Session 3
Master class | Quality for new data sources: Progress, challenges and directions for the European Statistical System

Fabio Ricciato[1]

Abstract

In this short contribution we reflect on the implications of (re)using privately-held data in Official Statistics from the perspective of "quality". Starting from the notion of "quality" as defined in the context of European statistics, we show that the regular production of Official Statistics based on privately-held data entails a combination of potential quality benefits and quality costs. Statistical authorities should carefully assess and select use-cases and data sources for which the cost-vs-benefit balance is positive. In the context of the European Statistical System, we highlight the factors that motivate the development of a common methodological framework, open-source tools and share infrastructures at the European level.

Keywords: Official Statistics; quality; non-traditional data sources; privately-held data.

1. Quality in Official Statistics

In the context of Official Statistics the term "quality" has assumed a peculiar meaning, wide in scope and embracing multiple dimensions. For the European Statistical System (ESS), the quality dimensions are defined in the European Statistics Code of Practice[2] (CoP) and in the Quality Assurance Framework[3] (QAF). Taken together, these documents define the self-regulatory framework that distinguishes *Official Statistics* from other sources of statistical information like, *e.g.* commercial statistics and other public statistics.

The notion of quality as defined in COP/QAF spans three areas, namely *institutional environment, statistical processes* and *statistical output*. This approach is motivated by the consideration that the characteristics of the final statistical figures (*What statistics* are produced) depend on the underlying production process (*How* they are produced), and the latter in turn depends on the surrounding production environment (*Who* produces them). For each of these areas, the COP and QAF define principles and indicators. These documents are at the same time *prescriptive and aspirational*: they set minimum conditions to be fulfilled, but also high-level objectives to be pursued in a perspective of continuous improvement.

The notion of *quality* in Official Statistics is very articulated, as we have just seen, but at the same time *dynamic*: it is indeed a continuously evolving concept. This is not surprising when one considers that the whole system of *Official* Statistics keeps evolving, and is itself embedded in an ever-changing societal context. The evolution of quality in Official Statistics is reflected in the temporal flow of norms at different levels: the EU Regulation 223/2009 on European Statistics was adopted in 2009, amended first in 2015 and then revised again in March 2024; the CoP was published first in 2005 with a second and third version in 2011 and 2017, respectively; the QAF was published first in 2011 and revised in 2019. It is natural

1 Fabio Ricciato (fabio.ricciato@ec.europa.eu), European Commission, Eurostat. The views and opinions expressed are those of the author and do not necessarily reflect the official policy or position of the European Commission.

2 European Statistics Code of Practice - revised edition 2017.https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf.

3 Quality Assurance Frameworkof the ESS-version 2.0.2019. https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V2.0-final.pdf.

to expect that both CoP and QAF will be revised again following the adoption of the new regulation on European Statistics.

2. New data sources in Official Statistics: quality benefits and costs

Official Statistics are traditionally produced based on a combination of primary statistical data, from censuses and surveys, and administrative records. Such *traditional data* sources are becoming insufficient to meet the growing demands and expectations by statistical users for more, better, richer, and timelier statistics.The ESS is now preparing to extend (future) production processes to leverage also other types of *new* "non-traditional" data sources, including data generated and held in the private sector, *i.e.* Privately-Held Data (PHD).

It is important to remark that PHD, like other new data sources, are set to *augment, not replace* traditional data sources (Baldacci *et al.* 2021). As discussed below, a new generation of statistical products can be developed based on the integration of non-statistical PHD with statistical data, going beyond what would be possible with each of the two in isolation.

Looking through the glasses of *quality* as defined in CoP and QAF, the perspective of (re) using PHD for Official Statistics brings opportunities but also major challenges.

