Statistical Journal of the IAOS 35 (2019) 589–603
DOI 10.3233/SJI-190584
IOS Press

Trusted smart statistics: Motivations and principles

Fabio Ricciato*, Albrecht Wirthmann, Konstantinos Giannakouris, Fernando Reis and Michail Skaliotis
European Commission, Eurostat 5, rue A. Weicker, L-2721 Luxembourg

Abstract. In this contribution we outline the concept of Trusted Smart Statistics as the natural evolution of official statistics in the new *datafied* world. Traditional data sources, namely survey and administrative data, represent nowadays a valuable but small portion of the global data stock, much thereof being held in the private sector. The availability of new data sources is only one aspect of the global change that concerns official statistics. Other aspects, more subtle but not less important, include the changes in perceptions, expectations, behaviours and relations between the stakeholders. The environment around official statistics has changed: statistical offices are not any more data monopolists, but one prominent species among many others in a larger (and complex) ecosystem. What was established in the traditional world of legacy data sources (in terms of regulations, technologies, practices, etc.) is not guaranteed to be sufficient any more with new data sources. Trusted Smart Statistics is not about *replacing* existing sources and processes, but *augmenting* them with new ones. Such augmentation however will not be only incremental: the path towards Trusted Smart Statistics is not about tweaking some components of the legacy system but about building an entirely new system that will coexist with the legacy one. In this position paper we outline some key design principles for the new Trusted Smart Statistics system. Taken collectively they picture a system where the *smart* and *trust* aspects enable and reinforce each other. A system that is more extrovert towards external stakeholders (citizens, private companies, public authorities) with whom Statistical Offices will be *sharing computation, control, code, logs* and of course final statistics, without necessarily sharing the raw input data.

Keywords: Officials statistics, trusted smart statistics, big data

1. A brief introduction to modern official statistics

The essential mission of official statistics is to produce a *quantitative* representation of the society, economy and environment for purposes of public interest, for policy design and evaluation, and as basis for informing the public debate. In other words, official statistics provides the society with "knowledge of itself" [1]. In modern states, this task has been carried out by Statistical Offices (SO), public institutions with legally guaranteed independence from other governmental bodies and private entities, established around the principles of statistical authority and protection of statistical confidentiality [2,3]. Historically, SOs have been in full control of the whole process chain, from the design and execution of data collection based on censuses and surveys, through the following stage of statistical production and then dissemination of the final statistical products. A *system of trust* was established through a consistent set of legal, organisational and technical provisions in order to ensure a high level of reliability and quality across the whole process. In a scenario where a single entity controls the whole workflow, *trust in data* (quality, veracity) and *trust in processing* (methodological soundness, principle of purpose) are delivered jointly.

Besides surveys, official statistics has made use of administrative sources, such as birth and death registers for demographic statistics. While in some countries they have always played an important role, other countries have started only recently to exploit administrative sources systematically for the production of official statistics [4,5]. Augmenting the statistical pro-

*Corresponding author: Fabio Ricciato, European Commission, Eurostat 5, rue A. Weicker, L-2721 Luxembourg. E-mail: fabio.ricciato@ec.europa.eu.

duction process by administrative data (in addition to survey data) led to important improvements in terms of timeliness, completeness and accuracy of the statistical products.

Differently from survey data, administrative data were designed and collected for different tasks, other than statistical production, and by different institutions other than SO. Still, the fact that administrative data were held within the public sector allowed them to be ingested by SO and included into the statistical production process within the same *system of trust* already in place. In the exploitation of administrative data for official statistics we can already identify the anticipation of some elements characterising the broader and deeper innovation spurred later by the "Big Data" paradigm.

2. The new datafied world

At the beginning of the new millennium, a compound of technological developments initiated a global process of digitalisation of the whole society. The key milestones were the building of the Internet and the World Wide Web, the advent of pervasive online social networks, the spreading of smartphones and other so-called *smart devices*, and more recently the development of the Internet-of-Things (IoT). Because of such technologies, we all live now in a world where almost every aspect of social, economic and physical interaction among individuals, organisations, objects or systems may be digitised – and actually *is* increasingly digitised. With digitalisation comes *datafication* (a term coined in [6]), i.e. every event or state, in the physical world and even more so in the cyber world, is readily encoded into *data* that are collected, exchanged, stored, processed, analysed and traded. The society, the economy and the physical environment have turned into new *data fountains* collected by a plethora of private and public entities acting as *data buckets*, as sketched[1] graphically in Fig. 1. The scales of volume, dimensionality, frequency, density and variety of such new data are many orders of magnitude higher than any conceivable data collection in the pre-digital era, motivating the adoption of the umbrella term *Big Data* to popularise this phenomenon.

The new scenario outlined insofar bears the question as to how SO should react to it. One possible option is simply: do nothing, stand still, ignore the new data *out there* and continue doing business-as-usual based on traditional survey and administrative data. The "stand still" option is probably the most risky for SO, considering the dual pressure of (i) increasing expectations from the users of official statistics (policymakers, researchers, media, citizens) in terms of timeliness, completeness and relevance of the statistical products; and (ii) increasing competition by other new potential providers of statistics, i.e. private companies offering alternative analytical products and figures. Furthermore, these new challenges add to the trends of declining response rates and decreasing budget that SO are facing since several years. One might argue that SO might survive doing business-as-usual, fenced by the *reputation* gained in the past (an intangible but important distinguishing soft asset not yet matched by potential private competitors) and, more concretely, by the current *legislation*. However, despite legislation and legacy reputation, the role of SO in the society might eventually weaken if their products and services do not keep pace with new needs and expectations and eventually drift towards obsolescence. In other words, reputation and trust cannot replace the quest for innovation.

The other strategic option is for the SO to embrace the new data, including those collected and held by the private sector, and leverage them to enrich and enhance the portfolio of official statistics products, so as to provide the society with a better "knowledge of itself". Recall that SO have the mandate to deliver high quality statistics, where *quality* refers to multiple dimensions as encoded in official documents like the European Statistics Code of Practice [3] or the UN Statistics Quality Assurance Framework [7]. Quality dimensions include relevance, accuracy, timeliness, punctuality, etc. Moving from the abstract principle formulation to more concrete target definition, we must recognise that what could be considered *timely* and *relevant* in a world of scarce data, is not necessarily acceptable today in a datafied world. For all these reasons, we agree with the view expressed in [1] that the adoption of new data for official statistics is not (only) a matter of opportunity, but a political obligation. It is not only a tactic for SO to survive as institutions, but rather a way to continue fulfilling their fundamental mandate and to

[1] The term *digital footprints*, and more recently also the term *digital crumbs* [1], have been used to signify to the fact that our behaviours are represented in some digital form. However, we prefer to adopt the analogy between data and water drops, that better captures the idea that data are not only generated but also collected, stored, traded and moved across entities. In other words the water-data analogy makes more immediate the fact that data, like any fluid, may *flow*, get mixed with other fluids and get processed.

