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

Time-to-market

Connecting readily available basic building blocks into processes

Quality

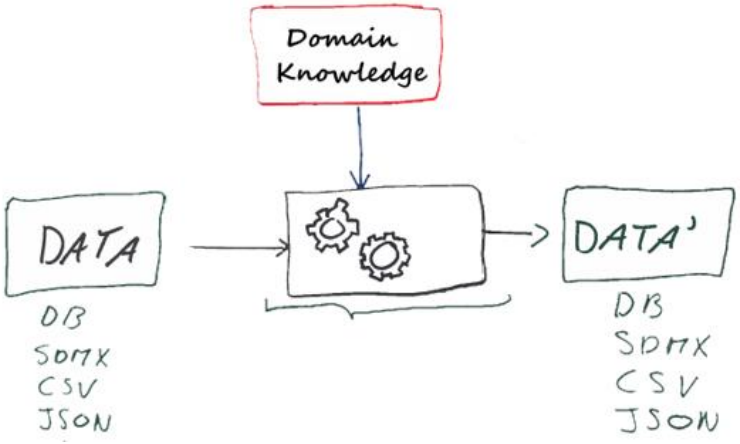
Use well-tested and proven implementations of generic methods

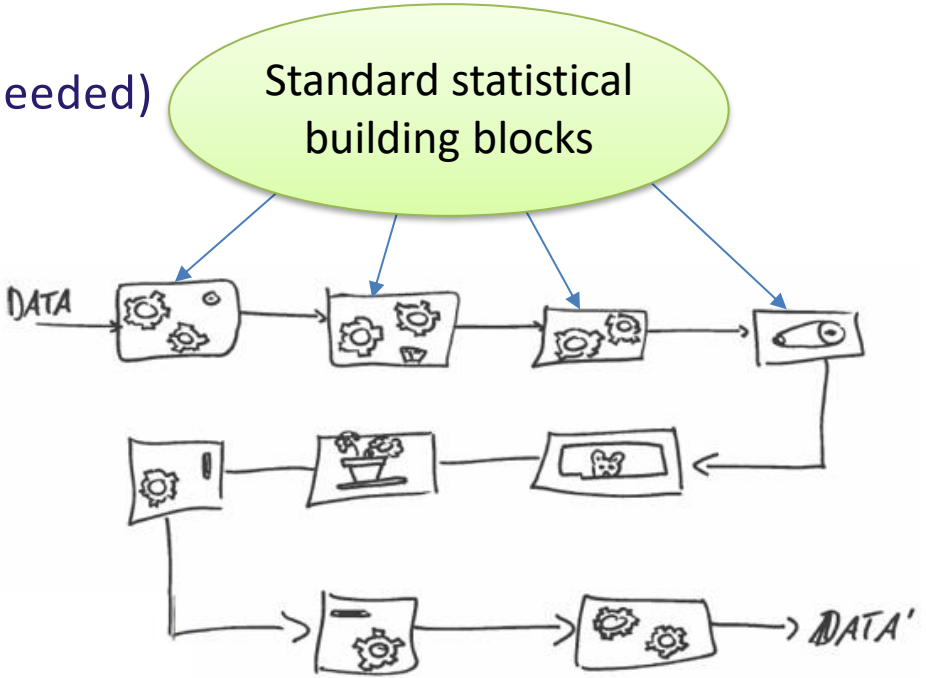
Standardisation

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

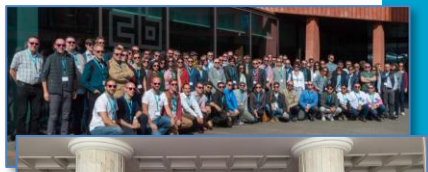
on YouTube

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 probabilisti… 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

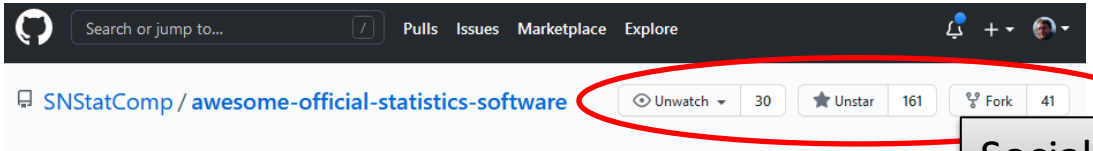
  R package GaussSuppression. Tabular data suppression using the Gaussian elimination secondary suppression algorithm.

