# ESAC workshop on

## Non-traditional (big) data sources and data science for official statistics

New methods, challenges for data quality, data governance and participation models

## Conclusions

Organisers: ESAC, EUROSTAT, ECB, STATEC

Wednesday, 20 October 2021

## Purpose

The workshop explored challenges and opportunities relating to diverse non-traditional data sources for official statistics, including crowd-sourced, remotely sensed and big data for Official Statistics.

It aimed to spotlight the topic of digital transformation in the context of the Conference on the Future of Europe. Data form an important part of the EU's digital transformation. Big data refers to collected datasets, which are so large and complex that they require whole new approaches, such as innovative statistical methods, artificial intelligence and data science methods that straddle statistics and computer science. These techniques generally collect data automatically and use new methodologies to analyse data and discover patterns for knowledge extraction facilitating rapid decision-making based on statistical and machine learning tools. The timeliness of analysis is an additional feature of these methods.

The workshop focused mainly on new participatory approaches to collecting data, applying data science techniques and producing statistical information with the help of a large group of people (crowd-sourced data). It also looked at methods for collecting georeferenced data using IoT sensors and earth observation (satellite imagery and remote sensing data) and techniques based on the datafication of human activities including computer-mediated communication (text mining and natural language analysis) and financial transactions (credit cards).

Aspects of data governance regarding the need for new principles covering data accessibility, quality assessment, and participatory roles for partners in these non-traditional data sources and official statistics production were also discussed.

## Target audience

The event has attracted interest of over 400 statistics stakeholders in Europe and beyond: renowned academics and senior European, national and local public policy makers, data entrepreneurs, representatives of the international and national banking sector, as well as students and other interested citizens.

In the light of the objectives of the Conference on the Future of Europe, it can be considered a great example of engagement for a common cause: using the digital transformation for enhancing the level and scope of evidence for informed policy making.

# PROGRAMME

## Wednesday, 20 October 2021

**8:40** Virtual event room opens

**8:50** **Opening addresses**

**8:50** Maurizio Vichi, Acting ESAC Chair, Sapienza University of Rome
**9:00** Mariana Kotzeva, Director General, DG ESTAT, European Commission
**9:10** Silke Stapel-Weber, Director General for Statistics, European Central Bank

**9:20** **European Statistics Day 2021 celebration**
Maria João Valente Rosa, ESAC Deputy Chair, NOVA University Lisbon

**9:30** **Panel 1: Agenda for shaping Europe's digital future using statistics, big data and artificial intelligence**

Moderator: Emanuele Baldacci, Director, DG ESTAT, European Commission

Panellists:
Maria Coduti, Data and Innovation Policy Unit, DG CONNECT, European Commission
Francesca Kay, Chief Information Officer, Central Statistics Office Ireland
Mirko Lorenz, Innovation Manager, Co-Founder and chairman Datawrapper Reasona
Caroline Willeke, Deputy Director General Statistics, European Central Bank

**10:45** ***Coffee break***

**11:00** **Advances in Data Analysis and Classification discussion paper:**
**Is there a role for statistics in artificial intelligence?**
*(https://link.springer.com/content/pdf/10.1007/s11634-021-00455-6.pdf)*

Moderator and presenter: Hans A. Kestler, Institute of Medical Systems Biology, Ulm University, Adalbert Wilhelm, Commerzbank Chair of Information Management, Jacobs University Bremen

Discussants:
David Dreyer Lassen, Prorector for Research, Copenhagen Centre for Social Data Science, University of Copenhagen
Chiara Osbat, Adviser, Directorate General Economics, European Central Bank
Vincenzo Spiezia, Senior Economist, Measurement and Analysis of the Digital Economy, OECD

**11:50** **Discussion**

**12:00** **Session 1: Recent developments in non-traditional data sources and new methods of data collection**

Chair: Serge Allegrezza, Director General STATEC

Speakers:
Dominik Rozkrut, President of Statistics Poland
*Access to privately-held data and other principles for non-traditional data sources*
Jacek Stankiewicz, University of Luxembourg; Ariane König, ESAC member, University of Luxembourg, *Can citizen science complement official data sources?*

