EUROPEAN COMMISSION
EUROSTAT

Deputy Director-General
Task Force Big Data

Doc. DDG.TF.BD.2018 04 19-20 03 – Reference Architecture.docx

ESS Task Force Meeting on
Big Data and Official Statistics
18 October 2018, 09:00 – 19 October 2018, 13:00
Eurostat

POINT 6

TRUSTED SMART STATISTICS – REFERENCE ARCHITECTURE

<blank>

Towards a Reference Architecture for Trusted Smart Statistics

Fabio Ricciato, Michail Skaliotis, Albrecht Wirthmann, Kostas Giannakouris, Fernando Reis

EUROSTAT Task Force on Big Data, 5, rue Alphonse 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, where traditional data sources (survey and administrative data) represent a valuable but small portion of the global data stock, much thereof being held in the private sector. In order to move towards practical implementation of this vision a *Reference Architecture for Trusted Smart Statistics* is required, i.e., a coherent system of technical, organisational and legal means combined to provide an articulated set of trust guarantees to all involved players. In this paper we take a first step in this direction by proposing selected design principles and system components that, as of the current state of play, we believe will be part of the final design. The goal of this contribution is not to propose a ready-made fully-fledged solution, but rather build awareness about the necessary elements (technological and not) and fuel the discussion with the relevant stakeholders[1].

1. A brief introduction to modern Official Statistics

The 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. For over a century, this task has been carried out by Statistical Offices (SO), public institutions with legally guaranteed technical independence from other governmental bodies and private entities, statistical authority and statistical confidentiality [1]. Since their establishment, SO have been in full control of the whole statistical production process, including the design and execution of data collection based on censuses and surveys. A *system of trust* was developed through a consistent set of legal, organizational 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) were 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, it was only more recently that administrative sources started to be explored systematically in most of the countries to augment official statistics production [2][3]. Such augmentation led to important improvements in terms of timeliness, completeness and accuracy of the statistical products. Differently from survey data, administrative data were

[1] *The views expressed in this paper are those of the authors and do not necessarily reflect the official views of the European Commission. Any potential errors, omissions and inconsistencies are the sole responsibility of the authors.*

designed and collected for different (administrative) 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 purposes 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 "smart devices", and more recently the development of the so-called Internet-of-Things (IoT). We now live in a world where almost any aspect of social, economic and physical interaction among individuals, organizations, objects or systems is digitised. With digitalisation comes *datafication* [4]: 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 "fountains of data". 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 term "Big Data" to popularize this phenomenon.

Given this new scenario, how should SO react to it? One possible option is to simply ignore the new data "out there" and continue doing business-as-usual based on traditional survey and administrative data. This "stand still" strategy 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., companies offering alternative analytical products and figures. The legacy *reputation* and *trust* gained by SO over many decades are intangible but important distinguishing "soft assets" that are not (yet) matched by potential competitors from the private sector. They might fence SO off competition to some extent, but they will

eventually become irrelevant if the products and services offered by SO 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 held by the private sector[2], and leverage them to enrich and enhance the portfolio of official statistics products. This would be beneficial for the whole society, as more timely and complete products are made available to citizens and policymakers, and at the same time reinforce the role of SO in the society. What it takes for SO to walk along this path?

3. Looking ahead: 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. To this aim, following the Scheveningen memorandum [5], the European Statistical System (ESS) launched in 2014 the Task Force Big Data to build methodological expertise on this matter. Such pioneering activities, collectively referred here as "Big Data for Official Statistics", have evidenced two main aspects. On one hand, new data 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. The traditional framework of official statistics processes, developed around survey and administrative data during the pre-digital era, cannot provide solutions to such challenges. A deep paradigm shift is required, but the new direction is up to us to define.

