Conference on New Techniques
and Technologies for Official Statistics
(NTTS 2019)

Closing speech by Director Kurt Vandenberghe

1. Introduction

Ladies and Gentlemen, dear colleagues,

It's for me an honour to deliver the closing speech of such an interesting and intensive three day event, which even covers a full week if you include the satellite events.

Don't expect me to say anything "new" to you, after the 100 presentations and keynote speeches. And don't expect me to even try to make a summary. That would not pay due tribute to the depth and value of the event.

I'll give you a few reactions from the perspective of DG Research and Innovation to the many subjects discussed over the last days.

Let me start by warmly thanking Eurostat for organising this conference. From a three-yearly event held in different places, it has now become a two-yearly rendez-vous in

Brussels. This is a sign that progress has accelerated and it shows the intention of bringing methodological discussions closer to policy making, notably at the EU level.

We live indeed in a world that is changing extremely fast. Every day we witness the emergence of new phenomena driven by the digital revolution or by sustainability concerns. The penetration by increasingly powerful digital technologies is ever expanding. By the end of 2018 the planet had more than 8 billion mobile subscriptions and 4 billion Internet users. A "Sharing Economy" of products, services and business models is emerging.

In the face of such fast and disruptive technological change, resource scarcity and societal and territorial inequalities, it is essential for Europe to rely on the "right" data for policy making that delivers prosperity and ensures sustainability.

Against this backdrop, we consider it essential to have a continuous look at the evolution of statistical techniques and technologies and be part of the debate and the work. Thank you, therefore, for the opportunity to give a few closing remarks.

2. Emerging new phenomena call for new approaches

Policy-making needs to be evidence-based and that requires timely and precise data.

Let me take the example of **research and innovation policy.**

Usually data - for example on investments, number of researchers or patents - are released with a few years' delay. As a result, the data show us the past and not the present, while R&I policy is about policies for the future. In a world that changes rapidly, there is a growing need for timely data.

Fortunately, new possibilities are emerging.

Big Data for example. The high speed of data production in the digital era gives us possibilities for timelier statistics and and for access to a greater volume and scope of data.

We have to be open to new, unconventional approaches. For example, the ETLANow project led by the Research Institute of the Finnish Economy recently released the first publicly available forecast model that uses "Google search queries" to forecast unemployment rates for the coming months.

Another avenue is through linking microdata from different databases. We will have to exploit this more - for example to

assess the impact of our EU funding programmes. In DG RTD we are currently exploring how to best link data from our funded projects to microdata on patents, bibliometrics and enterprises to assess the impact of Horizon projects.

Another example of how things are changing is the emergence of <u>Open Science and Open Access</u>.

Open Science is changing the way science is done, making research results and data available "for free" and in real time across a variety of platforms, including social media.

This implies that traditional indicators to assess research quality and impact, such as the number of articles in scientific journals, most likely no longer capture the full picture.

We need to look at other indicators, which are better adapted to this new open context; we call this *Altmetrics*. Altmetrics measure how often journal articles and scholarly outputs are discussed and used through, for example, bookmarks on reference managers like Mendeley or mentions on social networks such as Twitter.

3. Data needs for policy

As we want policy-making to be increasingly evidence-based, the **need for statistical data has grown in the past and is likely to grow further in the future**.

The Commission as a whole is making very intensive use of Eurostat data, which we exploit on an almost daily basis.

But the needs for data are growing and changing. Let me highlight a few aspects of such growing needs.

Data needs are so diverse that they cannot be satisfied via a single source anymore. For our analytical work, in addition to official statistics we use a variety of other sources. For research and innovation policy we use, for example, specialised sources on patents and publications.

This reflects a general trend of an **increasingly diverse data producer landscape**. This makes data provision more complex but also better addresses user needs, especially in relation to real-time data and linking data at the micro-level.

It is also interesting to note that statistical data are such an important content for media, that some have created their

own surveys and scoreboards, such as various university rankings. Yes, there are savvy stories in statistics.

An important question then arises: how can we make sure that **all data we use are of sufficient quality** if we go beyond official statistical sources.

We see a growing need for a quality assessment framework for data sources, with an important "labelling role" for official statistics, including Eurostat.

Another general trend that we should be aware of is the need for a **more holistic coverage** of different policy areas. While we now have fairly good data for example on financial inputs, we still lack data on the impact of such investments.

Such data are important: as public resources are scarce, we need to monitor the returns of public financing in terms of growth and societal well-being.

