

**Economic and Social Council**Distr.: General  
24 May 2019

English only

**Economic Commission for Europe**

## Conference of European Statisticians

**Sixty-seventh 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****New data sources, leading to new indicators  
and to new collaborations****Note by Statistics Netherlands***Summary*

With the start of the Center for Big Data Statistics in 2016, the scouting of new data sources has really taken off at Statistics Netherlands. This document presents thoughts and experiences on what it takes to get access, as well as models of cooperation in the ecosystem of Statistics Netherlands, infrastructure needed and consequences for official statistics. The paper also describes results obtained from these sources.

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.



## I. Introduction

1. Society is changing rapidly, leading to drastic changes in both availability of data and information needs. Many human interactions, devices and sensors generate huge amounts of data, offering opportunities to fasten up statistics, provide more detail, improve quality and lower administrative burden for both businesses and individuals. There are technical possibilities and a societal desire to process, combine and use these data leading to innumerable applications. Policy makers have a strong ambition to underpin new policy measures with timely and relevant data. The life cycle of policy measures tends to decrease. On the other hand, tension in society between the potential benefits of data usage versus preserving individual privacy seems to be growing.
2. The real challenge of the statistical world today is how to exploit the societal benefits of these new data sources for official statistics, thus remaining relevant and being able to address changing information needs in an objective way. For this challenge, Statistics Netherlands has launched the Center for Big Data Statistics (CBDS). At CBDS, we aim at maximizing the societal value of new data sources, mostly generated in the private sector, by working together in a data ecosystem with partners. We strongly believe in the possibilities of introducing those data sources into official statistics, as a valuable add-on to registry and survey-based statistics.
3. In this paper, we focus on the main hurdles of this challenge. How to get access to these new data sources, start new collaborations and combine all different interests involved? What are the effects of this for official statistics?
4. For this, we will first share our strategy on cooperation, our thoughts and experiences on what it takes to get access. Second, we will present models of cooperation in our ecosystem, the infrastructure that is needed and what change in mindset it takes to implement results into production. Last but not least, we plead for a needed increase of awareness among citizens on the usage of data for societal good. Finally, we will also share some results that can be obtained from these sources and collaborations.

## II. General strategy on cooperation and the data ecosystem

5. It is our ambition to enlarge the societal impact of our data, and therefore we support governmental organizations to become data driven. The societal issues these organizations have to deal with are more and more complex, like the energy transition, undermining crime, overcrowded infrastructures and social inclusion. Many organizations are involved, leading to a very fragmented data landscape. We believe ecosystems, coalitions and consortia are the future to gain insights in these complex phenomena.
6. For two reasons: due to budget cuts, the fast pace of technological changes and growing number of relevant data sources, co-creation is needed, and no single organization will be able to act merely on its own. On the other hand, to fully address the information need of governmental organizations, cooperation is needed with others because of limitations in our mandate. Examples of this are forecasts, scenario building, advice, etc. In this paper, we focus on the first reason, since the second one is still under development.
7. Although Statistics Netherlands has wide experience with cooperation with universities and knowledge and research institutions, cooperation with private companies is quite new. Within CBDS we have created an ecosystem focussing on co-creation. We work with more than 40 partners, have realised more than 20 projects with partners and have established some simple rules of engagement on the way which are outlined below:
  - (a) All partners should benefit. Partners in the ecosystem must understand each other's goals that need to be addressed in the cooperation. These can vary; we touch upon some of them later on in this article. Cooperation is with closed wallets, non-exclusive and focusses on joint development of (possibly paid) products or knowledge;
  - (b) For this reason, a tailor-made approach is needed, and no cooperation is or will be alike;

(c) Sharing. In this cooperation, (a combination of) data, knowledge or resources must be shared. Publications about the development are jointly made, intellectual property over the developed product is shared;

(d) Co-creation can lead to changes or additions in our statistical programme, or lead to new one-time or repeated indicators;

(e) Flexible working arrangements. Agreements are made in a simple memorandum of understanding, with limited number of conditions and an easy process of termination. Actual projects lead to more detailed agreements;

(f) For the developed product, a business model is designed;

(g) The Statistics Netherlands Act is (possibly amongst others) always applicable: results must be available to all under the same conditions, methodology needs to be transparent.

8. At this moment, Statistics Netherlands is exploring how this ecosystem can be expanded to cover the entire scope of Statistics Netherlands and not just big data.

