



The awesome list of official statistics software & FOSS principles

www.awesomeofficialstatistics.org

Olav ten Bosch

**Discoverability & Sharing sub-team meeting of the UNECE HLG-MOS Open-Source
project, 22-05-2024**

Contents

- What is this awesome list?
- Zooming out: what is the aim?
- CBS history, FOSS best practices
- Eurostat OS4OS principles (Eurostat OS4OS group)
- 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



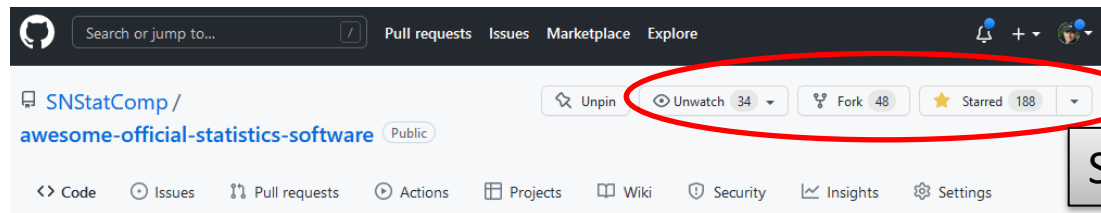
What is the awesome list?

Curated list of software for
official statistics



awesome

www.awesomeofficialstatistics.org



Social interactions, watch

Awesome official statistics software

Criteria

An awesome list of open source software for official statistics

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.

License



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

Open license

Contributors 17



+ 6 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 dot vanderloo at gmail dot com](#) or [olav dot tenbosch at gmail dot com](#) or tweet [@olavtenbosch](#) or [@markvdloo](#)

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 License

Excel [SDMX_Matrix_Generator](#). Excel-based visual SDMX and 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-ARGUS](#)

- CRAN 5.7.6 - 2 months ago license GPL-2

R package [sdcMicro](#). Disclosure control for statist

- CRAN 0.32.6 - 4 months ago

R package [sdcTable](#). Disclosure control for tabul

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 probability 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

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 pairs, comparing records, em-algorithm for estimation, also be used for pre- and post-processing for machine

- CRAN 0.4-12.4 - a year ago license GPL (>= 2)

R package [RecordLinkage](#). Implementation of the Fellegi-Sun

- CRAN 1.4.1 - 2 years ago license GPL (>= 2)

R package [StatMatch](#). Statistical Matching or Data F

- CRAN 0.6.1 - 24 days ago license GPL (>= 3)

R package [fastLink](#). Implements a Fellegi-Sunter procedure and the inclusion of auxiliary information. [Documentation](#)

- CRAN 0.9.12 - 13 days ago license GPL-3

R packages [stringdist](#). Approximate string matching (Levenshtein, Hamming, Levenshtein, optimal string alignment), q-gram (Jaro, Jaro-Winkler). An implementation of soundex

- CRAN 0.1.6 - 4 years ago license MIT + file LICENSE

Over 30 software packages, giving access to > 60 data providers
majority are R-packages

Access to official statistics (GSBPM 7.4)

- CRAN 0.6-3 - 7 months ago license GPL (>= 2)

R package [rsdmx](#). Access to data or metadata from statistical organisations through SDMX. The package contains a list of SDMX access points of various national and international

- CRAN 0.3.1 - 7 months ago license GPL-3

R package [readsdmx](#). Read SDMX into dataframes from local SDMX-ML file or from OECD.

- GitHub v2.14.0 last commit last wednesday license Apache-2.0

Python [sdmx](#). Python library that implements SDMX 2.1 to explore data from SDMX data and metadata and convert it into Pandas objects.

- CRAN 0.4.3 - 7 months ago license MIT + file LICENSE

R package [rjstat](#). Read and write data sets in the JSON-stat format.

- PyPI v2.4.0 license Apache License 2.0

Python [pyjstat](#). Read and write JSON-stat.

