



The awesome list of official statistics software & FOSS best practices

www.awesomeofficialstatistics.org

Olav ten Bosch

Eurostat Task Force on Trusted Smart Statistics, 10-03-2022

Contents

- What is this awesome list?
- Zooming out: what is the aim?
- FOSS best practices
- Wrap-up



What is this awesome list?

- When: born during the **UNECE SDE conference** april 2017 (The Hague)
- Why: because we needed something simple to **collectively remember useful software** in official statistics
- Who: initiated by SNStatComp, maintained by **statistical community**
- What: a **community approach** to knowledge management
- How:
 - Using the **[awesome concept](#)** on GitHub
 - A **public** list which started **simple** and continues to **grow**
 - Clear and simple **criteria**
 - **awesomeofficialstatistics.org**



What is the awesome list?

Curated list of software for
official statistics



awesome

www.awesomeofficialstatistics.org

The screenshot shows the GitHub interface for the repository 'SNStatComp / awesome-official-statistics-software'. A red circle highlights the 'Unwatch 34', 'Fork 48', and 'Starred 188' buttons. Below the repository name, there are navigation tabs for Code, Issues, Pull requests, Actions, Projects, Wiki, Security, Insights, and Settings. A callout box labeled 'Social interactions, watch' points to the highlighted buttons.

Social interactions, watch

master 1 branch 0 tags Go to file Add file Code About

Awesome official statistics software awesome

An awesome list of open source software for official statistics

Criteria

An item on this list is awesome because it is

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

We prefer software that is easy to install and use, has at least one stable version, and is actively maintained. [Contributions](#) welcome.

The screenshot shows the 'Contributors' section with 17 contributors. A red circle highlights the list of contributor avatars. A callout box labeled 'Working together' points to this section.

Working together

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](mailto:mark.vanderloo@gmail.com) or [olav dot tenbosch at gmail dot com](mailto:olav.tenbosch@gmail.com) or tweet [@olavtenbosch](https://twitter.com/olavtenbosch) or [@markvdloo](https://twitter.com/markvdloo)

License



This work is licensed under a [Creative Commons Attribution 4.0 International License](#).

Open license

Statistical disclosure control (GSBPM 6.4)

- Java application [μ-ARGUS](#). Tool to create safe micro-data files. See also the [casc page](#).
- Java application [T-ARGUS](#). Tool to protect statistical tables. See also the [casc page](#).
- R package [sdc](#)
- R package [sdc](#)
- R package [easy](#)
- R package [sdc](#)
- R package [Sma](#)
- R package [sim](#)
- R package [sdc](#)
- R package [syn](#)

Data integration and record linkage (GSBPM 5.1)

- R package [reclin](#). Functions to assist in performing probabilistic record linkage and record matching. It includes a distance function for comparing records, em-algorithm for estimating m- and u-probabilities, functions for record matching, and functions for record merging. It can also be used for pre- and post-processing for machine learning methods for record linkage.
- R package [RecordLinkage](#). Implementation of the Fellegi-Sunter method for record linkage.
- R package [fastLink](#). Implements a Fellegi-Sunter probabilistic record linkage model and the inclusion of auxiliary information. Documentation can be found on <http://www.fastlink.org/>
- R packages [stringdist](#). Implements approximate string matching. Supports various distance metrics (Levenshtein, Hamming, Levenshtein, optimal string alignment), qgrams (q- gram, q-gram), and edit distance. An implementation of soundex is provided.
- R package [fuzzyjoin](#). Join tables based on exact or similar matches. Allows for fuzzy matching and handling of missing values and inaccuracy.

Scraping for Statistics (GSBPM 4.3)

- Java application [URLSearcher](#). An application for searching ISTAT.
- Java application [URLScorer](#). Gives a rule based score to scrape websites.
- Node.js tool [RobotTool](#). A tool for checking (price) changes on websites.
- Python [Social-Media-Presence](#). A script for detecting social media presence in Poland.
- Python [Sustainability Reporting](#). A script for measuring sustainability reporting.
- Python [urlfinding](#). Software for finding websites of enterprises.