### III. Business models and motives

9. Data is the new gold. So, what would persuade other organizations, mainly from the private sector, to work with a statistical office? The keywords here are reputation, knowledge transfer and expanding markets. Collaboration can lead to a variety of business models, for example: a joint development on private party data in combination with data from Statistics Netherlands leads to products that can be offered to third parties against payment; a joint tender with a major international business service provider; orders for a ministry jointly with start-ups; investigating with various companies whether we can achieve paid services (licences for real-time streaming insights); analysing a private source for its usability for statistics (the company involved gains insight into the quality of that source and the application possibilities for its own primary process, or increases its knowledge on data science), using data of a particular company that boosts up its societal and trust image.

### IV. Getting access

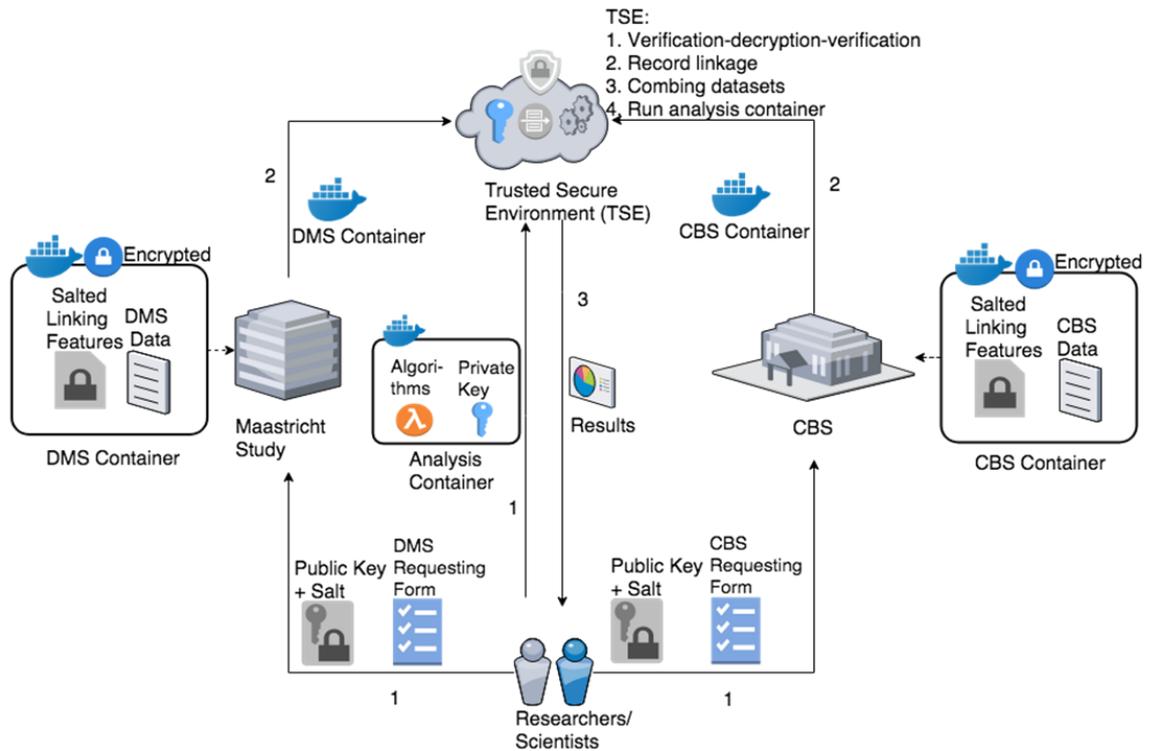
10. Although there might be willingness to cooperate, and there are general ideas about a valid business model, the trouble begins when the actual project starts. How is data shared? What terms do we need to agree on?

11. Most data sources contain sensitive data, making information exchange between organizations difficult due to applicable law and terms, and also due to agreements a company has made with its customers. To make data exchange possible, and also lower thresholds to increase the willingness to collaborate, we actively seek for safe ways in which multiple unbalanced sources, owned and or hosted by different organizations, can be combined. How can we share sources, in which a common key can be identified but with different data elements in each source, and where duplication, because of the size of the sources at hand, is not possible?

12. This is not only a technical and security challenge, but also from a privacy and organizational perspective. How should we organize data governance, how do we keep track on authorizations, on algorithms used for analysis to make sure that both security and privacy are secured?

