United Nations ECE/CES/2022/26

Economic and Social Council

Distr.: General
9 June 2022

English only

Economic Commission for Europe

Conference of European Statisticians

Seventieth plenary session
Geneva, 20-22 June 2022
Item 9 of the provisional agenda
Collaboration with private data providers

Collaborating with private data providers – experiences from Denmark

Prepared by Statistics Denmark

Summary

The document provides examples of data collaboration with the private sector in Denmark, concerning data from mobile network operators, scanner data and data from electricity meters. The paper touches upon legislative aspects of data provision, identifies issues and challenges that arise, and concludes with some recommendations how to ensure successful collaboration with private data providers.

The document is presented to the Conference of European Statisticians' session on "Collaboration with private data providers" for discussion.

I. Overview

1. National Statistical Institutes are constantly requested to inform about and monitor new phenomena that emerge in the modern world, much of the information requiring new approaches to producing statistics. At the same time, the NSOs are under constant budget constraints, which do not open for increased statistical production without some form of prioritisation. In this context, non-traditional data sources[1] are perceived as a possible means to expand statistical production without a significant increase in costs. Non-traditional data can be characterised by big volume, variety and velocity, but also as data not yet used in statistical production but with applications potential. Non-traditional data currently considered for potential production of official statistics include geospatial data, telecommunications data, data from electricity meters, AIS (Automatic Identification System) data, and citizen-generated data. The list of non-traditional data that can be used for production of statistics is constantly expanding and the statistical community is working on adapting new sources to produce more and more timely statistics. However, it is country dependent whether data can be categorised as non-traditional – some countries have broad access to various registers, whereas, for others, it can be a type of data not yet used, thus falling into the category of non-traditional. This paper will mainly focus on data held by the private sector but will also provide an example of data held by branch organisations that can be used in statistics.

2. There can also be various challenges for a concrete implementation of non-traditional data sources into statistical production: quality, continuity and adapting the data to current statistical needs, but also, how non-traditional data can be accessed. NSOs are subject to a wide range of quality demands, which naturally affect what data and in what manner data can be applied. Furthermore, non-traditional data is a valuable asset for the companies that hold it. The data is generally not publicly available, and to gain access, NSOs need to negotiate agreements or gain legal access rights. Data needs to be obtained in an identifiable form to enable statistical use of it – e.g. to enable linking to a random data set for estimation purposes, or for records to be linked to produce wider analytical value: but the published results need to be non-identifiable[2].

3. Long-term supply of non-traditional data requires the NSO to develop an ongoing and strategic relationship with the data custodians, resulting in contractual agreements or memoranda of understanding between the NSO and data custodians to guarantee the continuing supply of data, as well as early notification to the NSO for any planned change to the collection, processing and dissemination process of the big data.[3]

II. Legislative aspects of data provision

4. In official statistics, data collection may be undertaken voluntarily, as with household surveys, or may be collected under legal mandate. Business surveys and the census of population are typically collected with legal sanctions for refusal to provide information, and an associated legal requirement on the statistics office to keep such data secure. In the Nordic countries, legislation also allows the NSOs to access data from the governmental administrative systems free of charge. This access is a cornerstone in the Nordic statistical systems that to a very large extent build on data from administrative sources. This of course implies that NSOs are trusted as regards data security and confidentiality from the official side and, equally important, from the side of the public knowing that ‘what goes into the NSO’ will solely be used for statistical purposes.

5. The challenges on accessing data from private sector companies are more challenging. Firstly, data for private sector companies can be a source of income and the companies would

[1] We decided to use the term non-traditional data instead of big data. There are many overlaps between those two terms, however sometimes non-traditional data cannot be characterised as big data but can still be used in statistical production.

[2] Penneck: Confidentiality in an era of Big Data: an official statistics perspective.

[3] SM Tam, Van Halderen: The five V’s, seven virtues and ten rules of big data engagement for official statistics.

probably think twice before sharing the data free of charge on a regular basis. Secondly, without certainty that data will be stored in a secure manner, a fear of losing a competitive edge may arise. Thirdly, the multinational character of many companies implies that companies have a good insight into different sets of rules and 'willingness' to pay for data from different governmental agencies. That can entail that willingness to pay for data in one country can create a precedence for a company that consequently will require payments in other countries.

6. Continuity of data transmissions is also an important aspect that NSOs should keep in mind when signing a data cooperation agreement with the private sector. Preliminary experiences demonstrate that the private sector companies are willing to participate in pilot projects on the application of data to statistics, which can provide some (promising) results. However, this form of cooperation is often conducted only as a pilot study giving countries knowledge on and practice in the application of data held by the private sector, but still not opening for a production of more regular statistics.