How does it work?

Curated list of software for official statistics

awesome

www.awesomeofficialstatistics.org



Social interactions

master · Go to file · Add file · Code

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.

Criteria

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

official-statistics · gsbpm

Contributors 15

+ 4 contributors

The right to wear the badge

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.

Working together

Contributions

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

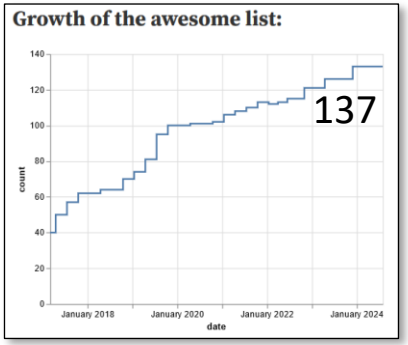
- The GitHub way: send us a pull request to add directly to this list.
- Add an item to the issue tracker issue tracker. (you need a GH account)
- Send an e-mail to `mark dot vanderloo at gmail dot com` or `olav dot tenbosch at gmail dot com` or tweet @markvdloo

License

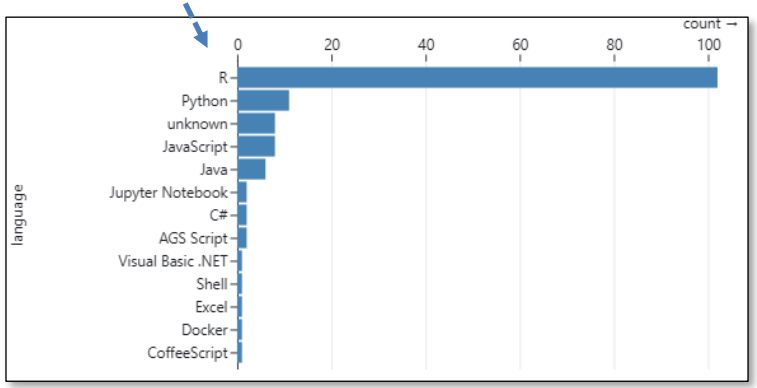
This work is licensed under a Creative Commons Attribution 4.0 International License.

