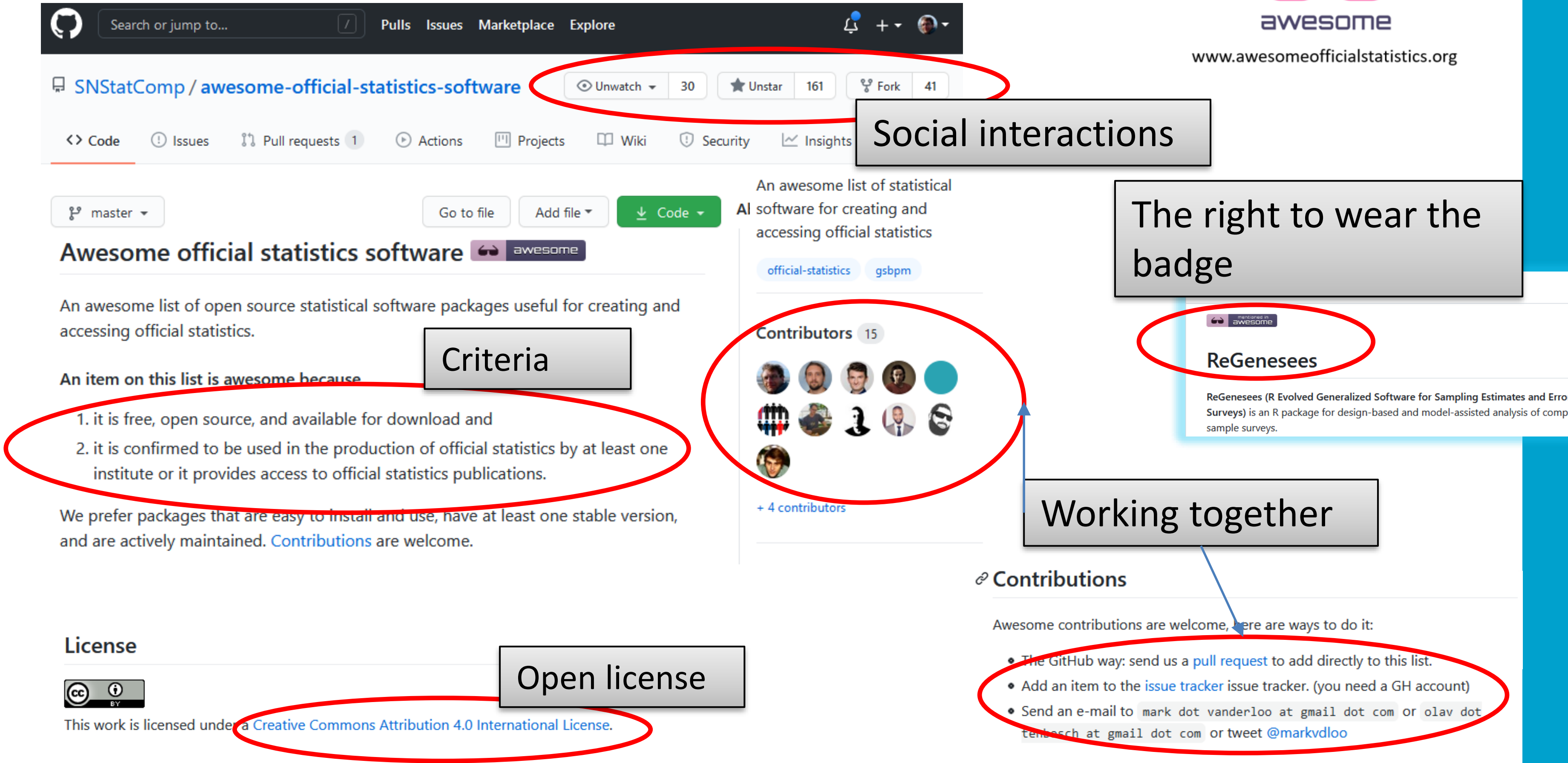
Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

What is the awesome list?

Curated list of software for official statistics

awesome

www.awesomeofficialstatistics.org



SNStatComp / awesome-official-statistics-software

Awesome official statistics software awesome

An awesome list of open source statistical software packages useful for creating and accessing official statistics.