- GitHub v0.2.8 last commit march 2023 license MIT

Java application [json-stat.java](#). Read and write JSON-stat. By Statistics Norway

- CRAN 0.2.5 - 2 years ago license CC0

R package [oecd](#). Search and Extract Data from the OECD

- CRAN 0.8.21 - 7 months ago license BSD_2_clause + file LICENSE

R package [sorvi](#). Finnish Open Government Data Toolkit

- CRAN 4.0.0 - 3 months ago license BSD_2_clause + file LICENSE

R package [eurostat](#). Tools to download data from the Eurostat database together with manipulation utilities.

- CRAN 0.22.5 - 3 months ago license EUPL

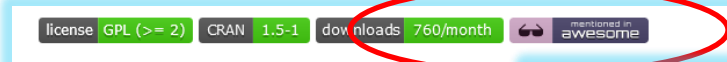
R package [restatapi](#). Search and retrieve data from Eurostat database, by Euro

- CRAN 2.1.4 - 5 years ago license GPL-3

The right to wear the badge

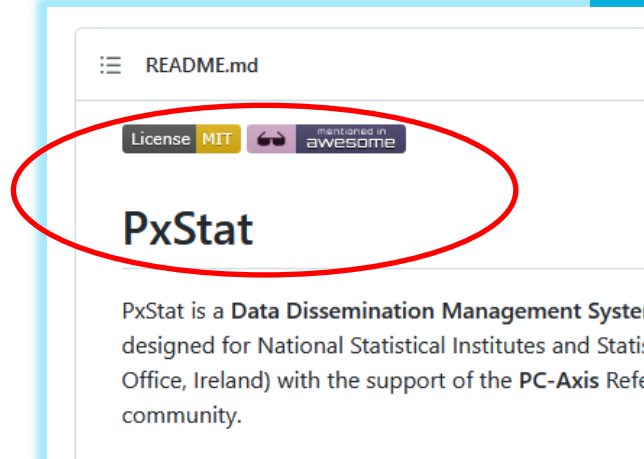
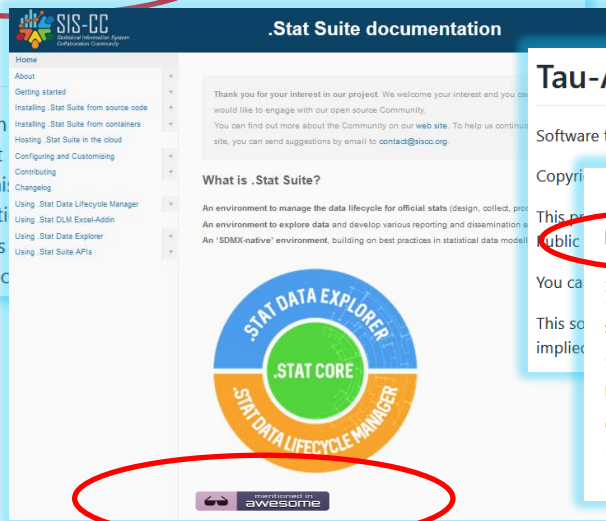
- The badge links to the list and improves findability:

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.



SamplingStrata

This package offers an approach for the determination frame, the one that ensures the minimum sample cost constraints in a multivariate and multidomain case. This genetic algorithm: each solution (i.e. a particular partition) is considered as an individual in a population; the fitness algorithm to calculate the sampling size satisfying prec



PxStat

PxStat is a **Data Dissemination Management System** designed for National Statistical Institutes and Statistics Office, Ireland) with the support of the **PC-Axis Reference community**.