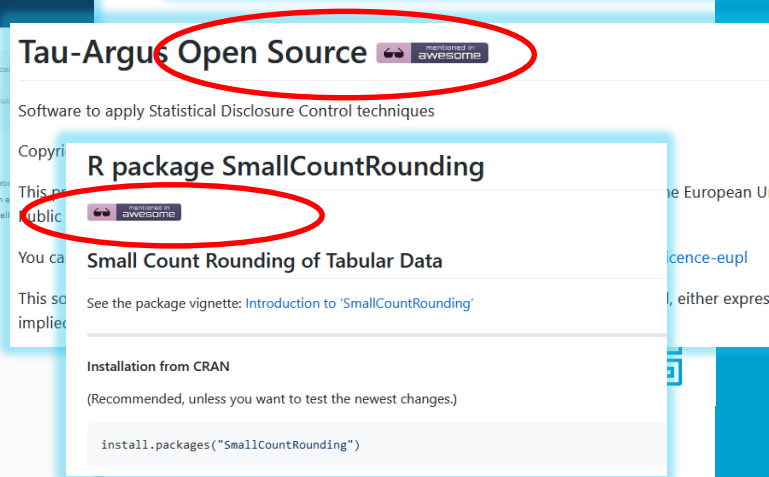
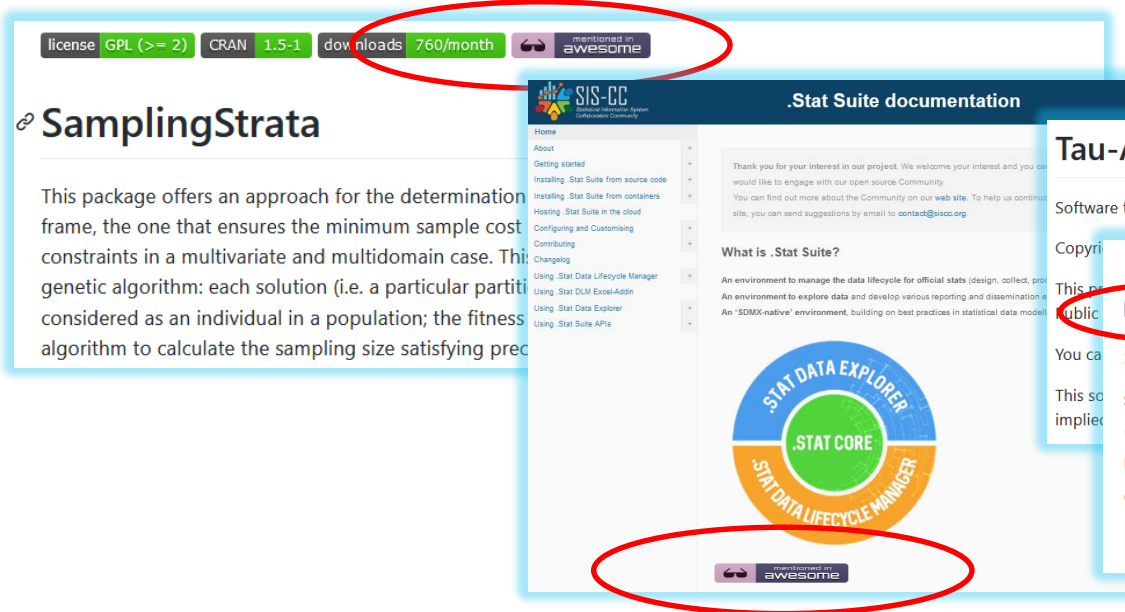
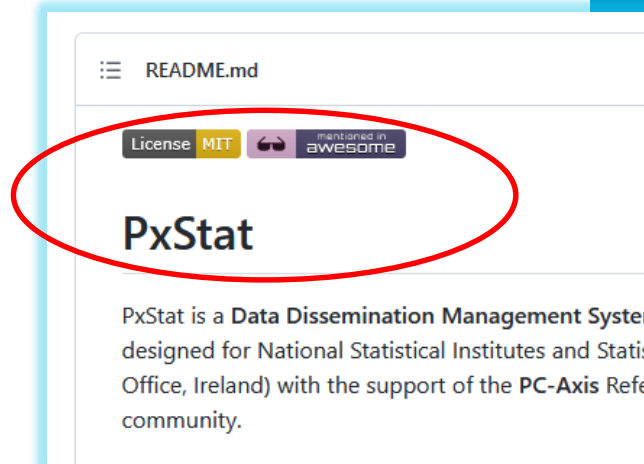
Access to official statistics (GSBPM 7.4)

- R package [rdsdmx](#). 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, World 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) cropland 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.

