Promoting reproducibility-by-design in statistical offices

S. Luhmann, J. Grazzini, F. Ricciato, M. Meszaros, K. Giannakouris, J.M. Museux and M. Hahn
European Commission – DG ESTAT (*name.surname*@ec.europa.eu)

Keywords: quality & trust; openness & transparency; sharing & reusability; reproducibility & auditability; participation & engagement; collaboration & co-creation; open algorithm & data; interactive computing platform; produser & citizen statistics; statistical literacy.

1. INTRODUCTION

Because policy advice is becoming increasingly supported by data resources, National Statistical Offices (NSOs) need to leverage the use of new sources of (small or big) data to inform policy decisions [1]. At the time where citizens' demands for more trust in the public institutions are growing, it underpins the movement towards more open, transparent and auditable (verifiable) decision-making systems [2]. Bearing in mind the reproducibility movement towards *Open Science* and best practices in the *Open Source Software* (OSS) community, it is expected that a greater openness, transparency and auditability in designing statistical production processes will result in improved quality of the analysis involved in decision-making [4] as well as increased trust in the NSOs [5].

Drawing upon the *Transparency and Openness Promotion* guidelines [6] and the *Reproducibility Enhancement Principles* [7] in computational science, as well as the recommendations to funding agencies for supporting reproducible research [2] and the various calls for open and transparent (data and) algorithms [8] in the field of statistics, we advocate for the following principles: ***Shared, Transparent, Auditable, Trusted, Participative, Reproducible,*** and ***Open*** (hereby referred to as *STATPRO*) to be adopted by NSOs. These principles build not only on existing and new sources of data, but also on new methodologies and emerging technologies, and advance thanks to innovative initiatives. With increased availability of *open data*, new developments in *open technologies* and *open algorithms*, as well as recent breakthroughs in *data scien*ce, it is believed that they can help improve current governance processes by enabling data-informed evidence-based decision-making and potentially reduce the bias, costs and risks of policy decisions [4]. In this regard, *Official Statistics* should be accompanied by access to the data analysed whenever possible, the detailed metadata information, the underlying assumptions (models and methods) and the tools (software) used to generate them [2]. In the context of a "post-truth" society, the *STATPRO* principles present substantial promises for *Citizen Statistics* and *e-Official Statistics*, *e.g.* for the interaction between NSOs, data users and data producers. Indeed, they make it possible to provide the public with the ability to both perform the analysis and repeat it with different hypotheses, parameters, or data, hence translating policy questions into a series of well-understood computational methods and scrutinizing the final decision.

In this paper, we further emphasize the need for *Official Statistics* to go beyond current practice and exceed the limits of the NSOs and the European Statistical System (ESS) to reach and engage with *produsers* – *e.g.* statisticians, scientists and citizens. Through the adoption of some best practices derived from the OSS community and the integration of modern technological solutions, the *STATPRO* principles can help create new *participatory models* of knowledge and information production. We illustrate this trend through Eurostat recent initiatives.

2. PROMOTING A CULTURE OF OPENNESS AND TRANSPARENCY *IN DA HOUSE*

Overall, greater openness and transparency contribute to a more holistic and extensive approach to production systems in NSOs with the development and deployment of high-quality statistical processes, as foreseen in the European Statistics Code of Practice. O*pen algorithms*, together

with *open data*, enable to track the whole decision-making process as well as its progress [8]. Besides existing practical aspects and technological considerations, the calls for transparent and defensible evidence-based data-informed policymaking [4] encourage the use of OSS in NSOs and other organisations producing statistics (*e.g.*, UN, OECD,…), so as to benefit from the many statistical libraries, advanced algorithms and innovative developments made available freely and openly. As an abundance of (data and) computing tools does not automatically guarantee good decision-making, the development and deployment of statistical processes needs (together with sound judgment) best practices, *e.g.* derived from the OSS community:

- Algorithms and methods must be available and accessible to anyone.
- Software must be available and accessible in open repositories[1]. OSS should be used preferably throughout the production process. Version control shall be adopted for all (collaborative or individual) code development.
- Algorithms and software should include adequate levels of documentation to enable independent reuse by someone skilled in the field. Best practice suggests that software include a testing suite that exercises and verifies its functionality.
- Details of the computational environment that help generate statistical outputs should be shared. Version control can also be introduced for environments.
- Workflow should also be tracked during the process and made available to others for scrutiny. End-to-end scripting of the production process should be carried out.
- Whenever possible, open licensing should be used for in-house software.