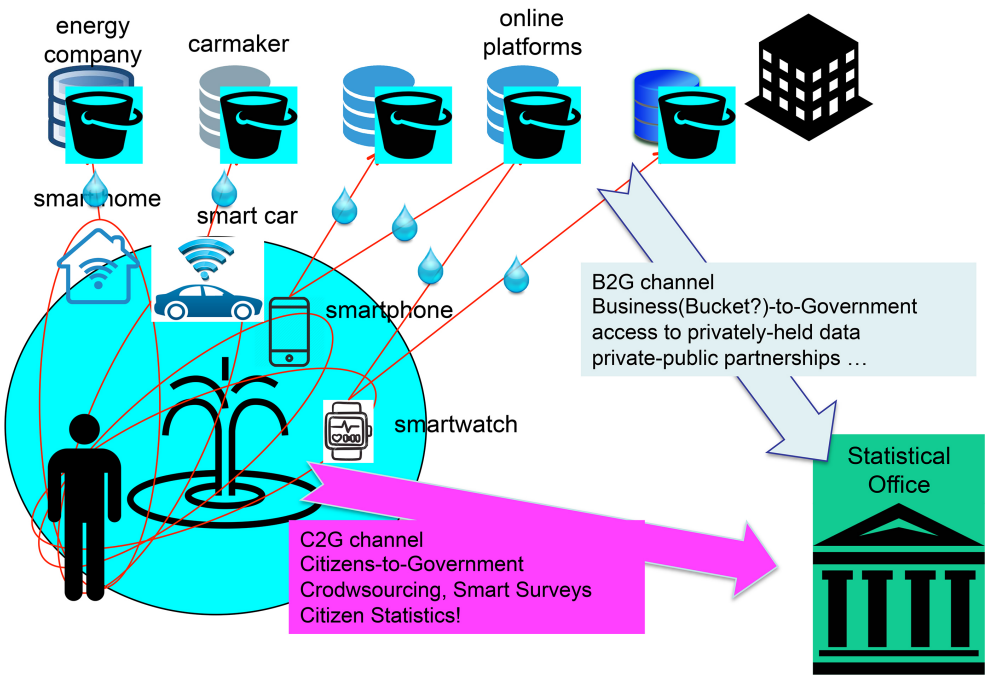

Fig. 1. Data subjects (fountains) and data holders (buckets) may both serve as data sources for Statistical Offices (SO). Trusted Smart Surveys represent an example of Citizen-to-Government (C2G) information flow where SO access the data directly from the data subjects. Accessing MNO data is an example of Business-to-Government (B2G) information flow where SO access the data held by a private company. In both cases, the goal of the Statistical Office is to extract statistical information from the data, not necessarily to acquire the raw data as such.

reassert their role in modern society. So the question is not any more about *Whether* or *Why* SO should embrace new data, but rather *How* they should do so.

3. From *Big Data for Official Statistics* to *Trusted Smart Statistics*

In the recent years, through several case studies, research activities and pilot projects, researchers and statisticians have demonstrated the potential of exploiting such new data sources for official statistics. In Europe, following the Scheveningen memorandum [8], the European Statistical System (ESS) launched in 2014 the Big Data Task Force to build methodological expertise on this matter. The ESS also launched in 2016 the first ESSnet Big Data [9] followed by a second one in 2018 [10]. Similar initiatives took place in other regions and at the UN level.

Such pioneering activities, collectively referred here as *Big Data for Official Statistics*, have evidenced two main aspects. On one hand, some of the new data sources offer an enormous potential in terms of timeliness, coverage, details and insightfulness. On the other hand, such big opportunity comes along with major challenges in almost any implementation aspect: methodological, technical, organisational and legal. Such initial activities led to recognition that the operational system of traditional official statistics, with its processes and practices developed around survey and administrative data, requires more than incremental adaptations to cope with the challenges posed by new data and to reap the opportunities offered by new computing technologies. Rather, a fresh new system-level view is required, calling for a more fundamental rethinking, at various levels, of how official statistics are produced. Such change may be more appropriately regarded as a *major evolution, rather than revolution* of the current official statistics model, considering also that (i) some aspects of the new model were already anticipated by the adoption of administrative data in census; and (ii) that the new model must necessarily guarantee continuity and seamless integration of old and new statistics. Having said that, what matters for our discussion here is that embracing new data sources – and new computation technologies, and new relationships with users of statistics and producers of data – entails a fresh and coherent system-level vision of future official statistics.

Why is this the case? First, new data sources are *not just quantitatively more or bigger* than legacy data

– as the popular but often misleading term *Big Data* might suggest – but (also) *qualitatively different* in almost any aspect. Second, they are generated by a completely new *data ecosystem* with very different actors, relations and dynamics than the ones at play in the context of legacy data sources. Third, the new data sources come along with new computing and processing technologies that were not available in the previous century. Fourth, all that is embedded in a new framework where awareness, perceptions, expectations, attitudes and behaviours by citizens and enterprises in relation to *data* – in their dual role of input data subjects and users of statistical outputs – are profoundly different from the pre-datafied era. In summary, what is entirely new is not only the *content in the data*, but also the *context around the data*. All that considered, it is clear that embracing "new data" for official statistics entails embracing a whole new context, and cannot be simplistically reduced to small incremental innovations of processes and practices. At the same time, legacy data sources and existing processes and practices are not there to be thrown away: new and legacy data sources have complementary roles and SO should aim to combine them in order to draw *a greater and better picture*, not to replace one with the other. Like in stereoscopic vision, where only the combination of different views by the two eyes allow to sense distance and depth, the diversity of measurements for the same phenomenon is itself a source of information, not redundancy.

Given the above requirements, we regard the evolutionary path towards embracing new data sources through the following car analogy: think of "data" as "fuel", and the system of processing data as the "engine". Modern official statistics is like a single-fuel car, with a legacy engine designed for legacy fuel. As SO set out to leverage a new kind of fuel to speed up our car and improve manoeuvring precision (before other cars take over), we claim that the legacy engine is not suited for the new fuel, and that minor tweaking and refitting the engine will not make it apt. Therefore, in our view SO need to build a second engine for the new fuel that will coexist with the legacy engine fed by legacy fuel. In other words, we cannot just refit the engine: we must refit the car towards a multi-fuel multi-engine vehicle. The new engine is basically what we call today *Trusted Smart Statistics* (TSS for short).

The transition from the notion of *Big Data in Official Statistics* to the concept of *Trusted Smart Statistics* embeds a *shift of focus from data sources to data systems*, and a change of perspective about innovation in official statistics from incremental augmentation towards a systemic paradigm change. The concept of *data system* is meant to signify an augmentation of the capabilities and role of data source beyond the mere generation of raw input data, to include also some degree of involvement in the data processing. The evolution of the debate is landmarked by the adoption of the Bucharest memorandum [11] by the ESSC in 2018, five years after the Scheveningen memorandum.

4. Design principles for trusted smart statistics

The operational definition of the TSS concept is a strand of ongoing work, initiated and led by Eurostat, that involves a continuous dialogue within the ESS members and with various external stakeholders, including private data holders, technology providers, academic communities, data protection authorities and other branches of the European Commission. While the authors' work is mostly focused on the European scenario, we believe that many of the concepts and ideas embedded in the TSS vision are relevant and applicable also to other statistical systems worldwide.