An item on this list is awesome because

1. it is free, open source, and available for download and
2. it is confirmed to be used in the production of official statistics by at least one institute or it provides access to official statistics publications.

We prefer packages that are easy to install and use, have at least one stable version, and are actively maintained. Contributions are welcome.

License

This work is licensed under a Creative Commons Attribution 4.0 International License.

About

An awesome list of statistical software for creating and accessing official statistics

official-statistics gsbpm

Contributors 15

+ 4 contributors

mentioned in awesome

ReGenesees

ReGenesees (R Evolved Generalized Software for Sampling Estimates and Erro Surveys) is an R package for design-based and model-assisted analysis of comp sample surveys.

Contributions

Awesome contributions are welcome, here are ways to do it:

- The GitHub way: send us a pull request to add directly to this list.
- Add an item to the issue tracker issue tracker. (you need a GH account)
- Send an e-mail to `mark dot vanderloo at gmail dot com` or `olav dot tenbosch at gmail dot com` or tweet @markvdloo

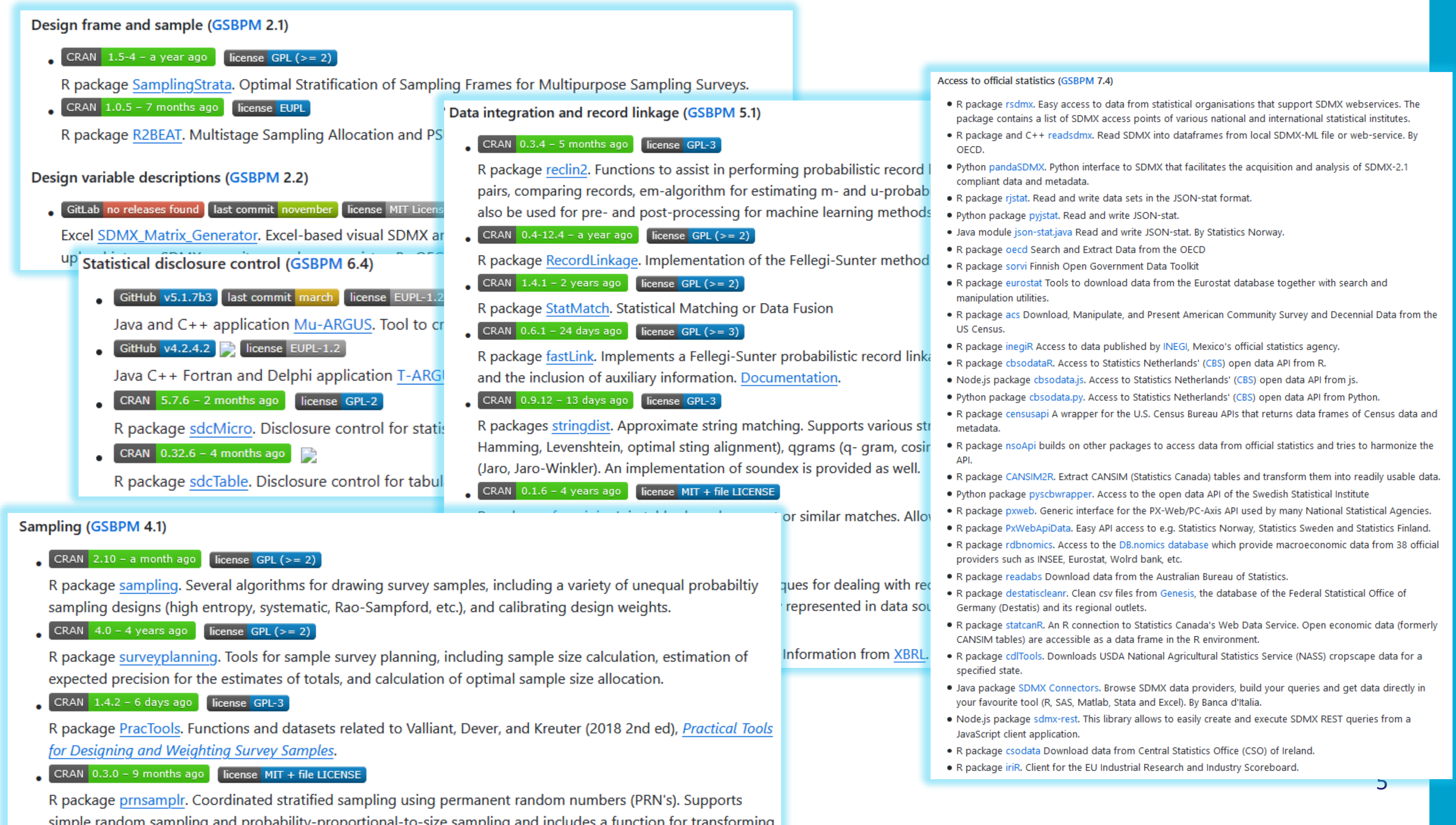
Social interactions

The right to wear the badge

Criteria

Working together

Open license



Design frame and sample (GSBPM 2.1)

- CRAN 1.5-4 – a year ago license GPL (>= 2)

  R package SamplingStrata. Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys.
- CRAN 1.0.5 – 7 months ago license EUPL

  R package R2BEAT. Multistage Sampling Allocation and PS

Design variable descriptions (GSBPM 2.2)

- GitLab no releases found last commit november license MIT Licens

  Excel SDMX_Matrix_Generator. Excel-based visual SDMX ar

  up

Statistical disclosure control (GSBPM 6.4)

- GitHub v5.1.7b3 last commit march license EUPL-1.2

  Java and C++ application Mu-ARGUS. Tool to cr
- GitHub v4.2.4.2 license EUPL-1.2

  Java C++ Fortran and Delphi application T-ARGU
- CRAN 5.7.6 – 2 months ago license GPL-2

  R package sdcMicro. Disclosure control for statis
- CRAN 0.32.6 – 4 months ago

  R package sdcTable. Disclosure control for tabul

