The awesome list of official statistics software

www.awesomeofficialstatistics.org

Olav ten Bosch

Eurostat Task Force on Trusted Smart Statistics, 10-03-2022

Contents

- What is this awesome list?
- Zooming out: what is the aim?
- Communities, repos, packaging, communication
- Best practices for FOSS in offstats
- 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***

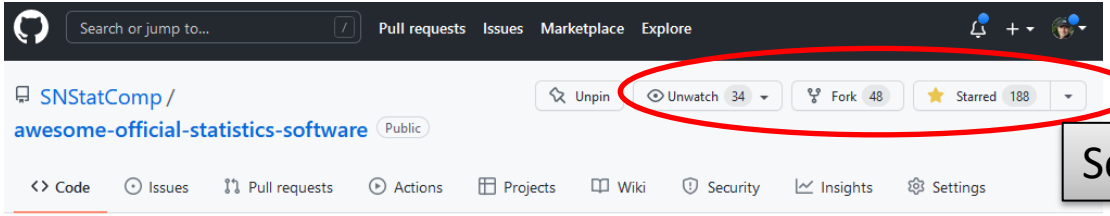
SNStatComp = Statistics Netherlands' Statistical Computing group

What is the awesome list?

Curated list of software for official statistics

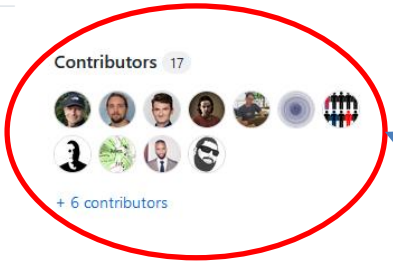
awesome

www.awesomeofficialstatistics.org



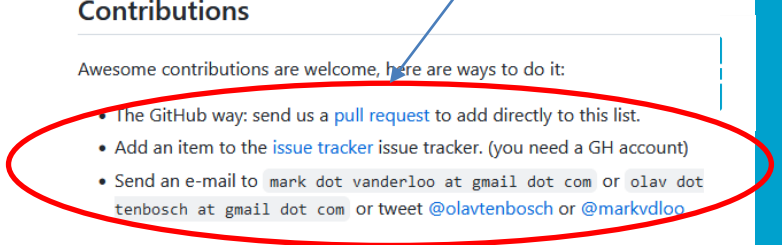
Social interactions, watch

Criteria



Working together

Open license



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 eas
- R package sdc
- R package Sma
  cross-classifica
- R package sim
- R package sdc
  assessment an
- R package syn
  safe to be relea

Data integration and record linkage (GSBPM 5.1)

- R package reclin. Functions to assist in performing probabilistic record linkage a
  pairs, comparing records, em-algorithm for estimating m- and u-probabilities, fo
  also be used for pre- and post-processing for machine learning methods for rec
- R package RecordLinkage. Implementation of the Fellegi-Sunter method for rec
- R package fastLink. Implements a Fellegi-Sunter probabilistic record linkage mo
  and the inclusion of auxiliary information. Documentation can be found on http
  /linkage.html
- R packages stringdist. Implements approximate string matching. Supports variou
  Levenshtein, Hamming, Levenshtein, optimal sting alignment), qgrams (q- gram,
  heuristic metrics (Jaro, Jaro-Winkler). An implementation of soundex is provided
- R packag
  inaccurat
- R Java M
  the same
- R packag

Scraping for Statistics (GSBPM 4.3)

- Java application URLSearcher. An application for searching
  ISTAT.
- Java application URLScorer. Gives a rule based score to scra
- Node.js tool RobotTool. A tool for checking (price) changes
- Python Social-Media-Presence. A script for detecting social
  Poland.
- Python Sustainability Reporting. A script for measuring sust
- Python urlfinding. Software for finding websites of enterpris
  Statistics Netherlands