13. At this moment, a Proof of Concept (PoC) with the Maastricht University is realized, in which a Trusted Secure Environment (TSE) is created for linkage and analysis. Docker containers and encryption facilitate that no involved entity has a complete overview of the data involved. Figure 1 presents a short overview of the Privacy Preserving Data Sharing (PPDS) infrastructure used. This infrastructure is regularly challenged with new commercial solutions.

Figure 1  
Short overview of the Privacy Preserving Data Sharing (PPDS)



14. At the moment, a Proof of Concept has been successfully realized, using medical data (DMS or De Maastricht Study is a regional cohort study on Diabetes type II). At the end of 2019, this infrastructure should be fully operational and useable as a service in collaboration with others. In addition to this, we experiment with the combination of secure multiparty computation and blockchain in the Techruption initiative (<https://www.techruption.org/usecase/privacy-preserving-analytics/>).

## V. Skills

15. Of course, appropriate resources and skills are needed to explore new big data sources. At Statistics Netherlands, luckily, we do have a good position on the labour market to attract new data scientists. The start of the Center for Big Data contributes to that and stimulates knowledge sharing with internships, PhD positions and secondments from other organizations. Next to that, the recently started CBS academy has the mission to develop data science skills among Dutch civil servants with a strong and ambitious curriculum. To make sure knowledge is spread to all departments at Statistics Netherlands, a Data Science Community was started. Data scientists can share issues, learn together, make guidelines for coding and documentation, work on business issues in internal hackathons, etc. Also, capacity building is part of each project where innovative beta products are transferred to production.

## VI. Results

16. A safe conclusion is that introducing new data sources into official statistics is not just about gaining access to data. It is about collaborations, finding mutual goals, jointly facing challenges, managed expectations and coping with consequences on quality and trends.

17. Below is a (limited) list of results, in which we also describe the data used and partners involved. Projects we run without partners, or where open data is used, are not mentioned. Results are always shared via [www.cbs.nl/innovation](http://www.cbs.nl/innovation).

- CBDS has acquired mobile phone data of one of the three major mobile phone operators and works together with a data provider of one of the remaining two. Results at the moment are visualizations of overcrowded city centers. Other applications, like mobility or (un)employment are under investigation.
- We set up a project with the data owner of all public transport chipcard data to dive into mobility to and from public transport stations.
- With a world leading navigation system provider, we benchmark their data with detection loops on Dutch highways, in order to gain insight in the usage of roads where no detection loops are present. The goal is to develop new indicators for lower public bodies on motives (who and why) behind mobility.
- For gaining insights in relations between companies, we work together with a large global provider of data about enterprises and thus strengthening our business register.
- With the Ministry of the Interior we investigate digital illiteracy by, amongst others, using data of large governmental platforms (MijnOverheid.nl). In this project we partner with the supplier of this platform.
- With the Ministry of Economic Affairs and Innovatiespotter, a start-up, we work on deepening details on innovative companies by using web scraping and text mining, also on companies with less than 10 full-time equivalents involved. This is planned to take into production this year
- The same goes for new models of calculating solar power generated by panels on individual houses, thus greatly speeding up information on this phenomenon.

## VII. Effects

18. Working in an ecosystem requires a different mindset. First of all, a network needs to be built and a proposition needs to be made, to manage expectations right. Second, before jumping into a project, thoughts are needed about what knowledge is already out there, with whom this project could be run best and under what conditions. These processes might take time upfront but can lead to much more efficient projects since each partner is used for its own unique selling points. Third, the developed product can have different results than the survey it aims to replace or enrich. If the re-use of data is deemed to be important, this might well lead to a paradigm shift within our statistical office. Do we still ask the right questions in our questionnaires? We must dare to recognize that indicators based on these new data sets also can represent society and (re)act accordingly.

## VIII. Future work

19. In the coming months, we aim at finishing the PPDS infrastructure, preparing legislation to guarantee access to mobile phone data, smart meter data and bank transaction data. We will also build and extend the data ecosystem, establish strong relations with both product development, account management and the CBS Urban Data Centers, and provide further rules of engagement.

## IX. References

P.M. van Meeteren, Dare to change in a changing society, Contribution to the Seminar on “Measuring What Matters” at the 2018 Conference of European Statisticians

S.M.M.G De Broe, P. Daas, P.M. van Meeteren, Big data to improve policy and decision making: The Experience of Statistics Netherlands, Contribution to DGINS Bucharest 2018

J. van Soest et al, Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data, European Federation for Medical Informatics (EFMI) and IOS Press, 2018

---