Data integration and record linkage (GSBPM 5.1)

- CRAN 0.3.4 – 5 months ago license GPL-3

  R package reclin2. Functions to assist in performing probabilistic record l
  pairs, comparing records, em-algorithm for estimating m- and u-probab
  also be used for pre- and post-processing for machine learning methods
- CRAN 0.4-12.4 – a year ago license GPL (>= 2)

  R package RecordLinkage. Implementation of the Fellegi-Sunter method
- CRAN 1.4.1 – 2 years ago license GPL (>= 2)

  R package StatMatch. Statistical Matching or Data Fusion
- CRAN 0.6.1 – 24 days ago license GPL (>= 3)

  R package fastLink. Implements a Fellegi-Sunter probabilistic record linka
  and the inclusion of auxiliary information. Documentation.
- CRAN 0.9.12 – 13 days ago license GPL-3

  R packages stringdist. Approximate string matching. Supports various str
  Hamming, Levenshtein, optimal sting alignment), qgrams (q- gram, cosir
  (Jaro, Jaro-Winkler). An implementation of soundex is provided as well.
- CRAN 0.1.6 – 4 years ago license MIT + file LICENSE

or similar matches. Allo

ques for dealing with re
represented in data sou

Information from XBRL.

Sampling (GSBPM 4.1)

- CRAN 2.10 – a month ago license GPL (>= 2)

  R package sampling. Several algorithms for drawing survey samples, including a variety of unequal probabiltiy sampling designs (high entropy, systematic, Rao-Sampford, etc.), and calibrating design weights.
- CRAN 4.0 – 4 years ago license GPL (>= 2)

  R package surveyplanning. Tools for sample survey planning, including sample size calculation, estimation of expected precision for the estimates of totals, and calculation of optimal sample size allocation.
- CRAN 1.4.2 – 6 days ago license GPL-3

  R package PracTools. Functions and datasets related to Valliant, Dever, and Kreuter (2018 2nd ed), *Practical Tools for Designing and Weighting Survey Samples*.
- CRAN 0.3.0 – 9 months ago license MIT + file LICENSE

  R package prnsamplr. Coordinated stratified sampling using permanent random numbers (PRN's). Supports simple random sampling and probability-proportional-to-size sampling and includes a function for transforming

