United Nations

ECE/CES/2019/26

Economic and Social Council

Distr.: General
12 April 2019

Original: English

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

The same role but changing function? The consequences of datafication for official statistics

Note by Eurostat

Summary

The document discusses the changing conditions of data access, processing and dissemination, and their impact on the functions of the statistical system. The authors assume that the mission of official statistics still holds, but the means and the conditions in which this mission is achieved will change. This will require a systemic approach where statistical offices will have to rethink their production and dissemination model.

The authors assume that statistical offices will not be able to centralize all data and production of statistics. The tasks of statistical offices could thus be enhanced and extended to providing guidance on the use, processing and communication of statistics. Statistical offices could take a leading role within government as a custodian of data used for statistical purposes. The paper also discusses the idea of certification of data, processes and outputs and the consequences for statistical offices. The increasing demand for high quality statistics satisfying specific needs and communicated just in time necessitates a pro-active and flexible attitude to inform different groups in society in public debates.

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. Technical developments in information and communication technologies initiated a global process of digitalisation of the entire society since the advent of the internet during the last decade of the previous century. Key enabling factors were the spread out of the internet and the world wide web as general information and application platform. With the development of mobile handsets, access to the internet was no longer limited to PCs having wired connections but enabled use of web services at any place at any time for any person.

2. Miniaturising these devices and putting more intelligence into them made them to multipurpose devices that accompany people 24 hours a day, seven days a week. At the same time, services and platforms were developed, which replaced physical activities or objects by their digital correspondent. Networking effects lead to a tendency of developing monopolies for specific applications, like searching or networking. Recently, more and more machines and appliances are equipped with information technologies (IT) to make them smarter. These “smart devices” are able to communicate with each other creating the so-called Internet of Things. Artificial intelligence is increasingly embedded into these devices or into systems that are created by the network of devices to increase their smartness. The final aim of this development is to develop a service to the users of the smart devices or smart systems instead of a good, e.g. instead of selling a car, companies are providing mobility.

II. Datafication as new paradigm

3. Digitalisation is the condition for datafication that transforms every activity and state into data. These data are collected, stored, processed, analysed and exchanged. The scales of volume, dimensionality, frequency, density and variety of these data are many orders of magnitude higher than data collections in the pre-digital era. In fact, expensive data collections of the past have turned into data as by-product of the datafication. Collecting data and providing information is not any more a quasi-monopolist activity of specific organisations, such as statistical offices, but can be performed by any organisation having the necessary skills to turn data into information.

III. Changing user habits and expectations

4. In addition to data as exhaust product, internet platforms collect data as part of their business model. In exchange for providing services or other incentives, users are providing personal or non-personal information that are re-used to develop new services but are also used for creating revenues, e.g. through advertising, to provide platform services for free to citizens. In addition to direct services to the users, internet platforms and applications employ additional incentives to engage users in a service. This could be gamification, where typical elements of game playing, such as point scoring or competition with other users are offered. Community award schemas is another example, where the creation of communities for achieving a common goal is encouraged, which rewards individuals and the community as a whole. An example could be encouraging energy saving activities within a neighbourhood, which could in addition be combined with gamification approaches. Public recognition could be another incentive to engage in a digital service and provide data.

5. Financial or pseudo-financial incentives are also employed to receive data from citizens. An example for pseudo-financial incentives is consumer cards award points, which can be used to buy goods. Sometimes, it is sufficient to promise the chance of winning prizes to convince citizens providing personal data. Citizens are at the same time consumer of services and producers of data. They are getting increasingly aware of these mechanisms and trade data for personal services or other incentives. Attitudes of citizens put in similar situations in different contexts, such as providing data for statistical offices, the public good, will change due to the described mechanisms, which they experience on a daily basis.

6. Personalisation is an important factor when providing services and encourage citizens to give personal data. Taking the example of providing data from personal health trackers to the vendor platform, the collection and provision of data are mostly done in the background,

via automatic data transfer. As compensation, users are receiving information on their health status, getting personalised behavioural advice and can compare themselves with other users or get the possibility of creating groups, e.g. for doing exercises together.

7. The widespread availability of information and services has changed user expectations. Information and services should be accessible just in time, 24 hours a day, seven days a week. Via search engines and online encyclopaedias, information is accessible at the time of demand. It is possible to buy any goods at any time of the day with delivery shortly after payment. The volume of information is adjusted to the demand. Its delivery is more or less immediate.

8. Automatic processing of data provides the possibility of almost real time information. The collected data can be processed, analysed and aggregated in very short time. The time lag between data collection and information provision can be drastically diminished; in some cases, real time information can be offered. E.g. in a smart energy grid, production and consumption of energy may be tracked in real time; traffic movements are collected and processed to provide information on traffic jams in real time and to make recommendations for optimized navigation.

