Statistics Coded – Storytelling through literate programming and runnable computing

L.Barrenada[*1], L.Bonamino[*1], R.Derayati[*1], E.Farias da Silva[*1], M.Girardi[*1], D.Gojsic[*2], S.Hadj Hassen[*1], K.Koehler[*1], I.Marinetti[*1], B.Querido[*1], F.Sheeka[*2], J.Davies[2], M.Meszaros[2], H.Lehtimäki[2], J.Grazzini[2]

[1] European Master in Official Statistics – [2] Eurostat, European Commission

Keywords: Do-It-Yourself Statistics; Statistical literacy; Participation & collaboration; Reusability & Reproducibility; Co-design & Co-creation; Data exploration & analysis.

1. INTRODUCTION

The production of knowledge lies at the heart of modern organisations. However, the value of any knowledge product hangs on its effective dissemination to present and future audiences. In the light of the recent debate around "alternative facts", "degraded public discourse" or "post-truth society", immediate call to actions to further improve the dissemination of *Official Statistics* (OS) to wide and diverse audiences is evident [1]. Storytelling is one way to communicate and interact with the public and improve the perception and awareness of as well as the trust in OS[1].

The so-called ***Statistics Coded***[2] introduced in this paper provide with ways to disseminate outputs that document and share statistical processes beyond just the data, but also including code, algorithms, protocols and workflows while also interactively dialoguing with the data. They help sharing best practices, learning from each other's experience, and adopt common methodologies, therefore fostering *Statistical Literacy* in the public.

Inspired by best practices in the *Open Source Software* [2], *Citizen Science* [3] and *Open Science* [4] communities, a participative and collaborative approach was adopted for the creation of the *Statistics Coded* in the form of a *Coding Lab* session with students**.** This initiative is rooted in principles of participation and transparency, enabling "others" to explore, scrutinise, reuse and contribute to statistical products, and spread knowledge as widely as possible. It presents substantial promises for *Citizen Statistics*, *e.g.* for the interaction between *National Statistical Institutes* (NSIs), data users and data producers.

2. TRADITIONAL DISSEMINATION OF OFFICIAL STATISTICS

While openness and transparency underpin the movement towards more inclusive and participative decision-making systems, NSIs have mostly addressed these questions under the sole angle of *Open Data*. It is indeed often believed that by opening up data to the people they aim at serving, NSIs instantaneously become more transparent and accountable. Opening up data obviously provides the opportunity to involve *external actors* (*i.e.*, actors outside the *ESS*[3], *e.g.*, statisticians, scientists, citizens, *etc*...) [5], but, alone, it does not automatically translate to public participation and engagement [6].

---

[*] These co-authors equally contributed to the work presented in this paper.
[1] https://ec.europa.eu/eurostat/cros/DIGICOM
[2] **https://github.com/eurostat/statistics-coded**
[3] https://ec.europa.eu/eurostat/web/ess