Access to official statistics (GSBPM 7.4)

- R package rsdmx. Easy access to data from statistical organisations that support SDMX webservices. The package contains a list of SDMX access points of various national and international statistical institutes.
- R package and C++ readsdmx. Read SDMX into dataframes from local SDMX-ML file or web-service. By OECD.
- Python pandaSDMX. Python interface to SDMX that facilitates the acquisition and analysis of SDMX-2.1 compliant data and metadata.
- R package rjstat. Read and write data sets in the JSON-stat format.
- Python package pyjstat. Read and write JSON-stat.
- Java module json-stat.java Read and write JSON-stat. By Statistics Norway.
- R package oecd Search and Extract Data from the OECD
- R package sorvi Finnish Open Government Data Toolkit
- R package eurostat Tools to download data from the Eurostat database together with search and manipulation utilities.
- R package acs Download, Manipulate, and Present American Community Survey and Decennial Data from the US Census.
- R package inegiR Access to data published by INEGI, Mexico's official statistics agency.
- R package cbsodataR. Access to Statistics Netherlands' (CBS) open data API from R.
- Node.js package cbsodata.js. Access to Statistics Netherlands' (CBS) open data API from js.
- Python package cbsodata.py. Access to Statistics Netherlands' (CBS) open data API from Python.
- R package censusapi A wrapper for the U.S. Census Bureau APIs that returns data frames of Census data and metadata.
- R package nsoApi builds on other packages to access data from official statistics and tries to harmonize the API.
- R package CANSIM2R. Extract CANSIM (Statistics Canada) tables and transform them into readily usable data.
- Python package pyscbwrapper. Access to the open data API of the Swedish Statistical Institute
- R package pxweb. Generic interface for the PX-Web/PC-Axis API used by many National Statistical Agencies.
- R package PxWebApiData. Easy API access to e.g. Statistics Norway, Statistics Sweden and Statistics Finland.
- R package rdbnomics. Access to the DB.nomics database which provide macroeconomic data from 38 official providers such as INSEE, Eurostat, Wolrd bank, etc.
- R package readabs Download data from the Australian Bureau of Statistics.
- R package destatiscleanr. Clean csv files from Genesis, the database of the Federal Statistical Office of Germany (Destatis) and its regional outlets.
- R package statcanR. An R connection to Statistics Canada's Web Data Service. Open economic data (formerly CANSIM tables) are accessible as a data frame in the R environment.
- R package cdlTools. Downloads USDA National Agricultural Statistics Service (NASS) cropscape data for a specified state.
- Java package SDMX Connectors. Browse SDMX data providers, build your queries and get data directly in your favourite tool (R, SAS, Matlab, Stata and Excel). By Banca d'Italia.
- Node.js package sdmx-rest. This library allows to easily create and execute SDMX REST queries from a JavaScript client application.
- R package csodata Download data from Central Statistics Office (CSO) of Ireland.
- R package iriR. Client for the EU Industrial Research and Industry Scoreboard.

Category "access to official statistics"

- Over 30 software packages, 28 R-packages
- Cran standard documentation used as starting point

rsdmx: Tools for Reading SDMX Data and Metadata

OECD: Search and Extract Data from the OECD

readsdmx: Read SDMX-XML Data

czso: Use Open Data from the Czech Statistical Office in R

rjstat: Handle 'JSON-stat' Format in R

readabs: Download and Tidy Time Series Data from the Australian Bureau of Statistics

sorvi: Functions for Finnish Open Data

eurostat: Tools for Eurostat Open Data

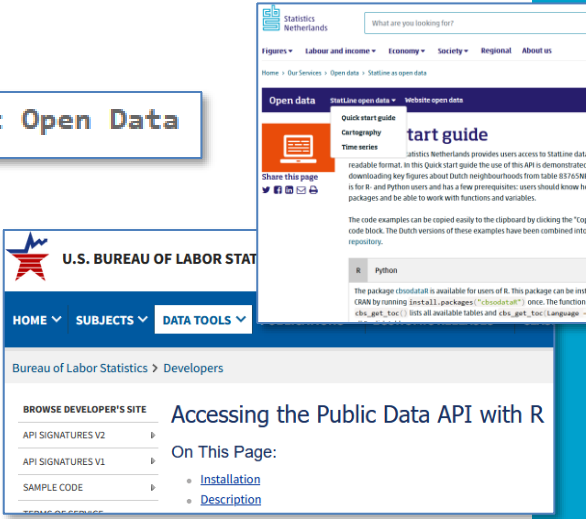
restatapi: Search and Retrieve Data from Eurostat Database

