Economic and Social Council

Economic Commission for Europe

Conference of European Statisticians

67th plenary session
Paris, 26-28 June 2019
Item 6 (a) of the provisional agenda
Emerging role of national statistical offices as offices for statistics and data
Session 1: Emerging data system opportunities and issues

Towards certification of new statistics of public interest: why, how?

Note from the National Institute of Statistics and Economic Studies (France)

Summary

>   This document presents the analysis and proposal to qualify new statistics of public interest in France. It describes the current situation in France regarding the qualification of statistics produced outside the National Statistical System, either public or private, and the issues and challenges related to their emergence as these statistics may complement but also sometimes challenge official statistics. The authors propose a three-tier mechanism, at the first level limited to the self-documentation of a statistic according to a predefined template, at the second level with the approval of such documentation, and at the third level consisting in the approval of the documentation as well as the methodology and outputs.

>   This document is presented to the 2019 Conference of European Statisticians seminar on "Emerging role of national statistical offices as offices for statistics and data", session 1 "Emerging data system opportunities and issues" for discussion.

I. Introduction

1.   In France, public statistics cover statistical outputs of statistical surveys or administrative sources, produced not only by the national statistical system (NSS) but also by public institutions or private bodies entrusted of public service missions. It is therefore potentially a very broad set of statistics, which constitute a public good, made available to all potential users.

2.   The context of open data and digitalization of the economy increases the scope of public or private new data sets and statistics. The production of data or statistics has thus become generalized to new actors who have their own data sources through their activity or who reuse public sources through the growing supply of open (raw) data. These new statistics can intersect the field of public statistics, providing additional insights or inducing noise if the messages are different, and also cover new fields of observation.

3.   This phenomenon is not new: for example, there have always been surveys, carried out by professional bodies, that have produced and published statistics widely reported in the media. What has changed with the digital revolution is the generalized capacity of actors of private and civil society sector to produce masses of data and to extract new statistics of general interest.

4.   Insee, in its strategy for 2025, has identified a prospective reflection aiming at "placing the public statistical service as data certifier". In a context where the phenomenon of "fake news" is becoming increasingly important, this approach seems all the more legitimate. How to develop a qualification function of these new statistics in order to give more visibility to those that bring added value? What would be the relevant field to cover? What are the possible methods for implementing this new function?

5.   To answer these questions, a mission of the Insee's General Inspectorate was requested, whose main findings are presented in this paper.

II. Overview of the situation in France

6.   The National Statistical System (NSS) has already taken up the need to qualify the statistics that underlie the public debate. Part of the statistical outputs of organizations entrusted with a public service mission is thus "labelled" by the Statistics Authority. This procedure ensures the respect of the principles of the European statistics Code of practice: professional independence in the design, production and dissemination of public statistics as well as the principles of objectivity, impartiality, relevance and quality of the indicators produced.

7.   Other statistics, produced by private actors or civil society organizations, are "calibrated" according to the term adopted by the National Council of Statistical Information (CNIS). However, this procedure is quite cumbersome: it includes not only the assessment of the documentation but also the validation of the methodology. As a result, the number of "calibrated" statistics remains very low compared to the development of data produced by the private sector or civil society organizations, while some of them receive a significant echo in the media or in public debates.

8.   The statistical system has already seized some of the opportunities offered by new public or private sources, rather than ignoring them and sticking to "business as usual" which would have entailed a clear risk of marginalization, as reputation and trust in the statistical system cannot replace the quest for innovation (Ricciatio et al., 2018). Many National Statistical Institutes (NSIs) have developed plans for using big data as inputs to their statistical production. Insee established an innovation lab (the SSPLab) in May 2018, as a resource and animation centre for applied research and experimental development to promote innovation and use of new data sources, technologies and data science methods. The SSPLab participates in the work of the Eurostat Task Force on Big Data. For example, the SSPLab has experimented the use of mobile phone data to produce population estimates.

