



Open-source software sharing through the awesome list of official statistics software

Statistics Netherlands

Olav ten Bosch, Mark van der Loo
NTTS, Brussels, 11-13 March 2025



Contents

- Why software sharing?
- What is the awesome list?
- Examples: data cleaning; access to offstats
- FOSS principles: development and adoption
- Wrap-up



Why software sharing in official statistics?

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

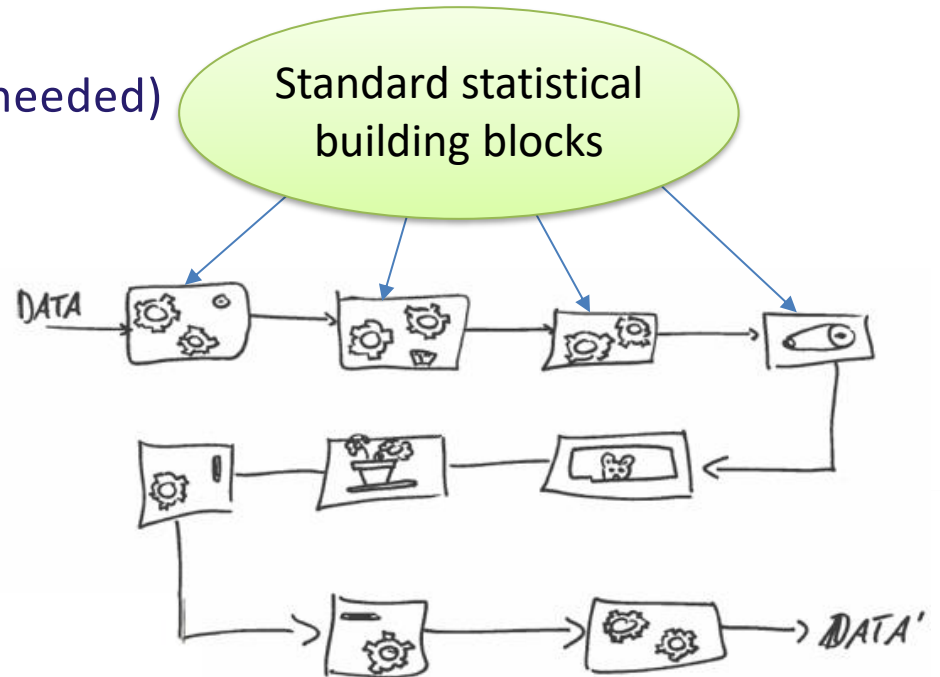
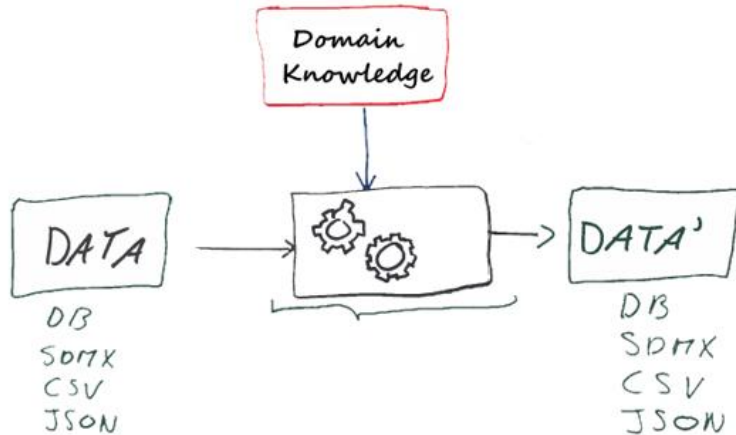
Standardisation

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



Concept

- Generic **methodological building blocks** for official statistics
- Configurable by domain (domain knowledge)
- Chainable
- Data in, data out (UI only where needed)



Communities, repos, package systems

- Software sharing is already happening
- Different communities have their own packaging platforms
- Package and document building blocks!
- Don't start a new one, just point to existing

Cran (R)
~ 19,020

Pip/Anaconda (Python)
~ 360,000

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

Julia general
registry (Julia)
~ 7,200



R Python Julia

R, Python, Julia: do you know them all?