inegiR: Integrate INEGI's (Mexican Stats Office) API with R

pxweb: R Interface to PXWEB APIs

csodata: Download Data from the CSO 'PxStat' API

danstat: R Client for the Statistics Denmark Databank API



tart guide

tatistics Netherlands provides users access to StatLine dat
readable format. In this Quick start guide the use of this API is demonstrate
downloading key figures about Dutch neighbourhoods from table 83765N
is for R- and Python users and has a few prerequisites: users should know h
packages and be able to work with functions and variables.

The code examples can be copied easily to the clipboard by clicking the "Co
code block. The Dutch versions of these examples have been combined int
repository.

The package cbsodataR is available for users of R. This package can be ins
CRAN by running `install.packages("cbsodataR")` once. The function
`cbs_get_toc()` lists all available tables and `cbs_get_toc(Language`

Accessing the Public Data API with R

On This Page:

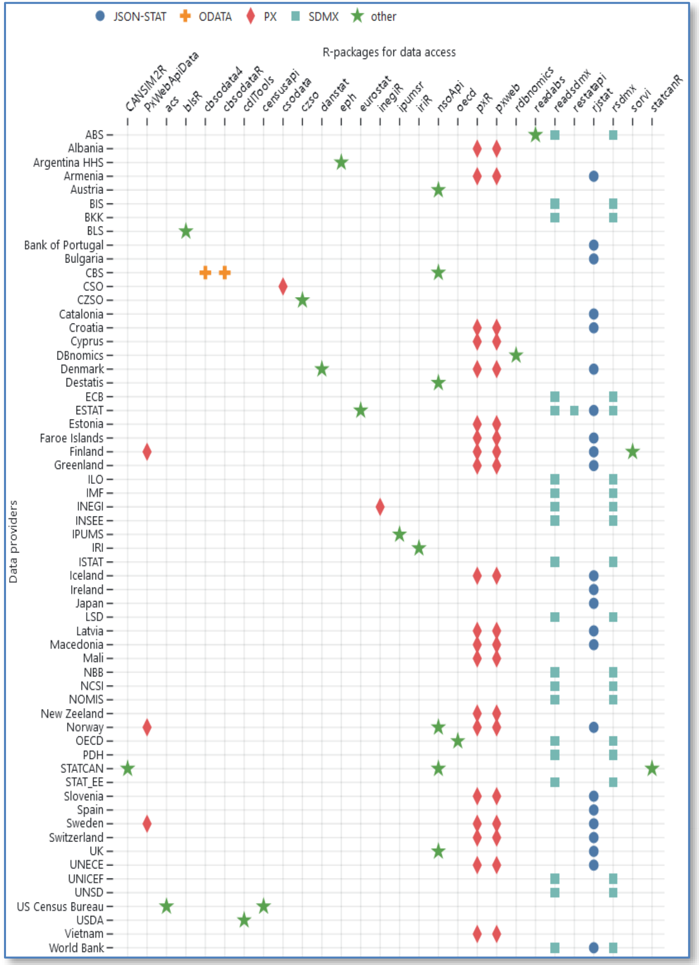
- Installation
- Description

“access to official statistics” R-software landscape (1)

- Matrix from docs, links to web pages and packages execution:

R-packages (28) x dataproviders (60) x standards (5)

