Open source software sharing through the awesome list of official statistics software

Olav ten Bosch, Mark van der Loo, Statistics Netherlands
o.tenbosch@cbs.nl, mpj.vanderloo@cbs.nl

1. INTRODUCTION

The use of mature open source software for the production of official statistics has many advantages. It saves costs, adds to quality, creates flexibility and may have a positive effect on standardisation. However, it may be difficult to know what software is already available, what their maturity is, and what can be used for which tasks. Hence, it is crucial to share knowledge on existing open source software solutions among statistical organisations.

In this presentation we explain the "*awesome list of official statistics software*" created in 2017 to share knowledge on open source software in official statistics. We describe its origin, its growth, its current status and some of the spin-offs of this list, such as the experiences flowing into the ESS principles on open source software. Moreover, we touch upon possible extensions and improvements.

2. THE AWESOME LIST OF OFFICIAL STATISTICS SOFTWARE EXPLAINED

Software sharing is not new. The exchange and reuse of statistical software among statistical institutes has proven to be valuable for long in domains such as statistical disclosure control, data editing, data collection, and statistical dissemination. In these domains a few well-known solutions have been used by many offices. However, these days the software landscape for official statistics is much more complex and dynamic. Numerous small software packages with specialized statistical functionality are being developed and continuously improved. This makes it much harder to maintain a good overview of the statistical software landscape and increases the risk of redeveloping solutions that already exist. Hence, there is a need to spread knowledge about the existence and use of generic official statistics software.

A popular way of maintaining such knowledge is the so-called awesome list concept [1]. In 2017, during the UNECE Statistical Data Editing Conference, this concept was adopted to create the so-called "awesome list of official statistics software" [2]. The initial goal was to remember the software presented at that specific conference, but over time the list was extended with many other software presented at conferences or suggested by members of the statistical community. This made the list essentially a community approach to facilitate open source software sharing.

The list reached 100 contributions by 2019, and currently lists 135 open source packages that are easy to download and install, have at least one stable release, and are used in statistical production in at least one Statistical Office. Packages that facilitate automated access to Official Statistics output are included as well. The list itself has an Creative Commons license, is developed in the spirit of open source and receives many contributions from collaborators internationally. Each item has a link to the software download and a short description. The items may

come from different communities, may have been developed in different programming languages and may be distributed via different package systems. Despite their different origin, every item on the list is accompanied with up to three badges showing the latest version, the last commit and the license. These are automatically derived from the packaging system metadata. Figure 1 shows some examples of entries on the list.

- GitHub v.2.1 | last commit june 2021 | license GPL-3.0

  Python Social-Media-Presence. A script for detecting social media presence on enterprises websites. By Statistics Poland.

- CRAN 1.1.5 – 7 months ago | license GPL-3

  R package validate. Data validation checks such as on length, format, range, missingness, availability, uniqueness, multivariate checks, statistical checks and checks on SDMX codelist. See Cookbook. By Statistics Netherlands.

- GitHub v2.2.5 | last commit last saturday | license EUPL-1.2

  Java application JDemetra+. The seasonal adjustment software officially recommended for the European Statistical System.

- GitLab v24.1.0 | last commit today | license MIT License

  Node.js and other .Stat_Suite. An SDMX-based platform to build tailored data portals, topical or regional data explorers, or lightweight reporting platforms. Documentation. By SIS-CC.

- CRAN 2.1.3 – 8 months ago | license GPL (>= 2)

  R package simPop. Simulation of synthetic populations from census/survey data considering auxiliary information.

Figure 1. Examples of items on the list

To give the user an indication of the use of each element on the list, it is organized according to the Generic Statistical Business Process Model (GSBPM). Figure 2 shows how the 135 items are distributed across GSBPM processes.