In the remainder of this section we sketch the main design principles and system components that collectively represent our current view of the TSS concept. In so doing, we do not mean to attribute to ourselves the novelty or special ownership of any single element of the "big picture" that we are going to draw. We acknowledge that the entire process of formulating a TSS view is continuously inspired and informed by ideas and debates that are taking place across different scientific disciplines and communities, and by pioneering projects and initiatives conducted inside and outside the official statistics domain, in different regions worldwide. Like many other intellectual, social and technological achievements, the development – and eventually the deployment – of the TSS vision is a collective process with several "fathers and mothers" that have individually contributed to start and/or advance some of its parts. Within such collective process, the goal of the present paper is to gather in a coherent vision the main elements underlying the *what*'s and the *why*'s of our view about TSS, and in this way contribute to the ongoing debate in the community of statistical experts and competent bodies – a discussion that however has already reached an important formal milestone in Europe with the adoption of the Bucharest memorandum by the ESSC [11].

As a preliminary step, it is convenient to establish some ground terminology. Unless differently specified,

we reserve the terms "data" and "information" to refer, respectively, to the available input and desired output of a generic computation instance [12]. We use the term "computation" in very general sense to refer to the sequence of instructions required to extract (compute) the desired output information from the available input data. Therefore, depending on the specific context, the term *computation* might be seen as a proxy for *program*, *algorithm*, *method*, etc.

4.1. Aim for output information, not input data

Statistical systems are facing a scenario of cross-domain data processing, where the input data are held by some entity in one (or more) institutional/administrative domain (input party, call it X) while the output information is of interest for another entity in a different domain (output party, call it Y). The key point here is that the output party is not interested in the input data as such, but only as a means to obtain the desired output information. This position does not mean that issues about quality and format of the input data should be disregarded by the SO and delegated to the external data holder. On the contrary, SO must seek to retain the highest possible control on the quality (and format) of the input data, or more precisely of the components therein that are relevant and necessary to extract high-quality statistics. However, we claim that exerting control on the quality of the data does not necessarily imply grabbing the input data directly. For instance, serious accreditation procedures should be devised in order to preliminary ensure that the input data meet acceptable minimum levels of quality, stability and transparency in order to be eligible for official statistics production (see e.g. [13]). If such minimum levels are not met, such data should not be considered for statistical production. However, we argue that the acceptable "minimum levels" of input data quality depend on the methodological capabilities of SO: to some extent, the application of more advanced methodologies may counteract or at least mitigate the impact of certain types of non-idealities in the input data, allowing lowering of the bar for acceptable "minimum levels" of data quality. In other words, the issue of input data quality should be considered jointly with that of methodological maturity.

Another important point is that, in general, there are multiple ways to let the output party Y obtain the desired output information from the input data held by X. Moving the whole input data from X to Y and running the whole computation in Y is only one of several possible strategies. Another strategy is to run the entire computation in X, and then move only the final output information to Y. In between these two approaches, the processing execution can be split between the two parties with some intermediate data exchange. In the context of this paper, Y is the SO and X is the *data holder* (a private company, another public institution or even an individual citizen). In the design stage all possible options should be openly considered so as to decide case-by-case the optimal split of computation between X and Y, considering the specific property of the data source – including what type of input data we are handling, what type of information is to be extracted, and what type of data holder we are interfacing with. In so doing, SO should be clear to themselves that the ultimate goal is not to grab the whole input data, but to obtain output information of the best possible quality. Controlling the quality of input data (by whatever means, possibly but not necessarily by direct acquisition) should be seen as a means to achieve the goal, not as a goal *per se*.

4.2. Clear separation between development and production

One important pillar of future TSS operations is the clear logical distinction between the two stages of *methodological development* and statistics *production*. The methodological development requires human experts to explore at least a subset of the input data (e.g. limited in space, time or from a sample of the target population) in order to devise the most appropriate computation method to extract the desired information (statistics) from the data at hand. The process of methodological development involves an interaction between the human experts and the data, and might lead to the discovery of data features that were not anticipated before, calling for an adaptation of established methodologies. When applied to new digital data sources, we claim that the methodological development stage should involve co-operation between professional statisticians and technology experts from the specific data domain. The final outcome of the methodological development should include a detailed and unambiguous description of the whole processing methodology, possibly in the form of a software program that can be executed fully automatically with no further manual intervention. The software program (and the associated code documentation) should come in addition, not in replacement of higher-level methodological description that is typically represented in the

form of human-readable manual. The software program, developed on the basis of (the exploration of) a subset of test data during the *development stage*, represents the core of the following *production* stage: it will be executed over the complete set of available data in order to extract the complete final statistics in a fully automated way.

By resorting to a software analogy, we may map the methodological development and the production stages to the processes of "writing the code" and "executing the code", respectively. The "code" represents a highly detailed and unambiguous representation of the whole data processing procedure, suited to be executed by machine(s) without human intervention. A well organised code, written in a very modular fashion and complemented by clear and complete documentation, enable a smoother transition of the code towards future versions to follow future advancements of available methods and/or our understandings of the input data. Therefore, we highlight the importance of developing high-quality code with an eye to future evolvability, a point that is elaborated later in Section 4.9.

In principle, separation between development and production, and full automatisation of the latter, might be applied also to traditional data sources, and some SO are indeed doing or attempting to do so at least partially. However, such elements are not as compelling for traditional data as they are for new data sources. Furthermore, in practice the strong legacy of established practices and processes that is in place for traditional data sources might slow down (if not completely prevent) the application of these principles to such data. Instead, the "weight of legacy" is not (yet) there for new data sources, for which SO can afford to take a blank slate design approach to both development and production stages. For these reasons, we expect that the advent of new data sources will boost the adoption of these principles in official statistics.

We also expect that the processing methodologies for new data sources will tend to be more complex and articulated – and therefore more costly to develop – than for legacy data sources. This is mainly because such data are often designed and generated for other different purposes, not for official statistics, and their generation process often includes technology-intensive aspects. Generally speaking, re-purposing such data for official statistics requires more efforts than what is needed for data that were designed specifically for the production of official statistics as primary purpose. In other words, we claim there is a general trade-off between the resources invested in the *ex ante* data design versus *ex post* data interpretation: the better you can design and control the data collection process, the simpler the following data processing methodology. If the level of *ex ante* control is zero, then the complexity of *ex post* data processing (including selection, preparation, cleaning etc.) is maximal. Therefore, from the perspective of SO, the cost saving in the data collection phase are partially offset by higher cost in methodological development

4.3. Pushing computation out instead of pulling data in

As the human intervention gets confined to the development stage, the production stage becomes fully automatic. This has several desirable consequences. For one, SO have now the option to *bring the code to the data* instead of bringing the data to the code. More in general, they may decide to allocate different computation tasks (execution of computation modules) to different machines (computers or networks thereof) in order to optimise the overall task distribution, also considering issues of machine ownership and administrative control.

