**Trusted Smart Statistics: A reflection on the future of (Official) Statistics**

Konstantinos Giannakouris, Eurostat, konstantinos.giannakouris@ec.europa.eu

Fernando Reis, Eurostat, fernando.reis@ec.europa.eu

Michail Skaliotis, Eurostat, michail.skaliotis@ec.europa.eu

Albrecht Wirthmann, Eurostat, albrecht.wirthmann@ec.europa.eu

Fabio Ricciato, Eurostat, fabio.ricciato@ec.europa.eu

**Abstract**

*The extended use of the Internet of Things (IoT) will eventually take big data to a whole new level and change the data landscape. Data capturing and processing capabilities coupled with analytical and statistical capabilities will be embedded in the smart systems themselves. Intelligence along the data life-cycle enhanced with cognitive processes will be essential components of future statistics. Algorithms will handle huge amounts of data at the limit of human capabilities for exploiting statistical data using traditional data processing methods. We call this smart statistics.*

*We identify smart statistics as the future role of official statistics in a world impregnated with smart technologies. Smart technologies involve real-time, automated, interactive technologies that optimize the physical operation of appliances and consumer devices. Statistics themselves would then be transformed into a smart technology embedded in smart systems.*

*However, statistics are only useful when they are trusted. In order to build trust into smart statistics the data life-cycle needs to be auditable, transparent, with guarantees of accuracy and privacy by design.*

*This paper provides a reflection on the future of official statistics in a hyper-connected world dominated by the IoT. It briefly outlines the concept of smart technologies shaping the future of statistics emphasising the need to embed trust in smart statistics under principles for transposing algorithmic transparency and accountability in smart statistics.*

**Keywords:** Trusted smart statistics, smart statistics, Internet of Things, Official Statistics, Smart technologies

# Smart technologies

Technologies are embedded in devices or more generally in systems in order that they operate as designed and behave as expected. Similarly, statistics are embedded in the policy decision and monitoring processes through designed iterative processes of data collection, data analysis, inference, modelling, interpretation and decision making. Extending the comment of Professor David Hand that "Statistics are the mirror through which we view our society"[[1]](#footnote-1), statistics may be considered as a technology enabling evidence based policy making and in the framework of IoT these statistics may be further extended.

Along the years, terms such as *smart*, *intelligent* and *adaptive* have been used for characterising materials [Hans Irschik et al., 2010], devices or systems that originally were capable of interacting with humans or modifying their behaviour to adapt to the environment. In the context of the IoT, these terms have evolved to describe the enabling technologies that are embedded into the smart devices or the smart systems and additionally embrace enhanced capabilities. For example, "smartness" continues to evolve by describing autonomous systems, capable of drawing conclusions from rules, capable of learning in the sense of using experience to improve performance, anticipating, thinking and reasoning about what to do next, ultimately with the ability to self-generate and self-sustain.

Smart technologies enable the exchange of vast amounts of data and are capable of handling them for analysis, inference and systems' operation. In this context, expressed as "the data deluge", the quantity of information in digital form in the world is soaring. This development can be explained by the term "Datafication", which means "taking all aspects of life and turning them into data" [Mayer-Schönberger, V. and Cukier, K. (2013)]. Consequently, if all aspects of life are becoming "datafied" that means that they can be measured and statistically analysed.

Smart devices, electronic networks and constant generation of data on all aspects of life and the environment will become an integrative component of how our societies and economies will function. Most if not all data in the course of the third decade of the 21st century is expected to be "organic", i.e. by-products people's activities, systems and things, including billions of low-end and affordable smart devices connected to the internet (i.e. IoT). In addition, the fourth industrial revolution (Industry 4.0) and the industrial IoT are transforming manufacturing operations, bringing smart technologies in the spotlight and automation to a higher level of integration of smart systems.

In order to make sense of the huge amounts of data produced within an IoT ecosystem, and ensure that their full richness is leveraged, we rely on algorithms. However, algorithms have evolved and will continue to evolve from rule-based set of instructions executed by computers to algorithms that can learn from data.

# Smart statistics: Smart technologies shaping the future of official statistics

Smart technologies are embedded into the IoT ecosystem. They are embedded in smart agents such as sensors, actuators, microcomputers, etc. and smart systems. Subsequently, they use data communications and interface technologies that allow information to be collected, tracked and processed across local and global network infrastructures. In practise, data flows between interconnected smart devices within broad smart systems, such as smart cities (e.g. smart transportation systems, smart grids) have been pioneering implementations of smart technologies for many years.