Differently from the past, new data sources are often held by private for-profit companies with a stake in their data. Such data are not only a by-product of other profitable processes, but represent themselves a competitive commercial asset. Consequently, issues of business sensitivity intermingle with user privacy, resulting in a richer set of confidentiality requirements. Additional challenges relate to the scale, complexity and peculiarities of such

[2] In this contribution we focus on eliciting statistical information from input data collected and held by private business companies acting as "data hubs". A complementary approach for the SO is to elicit information directly from citizens, through e.g. "smart surveys" and crowdsourcing initiatives targeting individuals and their end-devices (smartphones, wearable, home appliances etc.). The "business" and "citizens" channels are complementary within the TSS view: while only the former is in the scope of the present contribution, much of the discussion related to principles and components applies also to the latter.

data: they often require a first layer of domain-specific knowledge to be correctly interpreted, and infrastructure resources to be analysed.

Considering all technical and organisational aspects, it became soon evident that the traditional model of "*pulling data in*" – from data sources to SO – will not fit in the new scenario. Instead, we envision a model based on "*pushing computation out*" – from SO backwards towards the data source. This shift of focus *from sources to systems* lies at the core of what we call "Smart Statistics"[3].

The concept of Smart Statistics represents another step away from the traditional paradigm of official statistics production: not only the data sources, but also part of the processing procedure is external to the SO domain (ref. Fig. 1). This transition, coupled with the private nature of data sources, breaks the traditional system of trust that lies at the basis of the traditional production model for official statistics. In other words, while *the trust principles and objectives remain the same*, a different set of technical, organisational and legal tools are needed to ensure compliance with them in the new scenario. The problem to face is therefore the design of a coherent system for *Trusted* Smart Statistics (TSS for short).