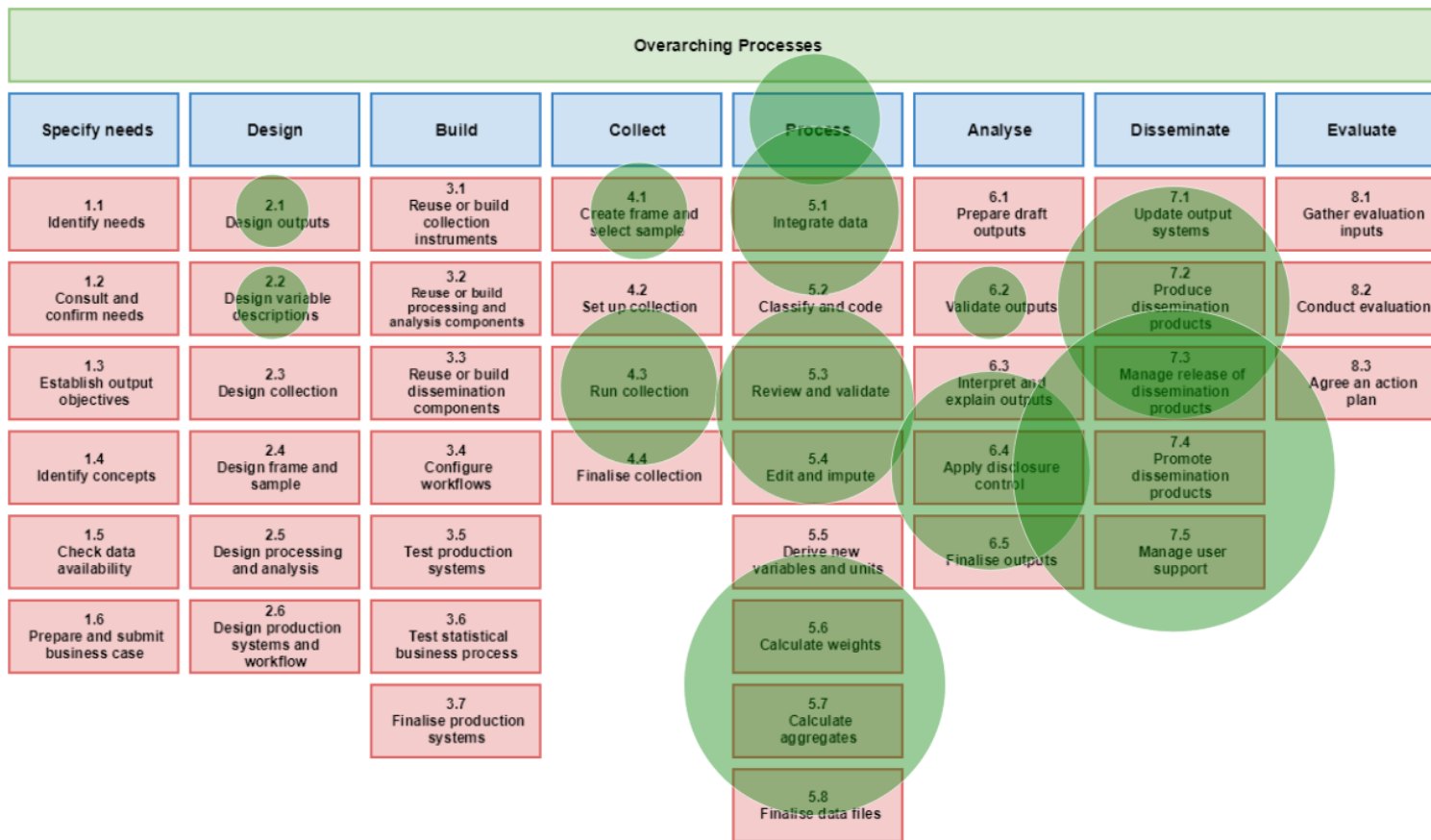
The right to wear the badge

- The badge links to the list and improves findability:

Wear the badge. Authors of software that is mentioned on this list gain the right to wear the [mentioned in awesome](#) badge on their website or GH repository. Please use the following code (or equivalent) to do so for your project.