Steve MacFeely, Director of Data and Analytics at WHO, Geneva, University College Cork –
*"You say you want a [data] revolution": A proposal to use unofficial statistics for the SDG Global Indicator Framework*

**13:05** **Discussion**

**13:15** ***Lunch break***

**14:15** **Session 2: Methodologies for the analysis of non-traditional data sources: challenges, opportunities and risks**

Chair: Corrado Crocetta, President of the Italian Statistical Society

Speakers:
Adam Csabay, Head of Suptech and Lubos Pernis, Head of Suptech Solutions Development at FNA
*Operationalising Alternative Data for Systemic Risk Analysis*
Diego Kuonen, Statoo Consulting & Geneva School of Economics and Management, Geneva University
*How to manage artificial intelligence and data science for decision making as a process of continuous improvement in the context of official statistics?*
Sophia Ananiadou University of Manchester & National Centre for Text Mining
*Natural Language Processing Methods for Search and Discovery*

**15:20** **Discussion**

**15:30** ***Coffee break***

**15:45** **Panel 2: New models of data governance and principles (accessibility, cooperation, proactivity, quality framework, modernisation,) for official statistics**

Moderator: Aurel Schubert, Honorary Professor, Vienna University of Economics and Business
Chair European Statistical Governance Advisory Board

Panellists:
Wieteke Dupain, Head of Knowledge, Research & Development, The European Social Enterprise Network Euclid Network
Eric Rancourt, Director General, Modern Statistical Methods and Data Science, Statistics Canada
Monica Scannapieco, Directorate for Methodology and Statistical Process Design, Italian National Institute of Statistics
Stefaan Verhulst, Co-Founder and Chief Research and Development Officer, The GovLab, New York University

**17:00** **Closing session**
Maurizio Vichi, Acting ESAC Chair, Sapienza University of Rome

## Conclusions

Opening the workshop, the Directors General of Eurostat and ECB Statistics underlined the importance to explore challenges and opportunities relating to diverse non-traditional data sources for official statistics, including participatory methods, remotely sensed, and big data.

The event also celebrated the 2021 European Statistics Day 2021 under the motto 'Statistics, a vaccine to protect democracy and combat the virus of disinformation'. ESAC's statement in that respect concluded that "the heart of our message for European Statistics Day 2021 is the need to defeat disinformation by strengthening the production and use of high-quality statistical information with the highest standards of transparency and accountability, as a vital means of protecting freedom and democracy".

The first panel of the morning concluded that there is a strong need to evolve and guide the digital transition, by means of tools of statistics and artificial intelligence, with the use of big data an, non-traditional sources of information. Panellists stressed that these approaches are fundamental to arrive at a digital Europe with data representing an infrastructure for everyone, close to the real needs of statistics stakeholders, with ethical and transparent methodologies, based on Statistics, Data Science and Artificial Intelligence (AI), with open and integrated administrative and privately-held databases, which can be reused in official statistics and scientific research activities.

The keynote session presented the seminal paper on the interplay between Statistics and Artificial Intelligence. The paper, drafted by a working group of the German Consortium in Statistics (DAGStat), shows that statistical methods must be considered as an integral part of AI systems, from the formulation of the research questions and the development of the research design, through the analysis and up to the interpretation of the results. Particularly in the field of methodological development, statistics can serve as a multiplier and strengthen the scientific exchange by establishing broad and strongly interconnected networks between users and producers of data and methodologies.

AI needs Statistical Science. In fact, AI has benefited greatly from research in computer science and statistics, and many procedures have been developed by statisticians. This has to be clearly understood by European citizens but also by policymakers. Statistics can help optimise data collection and its design (sampling design), while AI can be used to analyse data. Statistical Science must be used to assess the quality of the data and the evaluation of uncertainty by computationally intensive inferential methods. However, also Statistical Science needs AI. AI tools for data collection such as scraping and natural language processing or tools for decision-making are important to speed up the process of data production, analysis and decision-making. Timeliness is a general problem for statistics and AI can help in that respect.

