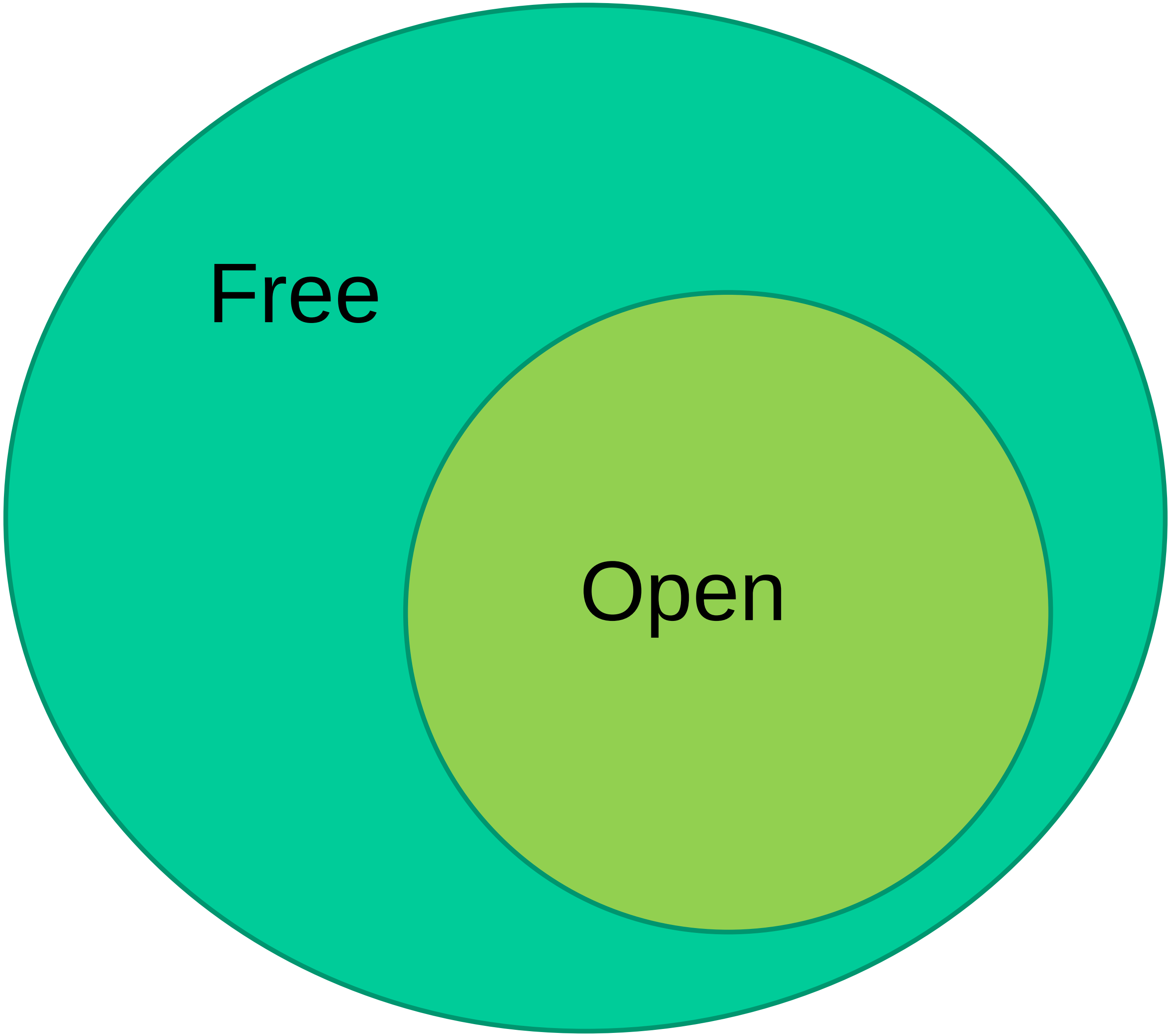
United Nations Economic Commission for Europe
Statistical Division

Free and Open Source Statistical Software

Steven Vale
UNECE
steven.vale@unece.org


Free
Open

Examples

Not free, not open – SAS

Free, not open – PC-Axis

Free and open – Tau-Argus

What about “shared source”?

= Free and open, but only within a specified community

Free and open is the ideal

but we should be pragmatic!

Should we also consider software that does not (yet) fully meet this ideal?

What should be open and shared?

- Algorithms and methods?
- Libraries?
- Code lists?
- Implementations?
- User interfaces?
- Complete packages? – “Plug and Play!”

Where to find free and open source statistical software?