In several practical cases involving new data sources (but not all) we expect that the most convenient split in computation execution is such that a large part of the computation is carried out at the source, i.e. at the premises of the data holder(s), and only intermediate data (more or less close to the final output) are passed from the data holder(s) to SO. In other words, the SO will be exporting (part of) the computation towards the source, instead of importing all input data inside. This approach is indeed among the key elements of the OPAL project that focuses on using private data for public good in developing countries (see [14] and references therein).

This model is attractive whenever the first stages of the computation chain results in massive reduction of data volume, e.g. through selection, aggregation and any other summarisation function. Running such initial functions at the source saves communication resources at both sides. This model is also attractive when dealing with confidential data that are business sensitive and/or privacy sensitive: reducing the amount of intermediate data that are passed to the SO mitigates the problem, and in some cases might completely resolve it. In any case, this approach is fully in line with the principle of *data minimisation* and *risk minimisation* that are the pillars of modern privacy regulations, and specifically of GDPR (see e.g. the brilliant discussion

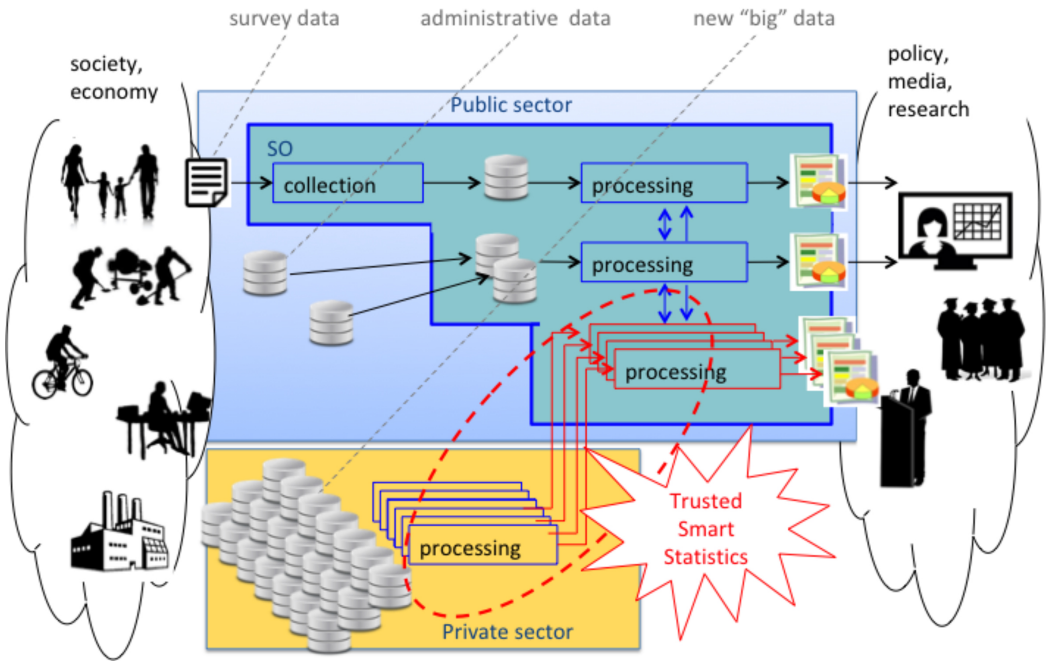

Fig. 2. Exporting part of the computation towards new data sources coexists with the legacy model of gathering (traditional) data in the statistical office.

in [15]). This model is appropriate for several (but not all) new sources of data, and we expect that the two models will coexist, as sketched in Fig. 2.

4.4. *Sharing control in development*

It is important to keep in mind that delegating the *execution* of the computation process does not imply delegation over methodological *development*. A direct consequence of splitting methodological development from production is that we can independently identify *who writes the code* from *who runs the code*. And as a complete methodology will consist of a chain of smaller modules, this principle can be applied to each individual module.

During the methodological development phase, one can design processes involving multiple stakeholders besides the SO to co-develop or at least approve the final code. For example, technology experts from the data holder(s) might be co-operating with professional statisticians from SO to design those parts of the processing chain that require more intensive domain-specific knowledge. More in general, selected key stakeholders might be called to approve the methodology (and code) that results from the development stage, so that each party can directly verify the respect of its legitimate interests and maintain direct (but non-exclusive) control over how data are used. For example, in the potential scenario where Mobile Network Operator (MNO) data are used by SO to extract spatio-temporal statistics about where people are and move from/to, it appears natural to foresee a code approval process that enables MNO experts to verify that what is extracted from the data (and how it is extracted) does not collide with their legitimate business interests, within the limits and framework of the applicable legislation. At the same time, involving qualified representatives from civil society associations (e.g. consumer associations or advocates of civil rights) in the approval process will reassure the public against potential risks of data misuse against the individual (e.g. personal re-identification), groups (e.g. selectivity bias) and the whole collectivity (e.g., social influence).

In other words, we claim that there are multiple key stakeholders that have a legitimate interest on how the data are used and for what purposes, including but not limited to those that hold the data (the MNO in our previous example) and those that have generated the data in the first place (e.g., citizens or companies). We acknowledge that engaging any additional stakeholder in the approval process *ex ante* involves some additional cost, also from the procedural and organisational point of view. However, such cost must be compared against the cost of alternative strategies (e.g. for acquiring their

trust by other means) including the hidden costs corresponding to risks (e.g. reputational).

4.5. Sharing control in production

Once that the code is written and approved by the selected stakeholders, technological solutions are needed to ensure that only the approved code will be executed on the data during the production stage. Moving from the centralised approach (where all data are gathered at the single place where all computation takes place) towards a decentralised scenario (where computation modules are moved towards the data) means sharing control over the code execution among all involved parties. For instance, if the code execution is split between the data holder and the SO then the final output can be only generated if both parties consent to the execution. In this way, *the data holder remains in full, non-exclusive control over what is done with the input data*. But the same applies to the SO. The point to be taken here is that both organisations remain in *full, but non-exclusive control* over the process.

One can extend this model to more than two parties (data holder and SO) and, similarly to what was done in the development stage for code approval, establish a computation process controlled by a group of defferent stakeholders. The configuration of the stakeholder group controlling the code execution does not necessarily coincide with the group in charge of code approval, although in many practical cases it can be expected that the two groups will have the same members (but this should be regarded as an option, not a constraint).

It is evident that *sharing computation* implies necessarily *sharing control* over code execution among the computation parties. Consequently, adopting any kind of Secure Multi-Party Computation technique (more on this in Section 4.6) implies that all involved parties are effectively sharing control over code execution.

However, sharing computation is a sufficient, not necessary condition to share control. In fact, methods may be devised to share control over code execution also with parties that, strictly speaking, are not taking part in the computation. Technological solutions can be identified to achieve this goal. For example, think of a computing platform (made of a single machine or multiple machines) that is enabled to run only binary code that has been jointly authenticated by a group of parties. All such parties are effectively sharing control over the code execution, regardless of who owns and administrates the machine(s). More in general, technological solutions can be identified that achieve the same results by combining existing technologies from the field of hardware security (e.g. Trusted Execution Environment [16]) and cryptographic tools for multi-party authentication in order to build a Multi-Party Controlled Trusted Execution Environment). It is however important to make a clear distinction between the system-level property that one aims to achieve – in our case, shared non-exclusive control over computation execution – and the technological solution that is put in place to achieve that property.

