



The awesome list of official statistics software: 100 ... and counting

Statistics Netherlands

Olav ten Bosch, Mark van der Loo, Alex Kowarik
uRos2020 2-4 December 2020

Contents

- What is the awesome list?
 - History, concept, status
- Zooming out:
 - What do we actually want to achieve?
 - Re-use / basic building blocks / communities
- uRos and the list
- Wrap-up



What is the 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 Statistics Netherlands' Statistical Computing group, SNStatComp maintained by the **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**



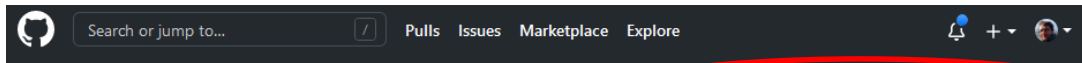
What is the awesome list?

Curated list of software for official statistics



awesome

www.awesomeofficialstatistics.org



Social interactions

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



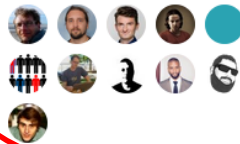
This work is licensed under a [Creative Commons Attribution 4.0 International License](#).

Open license

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

[official-statistics](#) [gsbpm](#)

Contributors 15



+ 4 contributors

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@gmail.com or olav.tenbergh@gmail.com or tweet [@markvdloo](https://twitter.com/markvdloo)

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 for data on the ISTAT website.
- Java application [URLScorer](#). Gives a rule based score to scraped data.
- Node.js tool [RobotTool](#). A tool for checking (price) changes on the internet.
- 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.

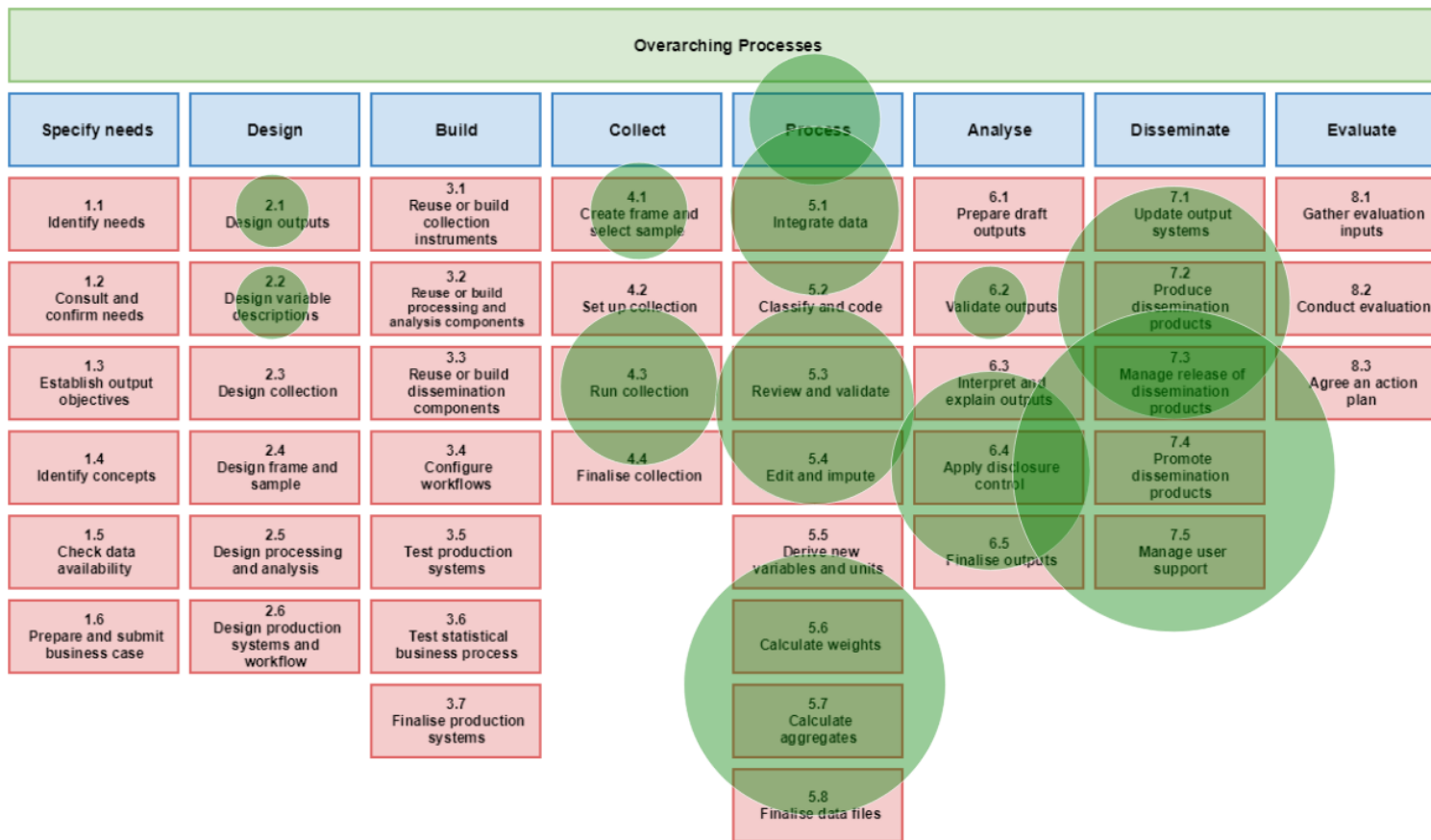
The screenshot displays a list of software projects on the 'mentioned in awesome' website. The projects shown are:

- SamplingStrata**: A package for determining the minimum sample size in a multivariate and multidomain case. It features a genetic algorithm. The badge for this project is circled in red.
- Tau-Argus Open Source**: Software to apply Statistical Disclosure Control techniques. The badge for this project is circled in red.
- R package SmallCountRounding**: A package for Small Count Rounding of Tabular Data. The badge for this project is circled in red.

At the bottom of the page, there is a circular logo for the **.STAT SUITE** ecosystem, which includes **.STAT DATA EXPLORER**, **.STAT CORE**, and **.STAT DATA LIFECYCLE MANAGER**. The badge for this project is also circled in red.

Each project entry includes a badge that says "mentioned in awesome" with a link icon. The badges for SamplingStrata, Tau-Argus Open Source, and SmallCountRounding are circled in red. The badge for the .STAT SUITE ecosystem is also circled in red.

Awesome list by GSBPM



Zooming out: what do we actually want?

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

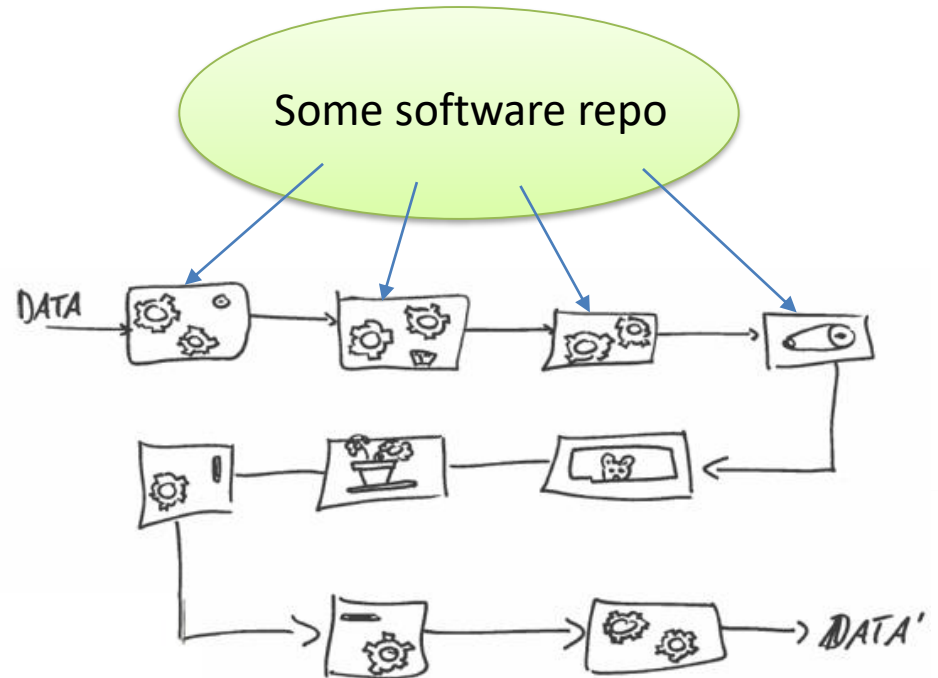
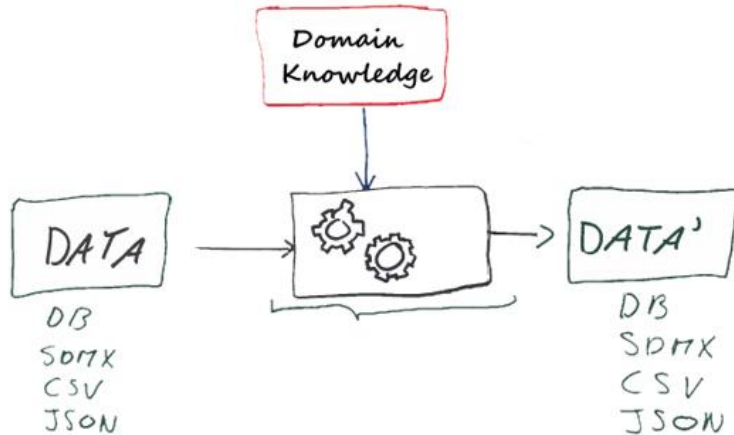
Standardisation