Tau-Argus Open Source

Software to apply Statistical Disclosure Control techniques

R package SmallCountRounding

This package is part of the European Union

Small Count Rounding of Tabular Data

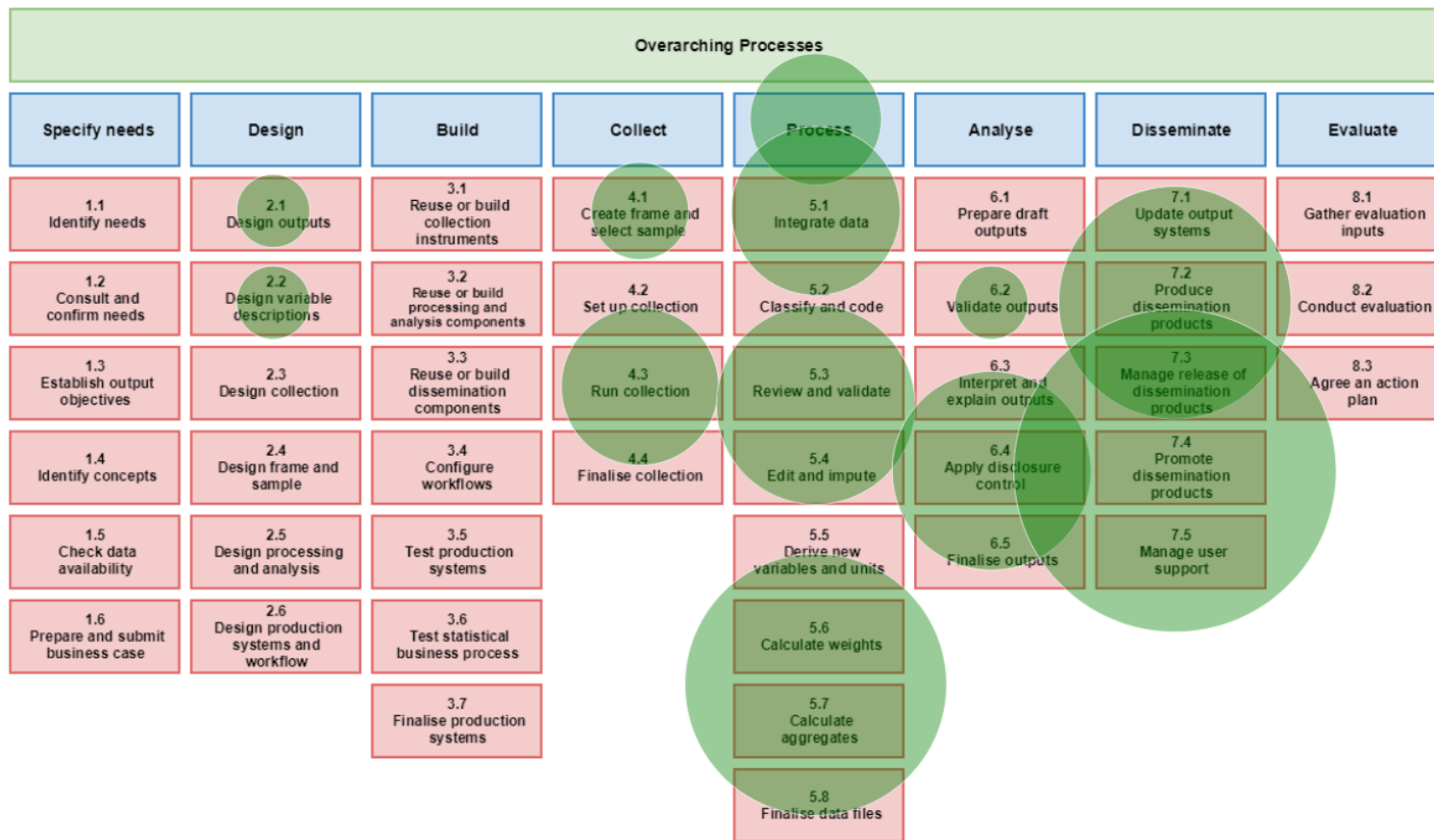
See the package vignette: [Introduction to 'SmallCountRounding'](#)