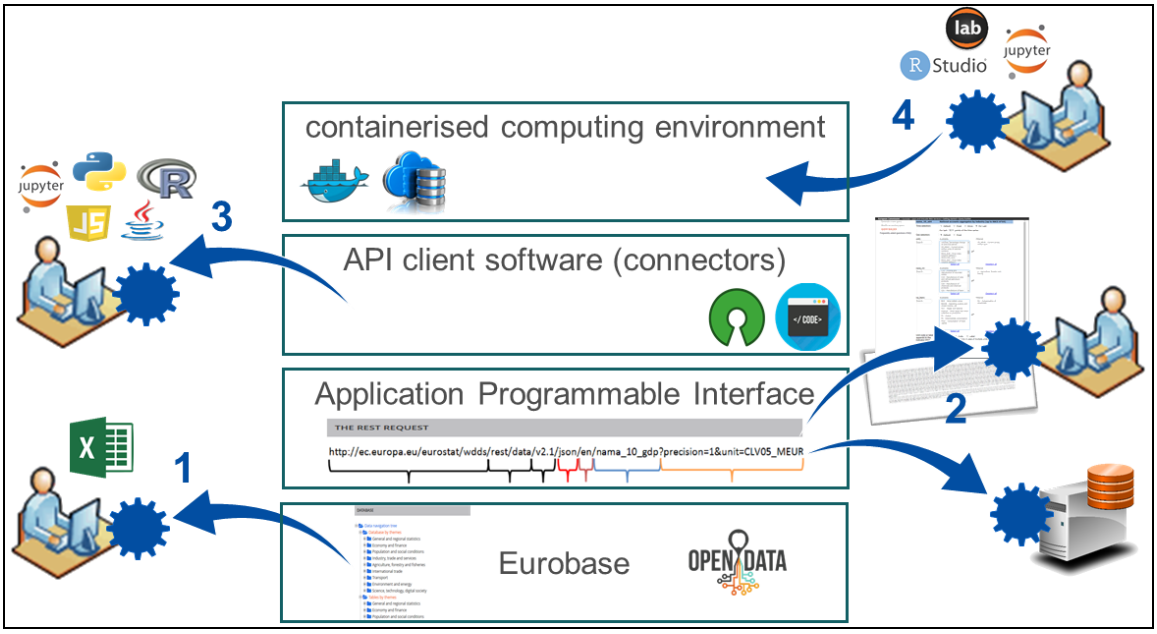


Figure A – *Statistics Coded*... explained. From top to bottom – 1. Users classically have to download the data from Eurostat online open database – 2. Through Eurostat API, it is possible to programmatically interact with the data – 3. Using programmable clients to the API, one can create computational narratives (which are both human- and machine-readable) of how (basic) statistics are presented and (simple) algorithms are constructed using Eurostat data – 4. To enhance the DIY storytelling experience, the *Statistics Coded* provide users with interfaces that enable them to access, discover and fetch Eurostat data. Through these interfaces, data can be dynamically loaded from the online database to a pre-built computing environment with all the necessary software dependencies to conduct their analysis along with the documentation describing how all the pieces fit together.

A substantial gap in the current communication system accompanies the growing need for interactive dissemination. In most disseminations of OS (including "interactive visualisations"), there is an overarching top-down ideology since statistical needs are largely determined by policymakers while the corresponding statistical products are designed and created by NSIs: the final users are only involved as the receivers of the data and services. Dissemination should go beyond traditional reporting and publishing, moving from uni- to multidirectional communication, namely:

- from **public understanding of statistics**: the idea behind public understanding of OS is to make statistics findings more accessible to non-statisticians. The *Statistics Explained*[4] articles and the recent *Regions in Europe*[5] publication are such examples,
- through **public engagement with statistics**: since unidirectional communication cannot solve alone the issue of detachment between OS and society, it is replaced by a dialogue between NSIs and external actors. In this regard, the *ESS forum on Experimental Statistics*[6] was an attempt to invite users to provide feedback on OS products and also engage with them at an early stage of the statistical production,
- to, ultimately, **public participation in statistics**: beyond information and discussion, the society actively participates in OS (through design, creation and reporting) since citizens "talk" to each other and with statisticians. *Produsers* (producers and users)

[4] https://ec.europa.eu/eurostat/statistics-explained/index.php/Main_Page
[5] https://ec.europa.eu/eurostat/cache/digpub/regions/
[6] https://ec.europa.eu/eurostat/web/ess/forum (to be phased out)

could work together during the whole design and implementation phases of OS in order to better align the products with the values, needs and expectations of society while safeguarding the sound methodology and comparability of OS.

Recent developments in open technologies and open algorithms can support the latter approach when involving all relevant stakeholders (including the public) and introducing new forms of co-design and co-creation of statistical products [7].

3. THE STATISTICS CODED AS AN EXAMPLE OF DO-IT-YOURSELF STATISTICS

Driven by the opening and sharing of all assets – not only data, but also methods, tools, and software – researchers nowadays increasingly use computational notebooks to combine – in the so-called *literate programming* paradigm [8] – text, code, and results in a single interactive and portable document [9].