Mark van der Loo
Statistics Netherlands

EMOS webinar, 24 March 2020

1

on [YouTube](#)



Awesome list of official statistics software

Uros

- Started at *UNECE SDE conference* 2017 (The Hague)
- a *community approach* to remember useful software
- A *public* list, clear and simple *criteria*
- awesomeofficialstatistics.org

2018
NL



2019
RO



2023
RO



2024
GR



Data integration and record linkage (GSBPM 5.1)

- CRAN [0.5.0](#) – 9 months ago license [GPL-3](#)
R package [reclin2](#). Functions to assist in performing probabilistic pairs, comparing records, em-algorithm for estimating m- and u also be used for pre- and post-processing for machine learning
- CRAN [0.4-12.4](#) – 2 years ago license [GPL \(>= 2\)](#)
R package [RecordLinkage](#). Implementation of the Fellegi-Sunter
- CRAN [1.4.2](#) – 6 months ago license [GPL \(>= 2\)](#)
R package [StatMatch](#). Statistical Matching or Data Fusion
- CRAN [0.6.1](#) – a year ago license [GPL \(>= 3\)](#)
R package [fastLink](#). Implements a Fellegi-Sunter probabilistic re and the inclusion of auxiliary information. [Documentation](#).
- CRAN [0.9.12](#) – a year ago license [GPL-3](#)

Statistical disclosure control (GSBPM 6.4)

- GitHub [v5.1.7b4](#) last commit [march](#) license [EUPL-1.2](#)
Java and C++ application [Mu-ARGUS](#). Tool to create safe micro-data
- GitHub [v4.2.5.2](#) last commit [august](#) license [EUPL-1.2](#)
Java C++ Fortran and Delphi application [T-ARGUS](#). Tool to protect st
- CRAN [5.7.8](#) – 8 months ago license [GPL-2](#)
R package [sdcMicro](#). Disclosure control for statistical microdata.
- CRAN [0.32.6](#) – a year ago license [GPL \(>= 2\)](#)
R package [sdcTable](#). Disclosure control for tabulated data.
- CRAN [1.0.7](#) – 2 years ago license [Apache License 2.0 | file LICENSE](#)
R package [easySdcTable](#). Provides an interface to the package [sdcTable](#).
- CRAN [0.9.0](#) – 2 months ago license [MIT + file LICENSE](#)
R package [GaussSuppression](#). Tabular data suppression using the Gaussian elimination secondary suppression algorithm.

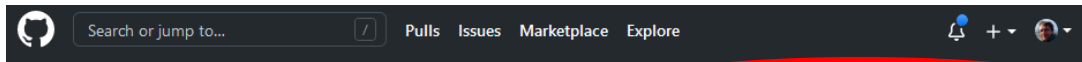
How does it work?

Curated list of software for
official statistics



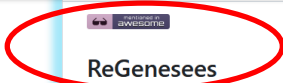
awesome

www.awesomeofficialstatistics.org



Social interactions

The right to wear the badge



ReGenesees

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

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

SNStatComp / awesome-official-statistics-software

Unwatch 30 Unstar 161 Fork 41

Code Issues Pull requests 1 Actions Projects Wiki Security Insights

master

Go to file Add file Code

Awesome official statistics software



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

Criteria

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.

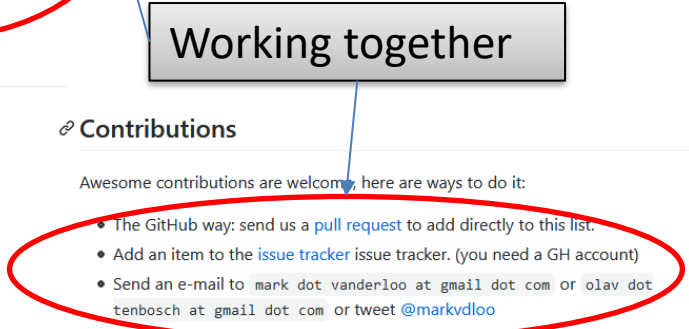
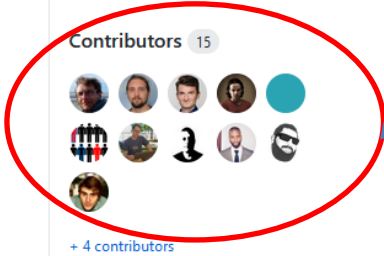
We prefer packages that are easy to install and use, have at least one stable version, and are actively maintained. [Contributions](#) are welcome.

