Statistical Open-Source Software for Official Statistics: State of Play and Future Directions

Statistics Netherlands

Olav ten Bosch, Mark van der Loo

Istat, Fourth Workshop on Methodologies for Official Statistics
1-2 December 2025

CC BY 4.0

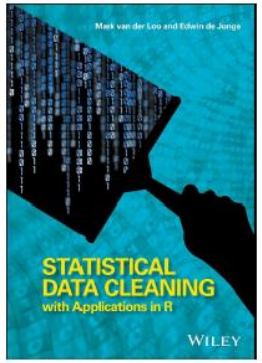
State of play

A brief history on FOSS at Statistics Netherlands

Period	Milestone
≤ 2009	R used only in research; first R package published
2010	R adopted as formal tool, first FOSS policy
2012	Python as formal tool
2014	Git as formal tool (succeeding SVN)
2017	**CBS starts awesomelist for OS software**
2018	CBS hosts uRos2018
2019/2020	New FOSS policy, Software Development Guidelines
2023	ESS FOSS principles
2024	IT starts getting interested in FOSS as users (K8S)
2025	**FOSS Principles Adopted at UNECE Conference of European Statisticians**

Example: 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
- ***simputation***: many different imputation methods
- ***rspa***: adapt numerical records to fit (in)equality restrictions
- ***reclin2:*** probabilistic record linkage and deduplication
- ***deductive***: solve errors based on control rules
- ***validatetools***: find inconsistencies and redundancies
- ***accumulate:*** advanced group aggregation
- ***synthesizer:*** turn (micro)data into synthetic data
- ***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**

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

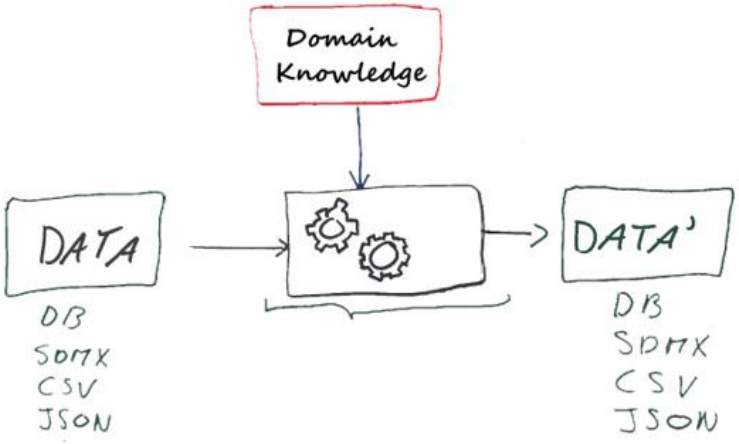
Use well-tested and proven implementations of generic methods