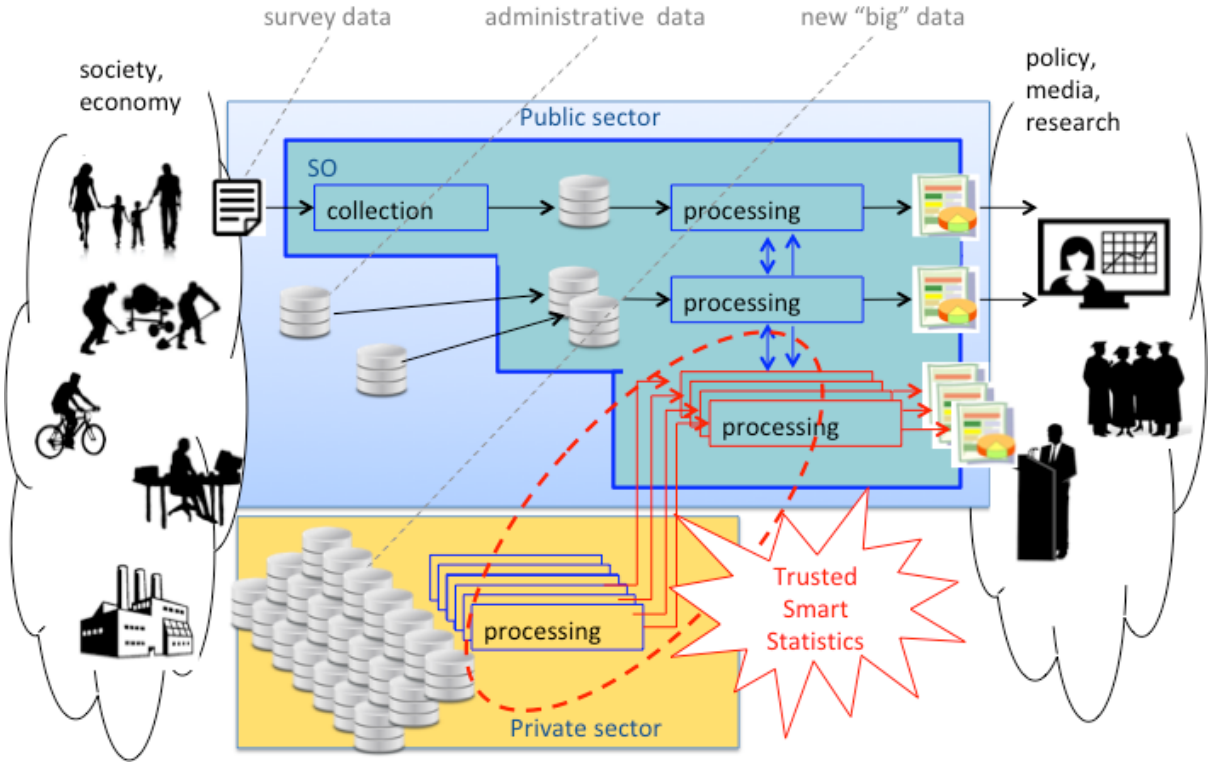

Figure 1

[3] The term is inspired by the notion of "smart systems" as systems that are endowed with the capability of *sensing and processing* information. This enables them to evolve from a collection of passive objects, operated by humans in order to achieve some goal (the light bulb is switched on, the car is driven), towards a composite active system delivering a service when needed (lighting, mobility). Along the same reasoning, Smart Statistics can be seen as a technology embedded into—or a service provided by—Smart Systems.

4. Design principles of a Reference Architecture for Trusted Smart Statistics

In order to take concrete steps towards the implementation of the TSS vision, a "Reference Architecture" is required, i.e., a coherent system of technical, organisational and legal means combined to provide an articulated set of trust guarantees to all involved players. This is a strand of on-going work, initiated and led by EUROSTAT that involves a continuous dialogue within the ESS, with National Statistical Institutes, and with various external stakeholders including private data holders, technology providers, academic communities, data protection authorities and other branches of the European Commission. Herewith we sketch the main design principles and system components that, as of the current state of play, we believe will be part of the final design.

As a preliminary step, it is convenient to establish some ground terminology and make explicit some general aspects about the fundamental notions of "data" and "processing". Unless differently specified, we reserve the term "data" and "information" to refer respectively to the *available input* and to the *desired output* of a generic computation instance [6]. The term "processing" refers to the sequence of instructions needed to extract (or compute) the desired output "information" from the available input "data".

Distinguishing between *input data* and *output information*. We are facing a scenario of "cross-domain processing", where the input data are held by some entity in one 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*. 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 output information to Y. In between these two approaches, we can split the processing execution between the two parties, and exchange some intermediate data. In our design we should openly consider all possible options.

Sharing computation, not data: the case for Secure Multi-Party Computation. When multiple input and output parties are involved, and input data confidentiality is an issue, we can resort to more sophisticated computational models based on Secure Multi-Party Computation (SMPC) [7]. Generally speaking, *under certain applicability conditions* SMPC technology can be used to elicit the desired output information while protecting the

confidentiality of the input data, including cases where the latter are partitioned across multiple input parties. At the core of the SMPC paradigm lays the notion of "sharing computation" instead of "sharing data". Privacy-preserving computation methods like SMPC unleash a paradigm shift that requires also a terminology change: from "sharing data" to "using data" across different domains.

SMPC techniques stem from the intersection between cryptography, computer science and distributed systems. During the last decade these technologies have matured making their way out from research laboratories into commercial products [8], and some of them are now available for adoption in production settings [9]. SMPC however should be seen as a technological component in a broader socio-technological system. It is not a solution *per se*, and likewise any other technology it must be casted into a framework of non-technological provisions (legal, organizational) to set in place the required "applicability conditions".

Distinguishing between *design* and *execution* of processing procedure. By "processing procedure" we identify the complete chain of computation steps that are needed to extract the desired output information from the available input data. The terms "programme", "algorithm", "statistical methodology" etc. may be used equivalently. The overall processing procedure is generally structured in a modular fashion as a chain of different components (e.g., data cleaning, imputation, inference, etc.). For each processing component we logically distinguish the *design* phase, typically performed by human experts, and the *execution* stage, carried out by machines. Such distinction is relevant especially when it comes to decide which organizational entity "controls" the processing procedure or parts thereof. In a complex scenario where multiple organizations (parties) are involved with different interests and roles, *design* and *execution* of the processing procedure should be seen as separate tasks, possibly assigned to different parties or groups thereof. In other words, those exerting control over the definition of the algorithm might not necessarily correspond to those in control of the physical computation resources (machines). Technological solutions (e.g., Trusted Execution Environment [10]) can be adopted in order to guarantee that what is executed (binary code) on the machines administered by one group of parties actually corresponds to what was designed and agreed-upon (source code) by another group. Such technological solutions give us an important degree of freedom for "engineering" the TSS architecture. We can clearly combine such technologies with SMPC, leading to an architecture where algorithm execution is distributed among the group of entities taking part in the SMPC infrastructure, while (experts from) the group of relevant stakeholders maintain control over the algorithm definition, with independent configurations of the two groups.