- The "Awesome List"
  - Started by Mark van der Loo and Olav ten Bosch – Statistics Netherlands (thanks for the slides Olav!)
  - https://github.com/SNStatComp/awesome-official-statistics-software

The Awesome list of Official Statistics Software

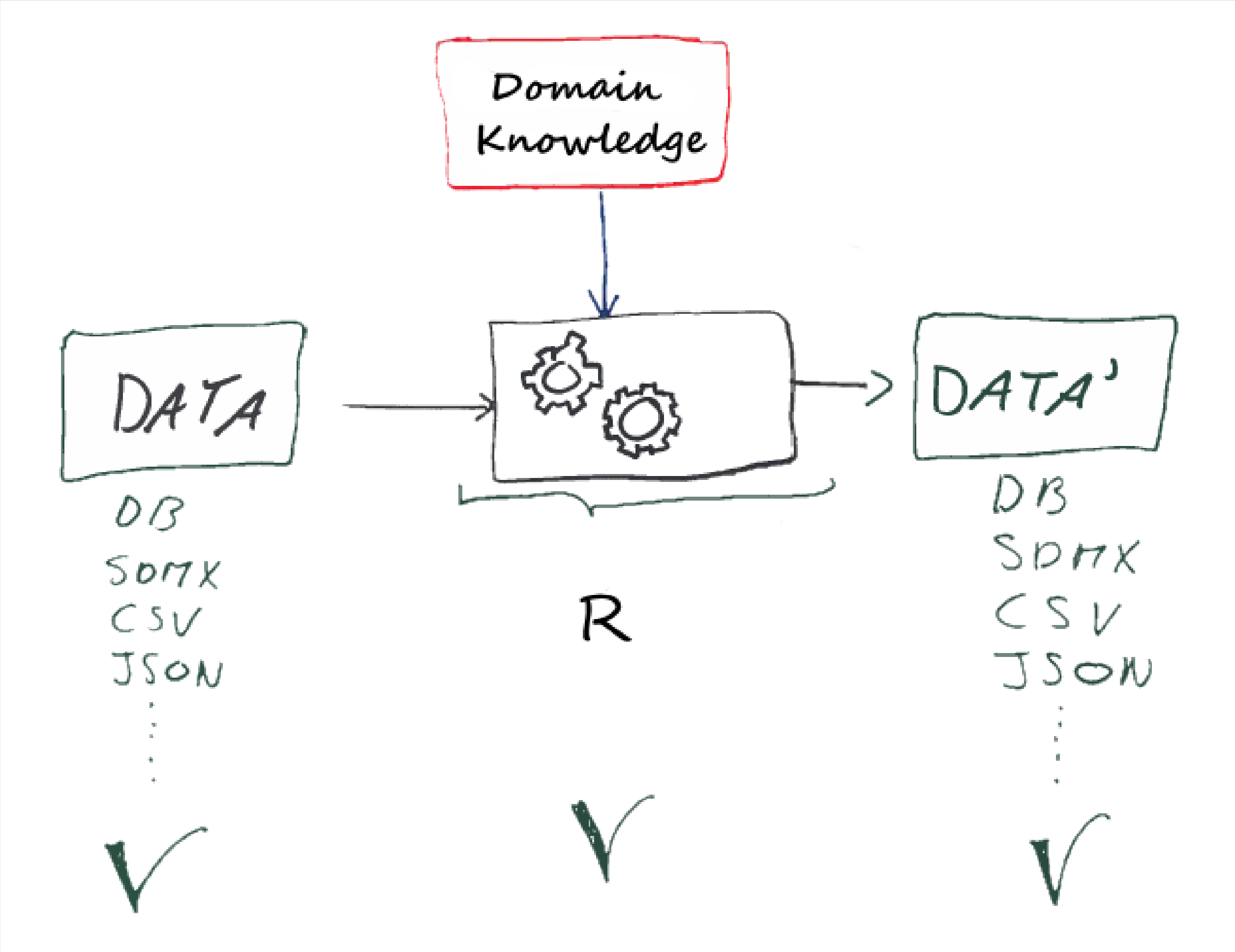
SCFE workshop, Wiesbaden

Olav ten Bosch, 7 July 2017 (screenshots updated Dec 2018)