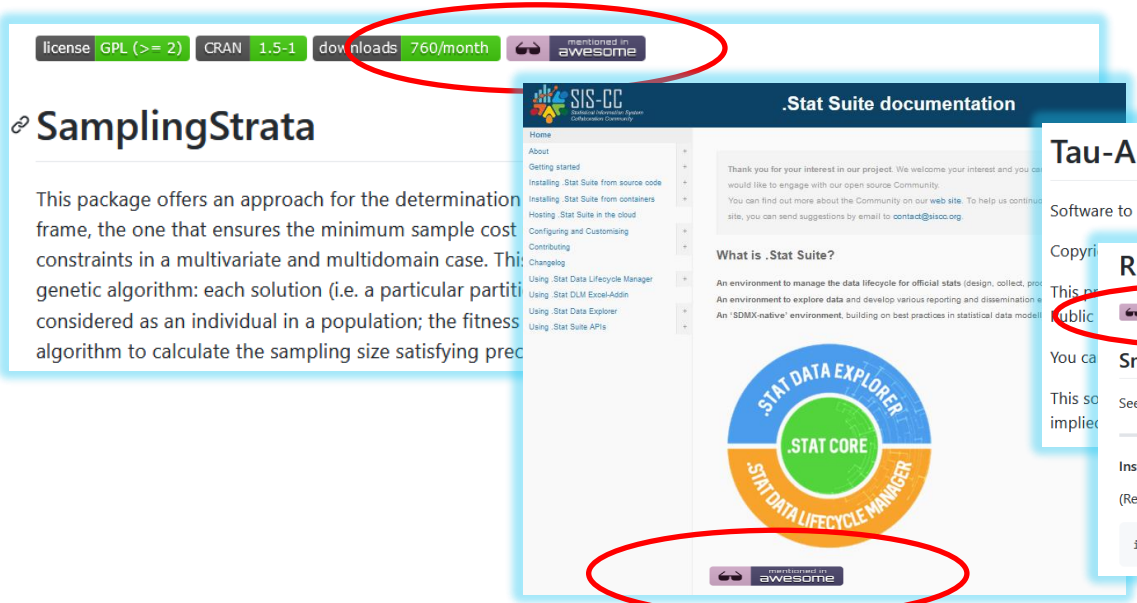
Access to official statistics (GSBPM 7.4)

- R package rsdmx. 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, Wolrd 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) cropscape 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.