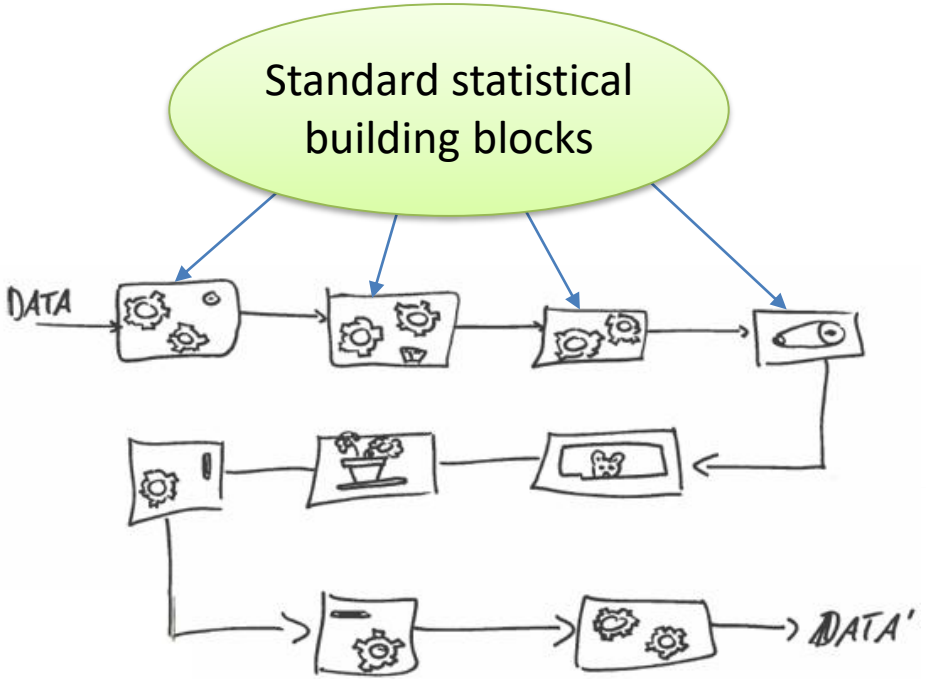
Standardisation

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





Awesome list of official statistics software

- Started in 2017 (The Hague)
- A ***community approach*** to ***remember useful software***
- A ***public*** list of links, clear and simple ***criteria***
- ***awesomeofficialstatistics.org***

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 r
  pairs, comparing records, em-algorithm for estimating m- and u-p
  also be used for pre- and post-processing for machine learning m
- CRAN 0.4-12.4 – 2 years ago | license GPL (>= 2)
  R package RecordLinkage. Implementation of the Fellegi-Sunter m
- 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 reco
  and the inclusion of auxiliary information. Documentation.
- CRAN 0.9.12 – a year ago | license GPL-3

Estimation and weighting (GSBPM 5.6 | 5.7)

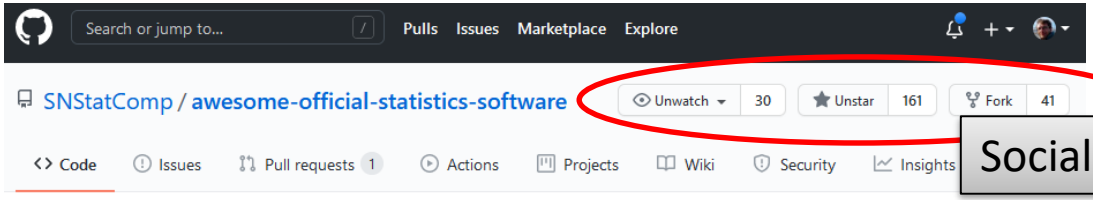
- CRAN 4.4-8 – 3 months ago | license GPL-2 | GPL-3
  R package survey. Weighting and estimation for complex survey designs, possibly under nonresponse. Also computes estimator variance. See also R package srvyr for integration with tidy tools.
- CRAN 1.2 – 4 years ago | license GPL-3
  R package hbsae. Small area estimation based on hierarchical Bayesian models.
- CRAN 0.8.0 – 2 months ago | license GPL-3
  R package mcmcsae. Small area estimation based on Markov Chain Monte Carlo simulation.
- CRAN 0.3 – 2 years ago | license GPL-3
  R package rsae. Small area estimation based on (robust) maximum likelihood estimation.
- CRAN 1.3.0 – 5 years ago | license GPL-2
  R package CalibrateSSB. Calculate weighs and estimates for panel data with non-response.
- CRAN 0.1.0 – 2 years ago | license GPL-3

How does it work?

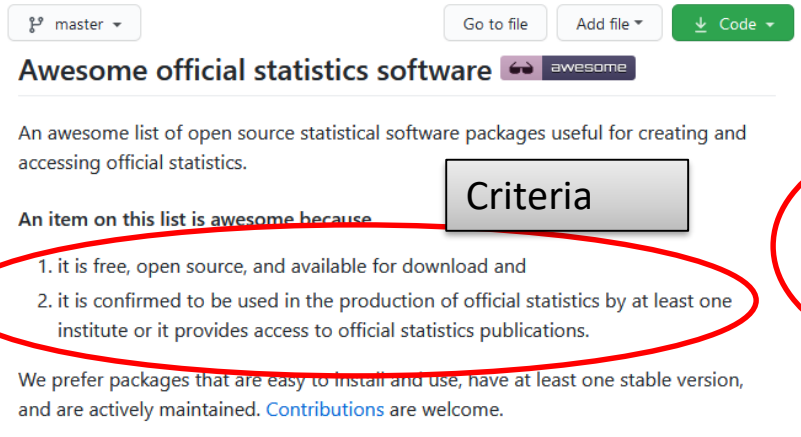
Curated list of software for official statistics