9. The internet does not know any authority that certifies correctness of information, or the quality of a good or a service. In order to create trust, bottom-up mechanisms have been created. Following the characteristics of the internet as network, service providers try to increase trustworthiness of information by linking with others nodes in the network. The number of links to an information source increases the relevance of a web search hit. Platforms collect information on product assessments by people who bought a specific product in the past. Some platforms further developed this concept to contrast reviews of sellers and consumer of services to create a “web of trust”. In these cases, reputation becomes a proxy for trustworthiness. Customers are encouraged to assess the quality of a good or service they consumed. The assessment of the quality often follows a spontaneous, non-systematic approach in terms of criteria that are used by a person to assess the product.

IV. Mission of official statistics

10. Statistical offices provide society with high quality statistical information on conditions and trends in society, economy and the environment. In order to achieve this mission, the statistical system has developed an environment rooted in legal frameworks.

11. In addition, trustworthiness in statistical data is created via ethical codes of conduct such as the European statistics code of practice and quality frameworks that systematically collect information (metadata) to describe the process and the statistical product. Audits verify the adherence to the principles of the ethical code. In case of the European Statistical Systems (ESS), an ESS IT security mechanism provides assurance and certification regarding IT security and peer reviews are conducted to establish compliance with the European Statistics Code of Practice. As data was a scarce and expensive commodity, main efforts were put on designing data collections in the most optimal way. The quality frameworks define and describe criteria, procedures and indicators for collecting and assessing the quality of statistical information products. In conclusion, the statistical system has been designed in a top down approach starting from its mission going down to the statistical products.

12. While this overall mission of official statistics is still holding, the means of delivering this mission will have to change, partially induced by the described changes in data collection, processing and information retrieval described above.

V. Consequences for official statistics

13. Official statistics has been criticized for losing its ability to accurately represent the world[1]. This phenomenon might also be related to changing habits of information retrieval

[1] William Davies (2017): How statistics lost their power; The Guardian; https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy

and changed expectations of citizens towards consuming information, induced by the advent of the ubiquitous internet. Along with changing habits goes a change in expectation. Users demand transfer of possibilities they experience in one domain to other areas of information retrieval and service provision. The changed expectations are also fed by the demand of data driven decisions, which ask for evidence based on information before taking decisions. This demand is present in the business sector but also holds for the government sector, especially for political discussions within the society. Data-driven decision-making or evidence-based policymaking will have huge impact on statistics as the demand for reliable, high quality data will rise when the demand is taken seriously.

14. The availability of new data sources and changing habits of using information is impacting the production side of statistics as well as the way statistical information is consumed.

A. Input side

15. Typically, the new data is not collected for the purpose of official statistics. Therefore, it has to be re-purposed in order to produce statistical data. Often these data are of a different granularity than the micro data that is treated by official statistics. E.g. it does not contain information on the health status of an individual but provides data on the blood pressure or heartbeat, coming from health trackers. In addition, the temporal frequency is usually much higher than microdata held by statistical offices. The data has to be analysed and translated into meaningful information suitable for official statistics. In contrast to the microdata used in official statistics, we call this data deep or nanodata.

16. The volumes of these data are much higher than the volumes of data processed in statistical offices. Using all or only a part of these data might go beyond data processing capacities of statistical offices. Therefore, these data should be pre-processed to turn them from nano to microdata to make them manageable and integrate them into statistical production process. An additional aspect of this pushing out of computation to the data holders covers data security. Centralising all possible information at one place increases dramatically the risk of data security breaches as the central data store will get very attractive to security attacks. In order to distribute the risk, data should be kept where it is produced and the technical means for data processing, aggregation, linking should involve privacy preserving technologies up to the point when outputs would not be personal any more but aggregates.

17. These principles shift the focus of data production from designing data collections to designing data processing and from accessing data to retrieving information, final or intermediate statistical information, which is the final purpose of statistical offices. The focus would shift from sharing data to sharing algorithms and methodology to produce statistical (final or intermediate) output. This approach would require close partnerships between data holders and statistical offices and the use of standard or agreed methodologies for processing data at the source.

18. In addition, mechanisms for ensuring a defined quality of the sources, scrutiny on the processing algorithms, preferably as open source, and assurance that the agreed programs are run on the agreed source data. Ideally, outputs should be usable not only by statistical offices but also by data holders to create a win-win situation. The creation of a win-win situation would drastically ease collaboration between data holders and statistical offices. However, there might be additional incentives apart from enabling data or information services.