Statistics Netherlands

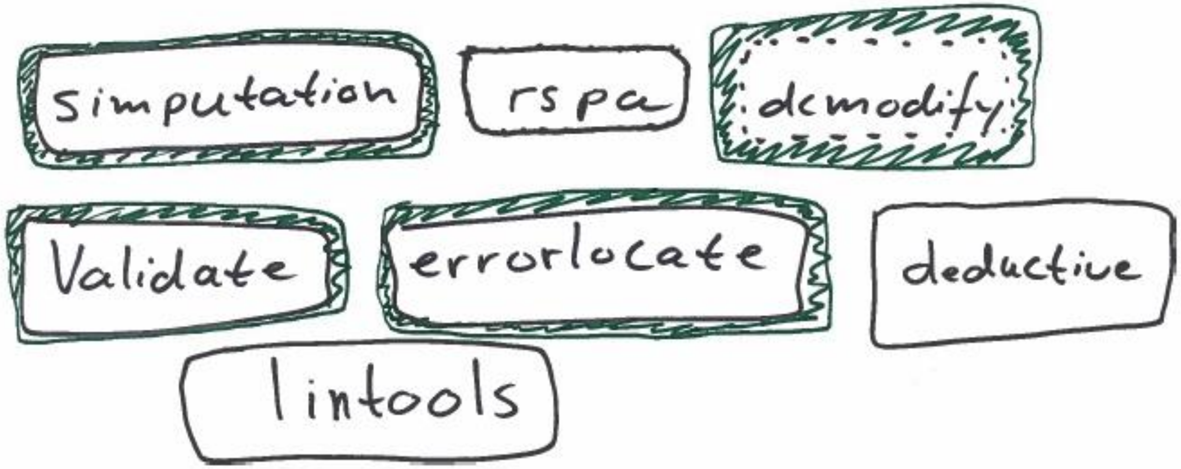
R tools for data validation, correction and imputation