awesome

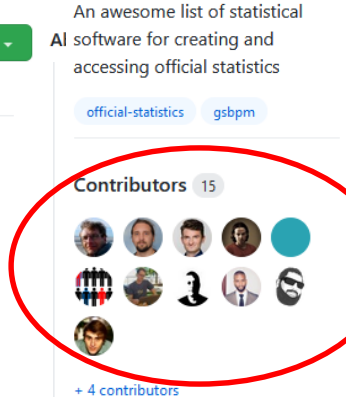
www.awesomeofficialstatistics.org



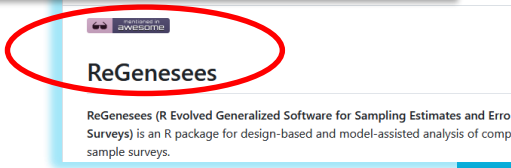
Social interactions



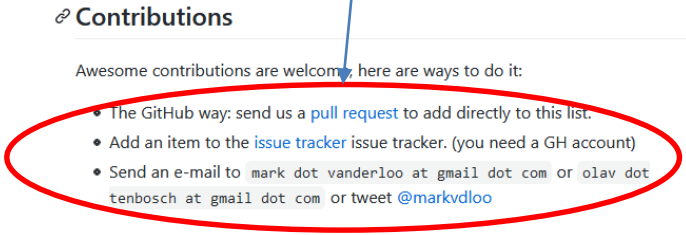
Criteria

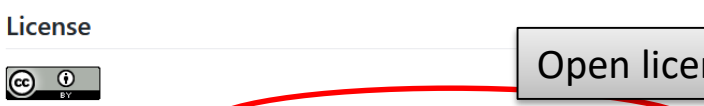


The right to wear the badge

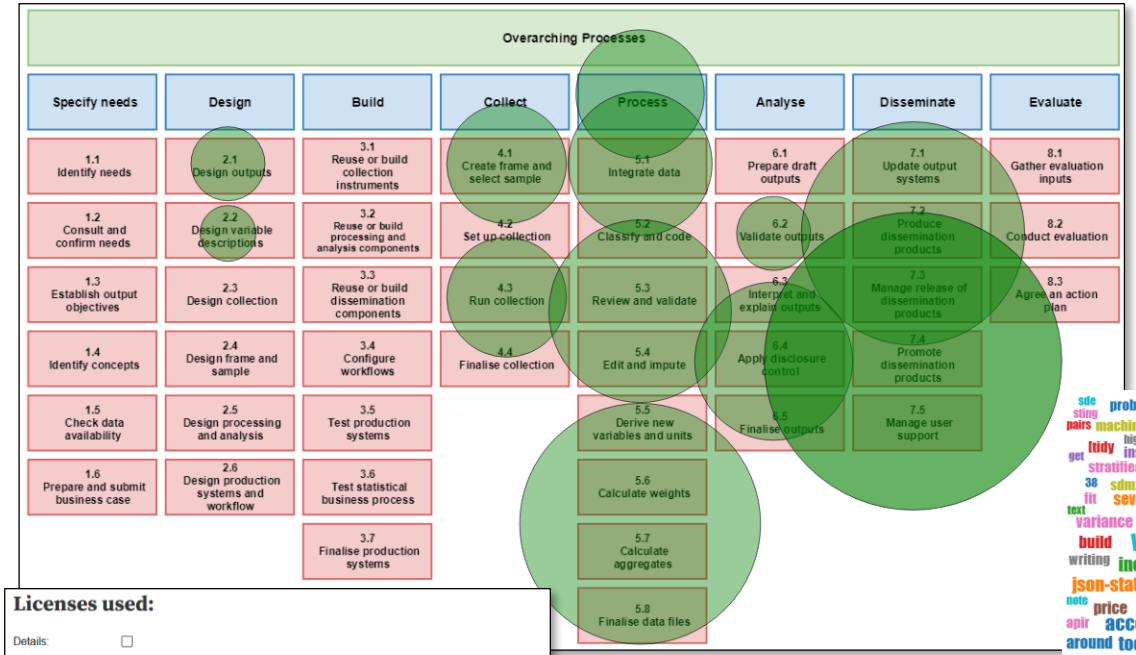


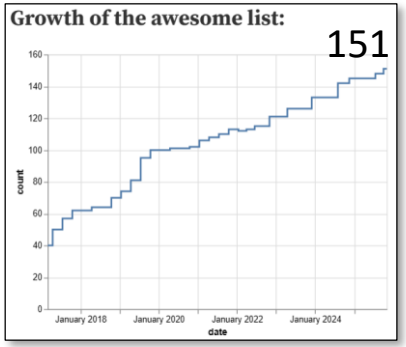
Working together

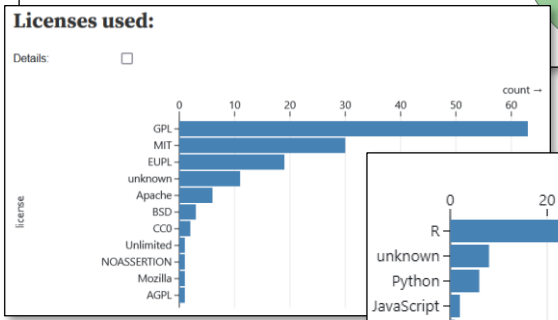


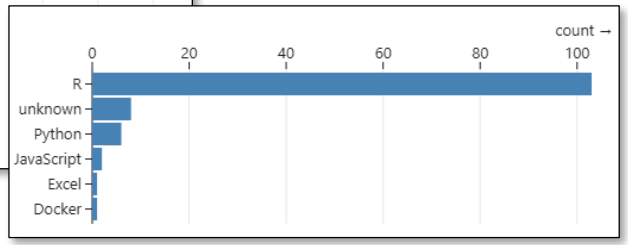


Open license










```
{
  "type": "R package",
  "name": "sdcSpatial",
  "url": "https://cran.r-project.org/package=sdcSpatial",
  "description": "Create privacy protected density maps from location data. Includes visu
  "gsbpm_name": "Statistical disclosure control",
  "gsbpm_number": "6.4",
  "license": "GPL-2",
  "licenseSimple": "GPL",
  "platform": "CRAN",
  "languages": [
    "R"
  ]
},
```


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

metadata/metadata.json

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

UNECE

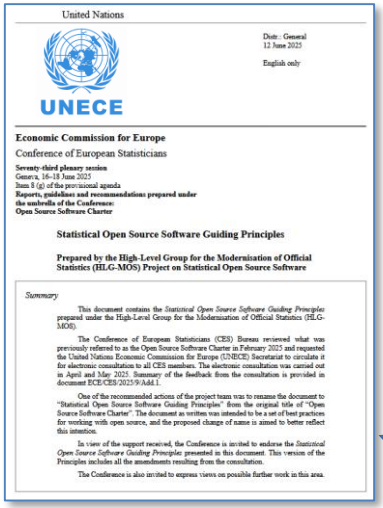
HLG-MOS Project on Statistical Open Source Software (2024)

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

UNECE Statistical OSS Guiding Principles

June 2025: endorsed by Conference of European Statisticians (CES)

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



United Nations

UNECE

Distr.: General
12 June 2025

English only

Economic Commission for Europe

Conference of European Statisticians

Seventy-third plenary session
Geneva, 16-18 June 2025
Item 8 (g) of the provisional agenda
Reports, guidelines and recommendations prepared under the umbrella of the Conference: Open Source Software Charter

Statistical Open Source Software Guiding Principles

Prepared by the High-Level Group for the Modernisation of Official Statistics (HLG-MOS) Project on Statistical Open Source Software

Summary

This document contains the *Statistical Open Source Software Guiding Principles* prepared under the High-Level Group for the Modernisation of Official Statistics (HLG-MOS).

The Conference of European Statisticians (CES) Bureau reviewed what was previously referred to as the Open Source Software Charter in February 2025 and requested the United Nations Economic Commission for Europe (UNECE) Secretariat to circulate it for electronic consultation to all CES members. The electronic consultation was carried out in April and May 2025. Summary of the feedback from the consultation is provided in document ECE/CES/2025/9/Add.1.

One of the recommended actions of the project team was to rename the document to "Statistical Open Source Software Guiding Principles" from the original title of "Open Source Software Charter". The document as written was intended to be a set of best practices for working with open source, and the proposed change of name is aimed to better reflect this intention.