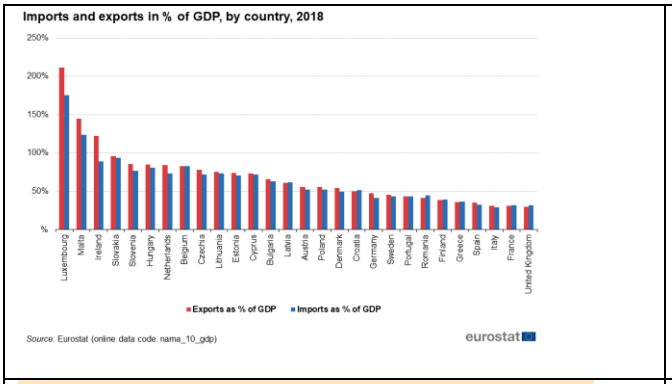
We adopt that same approach to create computational narratives of how (basic) statistics and (simple) algorithms are constructed to produce descriptive statistics based on Eurostat open data[7] (see Figure A). Practically, we extend the *Statistics Explained* articles into interactive computing documents that interleave explanatory text with executable code and simple visualisations. The *Statistics Coded* offer a narrative of any given statistical analysis that is both human- and machine-readable. By using programmable clients to Eurostat Application Programming Interface[8] (API), they provide users with interfaces in various programming languages (*e.g.*, Python, R and JavaScript) and different analytics environments (*e.g.*, Jupyter, R Markdown and Observable notebooks) that enable them to access, discover and fetch Eurostat data. Through these interfaces (*e.g.*, versatile interactive notebooks), data can be dynamically loaded from the online database to a pre-built computing environment (*e.g.*, lightweight virtualised containers) with all necessary software dependencies to conduct (automatically run, udpate and rerun) the study along with the documentation describing how all the pieces fit together.

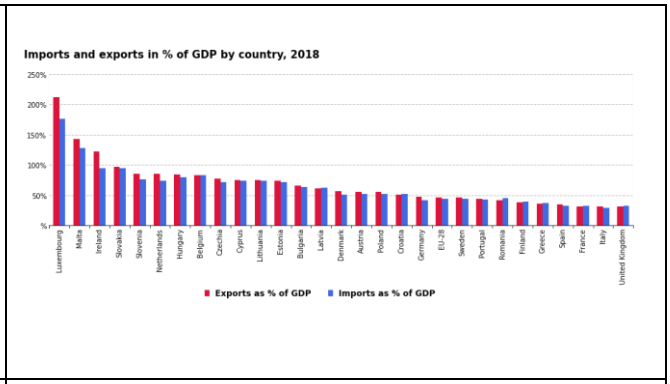
The interactive nature of the *Statistics Coded* makes it quick and easy to reproduce, try out and compare different approaches or parameters [10]. They provide the public with the ability to both perform the analysis and repeat it with different hypotheses, parameters, or data (translating OS into a series of well-understood computational methods), and scrutinize the final products. Advanced users can mine and analyse data to explore patterns, discover problems and link them to OS. Other users can fully reproduce experiments, by rerunning or tweaking previous data analyses, and judge for themselves.