The approach outlined so far of sharing (non-exclusive) control among stakeholders is fundamentally different from the traditional approach of relying on a Trusted Third Party (TTP). The two strategies are graphically sketched in Fig. 3, where each circle represents the logical control perimeter of a specific stakeholder. With the TTP solution, the interested stakeholder must *fully delegate* control to an external entity, namely the TTP, that lies *outside* the control perimeter of all stakeholders. Instead, we propose to consider solutions where process control lies *inside* the control perimeter of all stakeholders, i.e. at their intersection. While TTP is based on delegation of control, this model assumes full retention of (non-exclusive) control by each party. Therefore, every stakeholders must ultimately trust itself – and the technology that is out in place to implement this solution.

4.6. Leverage privacy enhancing technologies

In several cases the desired output information must be obtained by processing input data held by different data holders. If the input data are confidential and the data holders do not trust each other (e.g. because they are business competitors) or anyway don't have a legal basis for sharing data with each other, we must resort to computation models that do not require moving input data from one administrative domain to another. There are different strategies to do so.

In the simplest scenario, the overall computation process can be factorised into independent blocks that are run independently on the data sets of different holders, and each block returns intermediate aggregate data that are regarded as less sensitive from the business and/or privacy perspective, and therefore can be exchanged between parties. The different computation blocks may be run in parallel or sequentially. A possible example of sequential factorisation is given by a neural network that is first trained (partially) on the data set at P1, then the intermediate weights are passed

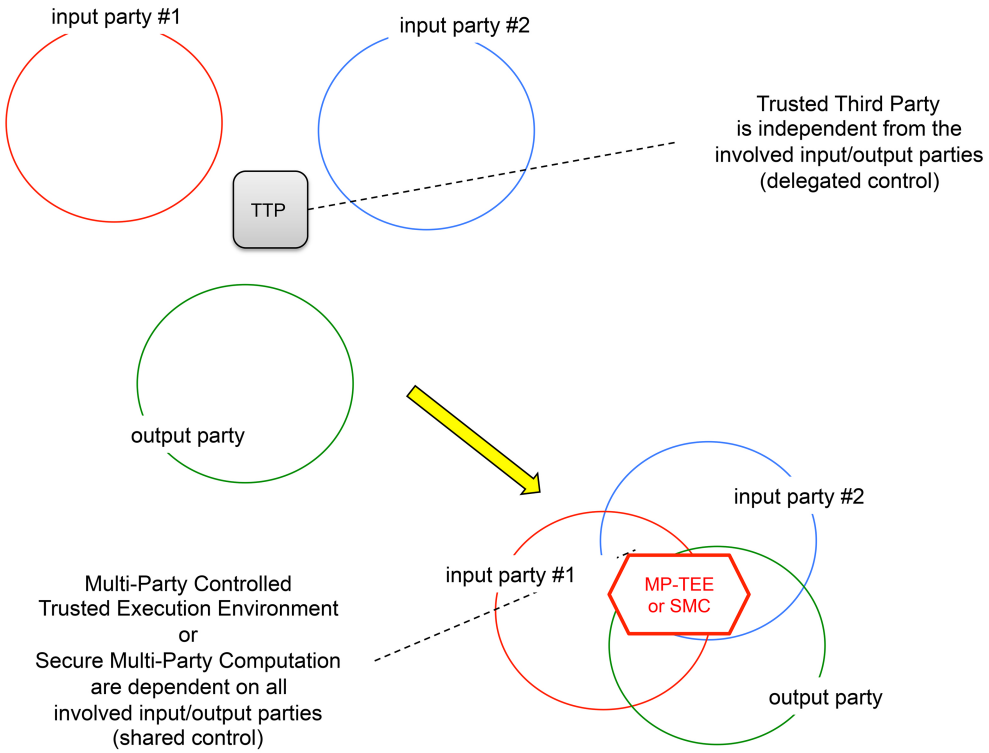

Fig. 3. Delegating control versus sharing control. The Trusted Third Party model (left) all parties must delegate control to an external entity. The technical solutions for Trusted Smart Statistics (like e.g. Multi-Party Controlled TEE or SMC) should instead aim to retain direct (non-exclusive) control among the key stakeholders

on to P2 to continue the training on the second dataset, and so on. In this case, only intermediate weights are passed from one party to the other, not the whole raw data. An example of parallel factorisation is given by the superposition of spatial density maps of mobile users produced by different MNO in the same country [17].

In the most challenging scenarios the computation process requires the exchange of data (raw input or intermediate data) that are still deemed to be confidential by their respective holders or owners. In this case, we must resort to methods that *transform the original data into some other form* before exchanging them between the computing parties. Such data transformation must be such to guarantee the respect of the following conditions *under the given operational scenario*: (i) the transformed data cannot be reverted back to the original data; (ii) the transformed data allow the computation of the correct output information. Regarding the latter point, note that the sequence of instructions required to extract the desired output from the transformed data may be different (and in general more expensive from a computation point of view) than what could be applied on the original data.

In other words, if we imagine the computation process as a path from input data to output information, data transformation represents a logical detour through an intermediate transformation of the data such that, from the transformed data, one can only move forward towards the final output but cannot go back to the original input *as far as the operational conditions foreseen at the design phase are respected*. Therefore, exchanging transformed data does not infringe the confidentiality of the original input data, that are never shared as such.

What we have just described is the essential working principle of so called Privacy Enhancing Technologies (PET), also known as Privacy-Preserving Computation Techniques (PPCT) – here we use the two terms as synonyms. This is an umbrella term for different methods and technologies that have recently emerged at the intersection of cryptography, computer science and distributed systems. Some of these technologies have considerably matured in the last decade and are making their way out from research laboratories into commercial products [18,19] and pilot projects in production settings [20]. Among PET/PPCT the sub-family of Secure Multi-Party Computation (SMPC) seem to

be particularly fit for applications in statistics, due to their lower complexity (relative to other kinds of PET/PPCT) and versatility for general-purpose computation. We remark that any PET/PPCT solution should be seen as one technological component in a broader socio-technological system. In other words, a particular SMPC platform cannot be regarded as a solution *per se*, but likewise any other technology it must be embedded into a framework of non-technological provisions and processes (legal, organisational) that collectively represent the *operational scenario*. Again, we highlight that the adoption of PET/PPCT is fully in line with the principles of *data minimisation* and *risk minimisation* that are encoded in GDPR (see e.g. [15]). This fact represents a solid stimulus for their practical adoption inasmuch such technologies mature to the point of becoming the (future) "state of the art" in protecting data confidentiality.

