Published at the Conference of European Statisticians, Seventy-first plenary session, Geneva, 22-23 June 2023
https://unece.org/statistics/events/CES2023

item 3: Moving towards open-source technologies – strategic and managerial perspective

Free and Open Source Software at Statistics Netherlands

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

Statistics Netherlands

Introduction

Statistics Netherlands has been involved in open source projects for more than a decade. In this paper we present a brief overview of some of the aspects of the introduction of open source at Statistics Netherlands, such as the choice for R and Python as standard tools and the creation of the "awesome list of official statistics software" to facilitate knowledge exchange among statistical organizations. Moreover we touch upon the contributions to FOSS from Statistics Netherlands' employees and our current FOSS policy. Finally we describe the ESS principles on OSS, which were recently launched as a result from a group of international experts in which Statistics Netherlands was heavily involved.

A brief history

Statistics Netherlands first approved the Free and Open Source (FOSS) tool R for use in production in 2010. Until then it was used only in some corners of the office as a research tool. Until then, production systems were mainly based on dedicated IT solutions, and `citizen automation' based in Office software, SPSS, and a few specialty IT products. The adoption of R was picked up in a bottom-up fashion by a group of professionals who organized user group meetings, courses, software development standards and who ran application management. Statistics Netherlands also adopted a FOSS policy that specified policies for using, contributing to, and developing FOSS products. Although formal support from IT was minimal, R quickly became popular under statisticians who started building many local production processes in R. Currently at least 50% of all jobs on the internal batch server for heavy processing involve R scripts. The adoption of R also lead to the development of production-ready R packages that are used internally but also released as FOSS products to the public. These packages were often developed under the umbrella of the research program by methodologists with a background in scientific programming, but local renewal projects and external funding also drove implementation and further development of the packages.

In 2012, Python was also approved for statistical production. Again, a bottom-up approach lead to user meetings and knowledge exchange. As the Python community is large and readily developed outside of statistics offices, Python courses could be bought externally, where extra parts specificly targeted at use at Statistics Netherlands were added. The uptake of Python has been slower than that of R. This probably has something to do with the fact that R as a language is closer to the world of statisticians whereas data scientists, who arrived later at the scene, often choose for Python. Currently, R is the recommended tool for statistical work on data, while Python is recommended for

orchestration and process management. Generally we see that Python and R have grown towards each-other in terms of functionality. Python started out as a general scripting language whereas R started as a language for working with data and statistics. The major strengths of Python over R, are or rather were, the ability to integrate with (web) services, and a strong tradition in machine learning and text mining. The strong points of R over Python were visualization, data processing and management, the strong statistical libraries (often based on academic work) and the best package management system for any FOSS tool around (CRAN). Apart from the difference in maturity on package management, R and Python can nowadays be considered roughly equivalent in many of the areas mentioned.

2014 saw the adoption of Git as a version control tool for non-IT programmers. Statistics Netherlands set up an internal Git server based on a FOSS tool called bitbucket. Git replaced the until-then used SVN version control system that was used mainly by scientific programmers or so-called 'tool-experts' with relatively deep knowledge of software engineering. Nowadays, a Git course is part of the standard curriculum of the internal educational department called 'CBS Academy'.

Similar developments occurred at other statistical offices [1]. Together with the front-runners in the international community, the authors of this paper started the so called *awesome list of official statistics software* [2] in 2017 during the UNECE conference on SDE that was hosted in The Hague[1]. This curated list of tools for statistical production reached 100 contributions by 2019, and currently lists 118 FOSS tools 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. The list is developed in the spirit of open source and receives many contributions from collaborators internationally. To give the user an indication of its use, the list is organised according to the Generic Statistical Business Process Model (GSBPM). Figure 1 shows how the 118 items are distributed across GSBPM processes.