- Standards: ● JSON-STAT ✚ ODATA ♦ PX ■ SDMX ★ other



https://observablehq.com/@olavtenbosch/access-to-official-statistics-from-r

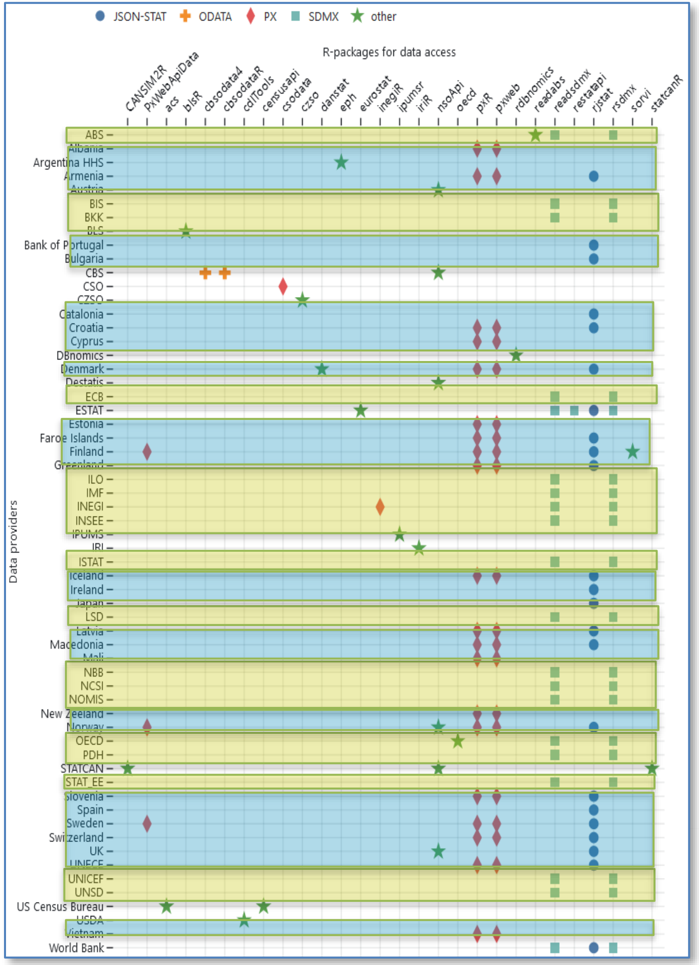
“access to official statistics” R-software landscape (2)

- Matrix from docs, links to web pages and packages execution:

R-packages (28) x dataproviders (60) x standards (5)

- Standards: ● JSON-STAT ✚ ODATA ♦ PX ■ SDMX ★ other

- JSON-STAT/PX v. SDMX: almost disjunct worlds
- ODATA: CBS only



https://observablehq.com/@olavtenbosch/access-to-official-statistics-from-r

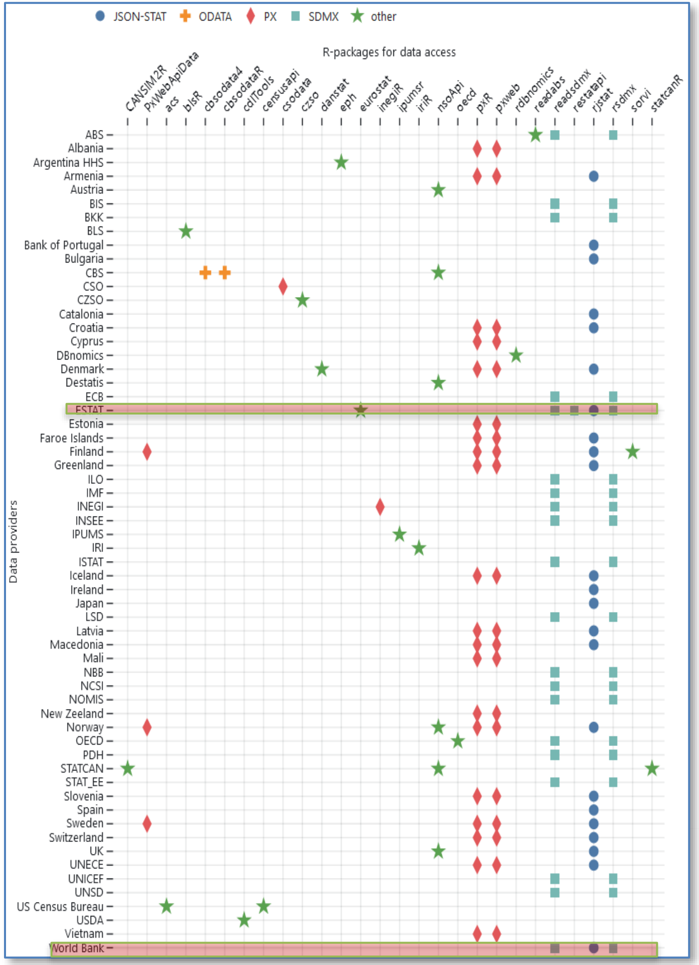
“access to official statistics” R-software landscape (3)