Similarly, we may think of smart statistics as being the future system of producing official statistics where essentially, data capturing and data processing capabilities coupled with analytical and statistical capabilities will be embedded in the smart systems themselves i.e. a data layer for statistics within smart systems.

The scale at which data will be captured makes the traditional model of collecting data from several data sources into a single place and then integrating them in order to produce official statistics not feasible. Current practices are not scalable to such degree and in addition technical, organisational and legal obstacles may not be bypassed by incremental adjustments but may require paradigm shifts. Essentially, we would require more interactive, distributed and responsive systems compared to centralised computing power and massive data storage. In addition, these systems would require domain-specific knowledge that would allow data to be correctly processed and interpreted.

The traditional model of *pulling data in* – from data sources to NSIs – will not fit in the new scenario. Instead, we envision a model based on *pushing computation out* – from NSIs to the data acquisition systems. This shift of focus from sources to systems lies at the core of what we call smart statistics.

Therefore, smart statistics requires an extension of the current paradigm of producing official statistics not only concerning novel data sources but also in the part of statistical processes that are moved outside the NSI domain.

Intelligence along the data life-cycle enhanced with cognitive processes will be essential components of smart statistics. Knowledge representation and the use of relevant algorithms should be part of the capabilities of smart statistics when using cognitive processes. Anticipating user needs ("ambient" intelligence of smart statistics), adapting data storage and processing, improving autonomously its algorithms may be considered as part of the intelligent processes that smart statistics will be able to realise.

For example, interactive processes would allow users to feed their needs for statistical information in terms of detail and even applying new alternative definitions and classifications (e.g. query the system for the number of residents in a region, taking as threshold 3 months living in a place instead of the standard 12 months).

# Trusted smart official statistics

The role of official statistics is twofold: 1) to accurately inform citizens in democratic societies about the situation of society and give them the means to assess the success the overall public policies; 2) to inform evidence based policy making.

In a post-truth era, *"relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief"* [Oxford dictionaries] official statistics become even more important and statisticians carry a responsibility in providing hard evidence for making evidence based policies.

In every possible aspect, official statistics need to be by definition trusted, both in what concerns the validity and accuracy of the outputs and also concerning how it handles individual personal data, i.e. protecting its confidentiality and respecting data subjects' privacy.

The regulatory framework [Official Journal of the European Union, 2016] for the protection of confidential information that is collected for the production of European and official national statistics is unambiguous. Moreover, complying with the vision and the mission of the European Statistical system, the [European Statistics Code of Practice] is based on 15 Principles covering the institutional environment, the statistical production processes and the output of statistics. From this point of view, official statistics offer in terms of trust the appropriate regulatory, institutional and organisational environment. Producers of official statistics are expected to maintain and reinforce their role a key providers of data4policy in digital world.

However, in the era of smart statistics, the level of detail provided by the IoT and the possibilities of predicting characteristics and information considered by individuals to be part of their private life pose a challenge for official statistics to maintain the trust of data subjects. *"Hence, end-to-end security between devices and the applications is of paramount importance for protecting the privacy of people's personal data across the different systems and technologies that are involved."* [Tragos et al.]

Moreover, the paradigm shift of *pushing computation out* – from NSIs to the data acquisition systems or to data holders bends the well-established system of trust that lies at the basis of the official statistics production until today. The new scenario requires a different set of technical, organizational and legal tools. The problem we have to face is therefore the design of a coherent framework for trusted smart statistics.

In this context it seems necessary to develop a reference methodological framework that will systematically address general and domain specific challenges. An additional step towards the direction of trusted smart statistics would certainly be the adoption of novel computational models for privacy preserving data usage. For example secure multiparty computation protocols preserve security properties and enable cross-domain data utilization and output delivery without moving nor sharing data across domains.

From the big data point of view, there have been several efforts to develop platforms that support crowdsourced citizen-science data. For example, this is feasible based on information that is collected and transmitted by wearable sensor technology that enables continuous seamless interaction with real-time information. Proofs-of-concept and the development of prototypes for using smart technologies in the data collection process have already been planned in mainstreamed statistics related to time use survey and household budget survey. In this wider aspect, citizen-science data produced by smart devices and smart sensors related directly (e.g. time use, well-being) or indirectly (e.g. a person's geographical location) to individuals' private information bring new types of privacy threats for the persons acting as data subjects.