[1] https://www.awesomeofficialstatistics.org

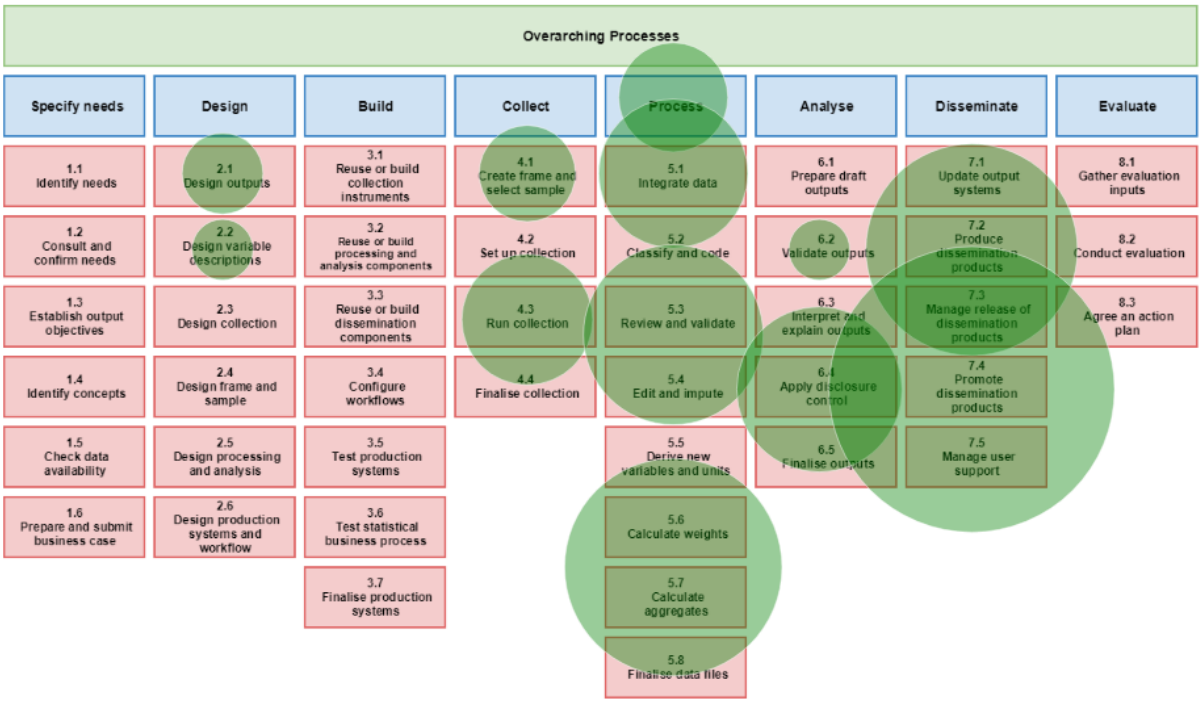

Figure 1: Awesome official statistics packages organized by GSBPM.

In 2018, Statistics Netherlands hosted the sixth installment of the *use of R in Official Statistics* conference (uRos) at the Hague. The conference sold out at 100 participants from over 40 countries. The event in The Hague marked the first occasion of uRos outside of its `home country' of Romania, and it has been traveling biannually since then, with the latest installment (2022) hosted by Statistics Canada (online).

Today, producing statistics without FOSS tools is unimaginable at Statistics Netherlands [3]. Indeed, the main renewal programs at Statistics Netherlands have chosen open source tools R and Python as the standard tools for production (rather than just being allowed, as before). The site dashboards.cbs.nl is fully developed with R Shiny, and the innovation program for the output is focusing on use and alignment of the free and open source .STAT suite managed by the OECD Statistical Information System (SIS) community[2] for dissemination instead of proprietary in-house develop solutions. Finally, open source tools are increasingly used for operations. In particular, many dashboards on top of our administrative system are implemented in R Shiny.