On the one hand, the motivation for considering PHD in the first place is rooted in the expectation that they will *increase the quality of statistical output*, by enabling the production of new, more, better, richer, and timelier statistics, mapping to the relevant CoP/QAF Principles11-14, namely *Relevance, Accuracy and Reliability, Timeless and Punctuality, Coherence and Comparability*. In the hypothetical scenario where the prospected new statistics (with comparable characteristics in terms of statistical output) were to be produced without recurring to PHD, the resulting cost and burden on respondents would be unbearable: in this sense leveraging PHD may be considered instrumental to *preserve the quality of the statistical processes* in terms of the CoP/QAF Principles 9-10, namely *Non-excessive Burden on Respondents* and Cost *effectiveness*.

On the other hand, bringing PHD into the statistical production requires finding new solutions to issues that touch almost all items in the CoP/QAF. Furthermore, some issues that are relevant for PHD go beyond the scope of the current CoP/QAF and require new extensions, for example in the direction of ensuring compliance with non-statistical legislation (*e.g.* telecom legislation for location data sourced from mobile networks or mobile phones); public acceptance for the secondary (re)use of highly granular personal data that were primarily collected for other purposes; sustainability of partnership models between statistical authorities and private dataholders; reproducibility, maintainability, explainability and sensitivity of softwarised statistical methods (Ricciato 2022). These are examples of issues and additional dimensions that may be expected to make their way into the future version of COP/QAF.

Taking a bird's-eye view over the quality opportunities vis-à-vis the quality challenges shows clearly that the (re)use of PHD in Official Statistics entails a cost-vs-benefit balance: the expected *quality benefits* associated to improved statistical output must be high enough to offset the *quality costs* incurred in fulfilling minimum conditions on almost all quality dimensions (*the gain must be worth the pain*).

It cannot be assumed that the cost-vs-benefit balance will be always positive: we must accept that in some cases the quality costs may prevail over the quality benefits. This simple consideration

should drive statistical authorities to be selective and assess very carefully the use-cases and data sources for which the anticipated quality benefits justify the prospective quality costs. In doing so, they should consider not only the initial costs required to *achieve* the stage of regular statistical production (*e.g.* research and experimentation; initial development of methodologies, software, and data interfaces; negotiation and establishment of partnership agreements with data holders; deployment of business processes) but also the operational costs to *sustain* regular statistical production (*e.g.* maintenance and continuous update of methodologies, software, and data interfaces; maintenance of partnerships and business processes).

Such cost-vs-benefit assessment, and the consequent selection of use-cases and data sources, should be conducted at different stages and start as early as possible in the statistical development path. Even a qualitative assessment should suffice to identify upfront those data sources for which the prospective quality costs, to achieve and/or sustain regular statistical production, are likely too high vis-à-vis the expected benefits. The technological maturity and market structure of the business sector where data are generated should be considered as key dimensions for the assessment. For instance, low levels of technological maturity and penetration, and high levels of market fragmentation and data heterogeneity, are elements that contribute to drive upwards several cost factors. Such analysis exercise should be preferably carried out in the perspective of statistical production at European scale, and should guide the allocation of resources and investments in methodological development and experimentation at the ESS level.

3. New data sources and new processes for new statistics

The secondary (re)use for statistical purposes of data generated primarily for nonstatistical purposes requires the establishment of new organisational processes. In many cases, the new processes will involve also the data holder(s) to some extent. The data holders must provide access to the data and to the associated meta-data, with modalities that need to be agreed with statistical authorities. But there is more information to provide: as statistical production extends to PHD, the (primary) data generation process enters the equation, therefore knowledge about the business and technological aspects that drive *how the data are produced, and therefore determine what information they carry*, becomes an essential component of the (secondary) reuse process in Official Statistics. Communicating such knowledge and information, called *para-data* in the ESS Handbook for quality and metadata reports[4], is required not only in the methodological development phase, to enable proper design and implementation of data processing methods, but also in the operational production stage. For example, anomalies, interruptions, and errors affecting the data generation process, hence the quality of the data that will eventually enter the statistical production pipeline, must be properly and promptly communicated by the data holder(s) to the statistical authority. Such dialogue will be unavoidably bi-directional: the statistical authority may detect unreported anomalies or implausible patterns, possibly based on comparison with other data sources, that are to be reported back to the data holder(s) to be correctly interpreted, possibly corrected or anyway mitigated. These examples imply that rules and roles on the side of both organisations, statistical authority and data holder(s), must be defined to ensure that events and incidents affecting statistical quality are properly detected and communicated. This requires the definition of agreed criteria (What information is relevant and should be communicated? What exactly should be considered "anomalous"?), functions and