Concept

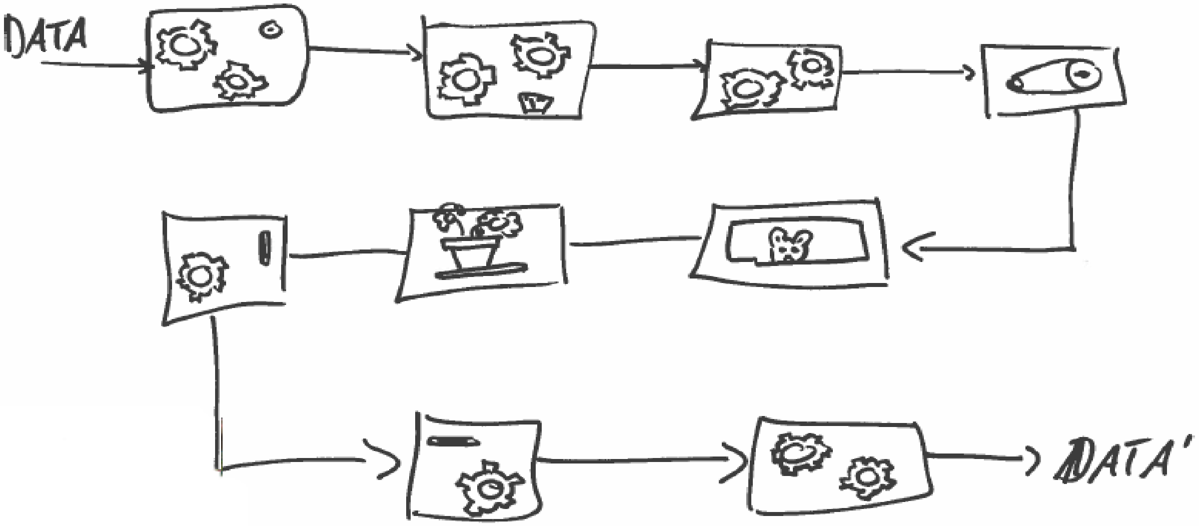


R tools for data validation, correction and imputation

Available Packages



Chainable

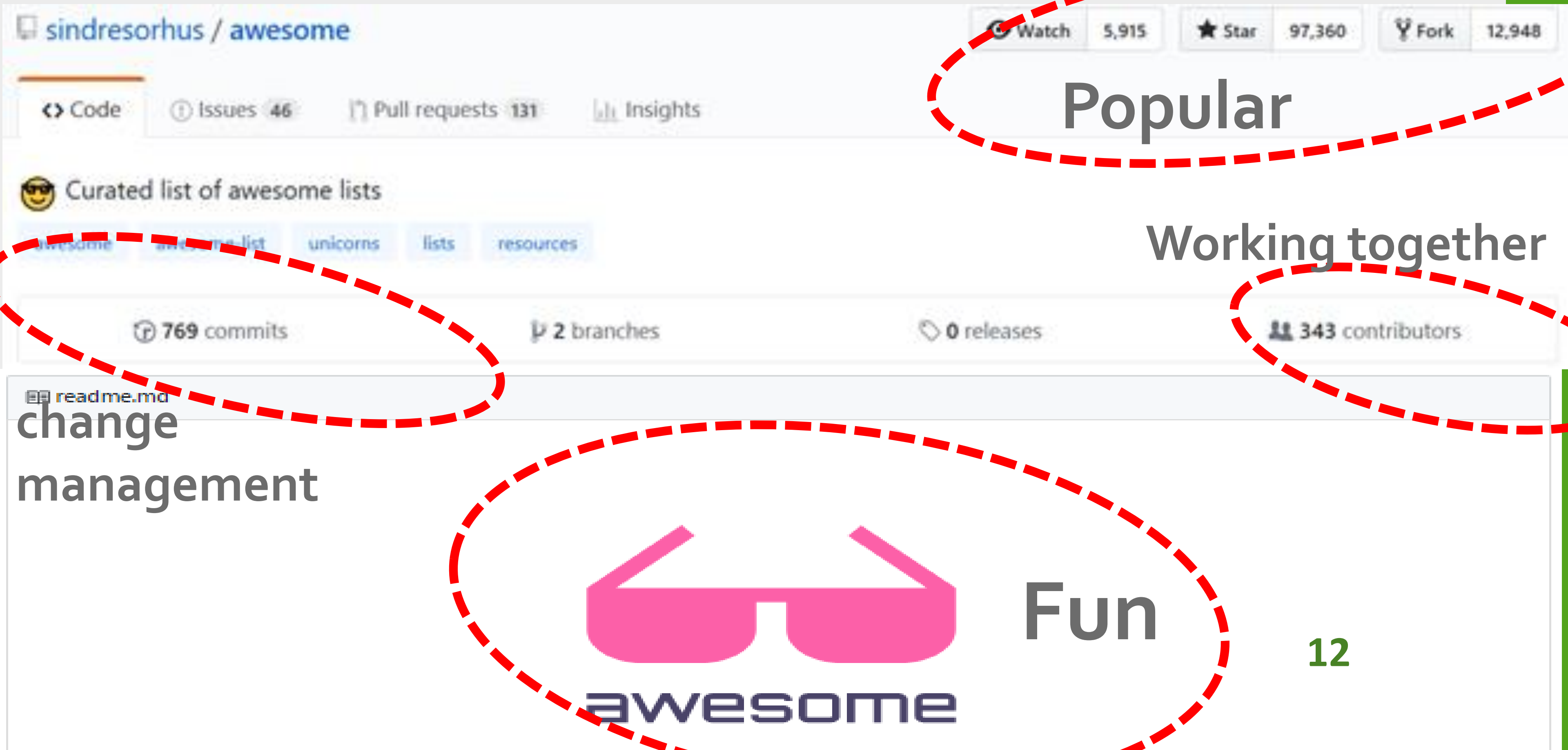


The awesome list

- When: born during the UNECE SDE conference April 2017
- Why: because it was not there and we needed something
- Who:
  - Started by Statistics Netherlands' Statistical Computing group (SNStatComp)
  - Open for everyone
- How:
  - Let's use common mechanisms from OS community => github and "awesome list" concept
  - let's start simple and let it grow
  - Use some simple criteria (free, available for download, used by at least one NSI, actively maintained)

What’s an awesome list?

– A curated list of “something”
– > 400 awesome lists, see
  https://github.com/sindresorhus/awesome