Contributions to FOSS

Researchers of Statistics Netherlands have contributed to FOSS projects since at least 2009. It is worth mentioning that contributions do not have to come in the form of code. In the open source community bug reports, contributed documentation, scientific papers, co-organizing meetings or conferences, or initiatives like curated software lists are strongly welcomed as well.

[2] https://siscc.org/stat-suite

Contributing back to the open source community in whatever form is beneficial for several reasons. On a technical level, releasing code has strong positive effects on software quality especially if it gains a user base (many eyes make all bugs shallow). From a strategic perspective it opens the possibility of collaborating with external developers who become part of the network of the staff of the institute. Contributing to documentation, bug reports, or otherwise turns users of open source software into experts in the software they use. The activity of writing a good bug report or documentation often yields profound understanding of a tool or its underlying methodology. Besides, it is often rewarding for staff to build their network and get recognition for their work outside of the context of the institute.

Currently, the research program has yielded a few dozen open source tools in the areas of data cleaning, statistical disclosure control, visualization, web scraping, URL finding, record linkage, fuzzy text matching and more (see references). Besides, staff of Statistics Netherlands contribute to FOSS-related conferences, are active in scientific committees of such conferences, and author scholarly articles in journals related to (statistical) open source software. One employee is currently a member of the editorial board of the R Journal. Statistics Netherlands is also actively involved in international FOSS-related groups in the area of official statistics, such as the FOSS initiative of the UNECE Blue-Sky Thinking Network[3] and the ESS informal group on FOSS which produced the ESS principles on Open Source Software that will be discussed below[4]. Such activities are both rewarding for individual employees but also contribute to a strategically valuable connection between Statistics Netherlands and open source communities.

Current FOSS policy

In 2020 Statistics Netherlands adopted a new FOSS policy. A renewal became necessary because of external and internal developments. External developments included adaptations in in The Netherlands's Competition Act, political developments including a Letter to the Parliament on releasing the source code of government software (2020)[5] and the normalization of FOSS outside Statistics Netherlands in the area of data processing ("open source has become the standard"). Internal developments were related to an updated enterprise architecture (IV architecture) and the way the use of FOSS has developed in Statistics Netherlands.

The current policy comprises four areas: Production and Operations, Building Tailored Solutions, Use for R&D, and Contributing to FOSS.

For statistical production and operations, the policy stipulates that FOSS software shall undergo the same selection process as for Commercial off-the-shelf (COTS) software. This includes, among other things, a test against the IV architecture, a check on the need for tendering, etc. Part of that test against the IV architecture is a test on the community, regarding matters such as activity and maturity, effort put into it, sponsors, etc. FOSS, for which an SLA is required from the user process, is

[3] https://statswiki.unece.org/pages/viewpage.action?pageId=261818141
[4] https://os4os.pages.code.europa.eu/pbbp/principles.html
[5] https://www.rijksoverheid.nl/documenten/kamerstukken/2020/04/17/kamerbrief-inzake-vrijgeven-broncode-overheidssoftware (in Dutch)

involved through an SLA via an external service provider that provides support for the relevant software. If an external service provider is not available, Statistics Netherlands can manage it itself if the risks are acceptable. The costs of this are included in the selection process.

In the area of building tailored solutions, FOSS is to be applied without changes. Derived work from FOSS is therefore realized in a separate code. Changes to the FOSS are made at source. FOSS source code is reproducible with, for example, a version control system. The policy also prescribes to use the most recent version of FOSS as much as possible.

Regarding R&D, there are no restrictions on the use of FOSS other than general security requirements (such as checking for viruses and malware). When the result of R&D is transferred to regular statistical production and business operations, any FOSS that is required for this must still comply with the applicable guidelines. Problems with FOSS that do not meet the guidelines for regular statistical production and management are resolved by the users themselves.

Regarding publishing FOSS, the policy states that all software newly developed by Statistics Netherlands that can be applied generically (not only usable by Statistics Netherlands) to support the statistical process is preferably published as FOSS. If possible, Statistics Netherlands publishes FOSS under a Permissive (e.g. BSD) license and if not possible under a Copyleft license (preferably EUPL 1.2).