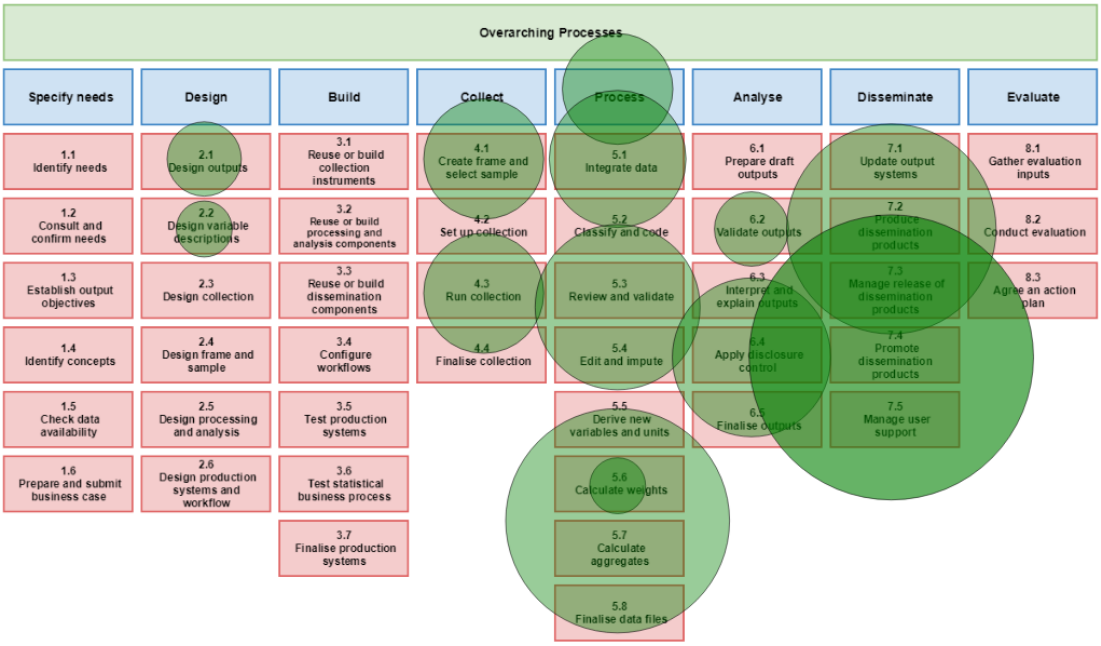

Figure 2. Software on the list by GSBPM process phase

The presentation of the list is derived from a machine readable YAML file. This information, including pointers to the package repos, in combination with package system APIs, allows for automatically deriving statistics on the official statistics software landscape. Figures 3 shows the distribution of programming languages of items on the list. The vast majority of items is written in R, which shows the excellent software sharing methods in this community. Figure 4 shows the licenses used on the list. GPL is the most popular license followed by MIT and EUPL.

