## SLS/STATEC joint event 17 May 2019

## "Preparing the Future: The Impact of Digitalisation on Official Statistics"

*The following text served for the opening address of Nico Weydert, president of the Luxembourg Statistical Society. It reflects his personal opinions and not the views of the Statistical Society.*

## Impact of Digitalisation on Official Statistics

First, let me thank STATEC to have associated the Luxembourg Statistical Society and invited me to give an opening speech for today's event.

Yesterday, there was an interesting workshop on mixed-mode surveys. The reasons to develop and rely on mixed-mode surveys are manifold. Official statisticians face a growing reluctance to respond to surveys. The nonresponse rate is a permanent threat for the quality of statistics. Moreover, the European Statistical System (ESS) also wants to reduce the response burden and to be cost effective. One possible way to do this, beside optimizing and modernizing traditional surveys, is to rely on new sources of data, as well as on new techniques. Today, a lot of the new data come from digital footprints we all leave in an ever-growing digital world.

### Digital footprints: manna from Heaven?

The Round table that follows, will present and discuss new tools and new data sources for National Statistical Institutes (NSIs), in other words, the impact of digitalization on official statistics. The question is: how will NSIs consider these changes, how will they prepare for the future, and will there be a new work organization as the title of the wrap-up session seems to indicate. A lot of questions and many are closely connected to the use of big data, artificial intelligence and machine learning. All this is quite recent. Viktor Mayer-Schönberger's and Kenneth Cukier's Big Data book came out in 2013, the same year where the statistical community took on board big data, at United Nations and European Union levels. Many discussions, analyses and pilot projects took place ever since. The DGINS conference last year wrapped all this up into Trusted Smart Statistics.

Let's acknowledge that statistics have been and are developing over time, with scientific advances and new technical possibilities. The changing social and political environment in turn asks more and different questions to statisticians. No surprise then, that big data, digitalisation, datafication has become a catchword in the strategic statistical discussions.

Let's come back to digital footprints. They do not only cover the trail of data we create while surfing on the web, although we leave a lot of information out there. I dare say that Google or any other browsers, or Amazon might often know much more on us than our closest family members. Beside surfing we have to add the footprints of email, Twitter, Facebook, of mobile phones and a growing number of digital devices of all sorts. These footprints are data, much data, big data. Some say that formerly, the statisticians had to make many efforts to collect data (they still have to do so), but now they are confronted to a deluge of data. Data is the bread for statisticians, and some consider big data as a sort of manna from Heaven.

What is big data? There are discussions on a precise definition. For our purpose, let's say it is a bulk of data, many variables, changing quickly, unstructured sometimes, it needs much storage and computer power and software to be processed. It is developing, it is embedded in an on-going process.

**The reaction of Official Statistics**

UNECE published in summer 2013 a key paper: What does Big Data mean for Official Statistics? A dedicated website of UNECE to Big Data was set up. Even a sandbox for those interested in experimentation was created.

In the ESS at the autumn 2013 DGINS conference there was the Scheveningen memorandum on Big Data and Official Statistics.