International collaboration: the ESS principles on FOSS

Starting 2nd half of 2022, statistics offices from multiple countries, together with Eurostat and OECD, decided to create the 'group on Open Source for Official Statistics (OS4OS)'. This group was set with the mandate to 'share and review current practices and lessons learnt on the use of OSS for statistical purposes, and look at what additional work can be carried out together in the ESS, including the governance and tasks at technical level'. Statistics Netherlands participated in this group and headed the subgroup on principles. Seven principles were formulated, that is:

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.

The full text of the principles can be found are available online[6], with for each principles the one-liner, a short statement to explain it, the rationale of the principle and the implications for NSIs and international organizations. These principles were not created out of the blue. They are based on many years' experience in joint open source projects in the official statistics community, many of which are listed on the Awesome list of official statistics software and discussed at conferences such as the "The Use of R in Official Statistics", 2015-2022 and UNECE ModernStats World Workshops 2018-2022. Moreover they are compatible with ESS and EU policies such as the ESS Code of Practice[7] and the EU-Open source strategy[8]. Finally they build on earlier work on best practices and strategies in as formulated in [4] and [5].

We think these principles can be a cornerstone for the further development of the open source community for official statistics. Ideally they could serve as a reference for creating the circumstances in which the statistical open source community can further grow in size and productivity. They do have implications for strategic and high management level participating in the CES conference, such as for example 1) supporting the primary choice for OSS solutions over proprietary solutions, 2) creating facilities for employees to work together 'in the open' with experts from other statistical organizations on OSS of any level of maturity, and 3) understanding and supporting employees in refactoring software solutions to 'give back' their improvements to the open source community the best way possible. As a consequence wide support for these principles from high-level management from statistical organizations could give the ESS OSS community a clear message that the open source work done over years and to be done in future does no longer rely on individuals but an integral part of the intended working of the whole ESS.

[6] https://os4os.pages.code.europa.eu/pbbp/principles.html
[7] https://ec.europa.eu/eurostat/web/quality/european-quality-standards/european-statistics-code-of-practice
[8] https://ec.europa.eu/info/departments/informatics/open-source-software-strategy_en

Conclusions

Using and contributing to open source software has been an essential part of the development of Statistics Netherlands over the last decade or so. Internally, open source software is used throughout statistical production and has recently entered the area of operations. Externally, open source software is used for example to power externally hosted thematic dashboards.

Contributing to open source software and its community has significantly enhanced published software products. It has also allowed staff members to build status and international network within open source software communities, which is also of strategic relevance for Statistics Netherlands.

Moreover, the open source way of working has lowered the barrier for collaboration internationally, within the official statistics community or broader to the point where that barrier has all but vanished. Similar trends can be seen in the world of open data and open (trained) models. All these trends are so promising, or have already become so valuable for official statistics that they deserve full support for the future.

References and further reading

[1] Kowarik, Alexander, and Mark van der Loo. "Using R in the Statistical Office: the experience of Statistics Netherlands and Statistics Austria." Romanian Statistical Review 1 (2018).

[2] Olav ten Bosch, Mark van der Loo, Alexander Kowarik, (2020), “The awesome list of official statistical software: 100 ... and counting”, The Use of R in Official Statistics - uRos202

[3] Van der Loo, M. P.J. (2021). “Computing in the statistical office”. Statistical Journal of the IAOS, 37(3), 1023-1036.

[4] Olav ten Bosch, TF-TSS meeting, March 10 2022. The awesome list of official statistics software & FOSS best practices, Zenodo: https://doi.org/10.5281/zenodo.7665189

[5] Proposal for ESS Open Source strategy implementation roadmap, Lehtinen et. al., Oslo 25-27 July 2022 https://i3s-essnet.github.io/Documents/2022/oslo/osos/os4os-oslo-document.html