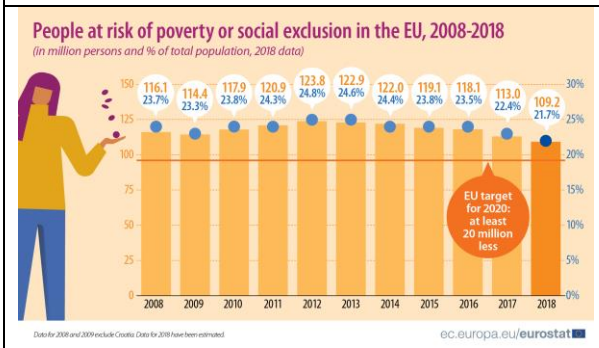
4. THE CODING LAB AS AN OPPORTUNITY TO FOSTER CITIZEN STATISTICS

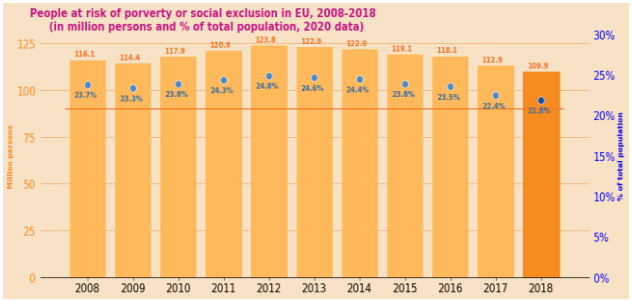
The production of *Statistics Coded* also aimed at creating some kind of *Citizen Statistics* capacities, by reaching out to the next generation of official statisticians (and citizen scientists), and promote the uptake of innovative technologies and methodologies. Practically, a collaborative and participative initiative was launched in the form of a *Coding Lab* session targeting non-professional volunteers with beginner to advanced programming skills. The *Coding Lab* actually targeted graduate students of *European Master in Official Statistics*[9] which is a network of 32 universities in over 20 countries supported to bring together producers of official statistics and academia.

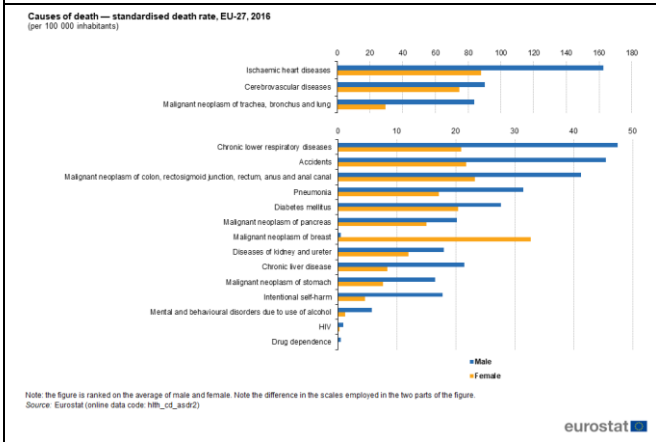
[7] https://ec.europa.eu/eurostat/data/database
[8] https://ec.europa.eu/eurostat/web/json-and-unicode-web-services/getting-started/rest-request
[9] https://ec.europa.eu/eurostat/cros/content/emos-explained_en