License



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

Open license



Steady grow 2017 -> 2025

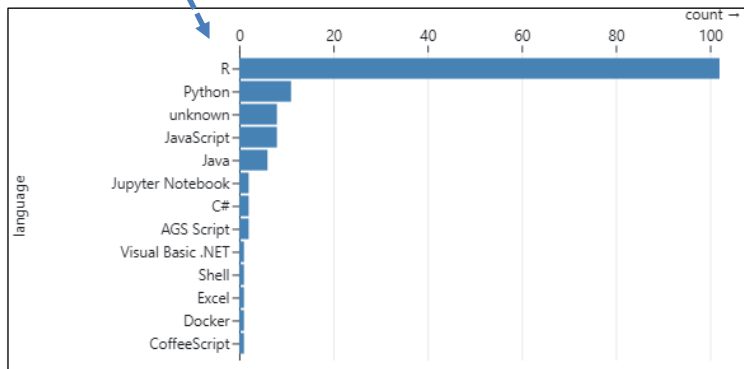
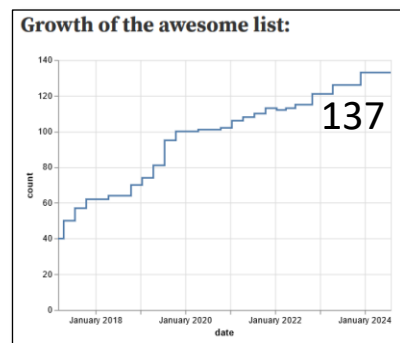
- Quite stable: growing to 137 entries (~+10)

- Goal:

optimal set of **mature** open source packages for official statistics

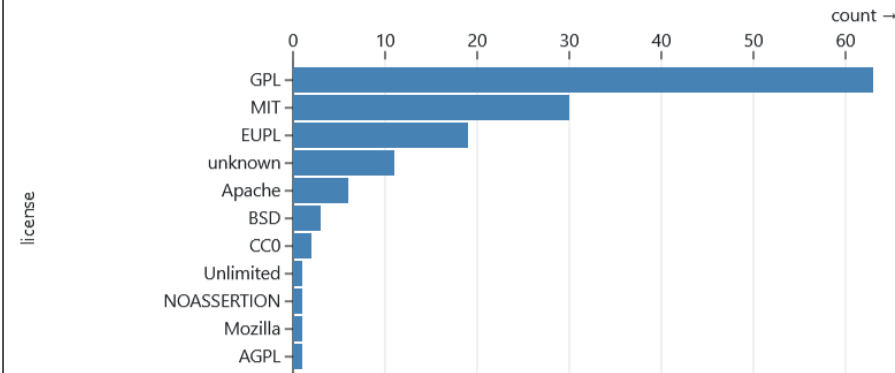
- **License** stats

- Majority is **R software**



Licenses used:

Details:



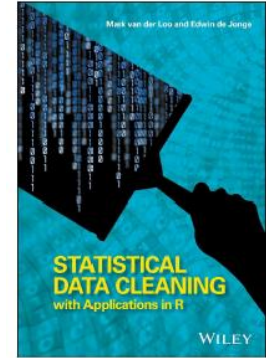
Awesome list by GSBPM

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			

Example 1: CBS R data cleaning ecosystem

- ***validate***: check data based on validation rules
- ***dcmodyfy***: change data based on ‘if-this-then-that’ rules
- ***errorlocate***: locate errors based on validation rules and mark them for correction
- ***simputation***: many different imputation methods
- ***rspa***: adapt numerical records to fit (in)equality restrictions
- ***deductive***: solve errors based on control rules
- ***validatetools***: find inconsistencies and redundancies
- ***accumulate***: advanced group aggregation
- ***lumberjack***: standardized logging



MPJ van der Loo and E de Jonge (2018)
Statistical data cleaning with applications in R
John Wiley & Sons, NY.

Used at CBS in social and economical statistics, agriculture, trade, education, environment, emmissions, income, shipping, STS, recreation, museums ... also used in other NSIs

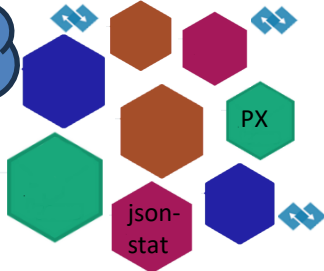
Example 2: access to Official Statistics

- > 50 software packages
- > 80 dataproviders
- > majority R-packages
- > Some use standards (SDMX, PX, JSON-STAT)

One 4 all?