In view of the support received, the Conference is invited to endorse the *Statistical Open Source Software Guiding Principles* presented in this document. This version of the Principles includes all the amendments resulting from the consultation.

The Conference is also invited to express views on possible further work in this area.

Document (pdf): Statistical Open Source Software Guiding Principles
Charter (website): UNECE Open Source Software Principles

Reflections on State of Play

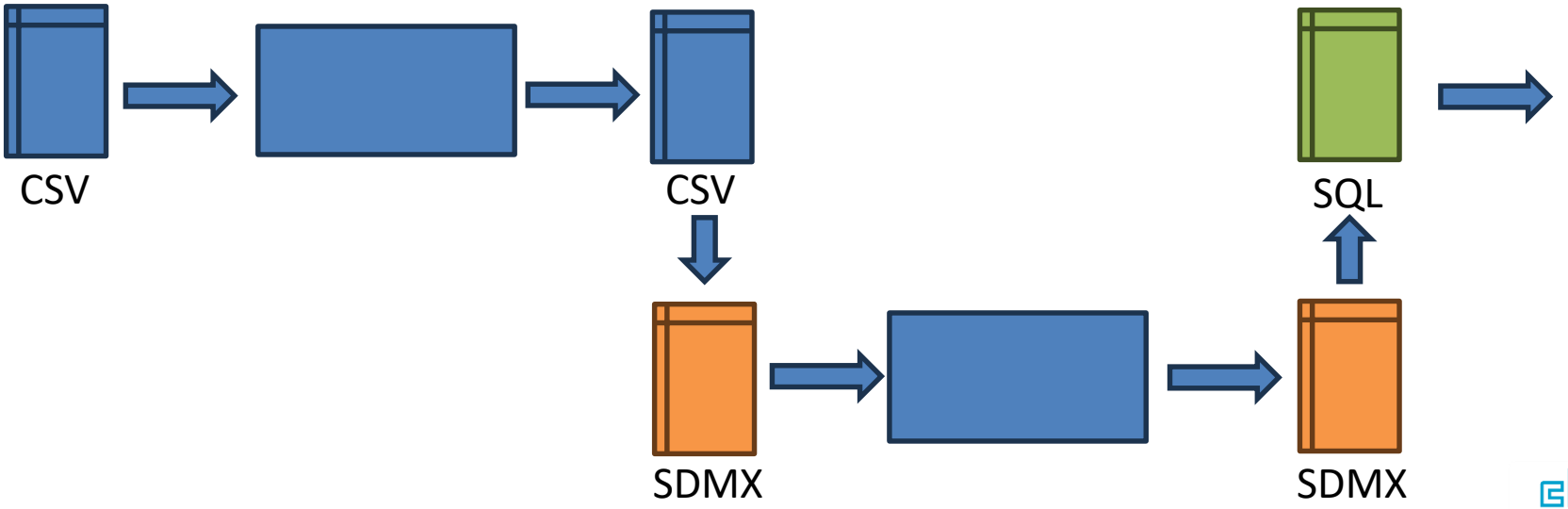
- Increased adoption of ***open source software*** in official statistics
- ***The awesome lists*** helps to know what's already there
- ***UNECE OSS guiding principles*** internationally adopted

However:

- Can we futher mature as a community?
- How to compose OSS-based processes in a correct way from a methodological and IT viewpoint?