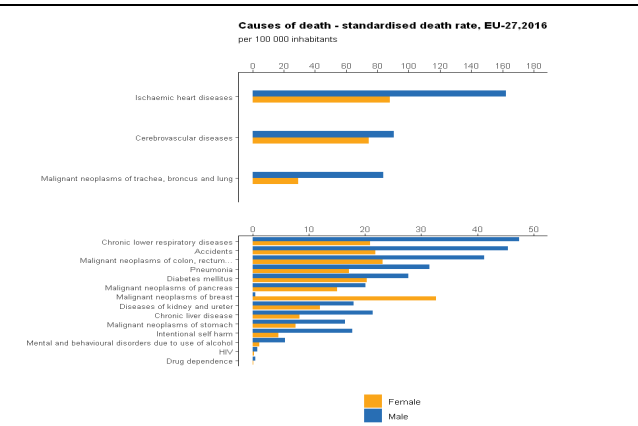


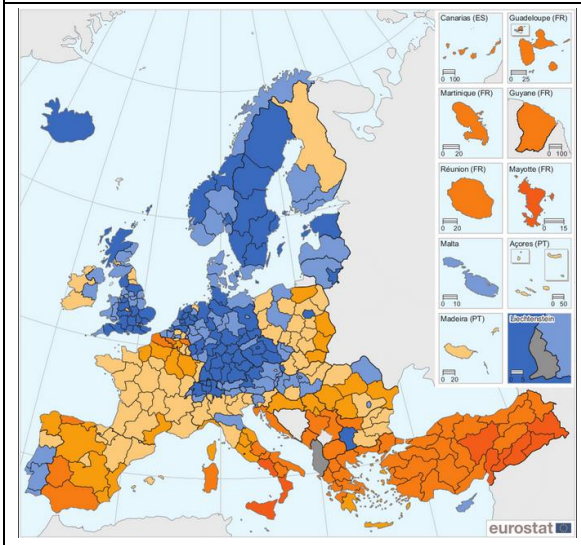


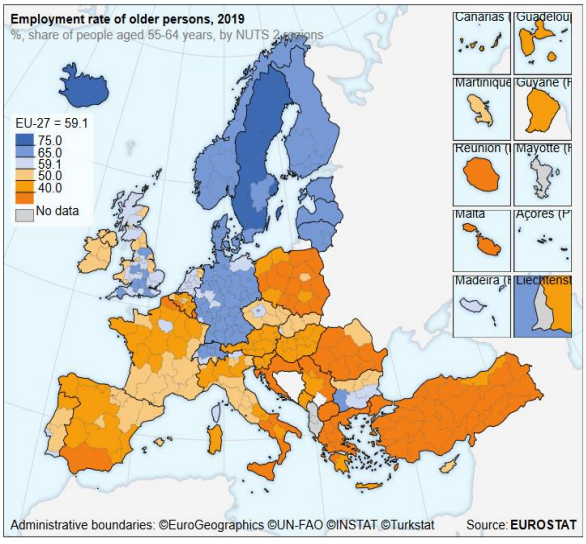












Figures B - Gallery of descriptive statistics from *Statistics Explained* (left) as reproduced in *Statistics Coded* (right). The source articles and the corresponding computational notebooks can be retrieved by clicking on the figures. More is available on the [Github project page](#).

A communication channel and a development platform were used to run the *Coding Lab*, allowing the volunteers to communicate with each other, expose their products and share good practices. The *Coding Lab* provided with a learning opportunity for state-of-the-art programming techniques and statistical developments. It was also the occasion for volunteers to collaborate with statisticians and get further acquainted with the way OS are produced and disseminated. As a result of this collective effort, new *Statistics Coded* artefacts have been generated, see Figures B.

Overall, *Citizen Statistics* shall not only be about collecting or generating data, it should also be a tool for engaging the public in co-design and co-creation activities in other phases of the statistical production, such as data analysis and interpretation, problem definition or dissemination of results, as illustrated by the *Coding Lab* session.

5. CONCLUSION

We introduced the *Statistics Coded* as a way to disseminate outputs that document statistical processes, including code, algorithms and workflows while also interactively dialoguing with data. Besides communicating transparently on methodologies and adhering to best practices, it can foster strategic partnerships when and where possible and help create new participatory models of knowledge and information production. While the importance of openness and transparency in statistical processes, and how these can be supported through open data, open source code and open algorithms has been already emphasized, reproducibility is further enhanced in this approach. The participative and collaborative approach to statistical production also emerges as an important way to innovate products and services traditionally provided in a top-down manner.

REFERENCES

[1] W. Davies (2017): How statistics lost their power – and why we should fear what comes next, *The Guardian*.

[2] M. van der Loo (2017): Open source statistical software at the statistical office.

[3] F. Heigl *et al.* (2019): Opinion: Toward an international definition of Citizen Science, doi: 10.1073/pnas.1903393116 .

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

[5] E. Kalampokis *et al.* (2016): Open Statistics: The rise of a new era for Open Data?, doi: 10.1007/978-3-319-44421-5_3.

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

[7] S. Luhmann *et al.* (2019): Promoting reproducibility-by-design in statistical offices, doi: 10.5281/zenodo.3240198.

[8] D. Knuth (1992): Literate Programming, *University of Chicago Press*.

[9] J. Grazzini *et al.* (2019): Delivering official statistics as Do-It-Yourself services to foster produsers' engagement with Eurostat open data, doi: 10.5281/zenodo.3240272.

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