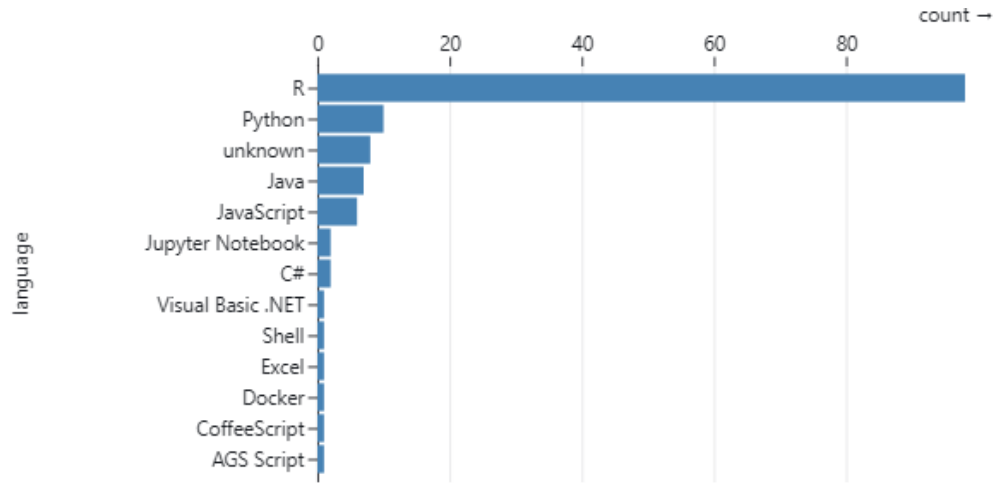

Figure 3. Programming languages used on the list

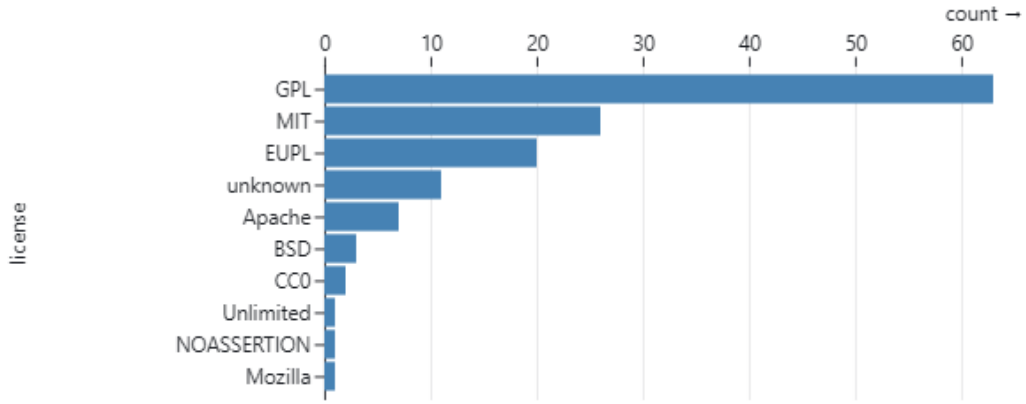

Figure 4. Licenses used on the list

3. THE LIST USED IN THE ESS PRINCIPLES ON OPEN SOURCE SOFTWARE

There have been some spin-offs of the awesome list initiative. First of all the experiences in the official statistics open source community have been written down into a list of best practices, which found their way into the *ESS principles on open source software* [3], put together by the 'group on Open Source for Official Statistics (OS4OS)', a number of statistics offices together with Eurostat and OECD. They put together seven principles that reflect the way of working of successful software projects found on the awesome list. For completeness we repeat the principles here:

1. **OSS by default**: in the production of official statistics we prefer the use of open source software solutions over closed software solutions. Moreover, we share our software solutions as open source.

2. **Work in the open**: We start our projects in the open from the beginning and clearly mark maturity status.

3. **Improve and give back**: We rather improve existing open source solutions than decide to create new solutions and we give our improvements back to the respective open source community.

4. **Think generic statistical building blocks**: In our open source work we strive for re-usable generic functional building blocks that support well-defined methodology in statistical processes.

5. **Test, package and document**: We test, package and document our open source software for easy-re-use.

6. **Choose permissive**: We choose the most permissive OS license possible for sharing our software.

7. **Promote**: We invest in promoting new developments or improvements on our open source software within the ESS community and where applicable in a wider context.

One notable other spin-off is the adoption of concept by the NTTS 2021 organisers to publish the links to all source code accompanying NTTS 2021 presentations.

Having a good overview of essential open source software also creates other opportunities. One of the largest subsections of the list is “access to official statistics”. The software in this category supports to access official statistics databases and has been used to identify FAIRness (Findable Accessible, Interpretable, Re-usable) of the official statistics landscape [4].

8. SUMMARY AND OUTLOOK

Ideas for extending the functionality are documented on GitHub itself containing ideas such as adding compatibility, maturity and popularity of the items. Download figures were previously added, but later removed as they are not representative across platforms and languages. A more fair indicator of popularity takes multiple aspects of use, maturity and functionality into consideration. All in all, we argue that the awesome list of official statistics software has proven to be a useful tool for sharing knowledge on existing open source software in official statistics and we hope it can play this role also in future innovation projects.

REFERENCES

[1] https://github.com/sindresorhus/awesome

[2] http://awesomeofficialstatistics.org

[3] https://os4os.pages.code.europa.eu/pbbp/principles.html

[4] O. ten Bosch, E. de Jonge, H. Laloli, To be FAIR, what is missing in Official Statistics?, COSMOS conference, Paris, march 2024