- Matrix from docs, links to web pages and packages execution:

R-packages (28) x dataproviders (60) x standards (5)

- Standards: ● JSON-STAT ✚ ODATA ♦ PX ■ SDMX ★ other

- Few data providers support multiple standards: ESTAT, WB



https://observablehq.com/@olavtenbosch/access-to-official-statistics-from-r

“access to official statistics” R-software landscape (4)

- Standards-oriented packages:

  *rsdmx, readsdmx, rjstat, px**

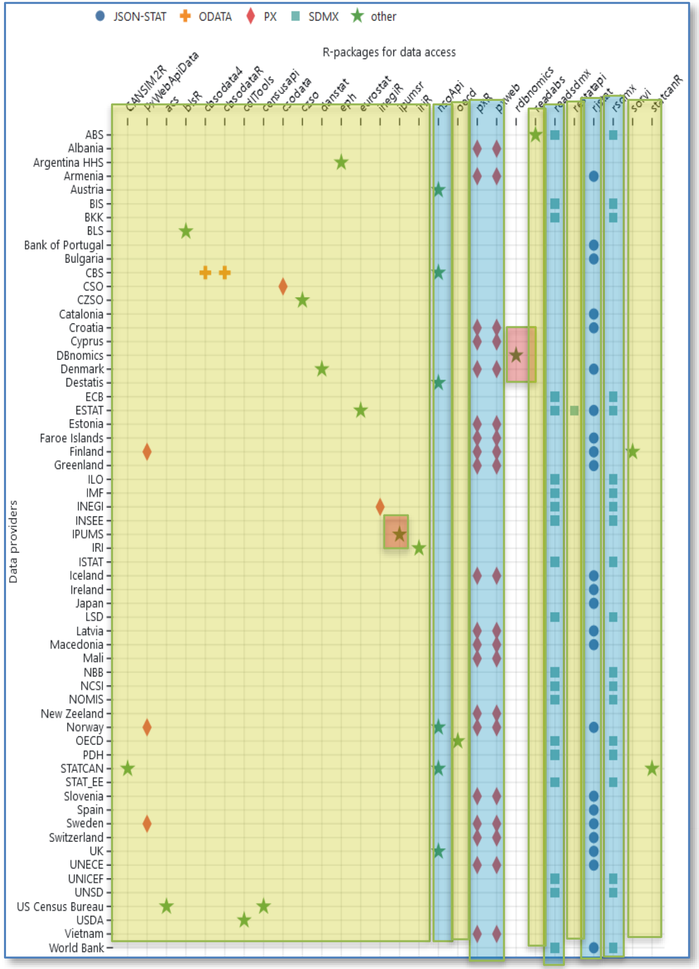
- Data provider-centric packages:

  inegiR, readabs, statcanR, eurostat

- Official statistics aggregator sites:

  rdbnomics: economic data

  ipumsr: census & survey data
  time&space harmonised



https://observablehq.com/@olavtenbosch/access-to-official-statistics-from-r

Features commonly offered

- ***endpoint hiding***:
  wrapping the preconfigured endpoint(s) in a R function within the package
- ***catalogue retrieval***:
  the ability to list the availability datasets on the endpoint(s)
- ***search***:
  the ability to search for datasets or within datasets on the endpoint(s)
- ***endpoint queries***:
  the ability to query for subsets on the endpoint(s) side
- ***local queries***:
- the ability to easily slice or filter the retrieved data on the client
- ***caching***:
  preventing unnecessary roundtrips to the endpoint(s) by caching results
- ***cartographic queries***:
  retrieve a (cartographic) map to be used with the data
- ***registry access****:*
  access to coordinated metadata in registries