4.7. Stepping up transparency and accountability

Transparency and accountability are among the pillars of Official Statistics. Traditionally, they are achieved through a composition of instruments and practices that collectively form a coherent "system of trust" (see e.g. the European statistics Code of Practice [3]). Key components of such system are e.g. the open publication of methodological manuals (for methodological transparency); regulations that establish *ex ante* what official statistics may and may not do with the data; a system of peer audits between statistical offices to cross-check and enforce compliance [21]. Such system of trust was designed around traditional data sources (survey and administrative data). As we move into the new world of Trusted Smart Statistics, we must accept the idea that the legacy system of trust might not suffice any more. As we aim at using more pervasive data about citizens and companies than was ever conceivable in the past century, we need to guard against new and higher risks (of data misuse, and reputation) internally and externally to the statistical system. As far as personal data are concerned, we must also consider that the relationship between *citizens* and *data* has deeply changed in the last years: as they produce and consume data in almost any action of daily life, citizens are now well aware of the value and impact of data, and of the associated opportunities but also risks, for the individual and the collectivity. Such new awareness comes with multiple implications for SO. On the positive side, SO have the opportunity to leverage new technologies – and the new *digital behaviours* that come with them – to improve the quality of their measurements, as discussed later in the next subsection. But at the same time, the increased awareness of the risks calls for a major strengthening of the system of trust that is at the foundation of SO mission.

To this aim, we should put in place new, stronger instruments in addition to legacy ones. We need to add *hard* technological safeguards in addition to legal provisions if we want to reassure external stakeholders (including the data holders and data owners) that misuse is not only *legally forbidden* but also *technically impossible*. In other words, we must seek to render data misuse unfeasible legally *and physically*. The adoption of PET/PPCT technologies discussed above goes in this direction. Additionally, we can leverage distributed ledger technologies (blockchain or, better, some more scalable variant thereof) to ensure that each and every query run on the data is logged and can be publicly accessed. In other words, in addition to *sharing computation* and *sharing control* we should learn to *sharing the logs* as well with the key stakeholders. The combination of PET/PPCT and blockchain-like technology is not an entirely new idea (see e.g. [22]) and the system of Official Statistics might learn from initial deployments in other domains.

When processing new data sources – that in many respects are more complex than traditional ones – it is often required putting in place sophisticated and complex processing chains. In this new context, publishing human-readable methodological manual remains important, but is not any more sufficient when methodologies becomes complex and articulated. We should move towards publishing the source code too – a rule that should not be a problem for code that is developed by public non-for-profit institutions like SO – and allow others to check, scrutinise, reuse and even improve it. The principle of algorithmic transparency is declared at the foundation of the OPAL (for OPen ALgorithm) project that focuses on using private data for public good in developing countries (see [14] and references therein). Going even further, we believe that SO should systematically ensure that all the software code along the whole data processing chain, specifically including also those involved at the first stages of initial data preparation (pre-processing, cleaning, selection, etc.), are made fully open and auditable by default. Statistics, like science, should be fully replicable as far as the methodological (computation) part is concerned [23,24]. Whenever proprietary code is used, alternative instruments should be put in place to safeguard methodological transparency, e.g. open

qualification procedures for closed-source code based on public benchmark dataset, certification processes, etc. On these aspects, SO don't have much *to invent*, but rather *to learn* from other disciplines and business domains where such procedures have been used for decades (e.g. safety-critical systems) and adapt them to their specific needs.

4.8. Engaging external stakeholders

SOs traditionally see themselves as the intermediate section of a linear pipeline, with *producers* of (input) data on the back-end and *users* of (output) statistics on the front-end. The data flow is seen as unidirectional: input data are pulled from the producers and final statistics are pushed ("disseminated", "communicated") to the final users.

Trusted Smart Statistics will evolve towards a model where SO will be less and less *in the middle* of a linear pipeline, and increasingly *at the center* of a more articulated network of relations with multiple stakeholders, many of which will be at the same time producers and users (or "prod-users") of information. We must take into account that citizens and other stakeholders have now a different perception of and attitude towards their *data*, and consequently new expectations.

In order to motivate and reassure potential data sources to make their data available with the best possible quality, SO will increasingly need to *give back* useful information derived from those data, and do so on an individualised level. One-way one-shot data flow needs to evolve towards a continuous bi-directional communication, i.e. towards *engagement*. This principle applies most prominently to citizens but also *mutatis mutandis* to other stakeholders, including private data holders.

Thanks to mobile technologies, nowadays it is common for individuals – in their role of *consumers* and *customers* – to maintain a continuous dialogue with multiple online platforms, service providers, retail fidelity programs, and alike. Some of them are also involved in so-called *citizen science* programs [25]. SO should learn from the experiences developed in other fields and initiate their own dialogue with individuals in their role as *citizens*. There is a lot to learn about citizen engagement from the fields of marketing and behavioural sciences, but also from the various *citizen science* programs. Inspired by the latter, SO have now the opportunity to develop a new strand of participatory statistics – or *citizen statistics* as termed in [26] – that is not limited to the acquisition of more and better data, but also extends to the participatory definition of *what* to measure and *how* to measure it. The potential impact goes well beyond eliciting more and better data from the citizen: by this way SO have the opportunity (if not the duty) to reassert the role and legitimacy of official statistics as a as a pillar of modern democracy, as brightly noted in [27] (see also [28]).

Building upon mobile technologies and capitalising the new behaviours of *digital citizens*, SO have the opportunity to evolve the way they interact (and engage) with citizens. The change is more profound than merely porting the old questionnaire from paper to screen. It's about redesigning the interaction model, taking into account the availability of new data but also the different perception of data. It's about implementing technological solutions to make data misuse legally *and technically* impossible. It's about communicating the value of statistics as a pillar of modern democracy. Eurostat has coined the term *Trusted Smart Surveys* to refer to the prospective evolution of the survey model in the direction outlined above [26].

4.9. New methodological frameworks

New data sources are different from traditional survey and administrative data in many respects. For one, they are more complex to understand and interpret, and subject to more frequent changes. This is reflected in the complexity of the analytic methods. The overall procedure to compute final statistics from raw input data consists of a thick sequence of functions and intermediate processing steps. The design of each module requires evaluation of multiple methodological options. Some specific modules, and in some case even the overall methodology architecture, requires domain-specific knowledge and cooperation with experts outside the traditional competence perimeter of professional statisticians (e.g. engineers, computer sciences). Since new data are generated from technological platforms and infrastructures that are subject to change – following the normal evolution of the underlying technology and user behaviours – some of the processing functions need updates in order to adapt to changes in the data generation process. In other words, new data sources might be *stable in terms of availability*, but non-stationary in terms of data model. Consequently, methodology needs to be revised and updated more frequently than for traditional data sources in order to track change in data, but also to incorporate new advancements in the methodology itself, e.g. new algorithms and methods.

The picture outlined insofar poses new challenges and requirements onto the methodological development process that are markedly different from those pertinent to traditional data sources. To address them, SO cannot merely develop new methodologies in the old way: they need to rethink the way that methodologies are developed. In other words, the methodological challenge must be approached from the meta-methodology level. The Trusted Smart Statistics paradigm needs to be built on new methodological frameworks fit for new data sources, and such frameworks must be designed to enable methodological agility, cooperative development and continuous evolution. SO already know how to handle data coming from different origins and changing in time: meta-data, data lifecycle, data lineage and provenance are familiar concepts in the statistical community. These concepts, conceived for *data*, need to be extended to *methods* (or equivalently *analytics*, *algorithms*, *software* etc.). SO must accept the idea that new methods (and their implementations) will not be static or quasi-static but rather *dynamic* objects. Likewise for data, we need to handle methods with finite lifecycle, with different origins, and that may move from one producer/user to another one.