7. Currently, there are not many examples of national legislation giving access to privately held data. In the Danish case, a mandate for accessing private data for the statistical office was introduced in a draft act on Statistics Denmark. However, it was taken out even before it was presented to the national parliament. The legislative mandate for accessing data from the private sector is also largely absent in other EU countries with a few exceptions. The access to data of private companies for the purpose of statistics is now being discussed at EU level. In February 2022, Eurostat formally started the process of revising the Regulation 223/2009 on European statistics, under the flagship title "European Statistical System – making it fit for the future". One of the main objectives of the revision is to introduce rules for accessing privately held data for European statistics. The plan is for the European Commission to publish a proposal at the beginning of 2023.

8. The following chapters will provide some examples of data collaboration with the private sector in a Danish context.

III. Mobile network operators (MNOs)

9. Mobile phone data is one of the fastest growing technologies in the world with global penetration rates reaching 90 per cent. Data is generated every time phones are used and have the potential to generate insights in official statistics. They can be used to compile official tourism statistics, population statistics, migration statistics, commuting statistics and employment statistics on border and seasonal workers.[4]

10. There are, however, some challenges in using mobile phone data for official statistics. There is usually more than one telecom service operator in a country (they are often multinationals) and the NSO will need to cooperate with all (or a clear majority of) providers to obtain data for compiling statistics.[5] Cooperation here requires a broad dialogue with data owners, which can also require different approaches to communication, as the companies vary.

11. As in many other countries, Denmark saw an increased use of MNO data during the pandemic. The data was used to measure movement ('trips'), which is essential to gauge the effectiveness of COVID-19 restrictions and to model the spread of COVID-19. However, the establishment of a data flow was legally challenging and the MNOs were very concerned about the data safety and with recovering the cost incurred. The data flow was stopped when COVID-19 was no longer deemed a threat to society, but at that point, the data was in fact not used much due to quality issues. This could point to the fact that the MNOs could be more willing to cooperate if a concrete and urgent situation develops.

[4] SM Tam, Van Halderen: The five V's, seven virtues and ten rules of big data engagement for official statistics.

[5] SM Tam, Van Halderen: The five V's, seven virtues and ten rules of big data engagement for official statistics.

IV. Scanner data

12. Scanner data is widely discussed as a potential source of statistics and currently it can primarily be applied in the calculation of consumer price indices, but other applications, such as consumer preferences, are also possible.

13. In Statistics Denmark, scanner data is currently used for consumer price indices building on input from several major supermarket chains. This is done to provide accurate and timely information on the prices of goods consumed by households. Since the data sources are extremely high dimensional as they consist of all individual purchases at the level of baskets, substantial aggregation occurs at the data provider. Based on experiences from other European NSOs, Statistics Denmark reached out to the data providers already back in 2010. After some years of testing, scanner data was implemented in the production of statistics in 2016 – first for food and beverages and later for other products frequently sold in supermarkets.

14. More concretely, in the inquiry to data providers, Statistics Denmark made a generic request for the data we were interested in. For the chains where scanner data was procured, point-of-sale data was provided. Furthermore, it was possible to procure additional helpful metadata, like shop location and the hierarchy of the products sold. After reception at Statistics Denmark, the files were loaded into a generic file storage system with user-controlled access. Further processing was also controlled by user-access, which helped to eliminate many potential concerns about data confidentiality that supermarket chains might have.

15. The long period between receiving the first files and implementing data in the consumer price indices was used for testing the data – both with regard to data contents and delivery reliability. For all the supermarket chains, declarations of intent were signed. They stipulated the possible use of data in the few statistics, where scanner data could be applicable. Furthermore, the declarations specified the frequency of the data deliveries (weekly), technical details and, most importantly, the commitment to inform Statistics Denmark three months in advance, if there were plans of any major revision of the data or the transmission that might imply changes to the processing or reception of data at Statistics Denmark.

16. The stability of data transmissions for all the chains is generally quite good. Nevertheless, two types of challenges can be distinguished. The ones that are non-planned and the ones that are planned.

A. Non-planned challenges

17. Two general types of non-planned challenges might arise. First, there might be issues with the content of the data. Before implementation of the data in production, Statistics Denmark learned to handle this type of error quickly requesting transmission of new files and receiving them equally swiftly. This type of error does not happen on a frequent scale any longer.

18. Second, updates of IT systems or IT errors at either Statistics Denmark or at the data providers might cause a small delay in the delivery of data. Normally, Statistics Denmark/ the data providers are able to locate these problems swiftly and generally within a week. This kind or error is more common than the first error mentioned above, but still very manageable.