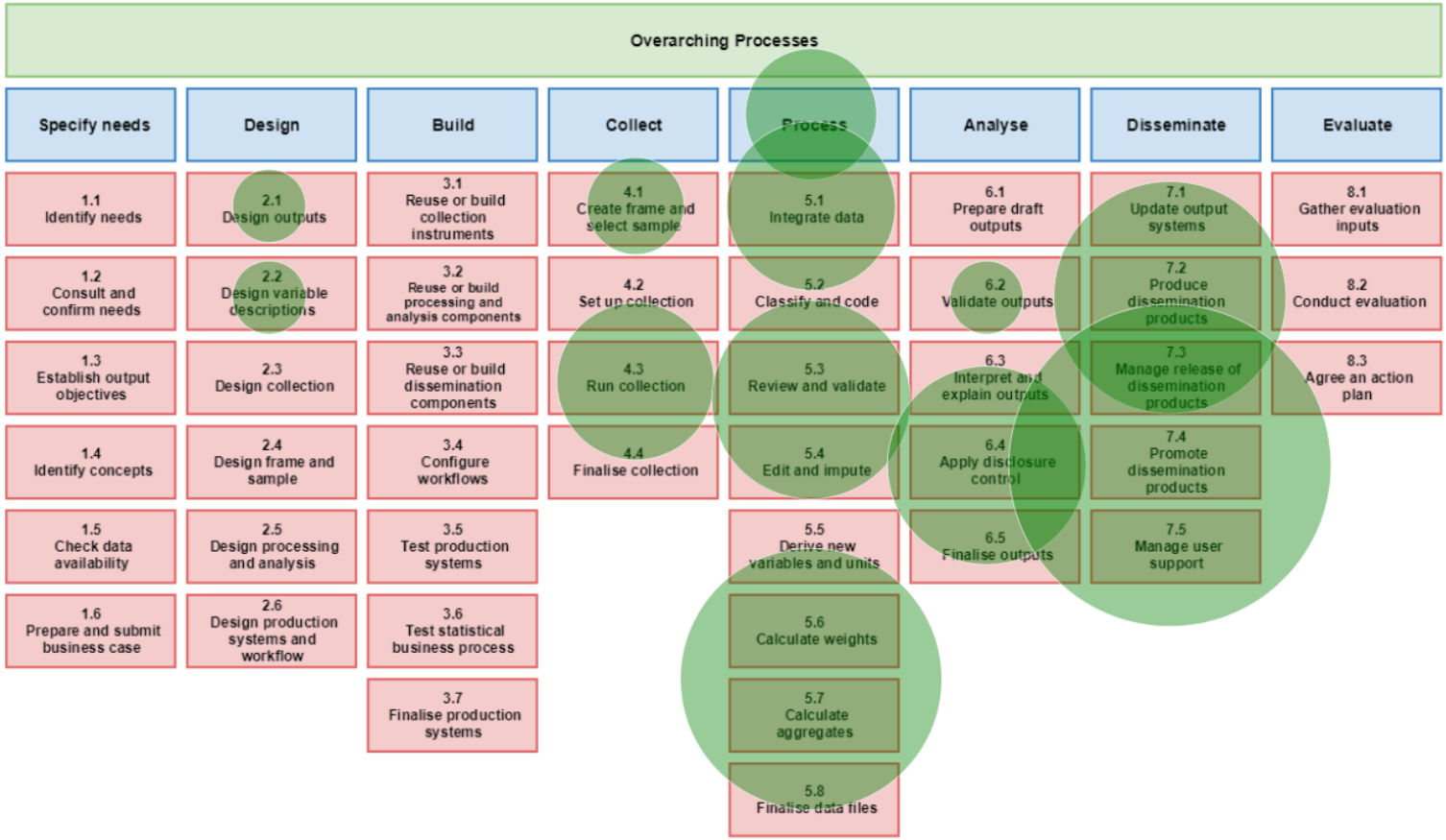
The right to wear the badge

- 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.



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

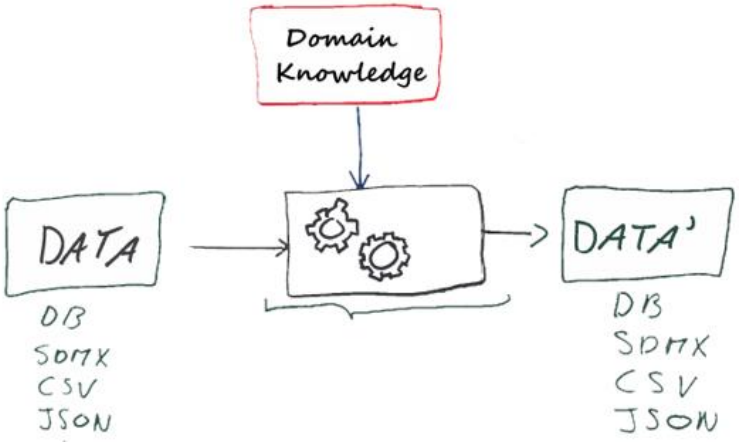
Use well-tested and proven implementations of generic methods