19. It will be necessary to use extensively incentive schemes to support retrieving data from private sources or from individuals. These incentives might be gamification, community engagement, public recognition, personalised services, monetary or pseudo-monetary reward for individuals. For enterprises this could be contribution to public good, information tailored to the company, enabling information services with the statistical office as third party which requires integrating data from competitors, e.g. getting information on how many companies are attacked by hackers without revealing which company is actually suffering from attacks, or joining roaming data to produce join statistical outputs. The introduction of freemium models might be a way of stimulating data markets.

20. In the context of company data, it is important to remind that the new data is in general not data informing on specific characteristics of the company but that data subjects are third parties. Therefore, the issue of response burden should be interpreted in a new way. In case of using incentives for data collection, providing data might not be perceived as additional response burden.

B. Processing of data within statistical offices

21. We assume that statistical offices would receive intermediate data products via privacy preserving computation channels that can be further processed to produce final statistical output. Often these new data can be used for different purposes, i.e. producing statistics in different domains, e.g. mobile network data can be used for enhancing population, migration, tourism, or balance of payments statistics. Therefore, the process has to be flexible to ensure multipurpose usage. At the same time, one statistical domain may profit from different input data, e.g. tourism statistics from surveys, mobile network data, traffic counters, etc. The architecture of the production infrastructure has to support receiving multiple inputs for producing multiple outputs.

22. In addition, one of these new data sources will most likely not be able to replace completely one traditional data collection but, together, the multitude of data sources will enhance different aspects of the final output product such as temporal or spatial granularity.

23. This new production architecture will increase the need for coordination between statistical domains within the statistical offices. Changes in the production process might have repercussions on multiple products. Opportunities of new data sources might be beneficial to various statistical domains.

24. Due to the technical characteristics of new data sources, the need for additional skills from engineering sector might increase in the statistical offices. In order to successfully negotiate and design the pushing out of the computation process, staff with specific expert knowledge in various domains of new data sources, e.g. communication, transport, electronic, engineers.

C. Output side

25. Previously we described the changing user habits when retrieving information from the internet. Statistical offices should pay more attention to the ways of communicating statistical information to different user groups, ideally providing them with information at the time of demand in the right quantities in an easy digestible way.

26. Users are expecting use of various modes of interaction to receive the requested information. Increasingly, people use digital assistants that interact with them via natural speech. This might create an expectation that information can be retrieved using speech recognition.

27. In general, the expectation on granularity of statistical information is increasing. Averages and sums aggregated on larger territories might provide a picture of the entire society but might fail in representing the reality or perception of reality of specific groups of societies and smaller regions. There is an increasing demand that statistics better reflects reality of smaller societal groups or parts of the territory.

28. Statistical information should be immediately at hand in formats and quantities that are suitable to the information need.

29. The perception of timely information will change due to immediate availability of information and emerging real time systems. People might no longer accept information that is more than one year old as recent information.

VI. Production of official statistics by third parties

30. The demand for statistical information is increasing considering the enlarged offer of digitally produced data and the paradigm of data driven solutions. New data sources will most likely not be able to replace existing data collections completely. Therefore, the approach of integrating new data into the process of producing official statistics will most likely not result in reducing of resources but in the best case doing additional things with comparable resources.

31. A mechanism for priority setting will continue to be necessary to decide on investments in specific areas. In addition, the focus of work will shift from collecting data to designing the process of distilling information from raw data. We assume that this process will consume most resources in the statistical production process in the future. This involves e.g. design of methodology, algorithms, IT infrastructures, framework for methodology, quality and metadata. These are elements, that would further be used for documenting data production processes and quality of final outputs assuming that the complete production process, or intermediate steps are performed by third parties. These elements would as well be suitable for introducing a certification process in which processes or outputs are certified for adhering to predefined frameworks.

32. A typical process for preparing certification could comprise the development of relevant frameworks, which would be best performed by technical specialists together with statisticians. Technical specialists such as engineers would provide specific knowledge on the data source, which is necessary to translate the source data into intermediate formats that can be the basis for statistical applications. The resulting frameworks would be used for documenting data processing and describing characteristics of the processing outputs. According to the desired level of product quality, certain elements could be selected for documentation. A third party, which could be different from a statistical office could certify the correctness of the documentation. Suitability of the resulting data for specific statistical applications could be assessed against its documentation.

33. Taking into consideration that most of resources would be consumed in the methodological design process and that statistical offices would perform the related activities, certification would not be suitable for enlarging considerably the overall offer of statistical information without increasing resources of statistical offices. In addition, it would be necessary to repeat the certification process because new data sources have the tendency to be volatile due to technological changes or changes in user behaviour. In conclusion, the introduction of a certification process would not liberate statistical offices from investing in human resources to perform the process.