4 ESS Handbook for quality and metadata reports – 2021 re-edition.https://ec.europa.eu/eurostat/documents/3859598/13925930/KS-GQ-21-021-EN-N.pdf.

policies (Who is in charge? Who shall communicate to whom?), interfaces and templates (How shall the communication take place?), etc. All these aspects must be encoded into a quality system designed specifically for PHD.

The deployment and execution of such processes will unavoidably consume resources and create additional burden on the side of the data holders as well as on the side of the statistical authorities. Receiving and processing information is not less resource consuming than preparing and transmitting it, and both the transmitter and the receiver share the common goal of minimising the amount of transferred information. Statistical authorities and data holders could cooperate to define communication processes that are not only effective but also efficient for each side, keeping the cost and burden down to the minimum possible level without jeopardising statistical quality.

4. Conditions for successful partnerships with data holders

The High-level Expert Group on facilitating the use of new data sources for Official Statistics states that the (re)use of PHD for Official Statistics should be based on fair and effective partnerships between businesses and statistical authorities, underpinned by a legal framework setting out clear requirements and safeguards for private data holders[5]. When the additional costs incurred by private data holders to enable data reuse for Official Statistics are substantial, they should receive financial compensations based on a fair reference cost model. Furthermore, on top of legislative and financial measures, the Expert Group recommends that statistical authorities put in place non-financial incentives to motivate data holders to cooperate with statistical authorities (for additional details see Eurostat 2022).

In the ideal scenario, the partnership model is designed in a way to let private data holders benefit not only from the act of cooperating with the statistical authorities (*e.g.* improved corporate reputation) but also from the statistical products (or by-products) deriving from such cooperation. If the data holders see their interest in that the final statistical product to which they contribute is of highest possible quality, then their cooperation efforts would go beyond the necessity to fulfill compliance requirements.

Translating this abstract goal (or wish) into concrete operational terms is admittedly very difficult, and in some cases devising a convincing system of incentives for the businesses will not be possible. However, statistical authorities should at least explore this direction and attempt to identify such set of incentives. The efforts may be successful in some business sectors. Hereafter we outline a possible approach for one specific kind of data, namely Mobile Network Operator (MNO) data, that may inspire similar reasoning for other business sectors.

Private businesses are already leveraging location data derived from the operation of mobile networks (MNO data) to deliver commercial statistics and "mobile analytics" services (terminology borrowed from Eurostat 2023). The success of this line of business should be seen positively by statistical authorities, as it sustains the business investments necessary to ensure MNO data availability in general (*e.g.* technical infrastructure, organisational processes), and specifically for reuse in Official Statistics. In other words, the use of MNO data for commercial analytics purposes is indirectly supportive of (rather than detrimental to) the prospective reuse

5 Empowering society by reusing privately-held data for Official Statistics – A European approach. Final Report of the Expert Group on facilitating the use of new data sources for Official Statistics, 2022. https://ec.europa.eu/eurostat/web/products-statistical-reports/-/ks-ft-22-004

of such data for Official Statistics. In the reverse direction, perhaps we can imagine a system where the public release of Official Statistics based on MNO data would not be detrimental but rather beneficial, at least indirectly, to the market demand for commercial analytics based on the same data. Such a hypothetical system would need to fulfill at least three necessary conditions.