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



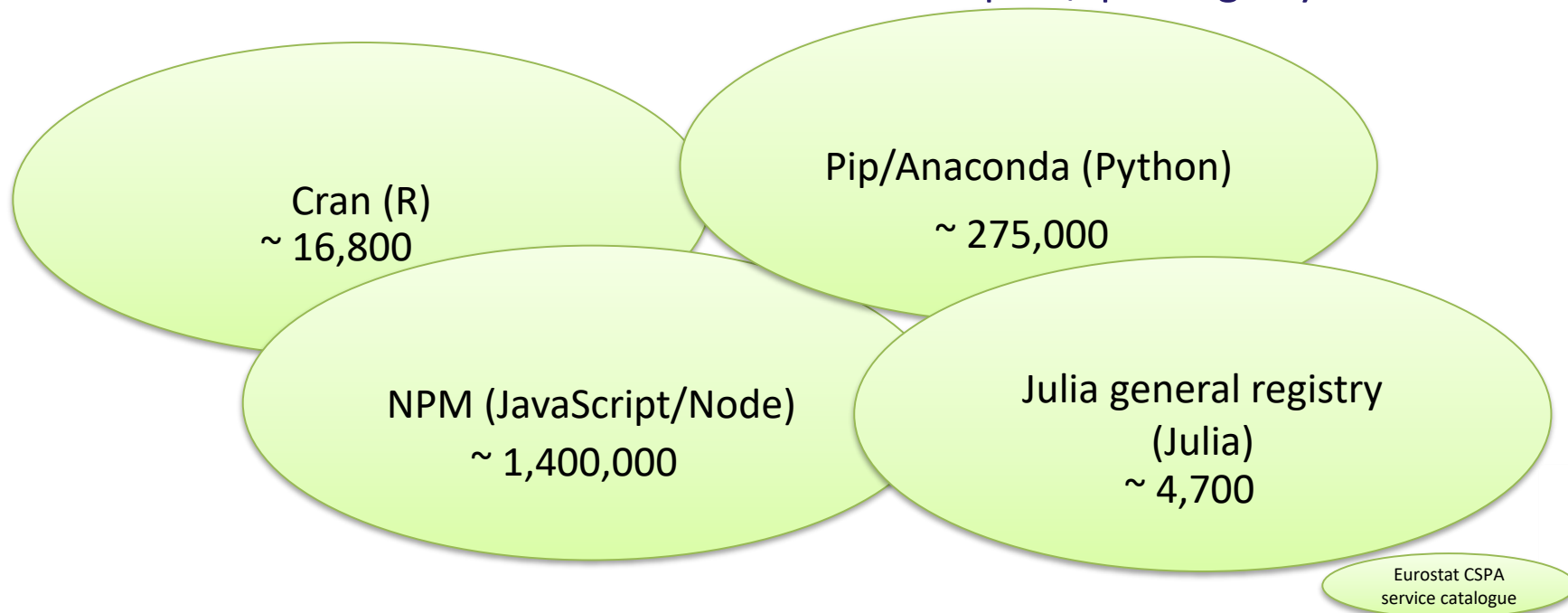
Basic building blocks

- 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

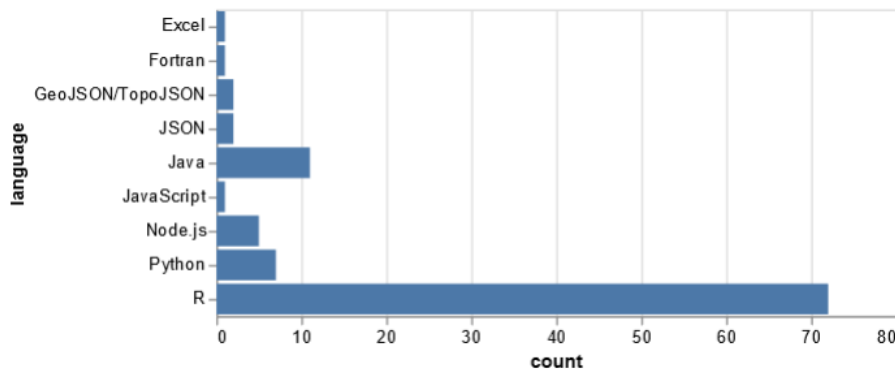
- Software sharing is already happening
- Different communities have their own repos / package systems



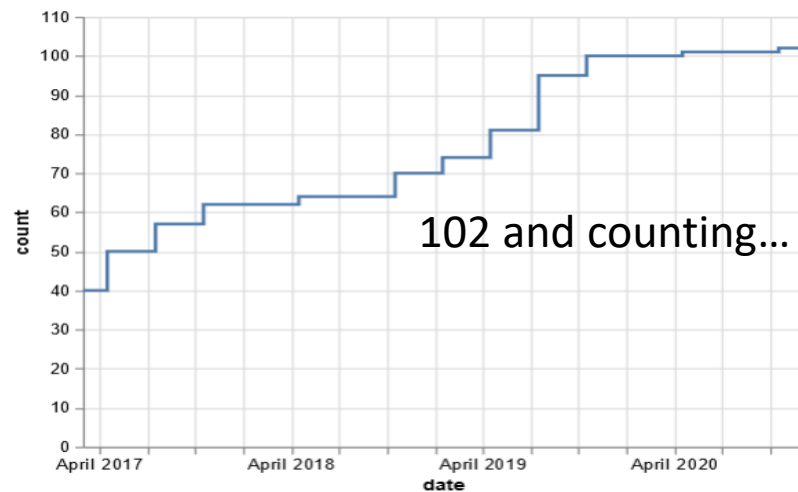
Awesome list status

- 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



Virtual ☹️

Some uRos2020 software not on the list (1)



Methods:

- [emdi](#) (2.0.1): estimates based on Area-Level models
- `mquantreg`: estimates generalized linear M-quantile regression models

Classifications:

- [klassR](#): classifications from StatsNorway API's (what about other API's?)
- `<code>`: Classifying nonprofit organizations by field of activity (any plans for a generic implementation?)
- SwissCheese: Balanced Imputation for Swiss Cheese Nonresponse

Lightning talks:

- [sdcLog](#): Utilities for SDC in Research Data Centers
- `<code>`: OCR unstructured annual reports (could it be of use for other NSIs?)
- Riot: R Input-Output Tools



Some uRos2020 software not on the list (2)