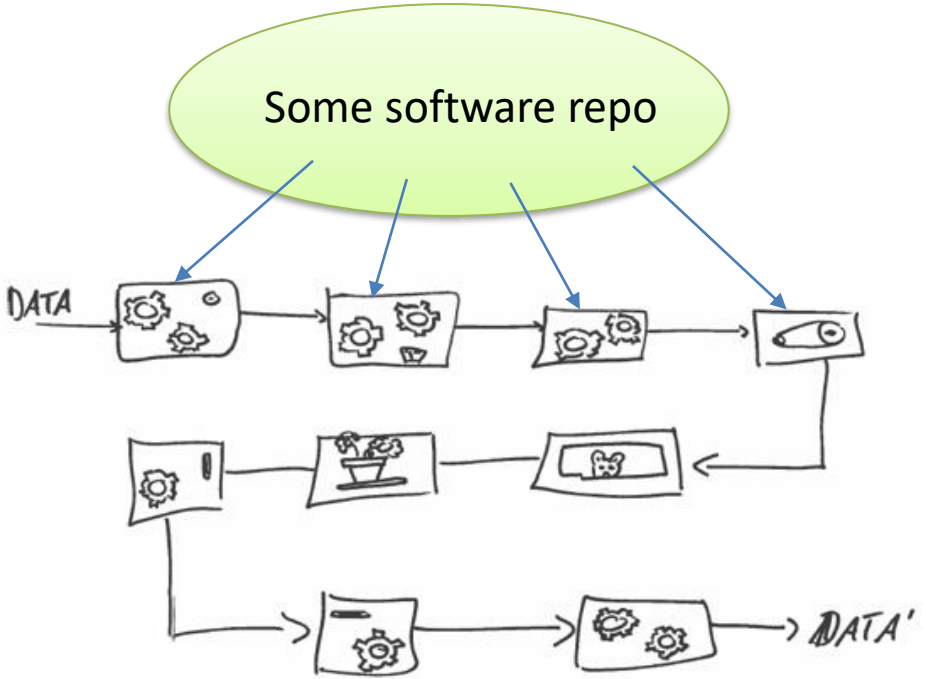
Standardisation

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

Basic building blocks

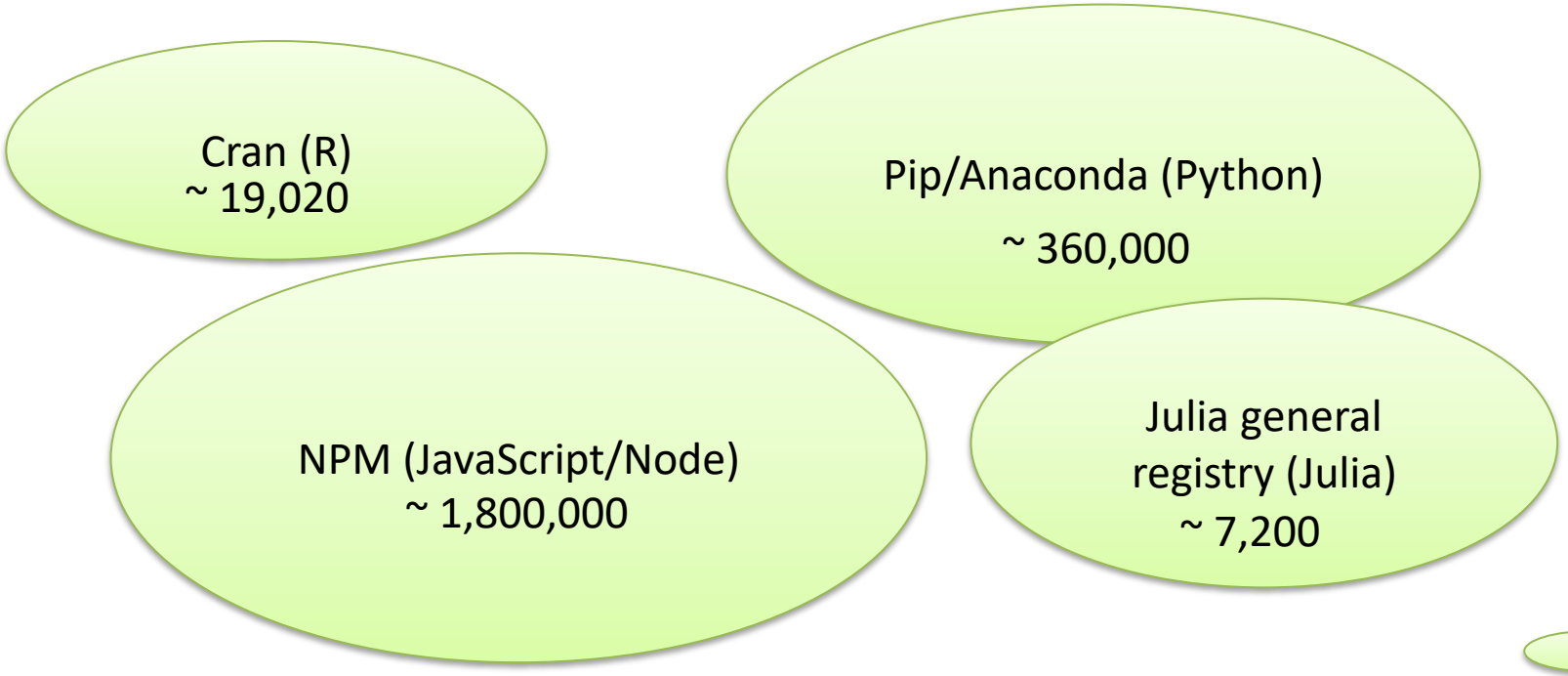
- The software landscape for offstats is getting more ***complex*** and ***dynamic***
- What are proven and succesful ***building blocks*** for offstats?
- Ideal scenario:
  - configurable per domain
  - chainable