First, Official Statistics based on MNO data should be sufficiently differentiated from commercial statistics. Businesses must be reassured that the publication by statistical authorities of Official Statistics based on MNO data will not cannibalise the market demand for mobile analytics offered on commercial terms. This is possible by differentiating the two lines of products along dimensions such as spatial and temporal granularity, timeliness, level of detail and variables, as argued already in Ricciato *et al*. (2018). Therein, the authors proposed to consider the possible analogies with the so-called "freemium" model that has proved successful in several business sectors, whereby making available to the public some "free" version of service or product does not reduce but rather increases the market demand for "premium" versions thereof. Along the same reasoning, the public release by statistical authorities of certain Official Statistics in aggregate form (*e.g.* average number of foreign tourists in mid-to-large towns during the previous quarter, delivered the next month) could potentially increase the appetite for more detailed analytics offered on commercial basis by businesses (*e.g.* the daily number of tourists and the number of nights spent in a particular town, disaggregated by the visitor's country of origin, delivered the next day). In other words, Official Statistics and commercial analytics would serve different purposes and would cover different segments of the information space.

Second, as elaborated by the ESS Task Force on use of Mobile Network Operator data for Official Statistics in their recent position paper (Eurostat 2023), Official Statistics based on MNO data must be based on a standardised and fully open *reference methodological framework*. Once developed and deployed operationally to serve statistical purposes, such methodological framework would then represent a natural standard also for the industry: producers of commercial analytics would have the opportunity to align (partly or fully) their basic definitions and methods to the standard ESS reference, and in this way increase transparency, comparability and credibility of their commercial figures towards their customers. This would not jeopardise their ability to compete among themselves in offering to potential customers on a commercial basis more advanced commercial analytics, *e.g.* based on proprietary improved methods and/or additional definitions. We tend to believe that also in this field, likewise other business sectors, a certain degree of standardisation is not detrimental but rather beneficial to market competition.

Third, within their information segment, Official Statistics produced by statistical authorities must deliver some added value going beyond what would be achievable by the commercial "mobile analytics" providers in the same area. The key added value may be increased accuracy through the integration of data from multiple MNOs and with statistical data. In fact, data from business companies typically refer to their specific customer bases that cannot be considered representative of the general population (as stated already in Baldacci *et al.* 2021). To counteract non-representativeness and bias (*e.g.* in population coverage or geographical coverage) of customer the base seen by each individual MNO, statistical authorities should aim to produce Official Statistics that integrate information sourced from multiple MNOs (an approach called "Multi-MNO orientation" in Eurostat 2023). Furthermore, they may integrate data from (multiple) MNOs with other kinds of non-MNO data, *e.g.* statistical data from ad-hoc sample surveys or censuses, to further improve stability and representativeness of the final figures. Such data integration can and must be done in full compliance with data protection legislation, possibly

leveraging advanced privacy-preserving technologies (Ricciato 2024), with no derogation to the established principle that statistical data cannot be used for non-statistical purposes, and without interfering with business competition dynamics among data providers (level-playing field). The final Official Statistics, produced and released publicly by statistical authorities, may then serve as reference for calibrating the commercial analytics developed independently by MNOs and their partner companies specialised in mobile analytic services (Eurostat 2023).

The perspective of combining data from multiple MNO with statistical data carries important strategic implications. In this scenario, statistical authorities would not be in the position of merely data *consumers* vis-à-vis the private data *providers*, but they would rather position themselves as *partners* of data holders, contributing with statistical data and methodologies to produce Official Statistics that are indirectly beneficial also for the contributing businesses. Moreover, they would further reassert the role of statistical data, in this case as a means to "fertilise" MNO data, and more in general the vast stock of so-called "big data", enabling the production of multi-source statistics that inherit the best of both worlds, namely the timeliness and richness of big data with the reliability and representativeness of statistical data.

5. The role of the European Statistical System

In the vision outlined insofar statistical authorities are required to address a number of important challenges along multiple dimension. The enterprise may overwhelm the resources, capacities and capabilities of any single statistical authority. The good news is that, for all these challenges, the terms of the problem are very similar if not identical for all European countries. Therefore, if a good solution can be found, it is almost certainly a good solution for all countries. We can identify two main reasons for this fortunate condition. First, the business and technological processes that generate PHD tend to be rather uniform across European countries (*e.g.* mobile network technologies do not change from one country to another). Second, contrary to administrative data that feature a certain degree of heterogeneity across different countries due to historical legacies in the development of national public administrations, dealing with PHD does not involve any historical legacy. Therefore, in the development of new methodologies for PHD, statistical authorities can enjoy the luxury of starting from scratch (clean-slate design).