Awesome list by GSBPM



Zooming out: what do we actually want?

Re-use

of software in official statistics

Costs

Develop once, use by many

Quality

Use well-tested and proven implementations of generic methods

Time-to-market

Connecting readily available basic building blocks into processes

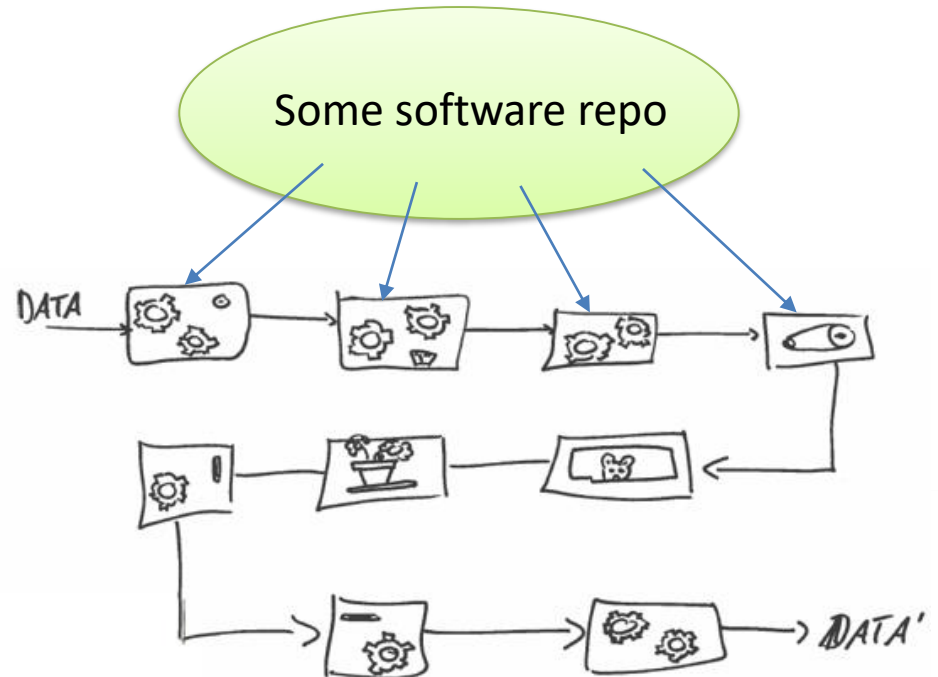
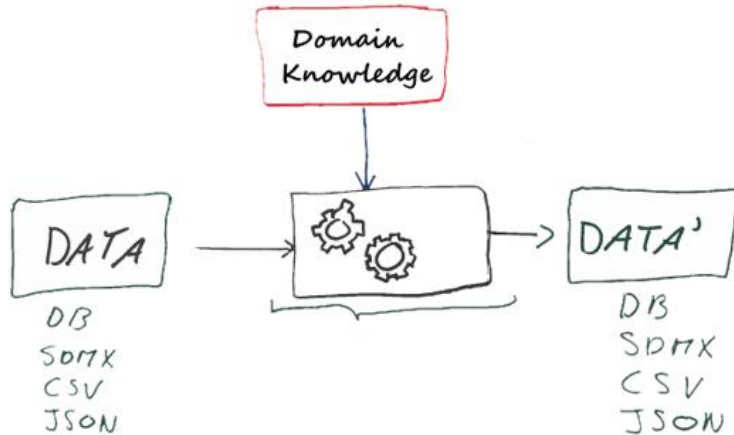
Standardisation

Using the same implementations of for common methods to standardise official statistics



Basic building blocks

- The software landscape for offstats is getting more **complex** and **dynamic**
- What are proven and succesful **building blocks** for offstats?
- Ideal scenario:
 - configurable per domain
 - chainable