sindresorhus / awesome

Watch 5,915 · Star 97,360 · Fork 12,948

<> Code · Issues 46 · Pull requests 131 · Insights

Curated list of awesome lists

awesome · awesome-list · unicorns · lists · resources

769 commits · 2 branches · 0 releases · 343 contributors

readme.md

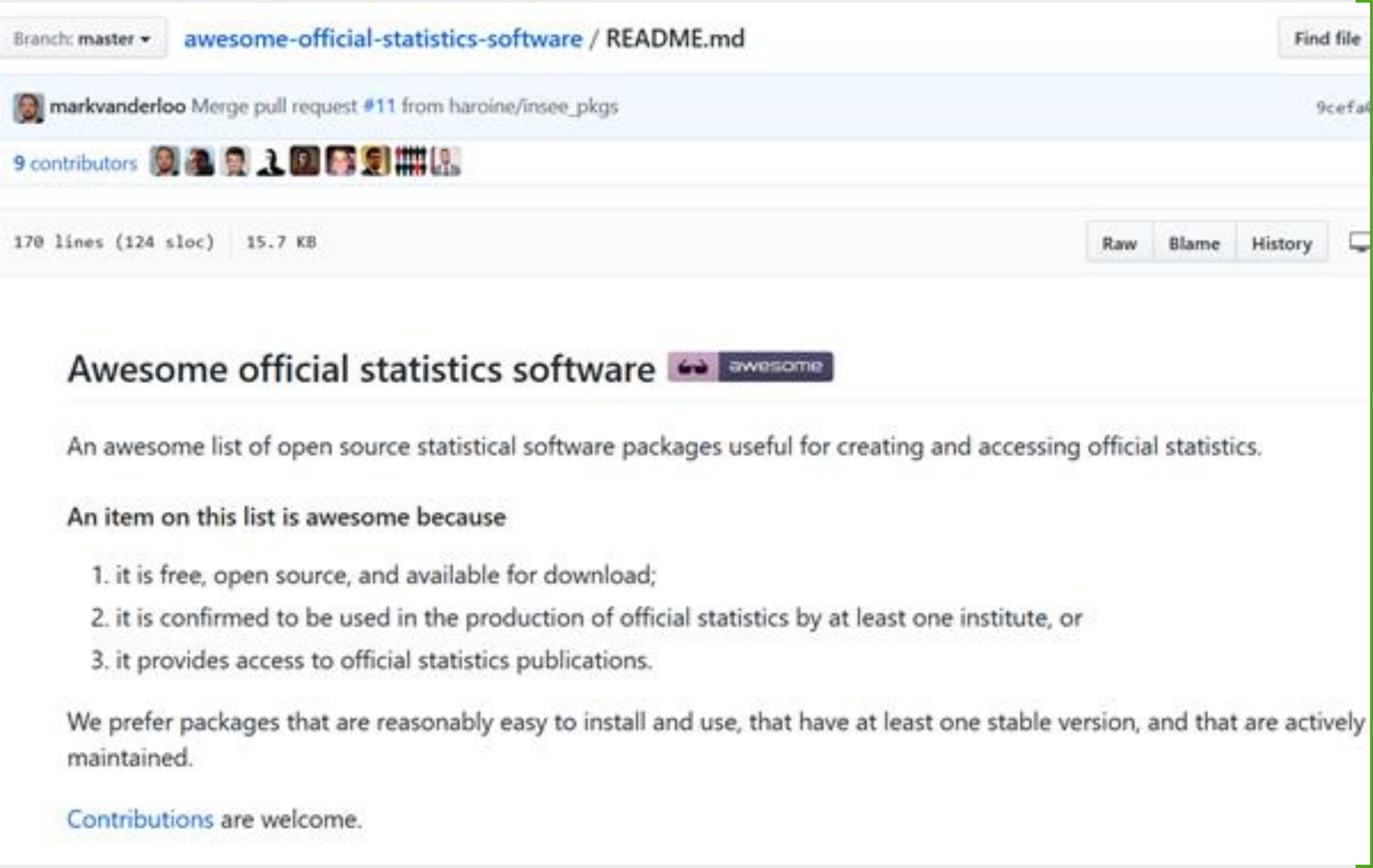
Popular

Working together

change management

Fun

Our awesome list (1)


Branch: master ▾
awesome-official-statistics-software / README.md
Find file
markvanderloo Merge pull request #11 from haroine/insee_pkgs
9cefa
9 contributors
170 lines (124 sloc) | 15.7 KB
Raw
Blame
History