Session 1 discussed recent developments in non-traditional data sources and new methods of data collection the access to privately-held data, as well as other principles for non-traditional data sources. Some of the principles considered more relevant were: the products of official statistics constitute public goods; relevant statistical authorities publish these products according to the fundamental principles of statistics; privately-held data acquired by statistical authorities can be used only for the purposes of production of official statistics and for scientific research; statistical authorities do not purchase data from the data holders.

The use of these data, in fact, improves official statistics and scientific research in all dimensions of quality such as timeliness, accuracy, and relevance. For example, information for all citizens and policy decisions could be clearly improved by having detailed data on mobility produced by the GPS navigation apps. Information on the consumption of goods and services, prices, and habits could improve with data from e-commerce companies; or from credit and debit card data.

A first very promising example of a non-traditional data source that can be developed for official statistics is suggested by citizen science where scientific work is undertaken by the general public, often in collaboration with or under the direction of professional scientists and scientific institutions. Thus, the public participation helps not only to collect data but also contributes to all phases of the statistics production, starting from the initial design. This public participation certainly allows also increasing the public's understanding of official statistics.

A second very interesting example for increasing the use of non-traditional data sources is the proposal of establishing recognised and mandated bodies, with the authority and competence to review non-traditional data sources, to assess whether these data can be certified as ‘official’ for specified purposes. Populating the SDG global indicator framework could be a clear example of the application of this approach. Obviously, these bodies should demonstrate adherence to the UN Fundamental Principles of Official Statistics and indicators should comply with the quality standards defined in the Statistical Quality Assurance Framework. This approach would allow official statistics to adopt a co-production model, take a more proactive role, exert some control and show leadership.

Session 2 debated methodologies for the analysis of non-traditional data sources: challenges, opportunities, and risks. In this context, the interesting the use of non-traditional data for systemic risk analysis was emphasised.

The idea of managing artificial intelligence and data science for decision making as a continuous improvement process in the context of official statistics is clearly very important and confirms that only the joint use of statistics, artificial intelligence, and data science will help official statistics to stay relevant also in the future. An exciting idea in this context is the establishment of a competence centre for AI.

The full use of digitalisation to deal with unstructured data is a challenge that official statistics must address. Some experiments have been conducted but they are limited to restricted cases. Activities to derive high-quality information from text or digitalised files, text classification, text knowledge discovery, or to enrich observed data with semantic information, are activities very useful and central for official statistics. The competence centre should also include a Natural Language processing infrastructure, as information in the form of natural language or text will increase over time.

It should also be noted that Natural Language processing is what allows ‘chatbots’ to understand people's messages in real time. Using this software for official statistics would be a realistic endeavour, as well as transforming unstructured texts into structured data, analysed in real-time with the scope to timely inform or help to make decisions.

The last panel of the workshop discussed the new models of data governance and principles for official statistics, as well as the need for new governance rules for non-traditional data such as big data. The panellists emphasised principles based on ethics,

necessity, and reality. They also stressed that the principles of official statistics are solid, and it is necessary to pave the way for new processes to implement these principles. The need to invest in experimental statistics has been also clearly underlined. Another interesting discussion examined how the access of official statistics to privately-held data should be regulated, with particular importance given to the principle of accessing data for multipurpose use. Finally, panellists debated how to create the necessary confidence of the users on statistics based on non-traditional data, as well as the need to include users in the definition of the production processes with direct participation, starting from the design of the production.

**Next step**

Another workshop, on the topic ‘Co-creation of an evidence-base on changes in human-environment interactions for statistics, policies and action’, is planned for first half of 2022.

It will build on the main insights of the workshop on ‘Non-traditional (big) data sources and data science for official statistics’ and explore how citizen science, in combination with geo-spatial data, may contribute to filling data gaps in official portfolios on human-caused environmental change and at the same time democratise data production processes.