Sharing control over processing procedure: consensus-based design. Let us start considering a simple scenario where the input and output parties correspond, respectively, to a private data holder (DH) and to the Statistical Office (SO). Both parties have legitimate reasons to maintain full control over the definition of the processing procedure. The SO has the responsibility to ensure that methodological quality is fully preserved and statistically sound methods are adopted across the whole processing procedure. The private DH needs to ensure that the output information extracted from its input data does not jeopardize its own business. By adopting a transparent consensus-based design approach both demands can be fulfilled. The processing procedure should be designed jointly, or at least agreed-upon, by experts from the two organizations, DH and SO. In other words, both parties can share full but non-exclusive control over the entire processing procedure. This strategy can be naturally extended to scenarios where multiple DHs are involved: only processing procedures (algorithms) agreed by all DHs and SO will be admitted to the execution phase.

The consensus-based approach implies that all involved stakeholders have full access to the exact representation of the processing procedure at very fine level of detail: ideally, they would be able to inspect the source-code of all software components. In some cases the processing procedure might involve proprietary components (e.g., algorithm modules) that cannot be publicly disclosed, still the SO should maintain the possibility to verify the methodological correctness and unbiasedness of said components. One possibility is to let the SO inspecting the proprietary source-code (or the parts thereof that are relevant for the statistical methodology) under non-disclosure agreement. Another possibility is to foresee a "qualification" protocol based on extensive tests conducted on benchmark input data (possibly synthetic), similarly to what is done in other domains where critical software is developed for dependable systems. The point to be taken here is that proprietary software and closed-source are not incompatible with the requirement that SO maintains full (non-exclusive) control on the behaviour of the processing components (algorithms) [11][12]. More in general, the adoption of computational methods in the TSS context should adhere to the principles of algorithmic transparency, reproducibility and accountability being developed in the scientific and policy domains [13] [14][15][16].

Potential involvement of certification authorities. In a setting where SO and private DH are cooperating to extract statistics from personal data, they are jointly responsible to ensure that the whole process complies with the applicable privacy regulations and ethical principles. Unclear points in the legal provisions translate into risks: the "liability risk" of incurring into penalties, in case of excessively permissive interpretation, and on the opposite

side the “relinquishment risk” of merely renouncing to carry out some potential analysis, in case of excessively conservative interpretation. In practice, such risks might represent overwhelming deterrents to the establishment of cooperation models between private data holders and SO. A possible way to overcome the problem is to establish a system of *certification* by independent authorities that have the legal accreditation and technical capability to certify compliance to legal provisions and ethical standard *for each specific processing procedure* already at design stage, before actual execution. Ideally, the certification process would be applied directly to the source code of the software implementing the processing procedure, and may involve automatic or semi-automatic compliance checking tools to lower the price of the certification services [17][18].

Besides removing the legal risk mentioned above, the compliancy verification conducted *ex ante* by independent certification authorities would strengthen protection of the privacy and ethical principles encoded in the applicable regulations, and ultimately contribute to increase transparency and *public trust* into the whole process.