Future directions and open question

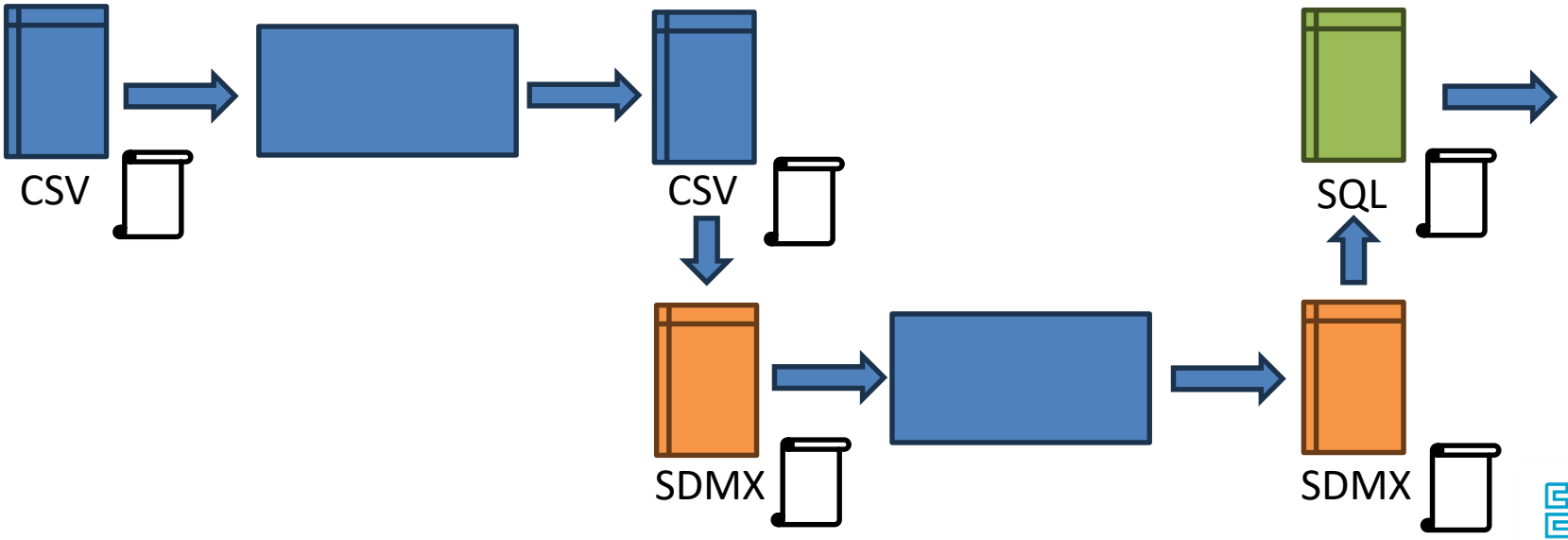
Technical compatibility



*Do components support compatible **data standards**?*

- CVS, JSON, SDMX, DB, R/Python dataframes, tool-specific formats

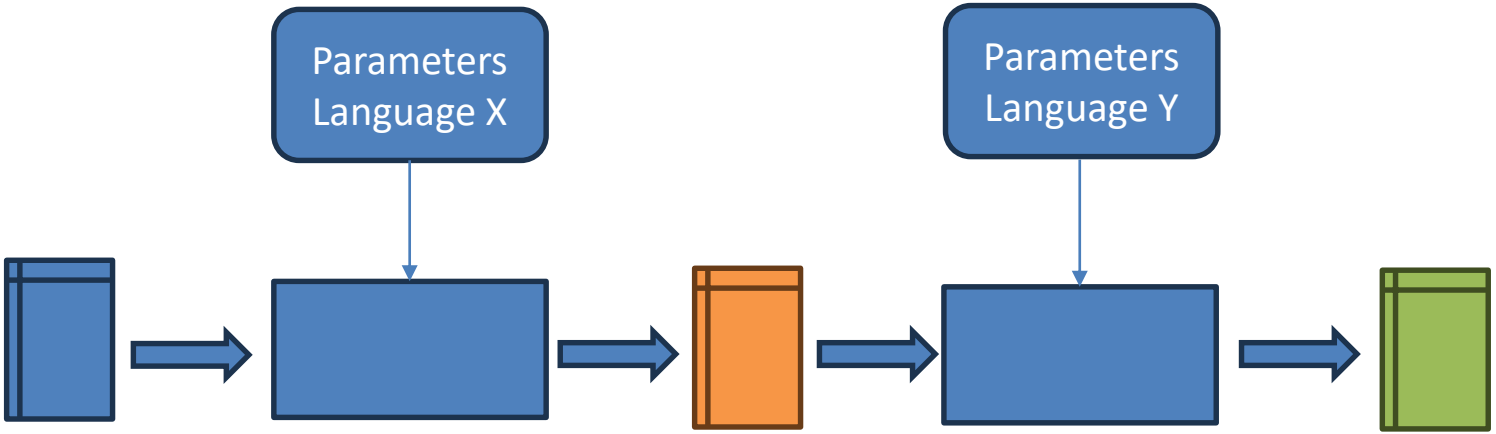