Installation from CRAN

(Recommended, unless you want to test the newest changes.)

```
install.packages("SmallCountRounding")
```

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

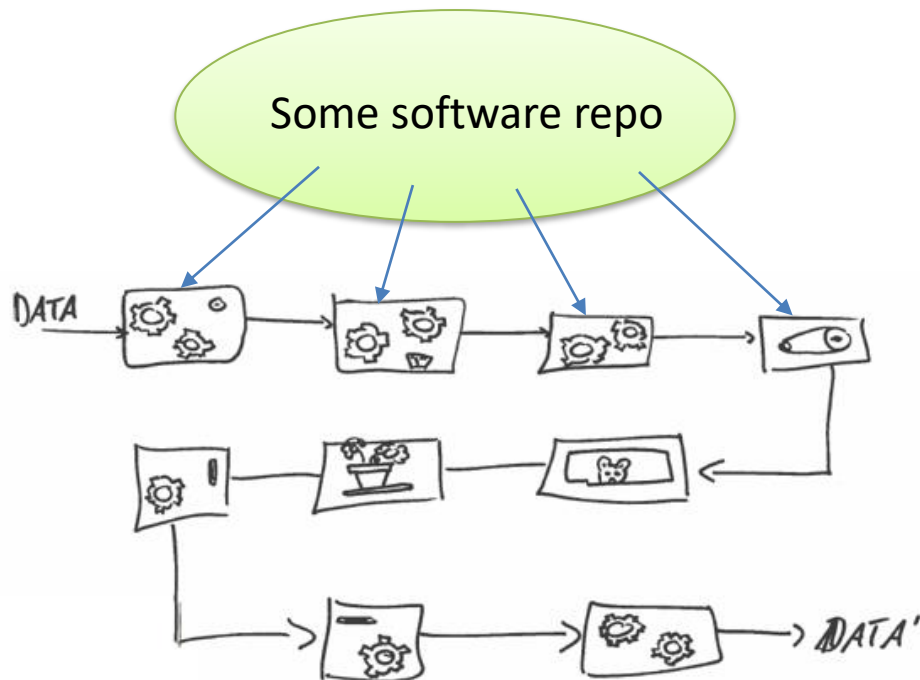
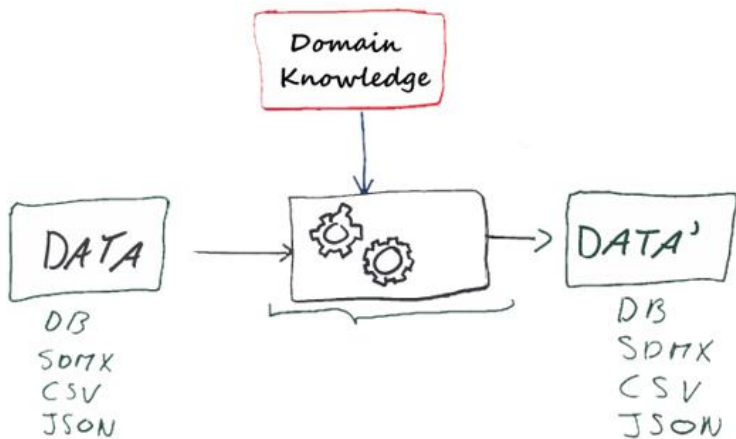
Standardisation