Such an approach reduces both the risk of having no-one available to maintain or update the system and the cost of the necessary testing and audit procedures needed for high-reliability software. Experience shows that OSS-based software solutions not only provide with an open, flexible and agile approach to immediate needs and legacy issues, but also address long-term problems and potential future software requirements for statistical production [2]. While it may not be possible to fully disclose, or even license, all proprietary software used in statistical processes, scripts designed to be executed by propriety software may still be openly licensed[2]. There are still broad benefits to code release, for example, allowing for code inspection, even if it cannot be executed [9].

3. FOSTERING SHARING AND REUSABILITY WITHIN THE STATISTICAL COMMUNITY

Within the ESS, the adoption of *STATPRO* principles should be encouraged. To better enable reusability across the statistical community, funding bodies should instigate [2]:

- Sharing of best practices: inside each NSO, and between NSOs, provide guidance about the best practices currently in use in the ESS, but also beyond, in particular in the scientific community. Internally support small-scale innovative initiatives, *e.g.* taking place for the development of new statistical products.
- Training: provide support for the development of appropriate courses to match future skills needs, and accompany potential corporate cultural changes.
- Provision of incentives: enforce shared software as full-fledged deliverables[3]. The use of mature and common OSS technologies should be encouraged, except when their use

[1] *E.g.*, **Eurostat** opens some code/software developed internally: https://github.com/eurostat, and distributes it under the European Union Public License whenever possible.

[2] *E.g.*, *PING* software for handling and processing **Eurostat** data in a SAS production environment.

[3] *E.g.*, in the recent call for proposals on "Multipurpose statistics and efficiency gains in production", **Eurostat** expressed its intention to evaluate positively *"proposals making use of free and open source technologies"* and

conflicts with other priority considerations for software implementation. Software solutions integrating modular and scalable features and supporting different target architectures shared within the ESS and exposed to a larger community – *e.g.*, under a public license – should be preferred.

These challenges support the ESS Vision 2020 regarding the sharing of software resources between NSOs. They should also be considered in the context of the various frameworks in place, *e.g.* the ESS Enterprise Architecture Reference Framework, the Common Statistical Production Architecture and the Generic Statistical Business Process Model. Besides, the adoption of OSS is in line with both the Open Source Software Strategy of the European Commission and the recommendations of the Interoperability Solutions for Public Administrations, Businesses and Citizens programme (ISA[2]) through the Sharing and Reuse Framework for IT Solutions.

4. ENCOURAGING CITIZEN STATISTICS AND EMPOWERING PRODUSERS THROUGH REPRODUCIBILITY AND PUBLIC SCRUTINY

The requirement of openness, transparency and auditability of decision-making systems is not automatically equivalent to publicising or accessing data, algorithms and software. In this aspect, the trend towards algorithmic decision-making – and automated data processing techniques – raises methodological and empirical difficulties that can lead to important biases that may be hard to identify [10]. Learning from the reproducibility movement in computational science [7][11], it is however already possible to provide the public with tangible insights into the workings of decision-making systems. In practice, today's technological solutions – *e.g.* flexible Application Programming Interface, lightweight virtualised container platforms, versatile interactive notebooks … – make the development and deployment of reproducible statistical workflows easy by supporting an approach where data and algorithms are delivered as portable, interactive, reusable and reproducible computing services[4]. This further highlights the importance of capturing and sharing data, algorithms and software as well as the computational components needed to "*generate the same results from the same inputs"* [11].