Another fundamental element for the data ecosystem to work is ensuring trust in the devices. In order for a device to use, besides the data it collects itself, data provided by other devices, and for the other device to share its data, they need to trust each other and the data they exchange. This is true not only between devices at the local level, but also between higher level systems and between levels, from the devices to the systems and vice-versa. That trust includes the consubstantial belief in the accuracy of the data, its authenticity (i.e. that it was not tampered with a specific purpose) and that the data will not be misused.

In order to build trust into smart statistics the data life-cycle needs to be auditable, transparent, with guarantees of accuracy and privacy by design. The latter term "… means that privacy enhancing mechanisms must be deeply rooted inside the IoT architecture. Furthermore, the solution should be such that every data subject should be able to give consent to the collection, storage and processing of their personal data for the particular known in advance purpose (consent condition and legitimate purpose condition[[2]](#footnote-2))." [Tragos et al.]

Moreover, we need principles that should address potential harmful bias in official statistics through the use of algorithms that will be embedded in the relevant smart technologies.

Beyond smart devices that are capable to run algorithms enabling the protection of security and privacy (locally, device software updating, data transmission, etc.), setting the standard for developing, producing and disseminating official statistics should inevitably include the design of principles for transposing algorithmic transparency and accountability in smart statistics. To that end, providers of official statistics may entrust the application of these principles to multi-disciplinary assessment boards. These boards would be responsible for ensuring implementation and assessment procedures under their remit are carried out in accordance to the principles.

We call this future vision of official statistics trusted smart statistics.

# Algorithms for decision making applicable in statistical operations

Shifting towards evidence-based decision-making requires that algorithms would be exploited to handle huge amounts of data at the limit of human capabilities for exploiting statistical data using traditional data processing methods. As aforementioned, algorithms will evolve from rule-based set of instructions to algorithms that can learn from data that have both a high number of attributes as well as a high number of observations.

In a wider conceptual context, when it comes to algorithmic decision making, Diakopoulos [Diakopoulos, 2015] characterises the function of algorithms in four broad categories: 1) prioritization, the denotation of emphasis and rank on particular information or results at the expense of others based on a pre-defined set of criteria; 2) classification, the categorization of information into separate "classes", based on its features; 3) association, the determination of correlated relationships between entities; and 4) filtering, the inclusion or exclusion of information based on pre-determined criteria.

In the context of smart statistics, Table 1 below – adapted from Diakopoulos - presents how algorithms may exert power through decisions that make in prioritizing, classifying, associating and filtering information. Although these "algorithmic decisions" would exclusively lead to choices for specific statistical operations there are a number of human influences embedded into algorithms, such as criteria choices, training data, semantics and interpretation.

Table 1: Functions for algorithmic decision making relevant to statistical operations

|  |  |  |
| --- | --- | --- |
| **Function** | **Decisions concerning statistical operations** | **Examples** |
| **Prioritization** | Selection of sampling units, estimating weighting parameters, …Selection of methods resolving selectivity issues | Selection of weather stations (meteorological spatial interpolation procedure) |
| **Classification** | Selection of stratification parametersClassification, clusteringIdentifying selectivity issues, … |  Classification of Global Trade Item Numbers (i.e. barcodes of products) into ECOICOP[[3]](#footnote-3) for the computation of the HICP |
| **Association** | Selection of statistical models, optimisation of model parameters, resolving collinearity, resolving autocorrelation of parameters over successive time intervals, … | Exploiting input from citizen science data, wearables |
| **Filtering** | Selection of sampling unitsFiltering web-scraped online job vacanciesDimensionality reductionFiltering high frequency data (selecting the frequency at which the signal is most relevant), detection of outliers | Selection of weather stations (meteorological spatial interpolation procedure)Online job vacancies  |

The functions that are presented in the context of smart statistics and algorithmic decision making for statistical operations imply that smart technologies have capabilities for processing quantitative and qualitative data produced within an extended IoT ecosystem[[4]](#footnote-4)

Algorithms are often described as black boxes with great difficulties to map input-output relationships particularly in fully automated smart systems that are capable of providing stream data capture and high speed real-time streaming analytics. Therefore, as mentioned above, it is necessary to design the principles for transposing algorithmic transparency and accountability in smart statistics and proceed with their implementation –either through the use of standards, by design or both. These principles should not ignore the necessary network security measures of the data sources (e.g. use of distributed ledger technology) within the extended IoT.