Awesome official statistics software

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

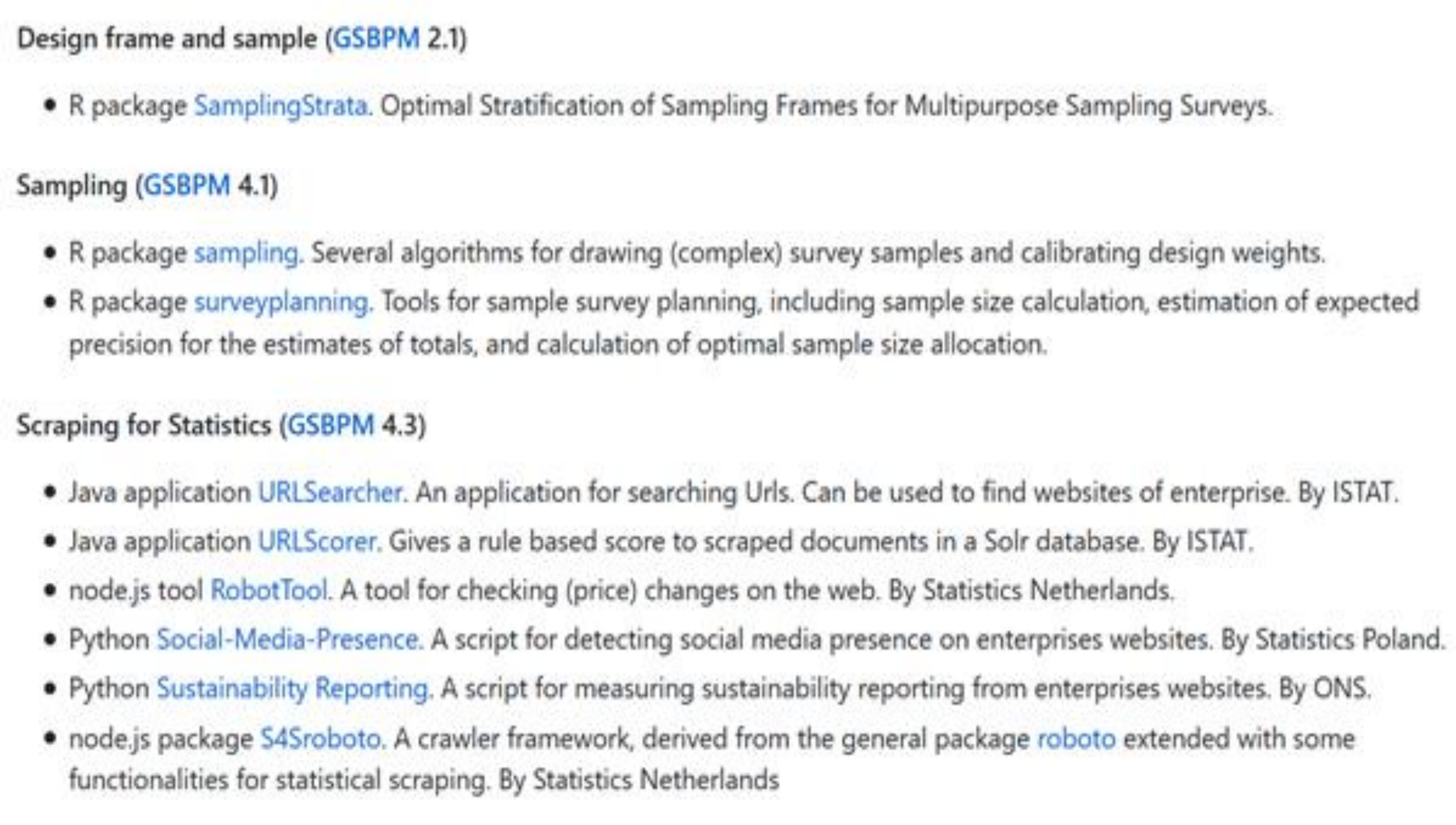
An item on this list is awesome because

1. it is free, open source, and available for download;
2. it is confirmed to be used in the production of official statistics by at least one institute, or
3. it provides access to official statistics publications.

We prefer packages that are reasonably easy to install and use, that have at least one stable version, and that are actively maintained.

Contributions are welcome.

Our awesome list (2)



Design frame and sample (GSBPM 2.1)

- R package SamplingStrata. Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys.