Semantic compatibility



*Do components support compatible **metadata standards**?*

- *both structural and reference metadata: DSD's, DDI, RDF, SIMS*

Compatibility on configuration level



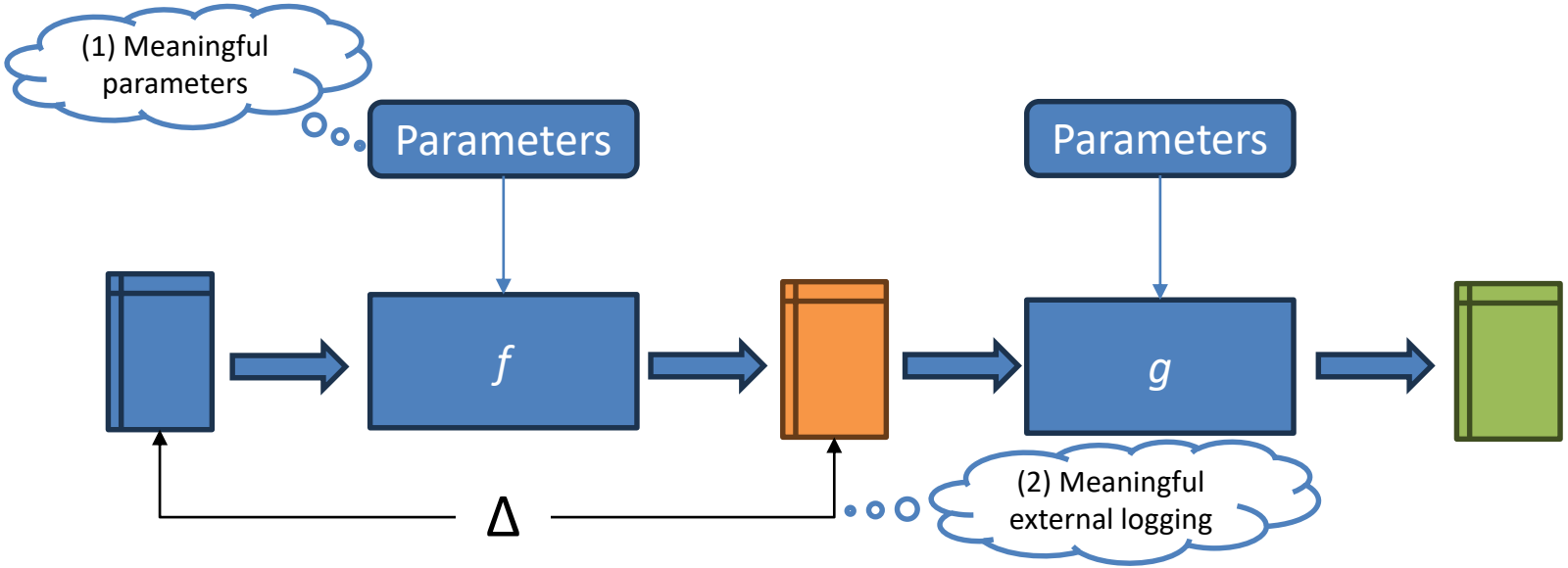
- *Do components use the same configuration concepts?*