Also in this direction, SO have to learn from other fields where such challenges have been tackled for decades, rather than (re)inventing things. For instance, the basic principles of Internet design, namely functional *layering* and structural *modularity*, might be serving as guiding principles also for the design of methodological frameworks for Trusted Smart Statistics [29,30]. The open-source movement and the community has shown the way to cooperative development of large and evolving software projects, and SO have the opportunity to capitalise on the experience matured in this field as well.

4.10. The smart and trusted cycle

New data sources are even more pervasive than traditional ones for data subjects (individual citizens, households, enterprises). Survey and administrative data represent general characteristics, corresponding to features that are static (e.g. birth date, gender) or slowly-varying (e.g. residence address, household composition, company size), coarsely aggregated summaries of individual behaviours and performances (e.g. monthly income, expenditures, number of trips). Some of such data are very sensitive (e.g., health information), but we argue that some kinds of new data sources are even more sensitive. New digital data capture the instantaneous behaviour at the level of individual events: every single transaction, purchase, encounter and even heartbeat can be recorded. Such data provide a even deeper view of how we behave, and in some cases they do so for a large fraction of the entire population. Therefore, we claim that the potential risks associated to data misuse are even higher for new data sources than for traditional ones , both at the individual level (personal identification) as well as at the level of the whole community (think e.g. to the Cambridge Analytica scandal). To guard against higher risks, we must put in place even stronger safeguards against the misuse of such data than what is in place for traditional data, including hard(er) technological instruments to enforce legal and ethical principles. Recent technological advancements play in favour of this trend, as discussed above for PET/PPCT technologies. In this way we can reassure data subjects, data holders and the whole society that their data will be used safely and transparently, and by this way we will be credible (and able) to access and make good use of their deep data. This circular reasoning is exemplified in Fig. 4.

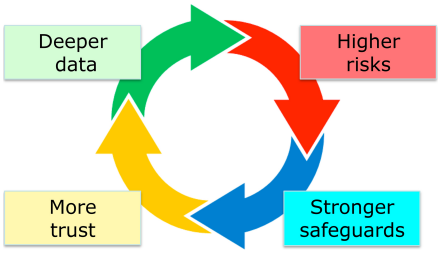

Fig. 4. The smart and trusted cycle.

5. Conclusions and outlook

Official statistics measures life, and when life is changing [official statistics] changes as well.[2] The advent of digital technologies has changed the world we all live in, has changed ourselves, and the change will unavoidable propagate to the world of official statistics.

The availability of new data sources is only one aspect of the global change that concerns official statistics. Other aspects, more subtle but not less important,

[2] Quote by Mariana Kotzeva, Director General of Eurostat at the opening speech of Tredicesima Conferenza Nazionale di Statistica, Rome, 4 July 2018 https://youtu.be/zeCRLizkKko.

include the changes in perceptions, expectations, behaviours and relations between the stakeholders. The environment around official statistics has changed: SO are not any more data monopolists, but one prominent species among many others in a much larger (and complex) ecosystem. This does not mean that the key role of official statistics to provides the society with a credible "knowledge of itself" has lost importance – on the contrary, this role is probably more critical now that many alternative (and less reliable) "knowledges" can be more easily proposed – but it does impact the way it delivers on that mission, i.e. its operational model. In other words, the reason *Why* the system of official statistics exists remains valid, but *How* it operates needs to be profoundly innovated.

What was accepted in the traditional world of legacy data sources (in terms of regulations, technologies, practices, etc.) is not guaranteed to be sufficient any more with new data sources. Trusted Smart Statistics is not about *replacing* existing sources and processes, but *augmenting* them with new ones. Such augmentation however will not be only incremental: the path towards Trusted Smart Statistics is not about tweaking some components of the legacy system but about building an entirely new system that will coexist (and eventually integrate) with the legacy one.

In this position paper we have outlined some key design principles for the new Trusted Smart Statistics system. Taken collectively they picture a system where the *smart* and *trust* aspects enable and reinforce each other. A system that is more extrovert towards external stakeholders (citizens, private companies, public authorities) with whom SO will be sharing computation, control, code, logs and of course final statistics, without necessarily sharing the raw input data!

The point to be taken is that we must *engineer a trust architecture* that reassures all key stakeholders. This is a pre-condition for credibly claiming the *right to use* (not necessarily to acquire) their deepest data. Designing and establishing the new trust architecture involves costs and investments, also because to this aim SO need to draw knowledge and expertise from other domains and acquire new skills. However, such costs are motivated by the higher value and quality of the knowledge that they will be able to produce, for the benefit of the whole society.

One might wonder whether such a complex construction is really needed, and not instead represent an excessive overdoing. Aren't there simpler alternatives to setting in place technologies and process or sharing computation, control, code, logs, etc. in order to avoid the plain sharing of data? One key point to be taken is that value as well as the risks associated to new data ultimately stem from their *use* (or misuse), not merely from their availability. Using data means processing the data and acting upon it. In other words, *value and risks* must be accounted jointly to the combination of *data*, *computation* and *action*. Preventing the collection of the data in the first place, erasing the collected data or anyway destroying their information content, in part or in whole, are simplistic and partial countermeasures against the risks that, however, diminish the value as well. All such strategies impinge exclusively on the *data* component, completely ignoring the *computation* component. Notably, all anonymisation techniques fall in this class as they ultimately destroy (by removal or randomisation) part of the information contained in the data in order to prevent the risk of personal re-identification. In so doing, they decrease somewhat the potential value of data use, and anyway do not protect against other kinds of risk, e.g. discrimination against large groups of people (that are not necessarily identified at an individual level) or social influence. In order to lower the risks (including but not limited to for personal re-identification) and at the same time preserve the value of data use, we must *shift the focus from the data component to the computation component*.

The principles outlined in this position paper collectively drive official statistics to regain their role as pillar of modern democracy. Leveraging new digital technologies to establish stronger participatory approaches will put SO in a pivotal role to exert a better *social control on data usage* – probably the only antidote against the risk of drifting data usage as a tool *for* social control.