- [cellKey](#) + Shiny GUI: for SDC using random noise
- [timeseriesdb](#): Time Series with R and PostgreSQL



Dissemination and visualization:

- [migraR](#): migration analysis (on [GH](#), plans for CRAN)
- [SSBtools](#)->HierarchyCompute: multidimensional hierarchical aggregations (VTL was too inefficient!)

R in the Organisation and in Production:

- [useSTAT](#): DevOps and R in Austria (generalizable?)
- [Persephone](#): OO wrapper around RJDemetra (on [GH](#))



Wrap-up

- www.awesomeofficialstatistics.org



A community instrument to *re-use* generic software for official statistics

- Please consider *contributing* your (R) software if:
 - it is used in the production (or provides access to) *official statistics*
 - it could be a valuable basic *building block* for others as well
 - it is reasonably *well-documented* and easy to *install*



Questions, ideas, suggestions



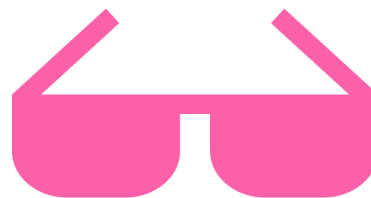
Olav ten Bosch

Mark van der Loo

o.tenbosch@cbs.nl @kobosch


mpj.vanderloo@cbs.nl @markvdloo

Curated list of software for
official statistics



awesome

www.awesomeofficialstatistics.org

Please  Star !