9.   Thus, national statistical systems reacted quickly to invest the field of big data and develop the use of it to produce official statistics. However, it cannot be ruled out that purely

private statistics derived from new technologies challenge official statistical production. More positively, private statistics could also complement official statistics, the issue being then to acknowledge them as worthy information among other sources.

10. For example, some private companies produce report on employment as a by-product of their activity of management of payrolls for private companies or use web-scrapping techniques to collect press clipping on layoff or hiring plans or local investments. Their business model underlying this activity is not to commercialize aggregated statistics but to seek visibility and credibility from the media in order to sell commercial products, such as detailed information for private clients.

11. However, the main producers of non-official statistics belong to the public sector: various observatories financed by public funds or local authorities' funds, for example regions or large cities. They produce statistics that can complement public statistics on a specific domain (either sectoral or geographical).

III. Issues and challenges

12. The first issue is to provide relevant information to the public by accompanying statistics produced with corresponding metadata. The first and most fundamental purpose of statistical metadata is to help users to interpret, understand and analyse statistics. The level of detail of metadata depends on the needs and skills of the users. For example, a journalist may not have the time to digest a large volume of detailed metadata and will need minimal information to avoid misinterpretation. Conversely, a social scientist may even want to question the assumptions made by the original producer of statistics and deduce new statistical results based on alternative hypotheses. The latter type of user should have access to all relevant assumptions and other circumstances in the collection, preparation and estimation of process data, as designed and operated by the statistical producer (Dippo, Sundgren, 2000).

13. Beyond metadata, what matters is the quality of data produced. The European Statistics Code of Practice (ESCoP) governs the production of the public statistical system in France and in the Member States. It ensures that statistics produced are based on the principles of quality: professional independence, relevance, quality of processes and products and are accompanied by the corresponding metadata.

14. Statistics produced outside the National Statistical System do not always provide the necessary metadata nor do they refer to quality commitments. Private companies present some elements of methodology description, but often incomplete. For example, the field covered, or possible fluctuations in the collection of information that may affect temporal comparability, are often missing. Even the concept, and the way to measure it, may be incomplete, leading to misinterpretation.

15. Statistical literacy is therefore of paramount importance for users to be able to understand how statistics are produced and what is their quality. For example, Radermacher (2018) argues that statisticians should pro-actively invest in developing "datacy" capabilities in society in order to educate citizens to make better use of statistics, official or otherwise.

IV. Proposed mechanism

16. In a context where the phenomenon of "fake news" is becoming increasingly important, journalists are confronted on a daily basis with the need to check the information available to them. In the French media, as elsewhere in the world, more and more newsrooms have now columnists dedicated to fact-checking. The basis of journalism is to cross-check sources and control their reliability (Léchenet, 2015). So, journalists prefer to conduct their research alone from existing sources rather than being "guided" by one of the stakeholders. Thus, NSIs must be careful in "qualifying" external sources, with the risk of being accused to protect their monopoly using a "name and shame" mechanism.

17. Therefore, the proposed mechanism is a three-tier process of qualification of statistics with a growing requirement gradient. The first level should require self-documenting the

statistical data according to a standardized template to be defined. The second qualification level should consist in the approval ("registration") of the documentation, on a voluntary basis. This does not mean approval of the methodology or outputs, just that the documentation is correctly filled out in order to provide the necessary information to users. The third level should imply an assessment of the quality and relevance of the output.

18.   This procedure would concern primarily producers of statistics entrusted with public service missions, potentially far more likely to adhere to such a scheme than private producers. The assumption is that this could play a positive ripple effect, even in the private sphere.