The availability of new data sources (also called "non-traditional data" or "big data" in the community of official statistics) is one important driver, but not the only one for Trusted Smart Statistics. We see clearly the potential value and opportunities of new data sources, but also the associated risks and challenges. In our opinion, exercising a fair amount of constructive critical thinking is absolutely necessary when considering new data sources for official statistics, avoiding the opposite but equally a-critical extremes of enthusiastic hype ("big data will solve all problems of official statistics!") and complete denial ("big data have no place whatsoever in official statistics!"). We acknowledge that not all new data may be accepted as appropriate sources for official statistics. At the same time, we are convinced that what is defined

as "acceptable" and "appropriate" depends on the operational model of official statistics and on the methodological capabilities of SO. For example, claiming that some data source X is of "insufficient quality" for the production of official statistics might either point to a deficiency in the data or to a deficiency in the methodology. Furthermore, such deficiency might not be unresolvable. There is a certain risk that professional statisticians, who are very well trained on traditional data but still relatively unfamiliar with some new kinds of data, might not fully and deeply understand all the sources of uncertainty at play in the latter. In such case, there might be an understandable mismatching between the methodology that is developed and applied on the new data (typically in some test or pilot project) and the ideal best possible methodology. Therefore, in the face of "poor output statistics" based on new data sources it is not always obvious whether the problem lies in the "poor input data" or rather in the "poor methodology". Similar arguments can be made in case the problem lies with the speed of change (rather than quality) of the data: is the input data changing too fast, or is rather the methodology development cycle too slow? All these considerations point in the same direction: the decision as to whether new data sources can be accepted or not for the production of official statistics comes logically after a serious and critical reconsideration of the production model for official statistics.

Finally, we remark that the relevance and applicability of the view expressed in this paper is not limited to the European context, and we hope the ideas gathered in this work would contribute to advance the discussion about innovation of official statistics on the global scale.

Acknowledgments

The authors are grateful to the two anonymous reviewers that have provide timely and detailed comments on a preliminary version of this contribution. Their critical and constructive remarks were precious to improve the final version of the present paper.

The views expressed in this paper are those of the authors and do not necessariy represent the official position of the European Commission.

References

[1] Letouzé E, Jütting J. Big Data and Human Development: Towards a New Conceptual and Operational Approach; 2015. DATA-POP Alliance White Paper, https//paris21.org/sites/default/files/WPS_OfficialStatistics_June2015.pdf.

[2] United Nations. Fundamental principles of official statistics. Official Resolution adopted by the UN General Assembly on 29/1/2014. https//unstats.un.org/unsd/dnss/gp/fp-new-e.pdf.

[3] European statistics Code of Practice – revised edition 2017. https//ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142.

[4] Wallgren A, Wallgren B. Register-based Statistics – Administrative Data for Statistical Purposes. John Wiley & Sons, 2007.

[5] UNECE. Register-based statistics in the Nordic countries; 2007. https//unstats.un.org/unsd/censuskb20/KnowledgebaseArticle10220.aspx.

[6] Cukier K, Mayer-Schoenberger V. The rise of big data. Foreign Affairs. 2013 May/June.

[7] UNSD. UN Statistics Quality Assurance Framework, 2017. https//unstats.un.org/unsd/unsystem/Documents-March2017/UNSystem-2017-3-QAF.pdf.

[8] Scheveningen Memorandum on Big Data and Official Statistics, 2013. https//ec.europa.eu/eurostat/cros/content/scheveningen-memorandum_en.

[9] ESSnet on Big Data 2016–2018. Final Technical Report; https://webgateec.europa.eu/fpfis/mwikis/essnetbigdata/images/f/f9/SGA-2_Final_technical_report.pdf.

[10] European Commission Eurostat/G6. Deliverable D5: Accreditation procedure for statistical data from non-official sources; https://europaeu/!HT93bY.

[11] European Statistical System Committee. Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics); 2018. http//www.dgins2018.ro/bucharest-memorandum.

[12] Zins C. Conceptual approaches for defining data, information, and knowledge. Journal of the American Society for Information Science and Technology. 2017; 58(4). doi: 10,1002/asi.20508.

[13] European Data Protection Supervisor. Preliminary Opinion on privacy by design. Opinion 5/2018, https//edps.europa.eu/sites/edp/files/publication/18-05-31_preliminary_opinion_on_privacy_by_ design_en_0.pdff.

[14] The OPAL project. https//www.opalproject.org/about-opal.

[15] European Data Protection Supervisor. Preliminary Opinion on privacy by design; Opinion 5/2018, https://edpseuropa.eu/sites/edp/files/publication/18-05-31_preliminary_opinion_on_ privacy_by_design_en_0.pdff.

[16] Sabt M, Achemlal M, Bouabdallah A. Trusted Execution Environment: What It is, and What It is Not. In: IEEE Trustcom/BigDataSE/ISPA, 2015. doi: 10.1109/trustcom.2015.357.

[17] Ricciato F, Widhalm P, Craglia M, Pantisano F. Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. Pervasive and Mobile Computing. 2016 May.

[18] Big Data UN Global Working Group. UN Handbook on Privacy-Preserving Computation Techniques, 2019. https//tinyurl.com/y3rg5azm.

[19] Archer DW, Bogdanov D, Pinkas B, Pullonen P. Maturity and Performance of Programmable Secure Computation. IEEE Security & Privacy. 2016; 14(5). doi: 10.1109/MSP.2016.97.

[20] Bogdanov D. et al. Students and Taxes: a Privacy-Preserving Social Study Using Secure Computation. Proc on Privacy Enhancing Technologies (PoPETs). 2016, https//eprint.iacr.org/2015/1159.pdf.

[21] Peer reviews in the european statistical system. https//ec.europa.eu/eurostat/web/quality/peer-reviews.

[22] Zyskind G, Nathan O, Pentland A. Enigma: Decentralized Computation Platform withGuaranteed Privacy, 2015. https//arxiv.org/pdf/1506.03471.pdf.

[23] Stodden V. The reproducible research movement in statistics. Statistical Journal of the IAOS. 2014; 30. doi: 10.3233/SJI-140818.

[24] Grazzini J, Lamarche P, Gaffuri J, Museux JM. Show me your code, and then I will trust your figures: Towards software-agnostic open algorithms in statistical production. In: New Techniques and Technologies for Statistics (NTTS) conference; 2018, https//zenodo.org/record/ 3240282#.XbsC-CVTPw6g.

[25] Kay SM. Citizen science for policy formulation and implementation. UCL Press, 2018.

[26] Ricciato F, Wirthmann A. Trusted Smart Statistics: how new data will change official statistics. In: Data4Policy conference, London, 2019. doi: 10.5281/zenodo.3066061.

[27] Ruppert E, Gromme F, Spilda U, Cakici B. Citizen Data and Trust in Official Statistics. Economie et Statistique/Economics and Statistics. 2018; 505-506. doi: 10.24187/ecostat.2018.505d.1971.

[28] Ruppert E. Different Data Futures: An Experiment in CitizenData. In: DGINS 2018 conference, 2018. http//www.dgins2018.ro/wp-content/uploads/2018/10/24-citizen-science-DGINS-Ruppert.pdf.

[29] Ricciato F. Towards a Reference Methodological Framework for processing MNO data for Official Statistics. In: 15th Global Forum on Tourism Statistics, Cusco, Peru, 2018. https//tinyurl.com/ycgvx4m6.

[30] Ricciato F, Lanzieri G, Wirthmann A. Towards a methodological framework for estimating present population density from mobile network operator data. In: IUSSP Research Workshop on Digital Demography in the Era of Big Data, Seville, 2019. https//europa.eu/!Xf83qG.