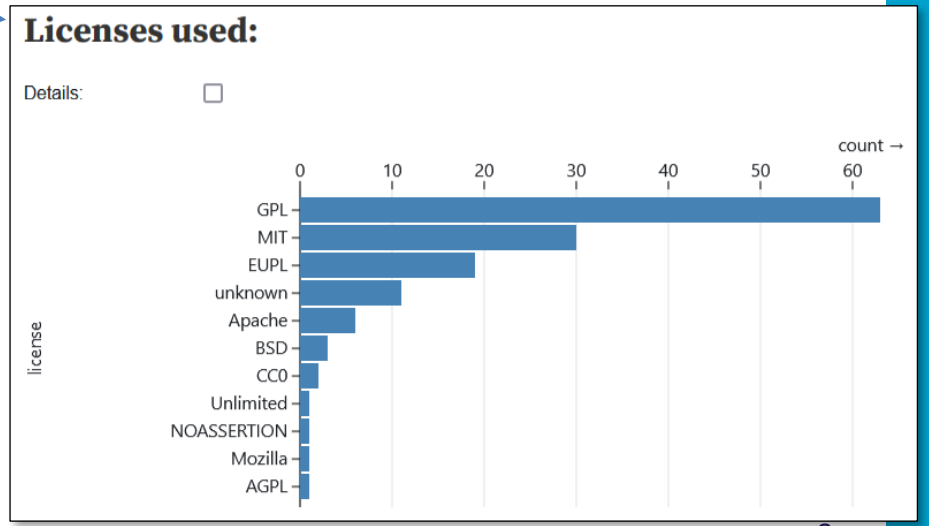
Open license

Steady grow 2017 -> 2025



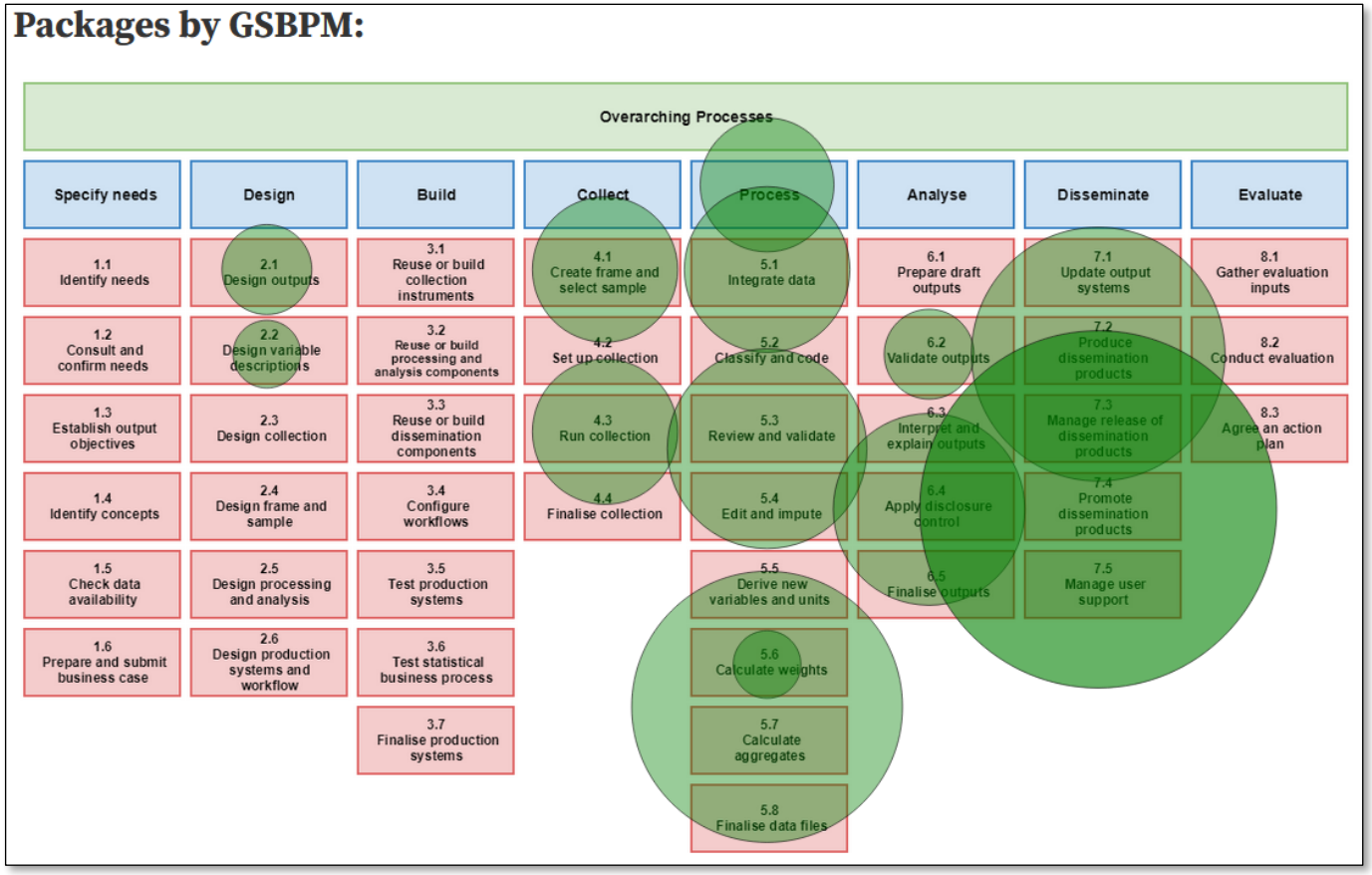
- Quite stable: growing to 137 entries (~+10)
- Goal:
  optimal set of ***mature*** open source packages for official statistics
- ***License*** stats
- Majority is ***R software***





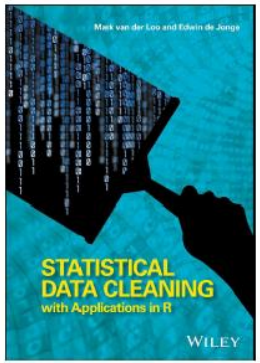
https://observablehq.com/@olavtenbosch/visualizing-awesomeofficialstatistics-org

Awesome list by GSBPM



Example 1: CBS R data cleaning ecosystem

- ***validate***: check data based on validation rules
- ***dcmodify***: 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

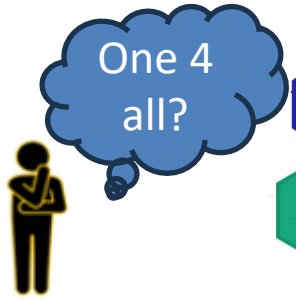


MPJ van der Loo and E de Jonge (2018)
Statistical data cleaning with applications in R
John Wiley & Sons, NY.

