United Nations ECE/CES/2019/39

Economic and Social Council

Distr.: General
15 May 2019

English only

Economic Commission for Europe

Conference of European Statisticians

67th plenary session
Paris, 26-28 June 2019
Item 2 (a) of the provisional agenda
New data sources – accessibility and use
Session 1: Accessing new data sources

Statistical quality and new data sources – lessons from the work of Organisation for Economic Cooperation and Development

Note by the Organisation for Economic Cooperation and Development

Summary

Organisation for Economic Cooperation and Development (OECD) in its data, analytical and quantitative work has increasingly made use of new sources for evidence – according to the latest count there have been nearly a hundred OECD projects of this kind over the recent past. Along with many opportunities, evidence based on new data sources also entails challenges by way of co-ordination of projects, exchange of knowledge, development of skills and not least adherence to quality standards. In many ways, OECD is a microcosm of what happens at the national level with similar opportunities and challenges, albeit on an international scale. This paper presents the main lessons that the Organisation has learned so far from its work to develop evidence from new data sources while respecting the need to ensure that its data and statistics remain of high quality and continue to command public trust.

This document is presented to the 2019 Conference of European Statisticians seminar on "New data sources – accessibility and use", session 1 "Accessing new data sources" for discussion.

GE.19-07912(E)
1907912

I. Background

1. The digitalisation of economies and societies has generated an explosion of new data, coupled with increasing computing capacity to exploit them. Many of these data have potential to generate new, timelier and more granular evidence and information for citizens, analysts and policymakers.

2. The international statistical community has actively discussed the challenges brought about by the digital era, including the emerging new institutional roles for national statistical offices (NSOs) in "ecosystems of data". Fora include the United Nations Statistical Commission (UNSC), the United Nations Global Working Group on Big Data for Official Statistics, the United Nations Expert Group on National Quality Assurance Frameworks, the ESSNet Big Data project, the OECD Committee on Statistics and Statistical Policy, and the Conference of European Statisticians. Significant steps have also been taken to cooperate and share resources internationally, in particular through the High-Level Group for the Modernisation of Official Statistics (HLG-MOS), the SDMX Initiative and the Statistical Information Systems Collaboration Community (SIS-CC) around the OECD's *DotStat Suite*. Each of these and other initiatives cover certain institutional, technical or data-related aspects of producing official statistics. In parallel, many initiatives have been taken at the national level.

3. While new sources of information can open the possibility of new streams of statistics in a range of areas, many of them also entail a fundamental change in the nature of statistical operations. Historically, NSOs have built most of their products in vertically organised, fully controlled lines of production based on surveys or censuses of households or firms and reporting from other institutions such as ministries, agencies or local governments. Other, secondary sources have also been present (examples include scanner data for price measurement or credit card information used in balance of payment statistics), but these were the exception rather than the rule.

4. With digitalisation, the balance is shifting towards increasing use of secondary sources and exploring new ways to interpret data through data science techniques and artificial intelligence (e.g. high-resolution satellite imagery, pattern recognition techniques).

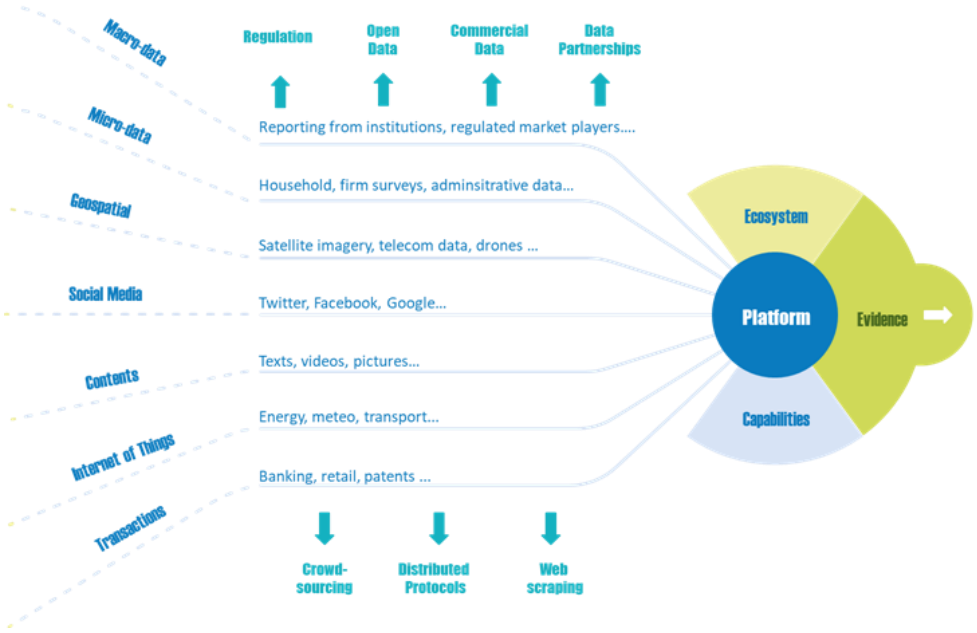
5. The new environment poses complex new questions about access to and use of data, at national and international level. For example, OECD analysts and statisticians have, over the past few years, made increasing use of non-conventional data sources in their work – indeed, about 100 such projects have been undertaken. As part of the OECD Smart Data Strategy, OECD is now reviewing the questions that have emerged around data access and data quality control and is in the process of developing a new data quality framework for the Organisation that will make explicit and detailed reference to new data sources. While this work is still at an early stage, we are already in a position to draw some important lessons about access to new data sources, how they blend in with more traditional activities, and how quality assurance of newly derived statistics and indicators can be tackled. The next section puts forward seven such lessons. The annex conveys the main features of the new statistical quality framework of OECD.

