A Reflection on The Re(Use) of New Data Sources for Official Statistics

Emanuele Baldacci
Eurostat, Director of Resources
Fabio Ricciato
Eurostat, Statistical Officer
Albrecht Wirthmann
Eurostat, Head of Unit Methodology; Innovation in official statistics

The essential mission of official statistics is to provide information about the society, the economy and the environment to citizens, policy makers, researchers and economic actors, enabling them to take decisions and form opinions based on facts and evidence. In a nutshell, official statistics provide the society with "knowledge about itself". Reliable, trustworthy, relevant and timely official statistics are among the enablers of a healthy democratic society.

Generally speaking, mandating the production of official statistics to specialised public bodies, i.e. statistical offices, impartial and independent from other public and private organizations, is a pre-requisite for ensuring trustworthiness of the statistical information. But what does the *production of official statistics* entails in practice?

In the pre-digital world, the only way to measure the society was "asking the people". In census and surveys, what to ask (i.e., questionnaires) and *whom to ask* (i.e., samples) were determined directly by the statistical offices and primarily for statistical purposes, with the aim of producing impartial and representative information at minimum cost. With survey and census data the entire statistical production process, from data collection through data processing and dissemination of the final statistics, is performed entirely by the statistical office.

Survey and census data were later complemented by administrative data, collected by other public bodies (central and local administrations) primarily for administrative purposes, and then re-used by statistical offices for statistical purposes. With administrative data, the data collection stage falls outside the statistical office but still remains within the public sector. Like survey and census data also administrative data are based on explicit declarations made by people (in the context of administrative procedures).

During the last two decades the digitalisation and *datafication* of our lives has changed many aspects of our behaviours. More and more of our actions, transactions and interactions are mediated by, directed to or at least sensed by digital systems and machines. Our online activities are natively embedded into the digital space, but also our physical activities leave digital traces via sensors and 'smart' devices. Each individual event in our lives – be it a purchase, a transaction, a movement, etc. – is now transformed into one or even more data points by some digital system(s). If 'micro-data' refers to the characteristics of an individual, the term 'nano-data' was proposed to refer to granular, behavioural data referring to individual events at sub-individual level. Private companies collect nano-data primarily for business purposes (e.g., for delivering services to their customers, or to gain more detailed knowledge about their needs and behaviours) and statistical offices are eager to reuse such data for producing more, better and timelier official statistics.

There is nowadays an increasing awareness that the data generated *by the citizens*, even if collected by private companies, have also a public value *for the citizens*. Such public value however is not being adequately ripped nowadays. Reusing such data for the production of better and timelier official statistics, that are then disseminated publicly and for free, is a way to "give back" part of the data value to the citizens themselves, to the extent that better statistics have the potential to lead to better policies and a better informed society.

The digitalisation of our lives has not only changed the way we produce data, but also the way we *consume* data and information. Citizens

and policymakers expect today to be informed in a more precise, comprehensive and timelier manner. They expect quantitative information (i.e., statistics) to describe an increasing number of phenomena. Official statistics must keep peace with such increased demand, and this is simply not possible by relying solely on traditional survey and administrative data. For statistical offices, plugging statistical production into the flow of machine-generated data is not only an *opportunity* driven by the availability of new data on the input side, but also a necessity in order to respond to the increased expectations and needs by the users of official statistics on the output side.

In most cases, new data sources will not replace entirely traditional ones but rather complement them. For example, some survey data may still be needed to calibrate and adjust the view offered by new data sources, e.g. to mitigate possible biases and incompleteness. In fact, the data from business companies typically refer to their customer base, that by definition a sub-population and cannot be considered representative of the whole target population, hence the need of making adjustments and calibration based on other auxiliary data. But a survey designed specifically to serve as auxiliary and complementary source for calibration purposes will be different (and likely less costly to implement) than a survey intended to serve as the sole primary data source. In other words, the opportunity to leverage new data sources involves a rethinking of the role of traditional data sources, not their complete dismissal.

Bringing new data sources into the process of statistical production requires solving a number of challenges and open issues. For some of them, the solutions will be different from those adopted in the context of traditional data sources.

The first issue concerns the legal and business aspects of data reuse. In most cases such new data are generated in the private sector. Whether such data represent a by-product or a central component of the business operations, they clearly have both a *private value for the business and a public value for the whole society*, and we need rules to govern their use in both domains. The European Commission is currently working towards a new legal framework as part of the European Digital Strategy. The reuse of private data for public purposes, or business-to-government (B2G) data sharing, will be focus of the forthcoming Data Act proposal that is currently under preparation by the European Commission[1]. For what concerns more specifically official statistics, Eurostat has recently taken the initiative to start a targeted revision process of Regulation 223/2009 on European Statistics with a view to adapt it to the new needs of official statistics. One prominent goal of such a revision is to enable the sustainable reuse of privately held data for the development, production and dissemination of official statistics in Europe. The ambition of Eurostat is to finalize the new Regulation by end of 2022.