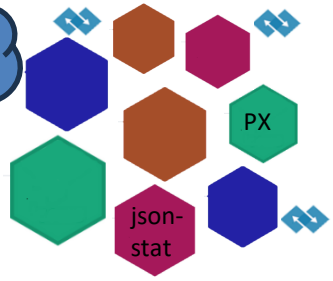
Used at CBS in social and economical statistics, agriculture, trade, education, environment, emmisions, income, shipping, STS, recreation, museums … also used in other NSIs

Example 2: access to Official Statistics

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

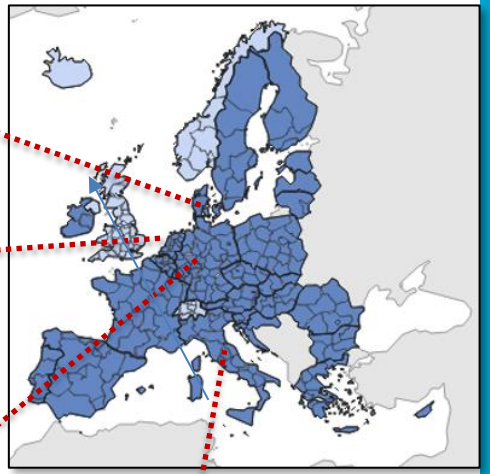


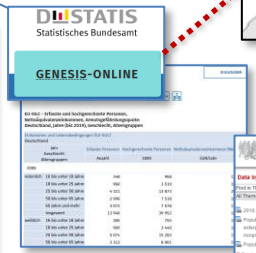
User



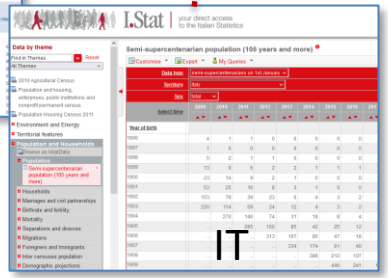
DK

NL





DE



IT

Experience => practices => principles

CBS historical FOSS timeline

Period	Milestone
≤ 2009	R used only in research
2010	R adopted as formal tool - 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

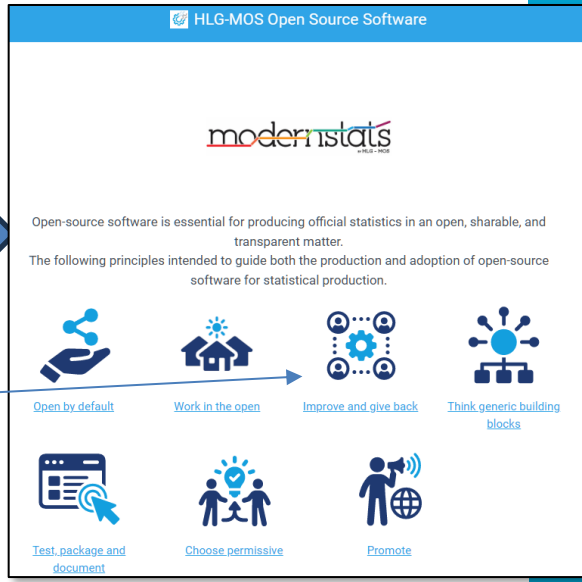
Don't copy existing solutions, ***use*** them, ***improve*** them and ***give back*** (PRs on repos)

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



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

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

UNECE HLG-MOS open source charter

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

2. Work in the open

Statement

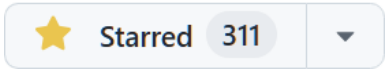
We start our projects in the open from the beginning and clearly mark maturity status.

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

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

Wrap-up



- 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/

Olav ten Bosch o.tenbosch@cbs.nl https://olavtenbosch.github.io
Mark van der Loo mpj.vanderloo@cbs.nl