# Principles for transposing algorithmic transparency and accountability in smart statistics

Inspired by the principles in the "Statement of algorithmic transparency and accountability"[[5]](#footnote-5) that was approved by the Association for Computing Machinery ACM U.S. Public policy Council and ACM Europe Policy Committee, the principles for awareness, access to algorithms, accountability, explanation and transparency, data provenance, data privacy and confidentiality, auditability, and validation and testing can be applied in the context of smart statistics and more particularly supporting algorithms used for decision making applicable in statistical operations for official statistics.

# Open issues, opportunities and challenges

It is obvious that due to the complexity of trusted smart statistics, it is required to develop a business architecture. In this long process there are open issues that should be still addressed in the general context of "trusted smart statistics" but more importantly in specific use cases and statistical domains.

* For instance, the future statistical information system in order to produce trusted smart statistics would require the identification and use of relevant standards, technological solutions and adequate methodologies that are expected to combine design and model based statistics, operate within hyper-connected societies, exploit multiple interoperable smart systems based on the IoT.
* Statistical concepts may need to be (re)defined, adapted to data systems, data availability and the capabilities to process the relevant data, in order to use a common denominator enabling the production of cross-country harmonised and comparable official statistics.
* An issue that needs to be resolved on a case by case basis would be how and at what level security and trust should be built into smart systems when used for the production of official statistics. A risk assessment should be carried out in order to identify, assess and avert potential faulty, biased or malicious algorithms or data sources.
* Close collaboration of statisticians with various stakeholders, engagement with data holders, the industry, subject-matter experts and international standardisation bodies is necessary for designing the future of official statistics based on trusted smart statistics. In particular, the issue of data holders should be deeply analysed based on their specificities with regard to trust and particularly the use of algorithms for decision making applicable in relevant statistical operations.

# References

Diakopoulos (2015) Algorithmic Accountability, Digital Journalism, 3:3, 398-415, DOI: 10.1080/21670811.2014.976411 http://dx.doi.org/10.1080/21670811.2014.976411

European Statistics Code of Practice (2011) http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-32-11-955 accessed on 13/12/2017

Irschik Hans et al. (2010) On the Use of Piezoelectric Sensors in Structural Mechanics: Some Novel Strategies ISSN 1424-8220 www.mdpi.com/1424-8220/10/6/5626/

Mayer-Schönberger, V. and Cukier, K. (2013) Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt

Official Journal of the European Union (2016) http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN and <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016L0680&from=EN>

Oxford dictionaries [https://www.oxforddictionaries.com/press/news/2016/12/11/WOTY-16 accessed in January 2018](https://www.oxforddictionaries.com/press/news/2016/12/11/WOTY-16%20accessed%20in%20January%202018)

Tragos et al. Securing the Internet of Things – Security and Privacy in a Hyperconnected World http://www.internet-of-things-research.eu/pdf/Building\_the\_Hyperconnected\_Society\_IERC\_2015\_Cluster\_eBook\_978-87-93237-98-8\_P\_Web.pdf p.191-192

1. Quoted by Prof Paul Allin, in "Holding a mirror up to society: some challenges in measuring national wellbeing and progress". Contemporary Issues in Statistics: a Conference Celebrating David Hand’s 65th Birthday, 18 September 2015, Imperial College London http://wwwf.imperial.ac.uk/~nadams/events/djh2015/invited.html accessed on 11/12/2017 [↑](#footnote-ref-1)
2. Concepts are relevant to the EU regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN> [↑](#footnote-ref-2)
3. European classification of individual consumption according to purpose (ECOICOP) [↑](#footnote-ref-3)
4. The term "extended Internet of Things data ecosystem" refers to an integrated system of smart devices and smart sensors that produce massive "machine generated" data in real-time or almost real-time. An extended IoT data ecosystem covers a wide range of diverse smart devices and cross-platform deployments of various embedded technologies. The integration of other human generated data cannot be excluded from an extended IoT data ecosystem. [↑](#footnote-ref-4)
5. ACM U.S. Public Policy Council (2017), ACM Europe Council (2017) on Statement on Algorithmic Transparency and Accountability. https://www.acm.org/binaries/content/assets/public-policy/2017\_joint\_statement\_algorithms.pdf accessed on 14/12/2017 [↑](#footnote-ref-5)