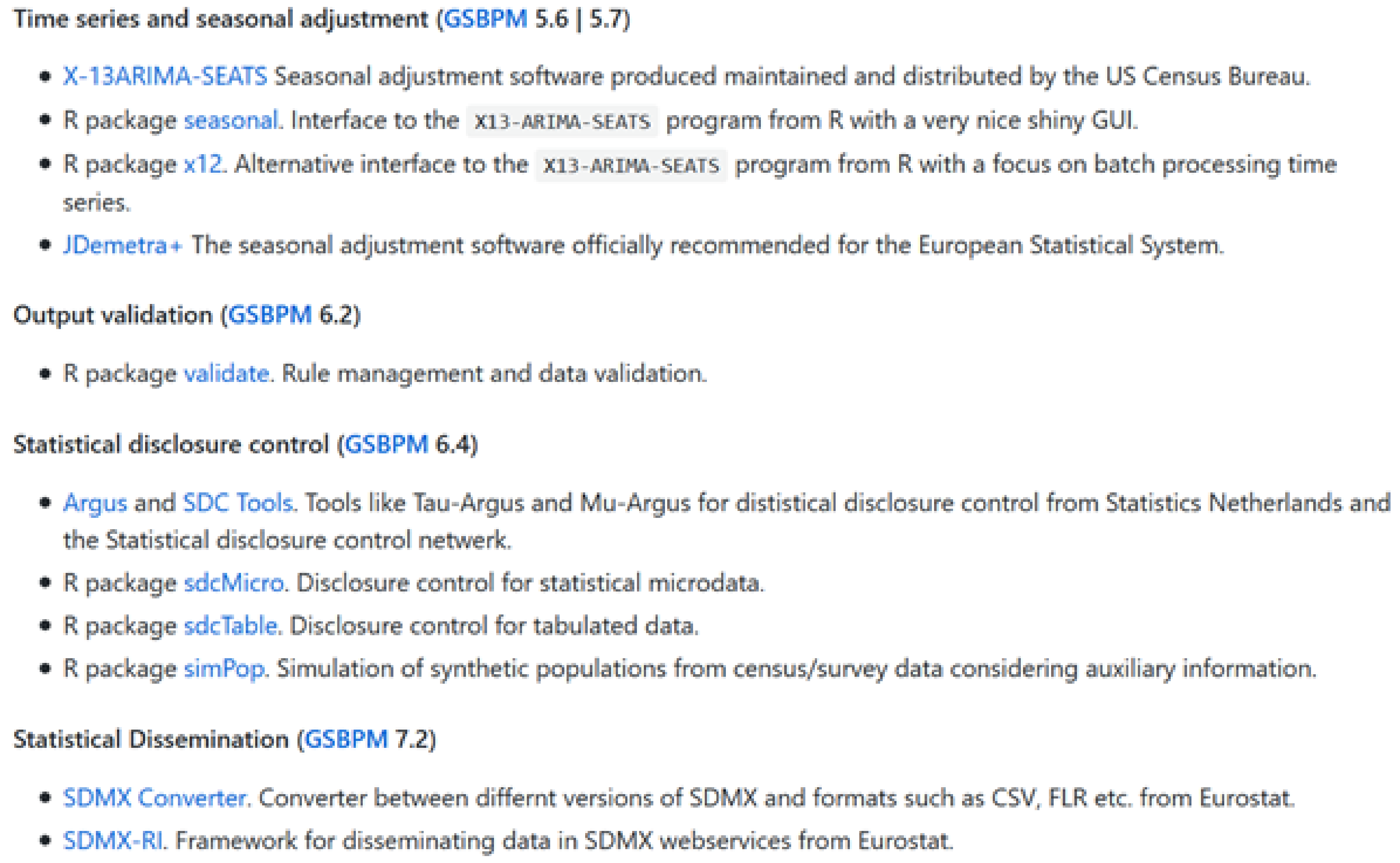
Sampling (GSBPM 4.1)

- R package sampling. Several algorithms for drawing (complex) survey samples and calibrating design weights.
- 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.

Scraping for Statistics (GSBPM 4.3)

- Java application URLSearcher. An application for searching Urls. Can be used to find websites of enterprise. By ISTAT.
- Java application URLScorer. Gives a rule based score to scraped documents in a Solr database. By ISTAT.
- node.js tool RobotTool. A tool for checking (price) changes on the web. By Statistics Netherlands.
- Python Social-Media-Presence. A script for detecting social media presence on enterprises websites. By Statistics Poland.
- Python Sustainability Reporting. A script for measuring sustainability reporting from enterprises websites. By ONS.
- node.js package S4Sroboto. A crawler framework, derived from the general package roboto extended with some functionalities for statistical scraping. By Statistics Netherlands

Our awesome list (3)



Time series and seasonal adjustment (GSBPM 5.6 | 5.7)

- X-13ARIMA-SEATS Seasonal adjustment software produced maintained and distributed by the US Census Bureau.
- R package seasonal. Interface to the `X13-ARIMA-SEATS` program from R with a very nice shiny GUI.
- R package x12. Alternative interface to the `X13-ARIMA-SEATS` program from R with a focus on batch processing time series.
- JDemetra+ The seasonal adjustment software officially recommended for the European Statistical System.

Output validation (GSBPM 6.2)

- R package validate. Rule management and data validation.