Communities, repos, package systems

- Software sharing is already happening
- Different communities have their own packaging platforms

Cran (R)
~ 19,020

Pip/Anaconda (Python)
~ 360,000

NPM (JavaScript/Node)
~ 1,800,000

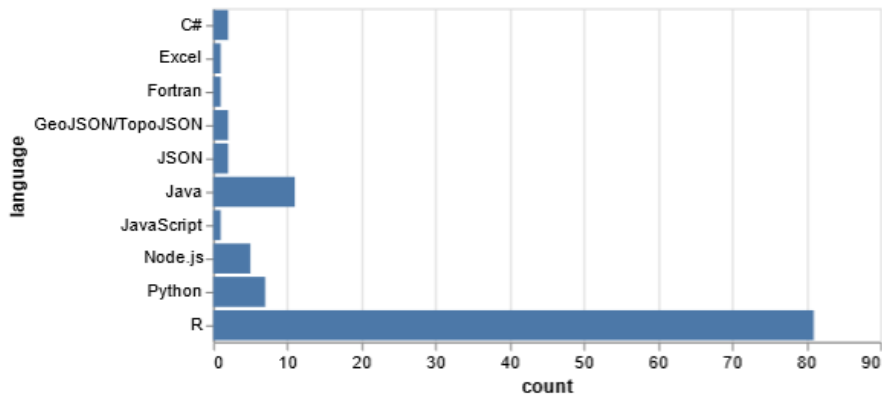
Julia general
registry (Julia)
~ 7,200



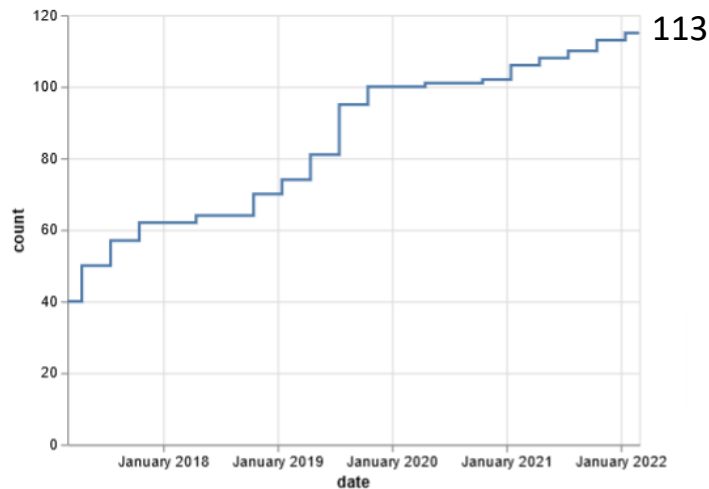
Awesome list status

- Bottom up approach
- Majority is R (now)
- Offstats *R community* more motivated towards sharing?

Packages by programming language:



Growth of the awesome list:



Awesome list promotions

- Unece SDE '17
- Unece SCFE '17
- uRos '18
- Unece SDE '18
- Estat Validation Grants kickoff '18
- uRos '19
- Unece modernstats World '19
- Unece modernstats '20 (virtual)
- uRos '20
- ICDSOS '21
- uRos '21
- TF-TSS '22



} **Virtual ☹️**



FOSS best practices

- ***Don't copy*** existing solutions, ***use*** them, ***improve*** them and ***give back*** (pull requests on repo's).
- Invest in making solutions ***re-usable*** based on ***generic functionality***.
- Don't start a new packaging platform, use monorepo and publish generic OSS on ***existing packaging systems***.
- Make ***simple*** and ***to the point documentation***. No docs > 100 pages but GH wiki or online tutorial.
- Nobody will use OSS software that is ***not known***. Invest in ***PR*** (possibly via awesome list)



Wrap-up

- Invest in re-use by **generalizing** software, **publishing** as open source on common OSS platforms and **sharing** among domain specialists

- www.awesomeofficialstatistics.org 
Spread the word and help maintain

Please ☆ Star 188 !

- Questions/ Ideas / suggestions:

Olav ten Bosch

o.tenbosch@cbs.nl

@olavtenbosch

Mark van der Loo

mpj.vanderloo@cbs.nl

@markvdloo