These considerations reinforce the motivation for ESS members to join forces and pool resources to address these common challenges collectively at the European level. For each PHD source, the methodology and quality frameworks can be defined, developed and maintained at the ESS level (and then implemented at national level). For aspects where national peculiarities require some degree of customisation, the necessary flexibility can be incorporated into the design of a common methodological framework. To the extent that the methodologies need to be implemented in software tools, the latter can be developed open-source and maintained at the ESS level (and then used at the national level). If some complex infrastructure is required, it may be designed, built and operated at the European level as a shared ESS infrastructure, and then used on-demand by ESS members.

The coordinated adoption of common methodological standards, shared tools and infrastructures that are developed collaboratively by the ESS members at the European level is not in contradiction with the choice by the individual statistical authorities to implement them (possibly with some customisations) at the national level, and in this way remain in direct control of the production processes.

The benefits of addressing these challenges at the European level are manifolds. First, it obviously prevents duplication of costs and efforts. Second, the *ex ante* adoption of common definitions and detailed methods greatly reduces, and possibly eliminates altogether, the need for ex-post reconciliation and assessment of comparability of the final figures. Third, when personal data are involved, the adoption of a common European methodological framework opens the possibility of defining the necessary data protection measures, and specifically the supplementary technical and organisational measures required by GDPR Art. 89, directly at the European level, through a dialogue with the European Data Protection Supervisor (EDPS) and European Data Protection Board (EDPB). Furthermore, adopting a common European approach (as opposite to heterogeneous national approaches) to the quality challenges posed by PHD may trigger positive reinforcement mechanisms (network effects). The following analogy illustrates the point: when some open-source tool becomes popular and gets used and developed by many entities, it will more likely attract further users and developers, in a cycle of positive reinforcement that will eventually *reduce the cost* and at the same time *improve the quality* of the software. Analogously, the adoption of common and open methodologies, tools and shared infrastructures by the ESS members could trigger reinforcement mechanisms vis-à-vis other actors, including other public entities and business companies, resulting in greater adoption and promotion of the same methodologies, tools and infrastructures, with a positive return for the ESS.

References

Baldacci, E., F. Ricciato, and A. Wirthmann. 2021. "A Reflection on The Re(Use) of New Data Sources for Official Statistics". *Indice. Revista de Estadística y Sociedad*, N. 83. http://www.revistaindice.com/numero83/p8.pdf.

Eurostat, ESS Task Force on the use of Mobile Network Operator (MNO) data for Official Statistics. 2023. "Reusing mobile network operator data for official statistics: the case for a common methodological framework for the European Statistical System - 2023 edition". *Statistical Reports*. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/eurostat/web/products-statistical-reports/w/ks-ft-23-001.

Eurostat, Expert Group on facilitating the use of new data sources for official statistics. 2022. "Empowering society by reusing privately-held data for official statistics - A European approach". *Statistical Reports*. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/eurostat/web/products-statistical-reports/-/ksft-22-004.

Ricciato, F. 2024. "Steps Toward a Shared Infrastructure for Multi-Party Secure Private Computing in Official Statistics". *Journal of Official Statistics*, Volume 40, N. 1. https://doi.org/10.1177/0282423X241235259.

Ricciato, F. 2022. "A reflection on methodological sensitivity, quality and transparency in the processing of new "big" data sources". In *European Conference on Quality in Official Statistics* (*Q2022*). Vilnius, Lithuania, 8-10 June 2022. https://zenodo.org/records/10246419.

Ricciato, F., F. De Meersman, A. Wirthmann, G. Seynaeve, and M. Skaliotis. 2018. "Processing of Mobile Network Operator data for Official Statistics: the case for public-private partnerships". In *104th DGINS Conference*. Bucharest, Romania, 10-11 October 2018. https://zenodo.org/records/10246468.