II. Seven lessons for new quality frameworks – so far

A. No one-size-fits all for new sources

6. While it is convenient to talk about "new sources" for evidence, these are highly heterogeneous in nature and so are the associated opportunities, challenges and required skill sets (figure 1). As a consequence, and notwithstanding some important general principles, there is no one-size-fits-all statistical policy or quality framework with regard to new sources. Quality criteria and good practices have to be tailored to each of the new sources.

Figure 1
New sources are increasingly heterogeneous



B. Quality assurance starts at the design phase of statistical production

7. The existing OECD Data Quality Framework has largely been conceived around collection, processing and dissemination of data from official sources, typically through statistical reporting. Here OECD will generally accept its members' data as submitted, since the members themselves are responsible through their own processes for its quality. With new, secondary sources much more emphasis needs to be put on quality assurance during the design and collection phases of statistical production. Quality assurance has also become more complex for data processing when multi-source statistical products are developed, for instance by combining geospatial data with population censuses. Although this point applies in particular to OECD and other international data operations, some aspects are equally relevant for NSOs. This concerns for instance machine learning (ML) as an emerging tool to generate and improve data. As Yung et al (2018) note: "*It is too early to develop a full quality framework for ML, but quality issues should be addressed. Traditional statistical quality frameworks assume that the data-generating process and further data processing steps are explicitly known. When applying ML methods, especially to 'found' big data or in multisource statistics, these assumptions are usually not valid. To guarantee quality, reproducibility and transparency, which are core values of official statistics, it is important to identify suitable quality indicators and performance metrics. The work on quality issues could either be a stand-alone effort or an extension of an existing quality framework.*"

C. Security, ethical and legal questions have moved centre-stage

8. Security, ethical and legal risks have moved centre-stage in quality management. Independently of new sources, digital security has gained tremendous importance and part of quality assurance now needs to be directed at assuring data safety and preventing any infringement of data integrity, for instance by defining and following protocols for treatment and storage of data. With the use of secondary sources, many more questions of a legal nature have also arisen; for example, whether and how websites can be scraped or how data collected from individuals can be used.

D. Bottom-up must meet top-down

9. Bottom-up must meet top-down: again, as a consequence of vastly different types of new sources, meaningful guidance for quality assurance has to be developed in close cooperation between area specialists and those who manage quality assurance processes and rules – otherwise important substantive or technical aspects may be left out or irrelevant criteria put forward. For instance, there may be established standards among specialists for coding geospatial data or accepted practices how to deal with websites whose contents is scraped. Closely related are new forms of statistical governance at OECD whereby specialists who draw on the same statistical source (e.g. geospatial data) or who use the same tools (e.g. *R* as analytical-statistical software) cooperate in designated communities of practice to exchange experiences and develop new practices, independently of where these staff members are working in the Organisation.

E. Towards "quality by design"

10. IT tools can help ensure "quality by design". Using shared, trusted, high-performance IT tools in statistical production can contribute to quality assurance. A case in point is the OECD's Algorithm Bank. As a virtual, easily accessible space, it supports project management, efficiency and quality by: (i) coding only once so that widely-used functions are readily available in a standard layout that can be directly called from various applications; (ii) more robust procedures through bug tracking, continuous testing and sharing between more users; (iii) harmonised methodology and validation against the OECD standard methodology.

F. Setting up a dynamic process

11. Quality assurance must be dynamic, and the framework and its good practices will need on-going review and development to reflect the fast-moving nature of digitalisation. When designing data quality frameworks in an increasingly complex environment, it is also helpful to identify the most important business risks. This helps develop general principles and focus the quality assurance effort on the biggest risks.

G. Defining a skills strategy to support quality assurance

12. The use of new data sources and their quality assurance requires specific skills. The OECD experience in matching staff skills with quickly evolving requirements is raising major human resource questions that must be approached strategically and in close co-operation with Human Resource Management, including:

- Is there a need for a new job family of "data scientists"?
- Which hiring strategies are needed to attract qualified staff?
- How can we develop cost-effective training plans, for instance through new forms of e-learning and partnering with private suppliers of training?