User



STATISTICS DENMARK

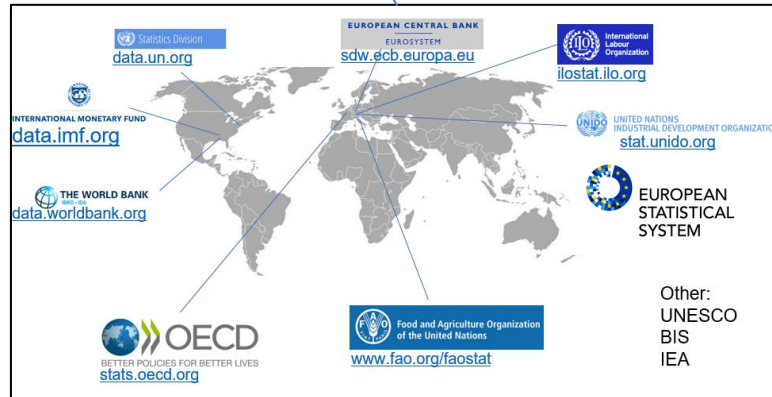
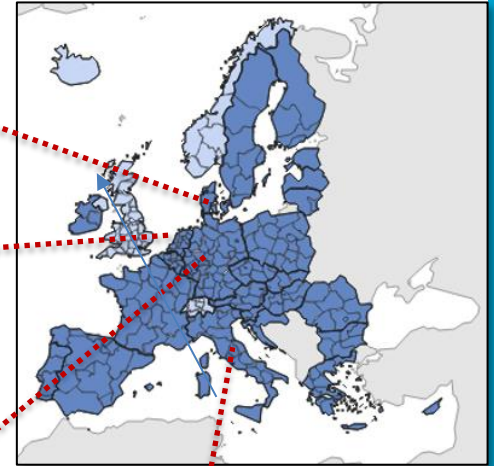
Labour force status in percentage by employment status, time, age and sex

Economic activity	Total	Men	Women
2020A	79.4	82.2	76.6
Age total	48.7	58.4	40.6
15-24 years	88.2	88.4	79.0
25-34 years	88.2	92.1	84.2
35-44 years	88.4	90.1	86.2
45-54 years	79.2	88.1	70.0

DK



NL



DE

GENESIS-ONLINE

Year	Value	Unit
2010	100	Index
2011	100	Index
2012	100	Index
2013	100	Index
2014	100	Index
2015	100	Index
2016	100	Index
2017	100	Index
2018	100	Index
2019	100	Index
2020	100	Index

I.Stat

Semi-supernumerary population (150 years and more)

Year	Value	Unit
2010	100	Index
2011	100	Index
2012	100	Index
2013	100	Index
2014	100	Index
2015	100	Index
2016	100	Index
2017	100	Index
2018	100	Index
2019	100	Index
2020	100	Index

IT



Experience => practices => principles

CBS historical FOSS timeline

Period	Milestone
≤ 2009	R used only in research
2010	R adopted as formal tool <ul style="list-style-type: none">- User group, courses, code/SD guidelines- FOSS policy- Application management- Research -> R packages
2012	Python as formal tool
2014	Git as formal tool
2017	CBS starts awesomelist for OS software
2018	CBS hosts uRos2018
2019/2020	New FOSS policy, SD guidelines (CIO-office)
2023	ESS principles on OSS

Don't copy existing solutions, **use** them, **improve** them and **give back** (PRs on repos)



EUROPEAN
STATISTICAL
SYSTEM

the group on Open Source for
Official Statistics (OS4OS).

18 NSIs + ESTAT + OECD

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

HLG-MOS Open Source Software

modernstats

Open-source software is essential for producing official statistics in an open, sharable, and transparent matter. The following principles intended to guide both the production and adoption of open-source software for statistical production.



Open by default



Work in the open



Improve and give back



Think generic building blocks



Test, package and document



Choose permissive



Promote

UNECE HLG-MOS open source charter

- ESS principles on OSS, derived from awesome list, adopted
- Under consideration for Conference of European Statisticians (CES) endorsement



[Open by default](#)



[Work in the open](#)



[Improve and give back](#)



[Think generic building blocks](#)



[Test, package and document](#)



[Choose permissive](#)



[Promote](#)

2. Work in the open

Statement

We start our projects in the open from the beginning and clearly mark maturity status.

Rationale

Many projects have the intention to publish results as open source but have difficulty deciding on the best time to do so. It might feel uncomfortable to put early ideas and rough implementation sketches on-line, but on the other hand sharing it too late prevents others from providing valuable comments and ideas or volunteering to work together on the project. To circumvent this dilemma we start publishing the results from the beginning and clearly mark maturity status.

<https://unece.github.io/OSS/>

Wrap-up



Starred

311



- Use the awesome list and help maintain!

www.awesomeofficialstatistics.org

- Experiences and practices found their way in to the UNECE HLG-MOS open source charter

<https://unece.github.io/OSS/>