The reuse of private data for public purposes, or business-to-government (B2G) data sharing, will be focus of the forthcoming Data Act proposal that is currently under preparation by the European Commission

From an operational point of view, the legal aspects of data reuse must be articulated with the technical aspects, i.e., translated into operational modalities defining how data held by one organization are accessed (i.e., reused) by another organization, and specifically by statistical offices. In many cases of practical interest, it will not be convenient, if at all possible, to move enormous amount of nano-data (or granular data) from the data holders to the statistical offices. In such cases, the input nano-data could be processed and aggregated, at least partially, at the premises of their respective data holders, with only the intermediate or

1 See the ESS position paper https://europa.eu/!KyUK9C and the ESGAB opinion concerning the forthcoming Data Act proposal https://europa.eu/!QQMbf4

final results passed to the statistical office. This approach reflects the idea that in statistical production *raw data are a means, not a goal*. Instead of moving raw data into the statistical office, it is often preferable to move methods and algorithms towards the point were data are generated. The statistical office will receive a lower volume of (aggregate) of data but with higher information content. This approach helps to minimise the risks.

Methodologies for new data sources need to be co-developed by multi-disciplinary teams where expert statisticians collaborate with specialists from the specific technological or business field relevant to the concerned data

The methods and algorithms by which the input data are processed, filtered and aggregated, including the part thereof executed automatically at the premises of the private data holders, should be fully open and standardised, so as to preserve methodological transparency and harmonisation across different data providers. Generally speaking, transforming machine-generated data into reliable statistics requires dealing with several sources of uncertainty (errors, bias, limits to temporal and spatial resolution, unknowns, etc.) that are specific to the system or technology from where they are generated, and do not always have an exact homologous in traditional human-generated data sources. For this reason, methodologies for new data sources need to be co-developed by multi-disciplinary teams where expert statisticians collaborate with specialists from the specific technological or business field relevant to the concerned data (e.g., telecom engineers for mobile network operator data). Compared to data that were designed primarily for statistics, the methodologies for processing reused data sources tend to be more complex. They need to cope with configuration-specific aspects of data generation that differ somewhat across different sources (e.g., certain characteristics of the location data generated by the mobile network depends on the configuration of the network infrastructure and on the characteristics of the mobile customer base that vary across operators). They must be adapted more frequently in order to adapt to changes in the data generation mechanisms that follow the physiological evolution of technologies and markets (e.g., improvement of mobile network technology and changes of mobile usage due to different tariff schemes). All these aspects contribute to increase the cost and the complexity of methodological development for the reuse of new data sources in official statistics. In other words, going from survey and census data towards new machine-generated data sources, the dominant cost component of statistical production shifts from data collection to methodological development.

Also for this reason, the European dimension becomes extremely important for the development of new statistical methodologies. Pooling methodological development at the European level is more *efficient* in terms of resources, skills and capacities, as one methodology developed jointly by the ESS at the European level can be applied (possibly with customisations and limited adaptations) to all Member States, but it is also more *effective* in terms of better harmonisation and consistency of results. This is possible also because machine-generated data tend to display less national specificities that, say, administrative data, since the underlying technologies tend to be uniform across countries (e.g., smartphones and mobile networks standards are global, not country-specific). In many cases, they are generated by systems and platforms with an intrinsic multinational dimension (e.g., online platforms, credit card payment systems). In some business sectors, the level of data heterogeneity (both in format and semantic) across different data providers, also *within* the same country, prevails over difference *across* countries. That means, all countries would have to face basically the very same methodological and technical challenges, reinforcing the convenience of pooling methodological development at the European level. There may be still national

specificities to be taken into account though, e.g. on behavioural aspects (e.g., the number of mobile phones used by each individual may vary across different countries) but such differences can still be addressed in the context of a single methodological process conducted at the European level by leaving room to flexibility and customisations.

From the discussion above it should be clear that reusing new data sources for official statistics has a cost. There are costs in developing new methodologies, there are costs in deploying IT infrastructures that implement these methodologies and there are organisational costs in establishing and implementing agreements between the statistical officers and data holders. Regardless of how they are distributed across the involved actors, these costs must be assessed against the benefits that more comprehensive, detailed and timely statistics will bring to the society. Such cost-vs-benefit analysis must be conducted for each kind (or class) of new data sources and periodically reassessed. As a matter of fact, some kinds of data sources are not yet sufficiently mature to be used reliably as a source of official statistics, or anyway involve costs that would be repaid by a proportionate benefit, but they may become appealing in the future. In other words, statistical system must be able to select pragmatically the subset of data sources for which *the gain is worth the pain*. Data sources generated by early stage technologies that are not yet sufficiently widespread would likely bring little benefit. Data sources generated by systems characterised by large degrees of heterogeneity and technological dispersion (many proprietary systems instead of few common standards) would likely involve too high costs. The most appealing data sources for official statistics are generated in mature sectors, with high penetration and low technological heterogeneity and the potential for reuseability in different statistical domains.

In conclusion, this short article has attempted to provide a coherent "big picture" of the issues that the European Statistical System is facing in order to move forward on the reuse of new data sources for official statistics. We must move in parallel on multiple fronts notwithstanding their interdependencies. Achieving an appropriate legal framework for data reuse is the most compelling *conditio sine qua* non future statistical production can benefit from certain kinds of new data sources. We as statisticians need to (co-) develop new methodologies for data that are very different from the ones we are familiar with, and in so doing we must seek the collaboration by experts from other domains. On each of these fronts, joining efforts at the European level will be key to success. ●