Processing procedure publicity and open-access by default. Whenever possible, the processing components adopted throughout the whole procedure should be made publicly accessible as open-source code. This approach has several important advantages. First, it increases the level of transparency of the whole setting towards citizens, and ultimately strengthens their *trust* into the whole process. This is especially important in those cases where the source data are highly pervasive, e.g. tracks of locations and activities by individuals. In other words, an *increased level of process transparency is due to compensate for an increased level of data pervasiveness*. Second, as demonstrated in other domains where open-source software is now common practice, openness comes with independent scrutiny by external experts, hence shorter cycles of identification and resolution of problems (e.g., bugs, security holes, methodological glitches) with positive benefit on the overall quality of the processing procedure itself. Third, it facilitates reproduction, adoption and exchange of components by other institutions (other SO, research institutions, other public organizations, etc.).

5. Conclusions and Outlook

In this contribution we have outlined the concept of *Trusted Smart Statistics* as the natural evolution of Official Statistics in the new *datafied* world, where traditional data sources (survey and administrative data) represent a valuable but small portion of the global data

stock, much of the latter being now held in the private sector. We have taken here a first step towards the definition of a *Reference Architecture* for Trusted Smart Statistics serving as basis for future implementation initiatives. The articulated set of interests and constraints originating by the involvement of diverse stakeholders adds complexity to the challenge. On the positive side, some recent technological developments, particularly in the field of privacy-preserving computation, play in our favour to provide instruments that were not available in the past.

The goal of this contribution is not to propose a ready-made fully-fledged solution, but rather to propose a selection of design principle and system components to build awareness and fuel the discussion with the involved stakeholders. In the continuation of this strand of work within the ESS, through an extensive dialogue with various external stakeholders, we aim to consolidate the view on principles and components and then produce a coherent initial proposal for a Reference Architecture for TSS by end 2019, with the perspective of launching initial pilots and/or proof-of-concept projects with selected private data holders by 2020.

The TSS construction is like a polyhedron with multiple facets, not all of them could be covered in the present contribution. Here we have focused on the use of privately held data for official statistics, i.e., on engaging with the private business. The closely related issue of public-private partnership models is discussed in a companion paper [19]. Another prominent facet of the TSS vision relates to the opportunity of *engaging citizens* in novel, smarter ways than traditional surveys: learning from data crowdsourcing and "citizen science" initiatives in other fields, and through the development of new "Smart Survey" concepts, there is room to develop an entirely new paradigm for "citizen statistics". We defer the discussion on this aspect to a future separate contribution, but we anticipate that several principles and components discussed in this paper – including the externalization of processing execution and SMPC – are entirely applicable in that context as well.

REFERENCES

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

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

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

[4] K. Cukier and V. Mayer-Schoenberger. The rise of big data. Foreign Affairs, May/June 2013. https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data

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

[6] C. Zins. Conceptual approaches for defining data, information, and knowledge, Journal of the American Society for Information Science and Technology, 58(4), 2017., doi: 10.1002/asi.20508.

[7] R. Cramer et al., Secure Multiparty Computation and Secret Sharing, Cambridge University Press, 2015.

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

[9] D. Bogdanov et al. Students and Taxes: a Privacy-Preserving Social Study Using Secure Computation. In Proceedings on Privacy Enhancing Technologies, PoPETs, 2016 (3), 2016.

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

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

[12] Guidotti, R. et al. A survey of methods for explaining black box models, 2018. arXiv:1802.01933.

[13] Association for Computing Machinery. Statement on Algorithmic Transparency and Accountability, 2017. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

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

[15] V. Stodden et al. Enhancing reproducibility for computational methods, Science, 354(6317), 2016. doi:10.1126/science.aah6168.

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

[17] S. Ranise, H. Siswantoro. Automated Legal Compliance Checking by Security Policy Analysis. SAFECOMP 2017 Workshops, LNCS 10489, 2017. DOI: 10.1007/978-3-319-66284-8_30

[18] P. Guarda, S. Ranise, H. Siswantoro. Security Analysis and Legal Compliance Checking for the Design of Privacy-friendly Information Systems. SACMAT'17. DOI: 10.1145/3078861.3078879

[19] F. Ricciato et al. Processing of Mobile Network Operator data for Official Statistics: the case for public-private partnerships. DGINS 2018 Conference.