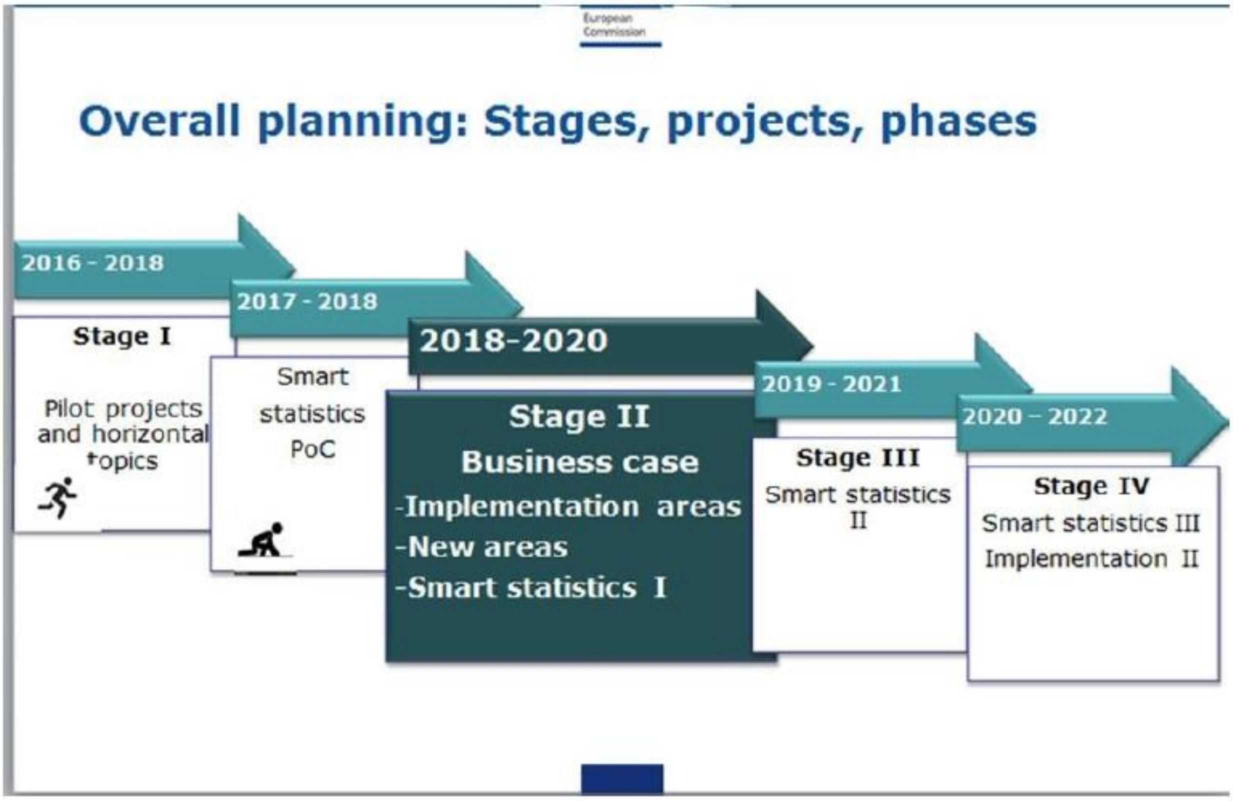
A Big Data task force was created and a Big Data Roadmap and action plan 1.0 established.

There were conferences, like NTTS, workshops.

There are ESSnets on Big Data. The projects concerned e.g.:

- webscraping of job vacancies, enterprise characteristics,
- smart meters used for compiling statistics on energy consumption, but possibly also on housing, household expenditure, environmental impact and energy production.
- Early estimates (GDP) and use of mobile phone data.

There is planning process:



Overall planning: Stages, projects, phases

2016 - 2018

Stage I

Pilot projects and horizontal topics

2017 - 2018

Smart statistics PoC

2018-2020

Stage II

Business case

-Implementation areas

-New areas

-Smart statistics I

2019 - 2021

Stage III

Smart statistics II

2020 – 2022

Stage IV

Smart statistics III

Implementation II

Some results of the ESSnets are very interesting. Promising projects are in the implementation phase, whereas new pilot projects have started.

Let me be a bit conservative or provocative: compared to the requests of the whole statistical programme, the possible new smart statistics are at their very beginning. All this is very exciting and interesting, but sometimes one might get the impression that there is a sort of general enthusiasm: “Let’s jump into the Big Data sea, where there are already some famous representatives from the private sector, let’s stay relevant and see what we can do”.