And we detect a growing need for **disaggregated data**, so that policies can be designed, evaluated and implemented at a more detailed and fine-grained level. Again to quote the case of research and innovation this would include

information on the impact of funding on researchers, research organisations and enterprises.

Here we face the general problem of confidentiality constraints. We are still trying to find an answer to that in the context of the **Multiannual Financial Framework** 2021-2027.

In a fast changing world, we need data on **emerging phenomena** to help improve policy intervention. Surveys and data collection systems need to have some built-in flexibility for covering such new phenomena.

Among the new phenomena we have to measure are even things like Big data and **Artificial Intelligence**. They help us to produce statistics, but we need also statistics on them, such as enterprises involved, investment, employment and impact. And data quality is an issue: we heard in this conference that 40% of apparent AI start-ups are not really AI start-ups.

And we also see a need for better **international coverage** of data, for more detailed **geographic breakdowns** (regional data), a wider coverage of **economic sectors** and better **comparability between countries**, datasets and time series.

In this context we also need more meta-data to be in a position to better assess the comparability and scope of data.

4. <u>Resource constraints/response burden as bottlenecks</u>

All this sounds very ambitious. It is only fair to recognise that **it is difficult to do more with fewer resources**. We realise we have to stay realistic as to what we can ask for in a world of resource constraints that affect the production of official statistics.

In recent years the European Statistical System (Eurostat and the Member States) has faced stagnating or even declining resources in terms of staff and budgets. We note that the System currently counts on a staff of 48 000 and an annual budget of about 3 billion €, both on a decreasing path in the last few years.

Yet, in such context, Eurostat has succeeded in increasing the number of published datasets by 15%, so productivity has grown! We know and appreciate Eurostat is also trying to find new and younger data users, via gamification, again boosting the cost benefit ration of statistical production.

We need to be aware that enterprises often complain about response burden from many lengthy questionnaires from statistical organisations and in addition from private entities, such as market research companies.

Private households are also concerned about confidentiality and privacy issues. The growing diversification of society makes information more difficult to collect via standard questionnaires.

5.   New techniques and technologies can help

Coming back to what I was mentioning before, we believe **big data** should help address the challenges the statistical system is facing.

While it is increasingly complex to collect data from households and enterprises via traditional approaches, the amount of data generated via digital technologies is growing exponentially. There are estimates that 90% of data in the world have been created in the last two years alone.

New techniques and technologies can also help to address the challenge of **bringing costs down**.

Currently the costs for European enterprises of answering to official business surveys amount to about € 700 million per year, while the costs to national statistical offices to produce business statistics increase the total to € 1 billion.

As regards household surveys, costs of collecting information for one household or person amounts to about € 100, or € 180 million per year for all Eurostat household surveys. What is costly is especially face-to-face interviews. A recent Eurostat paper estimated that if 30% of the data could be collected via web-interviews, costs could be reduced by 20%.

In many countries registers of businesses and the population help to contain census costs. An example for the explosion of census costs in the absence of registers is the US. The costs of the US population Census have doubled each decade since 1960, reaching 4.5 billion $ in 2000 and an incredible 13 billion $ in 2010, the most expensive census in the world. The growing costs of the US census triggered already in the 19th century innovations like "punch cards". New solutions like exploiting social media data are being sought for the future.

Coming back to **data timeliness**, there are surveys like the Community Innovation Survey, which is a valuable source for

data on innovation activities of enterprises, but which tends to have a time lag of 3-4 years.

Collecting more timely data on enterprise innovation via for example the scraping of enterprises' websites could help to reduce costs for enterprises and the statistical system. However, the corresponding techniques need to be refined to ensure representativeness of results and comparability. The **EU Horizon programme for research and innovation** is financing several studies on big data and we have heard about several activities during the last 3 days.

6. The Horizon programme

Talking about the Horizon programme, NTTS provides a very useful overview of results from ongoing projects and stimulates the preparation of new ones.

Horizon 2020, with a budget of about € 80 billion in current prices, supports not only new techniques and technologies for statistics but also many activities that generate, compile exploit and analyse existing data.

Among those activities, I would like to mention the calls for proposals that we have issued for the development of new

indicators or the yearly European Innovation Scoreboard and R&D Industrial Investment Scoreboards, but also the work with OECD on enterprise micro-data and public R&D support.

I also like to mention the "**MAKSWELL**" project to extend and harmonise indicators capturing the main characteristics of the "beyond-GDP" approach. There will be a satellite event on this tomorrow.