The dissemination of reproducible (and reusable) computational statistics platforms shall enable to engage actively and durably with the *produsers*' communities. Not only it provides them with the data, tools and methods to fully reproduce experiments [2][4][11], by (re)running or tweaking previous data analyses, it also allows them to "*judge for themselves if they agree with the analytical choices, possibly identify innocent mistakes and try other routes"*[13]. In a constantly evolving data ecosystem, this approach supports new modes of production of e-*Official-Statistics* since perfectly configured and ready-to-use computing environments can be distributed with any newly published Official Statistics to the public. There is generally, in an open environment, a set of positive incentives for useful modifications and dramatic improvements to be shared back for the benefit of the entire community [5]. Additionally, it also fosters statistical literacy – in the literate computing paradigm in which the computing is captured along with its motivations and results – and provides with a way to understand "*more epidemiology on how people collect, manipulate, analyse, communicate and consume data"* [13]. Overall, it offers the prospect of engaging the public in the co-creation, design, implementation, testing and validation of statistical products. Last, this will further support the

---

"submissions that explicitly promote the development of fully tested and documented source code and the sharing of reproducible and reusable software solutions".

[4] Beyond providing *"Do-It-Yourself"* methods and tools to access its datasets (*e.g.*, eurostat.js, java4eurostat and pyrostat on its github domain), **Eurostat** is also currently testing some prototypes of ready-to-use computing platforms (*e.g.*, PRost and happyGISCO). See also [12].

potential evolution of Official Statistics from an operation model based on data concentration towards an alternative model based on computation spreading [14].

5. CONCLUSION

The *STATPRO* principles proposed herein aim at introducing a new form of *statistical production and dissemination* that is aligned with the current best practices in other domains and can contribute to new *interaction models* between NSOs and *produsers* to engage in *data processing* for decision-making. They also support new *collaboration models* between NSOs and external partners – may they be individual citizens or institutions from the private sector – for *knowledge production* where each party has full visibility of how the data are processed. This is to say that algorithms and software can be developed in cooperation regardless of where the computation is physically executed.

REFERENCES

[1] European Statistical System Committee (2018): Bucharest memorandum on Official Statistics in a datafied society (Trusted Smart Statistics), DGINS conference.

[2] Broman K., *et al.* (2017): Recommendations to funding agencies for supporting reproducible research, Communication of the American Statistical Association.

[3] Grazzini J. *et al.* (2018): *"Show me your code, and then I will trust your figures"*: towards software-agnostic open algorithms in statistical production, in Proc. *Quality* conference, doi: 10.5281/zenodo.3240282.

[4] Hoces de la Guardia F. (2017): How transparency and reproducibility can increase credibility in policy analysis: A case study of the minimum wage policy estimate, PhD. dissertation, Pardee Rand graduate school.

[5] Wijnhoven A.B. *et al.* (2015): Open government objectives and participation motivations, doi:10.1016/j.giq.2014.10.002.

[6] Nosek B.A. *et al.* (2015): Promoting an open research culture, doi:10.1126/science.aab2374.

[7] Stodden V. *et al.* (2016): Enhancing reproducibility for computational methods, doi:10.1126/science.aah6168.

[8] Association for Computing Machinery (2017): Statement on algorithmic transparency and accountability, US Public Policy Council and Europe Policy Committee.

[9] Barnes N. (2010): Publish your computer code: it is good enough, doi: 10.1038/467753a.

[10] Pedreschi D. *et al.* (2018): Open the black box data-driven explanation of black box decision systems, arXiv:1806.09936.

[11] Beaulieu-Jones B.K. and Greene C.S. (2017): Reproducibility of computational workflows is automated using continuous analysis, doi: 10.1038/nbt.3780.

[12] Grazzini J. *et al.* (2019): Delivering official statistics as Do-It-Yourself services to foster produsers' engagement with Eurostat open data, in Proc. conference on *New Techniques and Technologies for Statistics*, doi: 10.5281/zenodo.3240272.

[13] Leek J. *et al.* (2017): Five ways to fix statistics, doi:10.1038/d41586-017-07522-z.

[14] Ricciato F. *et al.* (2019): Deep data and shared computation: Shaping the future Trusted Smart Statistics, in Proc. conference on *New Techniques and Technologies for Statistics*.