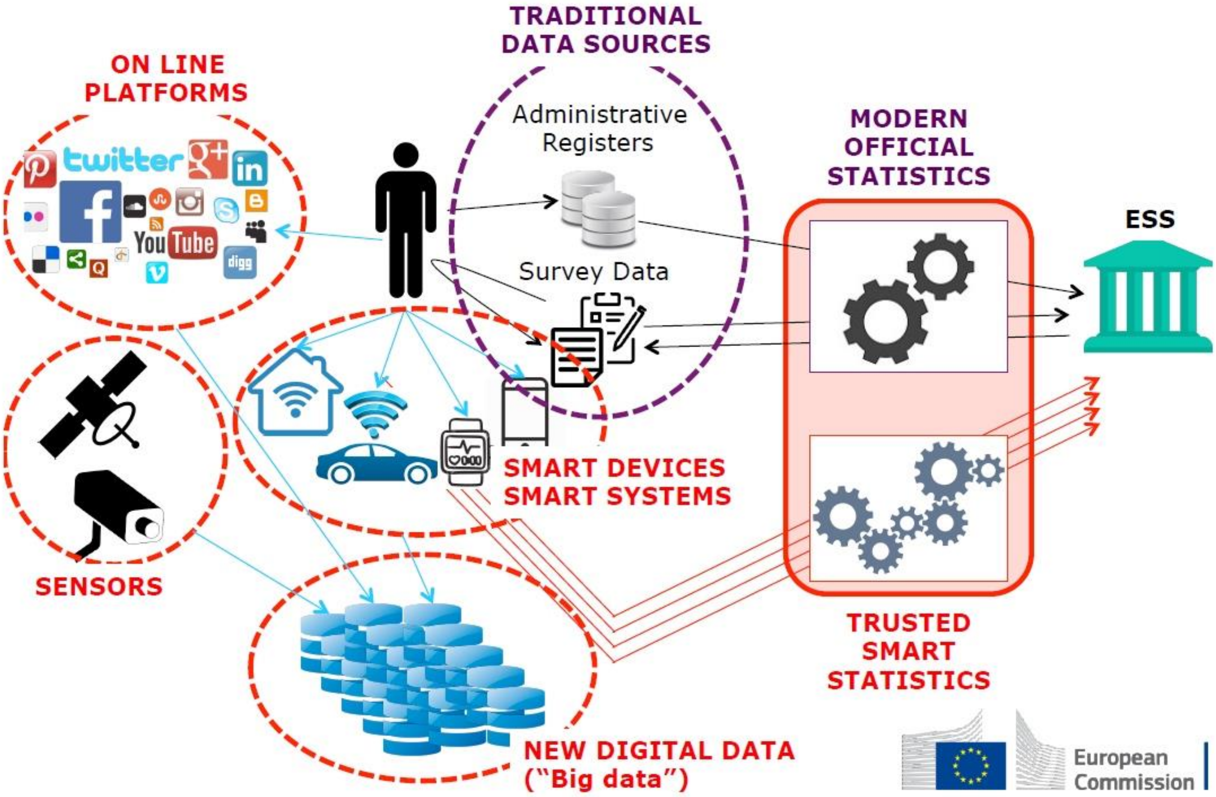
I would like to take a more cautious approach and make a step back and ask: What is the role of the NSI? Let me quote three CBS colleagues (Struijs, Braaksma and Daas) saying NSIs aim is “providing information on all important aspects of society an in impartial way, and according to the highest scientific standards.”

The question is of course, what are the important aspects of a society and who decides? In my understanding these aspects are decided in the EU statistical programme. At country level there might be some national statistics, but their part is generally small. The ESS statistical programme is a regulation of the European Parliament and the Council and the yearly implementation and financing is part of a Commission decision. There are of course political aspects that enter the programme as well as theoretical concepts, like national accounts, unemployment, that are operationalized through specific regulations in order to produce comparable EU statistics. What is important in my eyes, is the fact that all this is a top down process.

When looking at presentations on Trusted smart statistics, one might get the impression that there is a move away from this to a bottom up approach. We all know of course that many actors in the private sector, often owners of the data, are heavily treating, maybe torturing these data with new techniques for commercial, promotional and political purposes (remember Facebook data used by Cambridge Analytica). We all know that there are many important aspects, like the digital single market, climate change, biodiversity at risk, migration, growing inequality and Official Statistics have limited resources, so they must prioritize the work. From the point of view of the global statistical programme and the products provided, research in the use of Big Data can only represent a small contribution.

Which data sources can be used to provide the statistics foreseen in the programme? Can new sources be incorporated to produce statistics, or maybe used to produce new relevant statistics and at which frequency? What are these new relevant statistics? What will be the cost of using new sources and in which form can we use them to build Official Statistics? For the time being it seems that these new sources can only complement traditional sources.

At EU level there is a strategic vision, like the one presented at the NTTS 2019 conference:



Now there are arrows in all directions, but somewhere there seems to be much data that gets merged and mixed to produce Trusted Smart Statistics. Those who have worked on merging and matching know that this is a pretty tricky business unlike commercial promises of some software companies.