I encourage statistical institutions to check the Horizon Work Programme 2018-20 for further funding opportunities.

The successor of Horizon 2020, **Horizon Europe**, is proposed to have a budget of € 100 billion and will continue to provide support to emerging techniques and technologies for statistics and new data for better policy making.

At the same time, the **Horizon Programme** is a source of many data. It can tell a lot on how researchers and innovators in Europe behave and on what they deliver. The Horizon 2020 Dashboard is an example of how our microdata can be visualised and communicated to all stakeholders. We will continue to work provide close-to-real time data reporting.

The **interim evaluation of Horizon 2020** and the **impact assessment of Horizon Europe** were drawing on state-of-the-art methods to analyse the programme's data. New techniques and technologies will help us improve the way we monitor and evaluate the programme. We have designed a novel approach to monitor Horizon Europe along "Key Impact Pathways" that should help us measure the scientific, societal and economic impact of European funding. Also there we will integrate new techniques on the collection, processing and analysis of data.

7. <u>So how will the future look like?</u>

I recently came across a statement on a Eurostat website that in the course of the third decade of this century *"**Most if not all data is expected to be organic,** i.e. by-products of people's activities, systems and things, including billions of smart devices connected to the internet".*

In that new context there is a growing need to examine and make use of the potential of "**B-to-G",** business to government data transfer. This involves data from social media, mobile phones, Internet of Things, etc.

There should be a new role for statistical institutions, captured by the term "**smart statistics**". I quote from the same Eurostat NTTS related page: *"**Smart Statistics** can be seen as the future extended role of official statistics in a world impregnated with smart technologies."*

Finally there is the issue of **trusted smart statistics,** again with an important role for official statistics, ensuring not only the validity and accuracy of the outputs, but also respecting data subjects' privacy and protecting confidentiality.

Privacy and confidentiality are a growing concern and we need more research on techniques and technologies helping to avoid misuses of data on individuals and enterprises.

I guess what we will see in the coming years is, however, not one technique replacing existing ones, but a coexistence of and synergies between established and new data sources and techniques, of public and private ones, and of general and specialised providers that complement each other.

This will include traditional questionnaire-based surveys, and administrative data sources, alongside new techniques such as big data. While some of these sources will provide basic

structural information in high quality, others will provide more timely data on key trends. What will be increasingly important is to have rich meta-information and knowledge about the quality of these sources and to guarantee and create **trusted statistics**, including **trusted smart statistics**.

And in all of this we cannot forget the role that **people with the right skills** will play. We saw already in the last few years that there is a strong growth in Europe in the demand for big data analysts and for managers who know how to deal with big data.

This is only expected to grow further. To avoid a skills gap we will have to encourage young people to take up studies in these fields and educational institutions to provide corresponding courses. In the debate around "the future of work" (future technological change might endanger traditional jobs), there is one thing that is certain: the need for data analysts will grow further.

And I guess it is safe to say that they will be increasingly supported by **Artificial Intelligence.** Artificial Intelligence can help to make sense of increasingly large amounts of data, to check the validity and improve their quality, relieving

statisticians from routine tasks. Artificial Intelligence could help us analysing data with greater scope, scale and speed.

In fact, a lot of what I said before and what you have discussed during the conference relates – directly or indirectly - to artificial intelligence – although AI does not seem very prominent on the programme.

Paraphrasing Isaac Asimov's quote about computers, we could say 'I don't fear AI, I fear the lack of it'. And maybe we should especially fear a lack of a European AI. Europe needs to lead on AI and develop AI that respects European values and makes the lives of Europeans better.

The Commission is therefore increasing its annual investments in AI by 70% under the research and innovation programme Horizon 2020. It will reach EUR 1.5 billion for the period 2018-2020, and resources will grow further after 2020.

8. Closing remarks

I think I should stop now. You probably have heard enough about statistics and data by now. The saying goes that you can talk about everything, but not about an hour.

I don't know if somebody has already observed that today, Thursday 14 March, is the 140th anniversary of the birth of Albert Einstein.

I'm sure you know the anecdote of Einstein once writing on a blackboard: "*Not everything that counts can be counted and not everything that can be counted counts*."

If we count all the presentations we heard in the last three days, it's obvious that each of them counted. Let me say that we count on you all and us together to make us better understand, develop and use new techniques and technologies to improve data production and thus also policy making in the future.

Thank you very much for your attention!