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

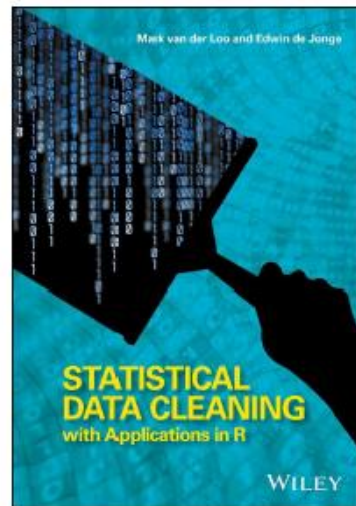


Some important CBS R-packages

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

R data cleaning ecosystem

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



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



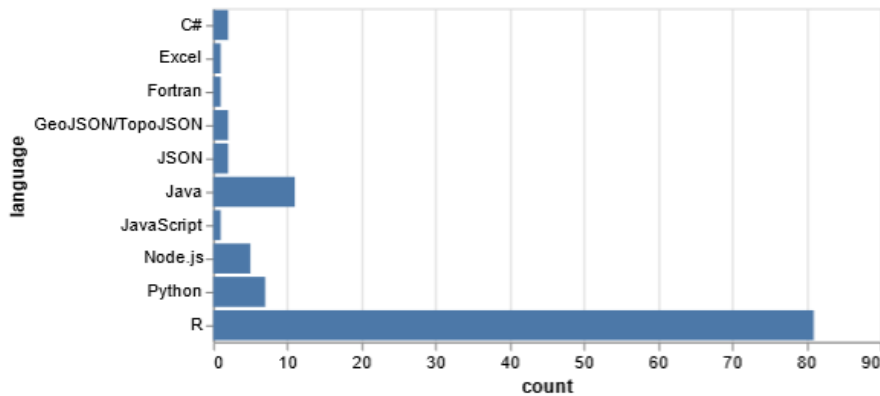
on [YouTube](#)



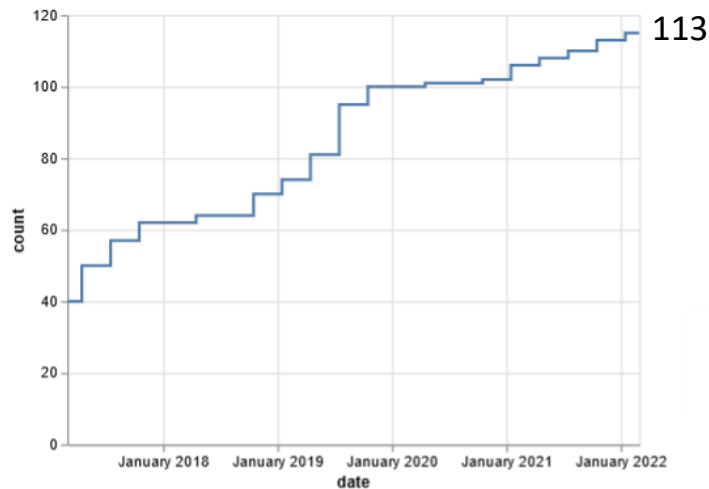
Awesome list status

- 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 ☹️



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

CBS FOSS best practices (2021)

- ***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)



ESS OSS Principles (Eurostat 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>

Related



Access to official statistics from R: an overview

Statistics Netherlands

Olav ten Bosch, Edwin de Jonge
uRos2023 12-14 December 2023



Based on awesome list....
Another time?



To be FAIR, what is missing in Official Statistics?

Statistics Netherlands

Olav ten Bosch, Edwin de Jonge, Henk Laloli
COSMOS 2024, 11-12 April Paris



CC BY 4.0

71st CES meeting

UNEP	STATISTICS
Statistics	71st plenary session of the Conference of European Statisticians
Fundamental Principles of Official Statistics	Statistics 2023
About us	22 June (14:30) - 23 June (17:30) 2023
Networks of experts	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
EC/ECE/2023/19 - Journey to using R - experience of the Central Statistics Office of Ireland	pdf
EC/ECE/2023/20 - Free and Open-Source Software at Statistics Netherlands	pdf
EC/ECE/2023/21 - "DIAPLA" - a cloud-based statistical production system and its implications for Statistics Norway	pdf
EC/ECE/2023/22 - Open-source solutions at Statistics Poland	pdf
EC/ECE/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



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
Mark van der Loo

o.tenbosch@cbs.nl
mpj.vanderloo@cbs.nl

@olavtenbosch
@markvdloo