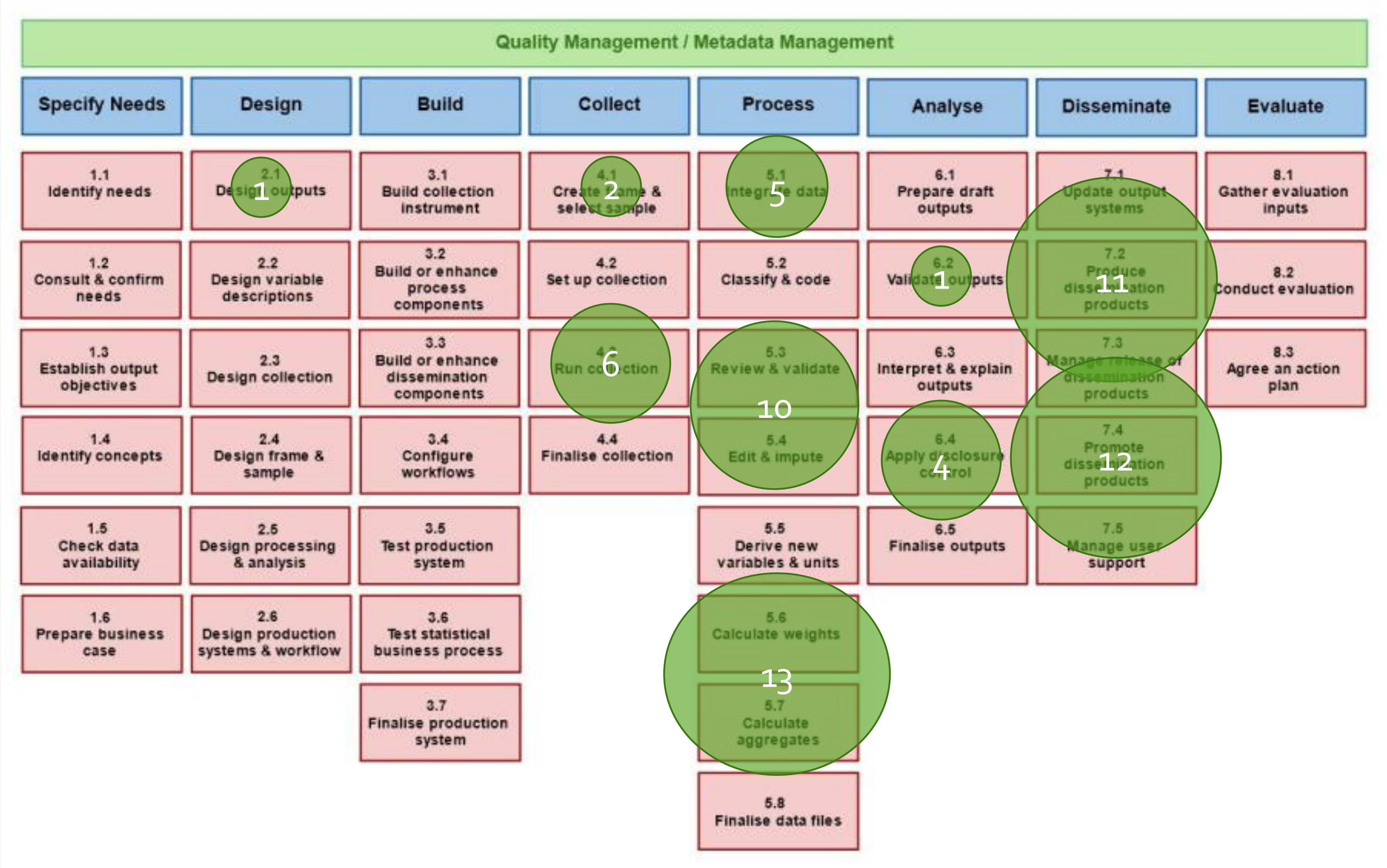
Statistical disclosure control (GSBPM 6.4)

- Argus and SDC Tools. Tools like Tau-Argus and Mu-Argus for dististical disclosure control from Statistics Netherlands and the Statistical disclosure control netwerk.
- R package sdcMicro. Disclosure control for statistical microdata.
- R package sdcTable. Disclosure control for tabulated data.
- R package simPop. Simulation of synthetic populations from census/survey data considering auxiliary information.

Statistical Dissemination (GSBPM 7.2)

- SDMX Converter. Converter between differnt versions of SDMX and formats such as CSV, FLR etc. from Eurostat.
- SDMX-RI. Framework for disseminating data in SDMX webservices from Eurostat.

The list and GSBPM



Quality Management / Metadata Management

Specify Needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Build collection instrument	4.1 Create frame & select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set up collection	5.2 Classify & 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 Build or enhance dissemination components	4.3 Run collection	5.3 Review & validate	6.3 Interpret & explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame & sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit & impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing & analysis	3.5 Test production system		5.5 Derive new variables & units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare business case	2.6 Design production systems & workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production system		5.7 Calculate aggregates			
				5.8 Finalise data files			

Join us !

https://github.com/SNStatComp/awesome-official-statistics-software

or google

“awesome official statistics”

- please !

- contributions welcome:

  GH pull requests

  markvanderloo@gmail.com, olavtenbosch@gmail.com

  @markvdloo

Where to find free and open source statistical software?

- CRAN
  - Comprehensive R Archive Network
  - Section for Official Statistics and Survey Methodology
    - Maintained by Matthias Templ, Statistics Austria
    - Over 130 entries – ordered by topic
    - https://CRAN.R-project.org/view=OfficialStatistics

What do you use?

Tips?

Experiences?

Demos?