Communities, repos, package systems

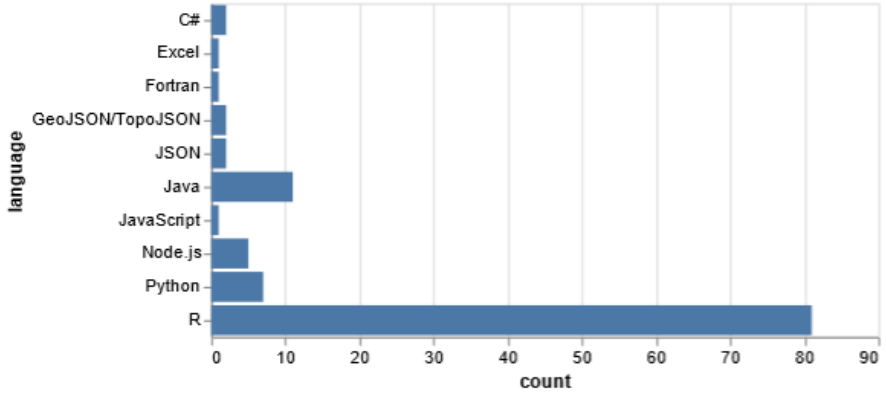
- Software sharing is already happening
- Different communities have their own packaging platforms



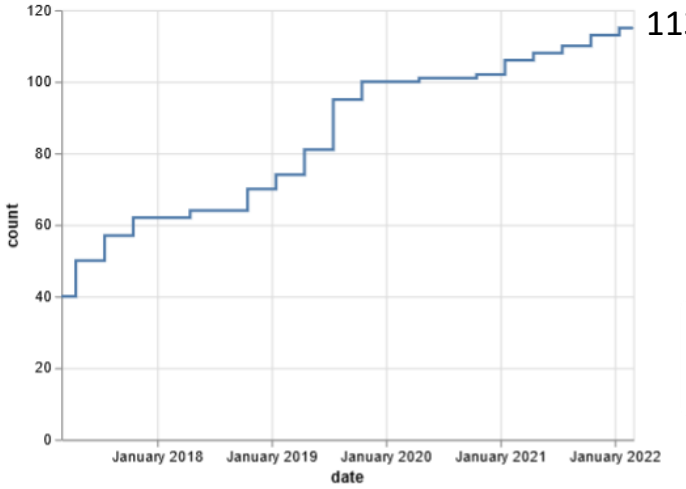
Awesome list status

- Bottom up approach
- Majority is R (now)
- Offstats ***R community*** more motivated towards sharing?

Packages by programming language:



Growth of the awesome list:



113

Awesome list promotions

- Unece SDE ‘17
- Unece SCFE ’17
- uRos ‘18
- Unece SDE ’18
- Estat Validition Grants kickoff ‘18
- uRos ‘19
- Unece modernstats World ’19
- Unece modernstats ’20 (virtual)
- uRos ’20
- ICDSOS ‘21
- uRos ’21
- TF-TSS ‘22

Virtual ☹

Best practices for FOSS in offstats

- ***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)

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	o.tenbosch@cbs.nl	@olavtenbosch
Mark van der Loo	mpj.vanderloo@cbs.nl	@markvdloo