B. Planned

19. All chains are requested to inform Statistics Denmark duly if they plan any major revision to their IT systems that involve risk of alterations in the delivery/content of data with more profound implications. Generally, these kinds of alterations happen for most chains, and though they often require several weeks of testing, Statistics Denmark is able to handle them because the information about the alterations comes in due time.

C. Going forward

20. With the signing of new contracts with the chains, Statistics Denmark is now able to use the data for general official use. This means that Statistics Denmark is now in the process of analysing which new kinds of data (especially granularity of point-of-sale data and the extra metadata) are needed. We are also in the process of investigating whether data can be procured from other types of data providers (an example could be petrol stations).

V. Electricity meters: data from branch organisations

21. Electricity smart meters provide high frequency data on consumption at each metering point. In Denmark, data from electricity meters is under the custodianship of Energinet, an independent public enterprise owned by the Danish Ministry of Climate, Energy and Utilities. Energinet owns, operates and develops the transmission systems for electricity and natural gas in Denmark. In 2021, it has upgraded all electricity meters to smart meters (3.8m). The smart meters read electricity consumption at least once an hour, which resulted in 42bn readings in 2021. All data is collected by a common data hub, which is also owned by Energinet and transmitted to Statistics Denmark on a daily basis only 8 days after the end of the reference day.

22. The rollout of smart-meters in Denmark in conjunction with the establishment of an efficient mechanism for consolidating the measurement across utility companies means that there is an extremely good infrastructure for using this data source for statistics in Denmark. The smart-meter data contains information that allows Statistics Denmark to link the electricity consumption into unique dwellings, and thus to connect data with businesses and households. This provides many potential uses of data in the statistical production, e.g. as a fast business cycle indicator, monitoring the activity within different industries and geographical areas. In addition to macro level statistics on electricity consumption, the granularity of the data has allowed for the creation of several new types of statistics, both in the domain of social and business statistics. For example, experimental statistics were created during the pandemic to gauge the extent of working from home or staying at home. Similarly, the lockdown of particular sectors of the economy (e.g. hairdressers and schools) were directly visible in the electricity consumption. Currently, this data source is used in the development of various new statistics that measure activities at construction sites and multi-location statistics that measure occupancy in homes and vacation homes using electricity consumption as one proxy.

23. Energinet has an obligation to make data available to researchers, which it does in collaboration with Statistics Denmark. Statistics Denmark acts as a data processor for Energinet and makes data available for the researchers on a research service scheme. The cooperation between Energinet and Statistics Denmark is agreed in a contract, which also allows Statistics Denmark to use data in the statistical production. In the initial phase of the collaboration, a number of clarifying meetings were held between Energinet and Statistics Denmark. The purpose was to agree on the technical and contractual conditions. Now, the contact is on an ad hoc basis, primarily by phone and primarily to clarify technical issues related to the data transmission.

24. However, it has to be noted that this statistical production is only carried out on an experimental basis.

VI. Conclusions

25. The experiences described above lead to the following observations and conclusions:

26. Written contracts (or agreements) are an important element of a successful collaboration with private data providers. They can either take a binding or non-binding form. Even though preference is given to binding contracts, they can be harder to achieve. Contracts should include details of what data is to be transmitted between a private data provider and an NSO, its formats, frequency and other technical specifications. It is useful if contracts

clearly state who the contact persons are with the respective parties. This can help solve possible challenges in a more immediate way.

27. Besides contracts, an ongoing dialogue is a very important part of the collaboration. Here both the NSO and data provider need simple access to their counterparts in order to ensure smooth cooperation and discuss possible challenges beforehand and, if possible, in an informal way.

28. NSOs must be very strict regarding compliance with data security and confidentiality. This is both crucial for the production of regular statistics, but also for production of statistics with the use of data from the private sector providers. Here the trust is central to any collaboration and can be built through ongoing dialogue with private data owners on the approach to data security and compliance with existing legislation and best practice.

29. NSOs should outline how possible delays in data transmissions are addressed. Various things can cause delays, often things that are beyond the control of individual employees involved in data transmissions. Even unintended delays in data transmissions from private data owners can influence the timeliness of the publication of statistics – this aspect can be difficult to explain to the users.

30. On the national level, legislation is not giving the NSOs access to privately held data for the majority of countries. This is naturally challenging for a more regular introduction of the privately held data into statistical production as many important aspects of data transmissions cannot be ensured. For the EU countries, the access to privately held data is now under discussion on the EU level. The development will be very interesting to follow, as possible access to privately held data can go against existing national legislation.