The merging or blending of data raises also confidentiality and data protection issues. Although these issues are recognized, the rush for smart statistics seems just to make vague declarations of intentions. Let me quote the smart statistics project of Eurostat "In addition, it will analyze horizontal issues, such as security, privacy, algorithmic transparency, or IT." This is nicely said but as statisticians we have to ask ourselves if we want to go towards such a "1984" direction.

Another issue of digitalisation is the speed of information. Smart statistics should arrive at high speed, like parcel delivery nowadays. This questionable from an ecological point of view and also in the field of certain statistics e.g. GDP forecasts in a small open economy like Luxembourg.

**Access, Continuity of access and Cost of information.**

Access to Big Data is also an issue. There was a workshop on the use of mobile phone data in these premises a few years ago. Access was also discussed, and I remember one telephone company saying. OK we can provide you access to the data, but the cost is 1 million € per year. This is a real issue: access to privately owned data! It's not enough to enshrine in law that official statistics have access to all needed databases. It is known that official statistics sometimes even encounter difficulties to access publicly owned data. OS try to find a workaround to this, saying that one might imagine partnerships, PPP, but do official statistics have the necessary power or weight in such a partnership? The continuity of access and the generalization throughout EU countries is of course also an important issue.

Let me shortly come back to the use of mobile phone data**.** We know some possible applications in the field of statistics like tourism and even in balance of payments. The example of using mobile phone data to estimate the visitors at the Belgian coast at an Easter weekend is often mentioned as a good example of the capability to deliver real time population estimates, faster than population registers. For me the question is: do official statistics need to estimate the population daily. This could anyway be done with good population registers that include far more reliable classification variables. For Luxembourg it might nevertheless be interesting to investigate commuter movements based on mobile phone data. A couple of years ago, POST company seemed to be open to help investigating possible use of their data. Could I suggest that STATEC launches a research project together with partners on this issue. Could such a research project be funded by FNR?

Let's come back to access issues. STATEC managed to incorporate scanner data in the computation of the consumer price index. This is great and STATEC recognizes the highly valuable contribution of those who provide the data. But it seems difficult to extent the coverage of scanner data and specially to incorporate more categories of products.

There is indeed a cost for private companies to provide such data. Moreover, if we consider big data, the size and property rights might be such that private companies would have to run procedures on their IT systems to produce statistical results. It's fine to claim that all this should be transparent. But what will be the incidence of pushing computation out or sharing computation with the private sector on trust in statistics? Beside confidence aspects of the public, the cost of such joint ventures should not be neglected.

I could imagine that OS will have to foresee financial resources in the future for getting access to privately owned data. But then what about physical persons selected in a sample. Many complain that responses to questionnaires often take a lot of time. For the Household Budget Survey and EU-Survey in Income and Living Conditions STATEC foresees a financial compensation. Yesterday we heard of the benefits of incentives on the response rate. Could this be a way for STATEC in the future?

**Statistical literacy**

Let me now come to a last challenge of big data: to handle such data official statistics need to hire persons with a strong mathematical background, with solid knowledge in statistics and also computational skills. Roughly speaking, official statistics need to recruit more data scientists or to set up interdisciplinary units with people having the appropriate skills and capabilities to work together. The ESSnets are a step in the right direction of sharing knowledge and building synergies. But more has to be done and one further challenge for NSIs in this field is the competition with the private sector that might be capable of offering more attractive salaries than the public sector.

Let me **conclude**, by saying that official statistics should in my eyes concentrate on their core business producing high quality trusted statistics. There is trust in official statistics and there is the knowledge to use sophisticated statistical tools. In certain, still limited fields statistics might be produced or be complemented today by new data sources provided they prove to be efficient from a scientific and economic point of view. But one should not rush to fast for Trusted Smart Statistics. The Commission at political level wants to reach the objective of advancing Europe as leading region in Internet of Things products and services. This is fine but will be hard to achieve. It reminds me a little bit of the over-ambitious Lisbon strategy.

Big data might lead to new statistics but the way to it should be an appropriate combination of top down and bottom up approaches. By the way, the same goes for machine learning. NSIs should jointly try to build the appropriate knowledge related to data science in their system. Yes, we might be at the eve of something new. So, let's consider it properly. Even if my position is a cautious, a bit provocative one, I would not like to do like Antoine-Augustin Cournot who rejected Alphonse Quetelet's "average man" as a physical monstrosity.

Nico Weydert

President of the Luxembourg Statistical Society