Future research: Packages x endpoints x standards x features

Considerations

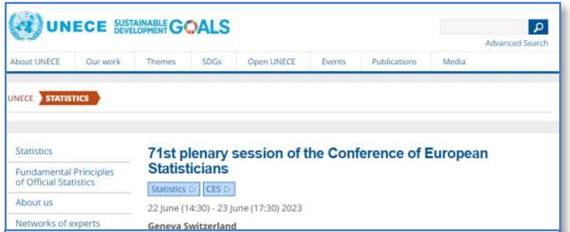
- Offstats landscape grows towards ***standardisation: SDMX, JSON-stat, PX***
- R user can choose from (at least) ***28 packages***, each offering specific functionality to ***60 data sources*** in total
- Many NSIs offer ***targeted R-packages*** for data and metadata access
- There is ***no 'one-for-all' R-package*** that provides access to all official statistics data providers
- ***Common features*** identified

- Dream: ***one generic R-packages*** to all official statistics from all dataproviders supporting maximum FAIRness

FAIR: Findable, Accesible, Interoperable, Reusable

International

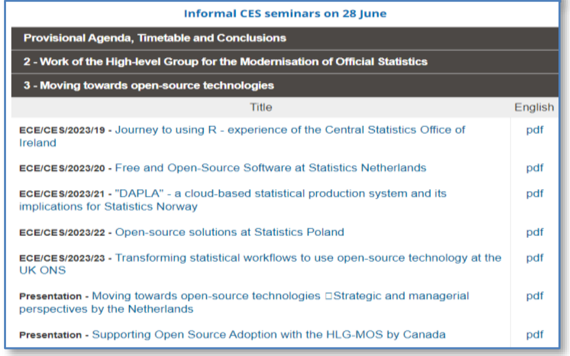
71st CES



71st plenary session of the Conference of European Statisticians

22 June (14:30) - 23 June (17:30) 2023

Geneva Switzerland



Informal CES seminars on 28 June

Provisional Agenda, Timetable and Conclusions

2 - Work of the High-level Group for the Modernisation of Official Statistics

3 - Moving towards open-source technologies

Title	English
ECE/CES/2023/19 - Journey to using R - experience of the Central Statistics Office of Ireland	pdf
ECE/CES/2023/20 - Free and Open-Source Software at Statistics Netherlands	pdf
ECE/CES/2023/21 - "DAPLA" - a cloud-based statistical production system and its implications for Statistics Norway	pdf
ECE/CES/2023/22 - Open-source solutions at Statistics Poland	pdf
ECE/CES/2023/23 - Transforming statistical workflows to use open-source technology at the UK ONS	pdf
Presentation - Moving towards open-source technologies □Strategic and managerial perspectives by the Netherlands	pdf
Presentation - Supporting Open Source Adoption with the HLG-MOS by Canada	pdf

- IE, NL, NO, PL, UK, CA

EUROPEAN
COMMISSION

Brussels, 21.10.2020
C(2020) 7149 final

COMMUNICATION TO THE COMMISSION

OPEN SOURCE SOFTWARE STRATEGY 2020 – 2023
Think Open

EU OSS strategy

ESS principles on OSS (OS4OS group)

1. OSS by default
2. Work in the open
3. Improve and give back
4. Think general statistical building blocks
5. Test, package and document
6. Choose permissive
7. Promote

https://os4os.pages.code.europa.eu/pbbp

Wrap-up

- www.awesomeofficialstatistics.org Please
☆ Star 239
!
  - Spread the word and help maintain!
- Meta analysis on awesome list “Access to offstats”:
  - ***standardisation***: SDMX, JSON-stat, PX
  - ***Common features*** identified; often reimplemented
  - ***No*** ‘one-for-all’ R-package for access to ***all*** offstats data
  - Lets work ***together*** to improve access to offstats from R

Extended abstract: https://olavtenbosch.github.io/pdf/2023_Uros_tenBosch_Access.pdf

Follow-up paper to be published at: http://cosmos-conference.org/2024/

Olav ten Bosch o.tenbosch@cbs.nl
Edwin de Jonge e.dejonge@cbs.nl