Granularity of Statistical Building Blocks

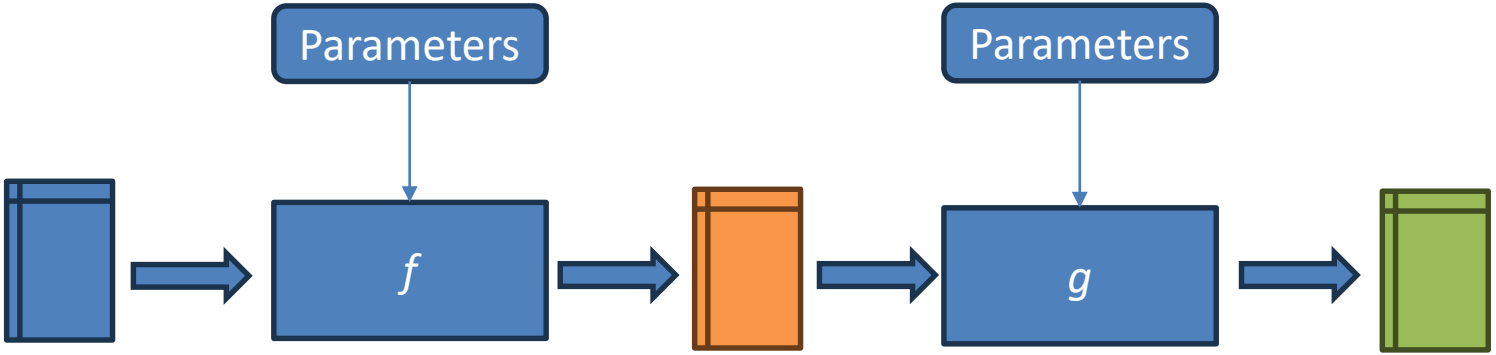
A statistical component should be:

(1) large enough such that subject matter experts recognize its function as a statistical task, but

(2) small enough such that its action upon a dataset can be understood by comparing output with input.

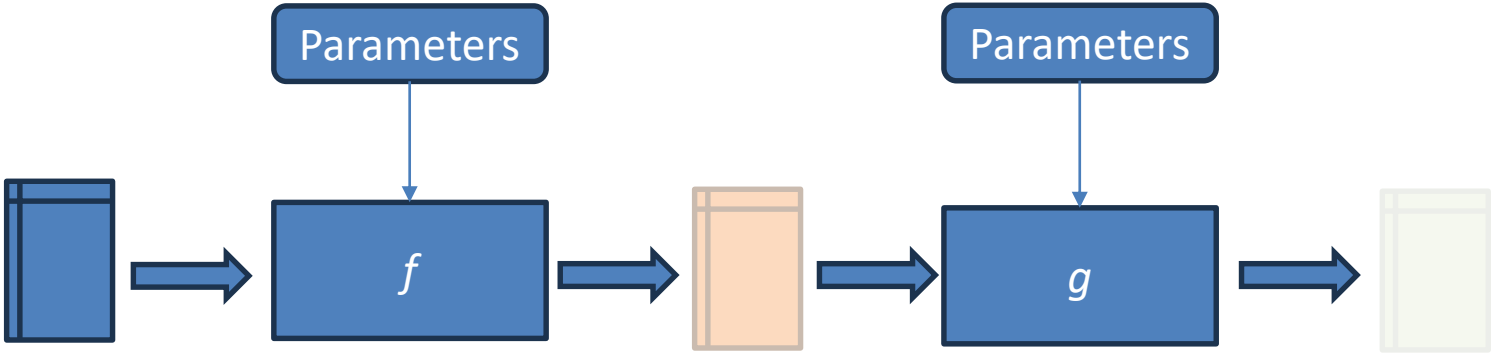


Truly independent software modules



- *Updating building block f should not force one to update or reconfigure g*
- *Updating parameters of f should not force one to update parameters of g*

Propagation of uncertainty in pipelines



- *Uncertainty of the whole chain depends on uncertainty of preceding building blocks*
- *How do we gain insight into the uncertainty in results automatically when configuring a system?*

Further progress

- What about (Gen)AI? should we share official statistics ***models***?
- Should we create an ***“awesome_verse”***? set of ***mature*** packages
- We need some kind of combined ***maturity score***
  could be based on downloads, community activity,
  documentation, test coverage, number of mentions, etc.
- What about rOpenSci? https://stats-devguide.ropensci.org
- Challenge: who uses what:
  “State of open source in official statistics”?

Wrap-up

1. We all have come a long way:
    1. Increased adoption of OSS in official statistics
    2. Awesome lists helps to know what's there
    3. UNECE OSS guiding principles internationally adopted
2. There is still much to do:
    1. Composability: from OSS building blocks to OSS process chains
    2. How to mature as a community

Statistical Open Source Software for Official Statistics: State of Play and Future Directions

Olav ten Bosch, o.tenbosch@cbs.nl , Statistics Netherlands

Mark P.J. van der Loo, mpj.vanderloo@cbs.nl , Statistics Netherlands and Leiden University

Questions, ideas, suggestions welcome!

To appear in SJIAOS (2025)