Annex I

I. The emerging revised OECD data quality framework

A. Reiterating core values…

1. For over a decade, OECD has relied on its Data Quality Framework (revised in 2011) as a reference for assessments of the OECD statistical products. Underlying this framework is a set of core values derived from the Fundamental Principles of Official Statistics adopted initially by the United Nations Economic Commission for Europe, and subsequently endorsed by the United Nations Statistical Commission in April 1994, and further developed in the 2015 OECD Council Recommendation on Good Statistical Practice. In addition, OECD has endorsed the Principles Governing International Statistical Activities, and the OECD statisticians are expected to commit to carrying out their work according to the International Statistical Institute declaration on professional ethics. The core values put forward in these references remain fully relevant to the OECD statistics and data work.

B. …and building on established quality attributes…

2. While well-established quality criteria (relevance, accuracy, reliability, timeliness, punctuality, accessibility, clarity and interpretability, coherence and comparability) have stood the test of time and remain relevant, some of them have become even more important than they used to be: in a "post-truth society", this includes accuracy and reliability along with timeliness, punctuality, coherence and comparability. As mentioned earlier, relevant legal and security dimensions must also be considered. The revised quality framework is now looking at i) intrinsic data quality, ii) timeliness, iii) accessibility, iv) interpretability and v) security.

C. …to provide good practices…

3. In a very pragmatic way, seven types of new data sources can be distinguished from the current OECD experience:

- Microdata from commercial sources: in particular enterprise-related (ORBIS, AMADEUS, BANKSCOPE, etc.) without confidentiality restrictions.
- Administrative and survey microdata: data on individual firms or persons from official sources, including those derived from administrative records, households' or firms' surveys, with confidentiality restrictions.
- Data obtained by using automated web scraping techniques or bots enabling large scale automatic data collection from websites and transforming them into useful data through various curation techniques, including text mining.
- Data for geospatial analysis, making use of satellite imagery, but also multiple other data sources with geolocalised information: crowdsourced, geolocalised information (such as OpenStreetMap), or geolocalised digital services data (such as mobile operators).
- Data from social media and, more generally, generated by GAFAs and large digital platforms (sentiment analysis based on tweets or Google trends, longitudinal analysis based on LinkedIn data, Facebook-managed surveys on specific populations, etc.).
- Transaction data (electronically registered and typically privately-owned data such as data from electronic reservation systems, credit card transactions, retail transactions, or electricity consumption).
- Data derived from crowdsourcing, polling and web surveys: data on a task or project collected by enlisting the services of a large number of individuals, via the Internet.

4. Two well established types of data sources remain fully relevant in the OECD work (and can be further broken down if needed). These are:

- Official data transmitted by national statistical organisations (NSOs and other members of national statistical systems)
- Qualitative ("policy") information collected through surveys by OECD from its members.

D. … across the various stages of statistics and data production

5. A key reference for the structure of the OECD Quality Framework is the Generic Statistical Business Process Model (GSBPM) that describes and defines the set of business processes needed to produce statistics in NSIs and International Organisations. GSBPM can also be used for integrating data and metadata standards, as a template for process documentation, for harmonizing statistical computing infrastructures, and to provide a framework for process quality assessment and improvement. (GSBPM V5.1, Introduction).

6. GSBPM distinguishes eight phases of statistical production: i) specify needs, ii) design, iii) build, iv) collect, v) process, vi) analyse, vii) disseminate, viii) evaluate. While GSBPM constitutes an accepted standard, it can be implemented flexibly and for the OECD internal purposes, a simplified version is proposed:

- Design
- Collect
- Process and analyse
- Disseminate and archive
- Evaluate.

7. A *q-matrix* (quality matrix) that traverses the five stages of statistical production (design, collect, process and analyse, disseminate and archive) with five quality objectives (intrinsic data quality, timeliness, accessibility, interpretability, security) provides the basic structure.

8. This structure will:

(a) Provide guidance to tools, references and indications of good practice that form a "checklist" for the data developer and statisticians. Guidance would typically be provided by type of data source and should be as concrete and specific as possible, addressing the most important risks and avoiding general statements, and it should be embedded in tools as much as possible to facilitate enforcement ("quality by design");

(b) Structure the contents of quality assessments of statistics and data. Such assessments would typically relate to particular statistics, data outputs or output areas.

9. "Sources" comprise both established sources – in particular official data that is transmitted to OECD from NSOs or other institutions that form part of countries' statistical system – and new sources.

10. The full quality framework is set up as a living and linked document that can be completed gradually and be modified as necessary.

Figure 2
Good practices are provided across stages of production