19.   The CNIS would first decide on the admissibility of the application for assessing the documentation of a statistical output, according to its relevance regarding public information. Following this notice of admissibility, the assessment of the documentation would be entrusted to a technical committee. This committee would include experts from inside and outside the National Statistical System. This instruction involves a variable cost, borne mainly by the committee: the approval of the documentation of a statistic may require back and forth with the applicant. To reduce this cost, it is necessary to clarify the standard and clarify the conditions of its use. The opinion of the committee would relate only to the quality of the documentation, without endorsement of the quality of the product.
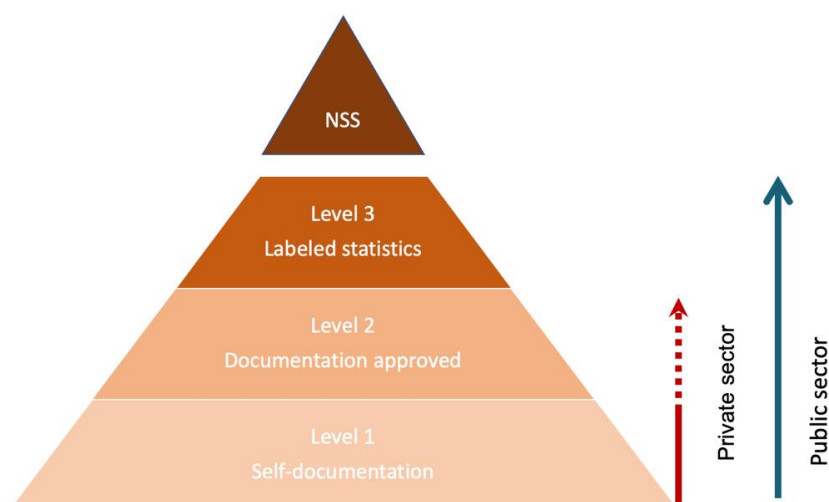
20.   This mechanism implies the definition of a template for documenting a statistic. To set the standard, it is proposed to establish a working group composed of official statisticians and experts outside the NSS, as well as potential producers from private and public sector, with the objective that the template provides the relevant information to be delivered to the users.

21.   The approval of the documentation would be granted for a limited, renewable period, and an online repository would describe the list of approved statistics.

22.   The third level of qualification corresponds to the labelling of public statistics: it would therefore be reserved for statistics produced by organizations fulfilling public service missions. It currently involves an assessment of the relevance of the output and the methodology used, as well as the verification of the compliance with the main principles of the European Statistics Code of Practice. As Insee and the Statistical departments of ministries, which form the NSS, are in relation with many partners producing public statistics, members of the NSS could play a dynamic role to accompany the labelling of sources of their partners. This would add a bottom-up approach, which could lead to extending the scope of labelled statistics.

Figure 1
Three-tier mechanism for qualification of statistical sources

23.   However, the approval procedure would remain inoperative if potential candidates are not aware of it and if users do not differentiate between different levels of quality. The pyramid made up of increasing levels of quality requirements, with the NSS at the top and its commitment to comply with the European Statistics Code of Practice must be presented in a visual and educational way.

24.   Finally, one of the ways to control or limit the audience of non official statistics is to communicate more convincingly on our commitments and achievement in terms of quality. It is therefore recommended to develop a more user-oriented communication on statistical quality, in particular on the Insee website. For example, the United Kingdom Statistics Authority has adopted user-oriented communication by organizing its code of practice for statistics into three pillars (trustworthiness, quality and value) so that users are clearly aware of it.

V.   Conclusions and recommendations

The CES seminar could exchange on the various national situations and on the opportunity to have a common approach to the qualification or certification of non-official statistics.

VI.   References

Dippo, C, Sudgren, B. (2000). " The Role of Metadata in Statistics", paper presented at the *International Conference on Establishment Surveys II*, Buffalo, New York. https://www.bls.gov/osmr/abstract/st/st000040.htm

Léchenet, A. (2015). Le journalisme de données. Synthèse des débats. Cafés de la statistique, soirée du 14 avril 2015. Société Française de statistique.

Radermacher, W. (2018). Official statistics in the era of big data opportunities and threats. International Journal of Data Science and Analytics. 6:225-231. Springer International Publishing AG.

Ricciato, R, et alii. (2018). Towards a reference architecture for trusted smart statistics. Eurostat Task Force on Big Data. Papier présenté à la Conférence européenne sur la qualité dans les statistiques officielles 2018,

———————————