UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE 10 March 2013
CONFERENCE OF EUROPEAN STATISTICIANS

WHAT DOES "BIG DATA" MEAN FOR OFFICIAL STATISTICS?

At a High-Level Seminar on Streamlining Statistical Production and Services, held in St Petersburg, 3-5 October 2012, participants asked for "a document explaining the issues surrounding the use of big data in the official statistics community". They wanted the document to have a strategic focus, aimed at heads and senior managers of statistical organisations.

To address this requirement, the High-Level Group for the Modernisation of Statistical Production and Services (HLG) established an informal Task Team* of national and international experts, coordinated by the UNECE Secretariat. The HLG is proud to release this paper to the official statistical community, and welcomes feedback on the content and recommendations.

This paper, and other information about the work of the HLG, can be found online at: http://www1.unece.org/stat/platform/display/hlgbas.

* The members of this Task Team were: Michael Glasson (Australia), Julie Trepanier (Canada), Vincenzo Patruno (Italy), Piet Daas (Netherlands), Michail Skaliotis (Eurostat) and Anjum Khan (UNECE)

I. Introduction

1. In our modern world more and more data are generated on the web and produced by sensors in the ever growing number of electronic devices surrounding us. The amount of data and the frequency at which they are produced have led to the concept of '*Big data*'. Big data is characterized as data sets of increasing volume, velocity and variety; the 3 V's. Big data is often largely unstructured, meaning that it has no pre-defined data model and/or does not fit well into conventional relational databases. Apart from generating new commercial opportunities in the private sector, big data is also potentially very interesting as an input for official statistics; either for use on its own, or in combination with more traditional data sources such as sample surveys and administrative registers. However, harvesting the information from big data and incorporating it into a statistical production process is not easy. As such, this paper will seek to address two fundamental questions, i.e. the *What* and the *How:*

- *What* subset of big data should National Statistical Organisations (NSOs) focus on given the role of official statistics, and;
- *How* can NSOs use big data and overcome the challenges it presents?

2. The What is illustrated by a paper on big data for development by the United Nations Global Pulse[1]:

[1] http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf

(p.6) At the most general level, properly analysed, Big data can provide snapshots of the well-being of populations at high frequency, high degrees of granularity, and from a wide range of angles, narrowing both time and knowledge gaps. Practically, analysing this data may help discover what Global Pulse has called "digital smoke signals" - anomalous changes in how communities access services, that may serve as proxy indicators of changes in underlying well-being.
(p. 12) Official statistics will continue to generate relevant information, but the digital data revolution (p. 9 - "the digitally trackable or storable actions, choices and preferences that people generate as they go about their daily lives") presents a tremendous opportunity to gain richer, deeper insights into human experience that can complement the development of indicators that are already collected.

3. Big data has the potential to produce more relevant and timely statistics than traditional sources of official statistics. Official statistics has been based almost exclusively on survey data collections and acquisition of administrative data from government programs, often a prerogative of NSOs arising from legislation. But this is not the case with Big data where most data are readily available or with private companies. As a result, the private sector may take advantage of the Big data era and produce more and more statistics that attempt to beat official statistics on timeliness and relevance. It is unlikely that NSOs will lose the "official statistics" trademark but they could slowly lose their reputation and relevance unless they get on board. One big advantage that NSOs have is the existence of infrastructures to address the accuracy, consistency and interpretability of the statistics produced. By incorporating relevant Big data sources into their official statistics process NSOs are best positioned to measure their accuracy, ensure the consistency of the whole systems of official statistics and providing interpretation while constantly working on relevance and timeliness. The role and importance of official statistics will thus be protected.

II. Definitions

4. A definition of Official Statistics is provided by the Fundamental Principles of Official Statistics, Principle 1[2] offers the following:

Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation.

5. Behind official statistics, there is the notion that official statistics are statistics that describe a situation, provide a *picture* of a country, its economy, its population etc. When Big data is used as an additional source of information this *picture* needs to be considered.

6. Big data can be defined as a variant of the definition used by Gartner[3]:

Big data are data sources that can be –generally– described as: "high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making."

[2] http://unstats.un.org/unsd/methods/statorg/FP-English.htm
[3] http://www.gartner.com/it-glossary/big-data/

III. Sources

7. The Big data phenomenon makes one realize that our world is now full of data. This cannot be ignored, hence the interest for official statistics. So far there are mainly two different ways that NSOs and International Organizations (IO) produce data: sample surveys and from administrative data sources including registers. The question that needs to be addressed is: how can big data help measure more accurately and timely economic, social and environmental phenomena?

8. In general, large data sources can be classified as follows:

a. Administrative (arising from the administration of a program, be it governmental or not), e.g. electronic medical records, hospital visits, insurance records, bank records, food banks, etc.
b. Commercial or transactional: (arising from the transaction between two entities), e.g. credit card transactions, on-line transactions (including from mobile devices), etc.
c. From sensors, e.g. satellite imaging, road sensors, climate sensors, etc.
d. From tracking devices, e.g. tracking data from mobile telephones, GPS, etc.
e. Behavioural, e.g. online searches (about a product, a service or any other type of information), online page view, etc.
f. Opinion, e.g. comments on social media, etc.

9. Administrative data is one of the main data sources used by NSO's for statistical purposes. Administrative data is collected at regular periods of time by statistical offices and is used to produce official statistics. Traditionally, it has been received, often from public administrations, processed, stored, managed and used by the NSOs in a very structured manner. Can one consider administrative data "Big" in accordance with the definition given above? For the moment the response would be, probably not. Administrative data can become "Big" when the velocity increases, e.g. using extensively administrative data where data is collected every day or every week instead of the usual once a year or once a month.

IV. Challenges

10. The use of Big data in official statistics presents many challenges. This section will discuss the main challenges that fall into the following categories:

a. Legislative, i.e. with respect to the access and use of data.
b. Privacy, i.e. managing public trust and acceptance of data re-use and its link to other sources.
c. Financial, i.e. potential costs of sourcing data vs. benefits.
d. Management, e.g. policies and directives about the management and protection of the data.
e. Methodological, i.e. data quality and suitability of statistical methods.
f. Technological, i.e. issues related to information technology.

Legislative

11. Legislation in some countries (e.g. Canada) may provide the right to access data from both government and non-government organizations while others (e.g. Ireland) may provide the right to access data from public authorities only. This can result in limitations for accessing certain types of Big data described in paragraph 8.

12. It is recognised[4] that:

> *The right of NSOs to access admin data, established in principle by the law, often is not adequately supported by specific obligations for the data holders*.

13. Even if legislation has provision to access all types of data, the statistical purpose for accessing the data might need to be demonstrated to an extent that may be different from country to country.

Privacy

14. Definitions may vary from country to country but privacy is generally defined as *the right of individuals to control or influence what information related to them may be disclosed*. The parallel can be made with companies that wish to protect their competitiveness and consumers. Privacy is a pillar of democracy. The problem with Big data is that the users of services and devices generating the data are most likely unaware that they are doing so, and/or what it can be used for. The data would become even bigger if they are pooled, as would the privacy concerns.

Financial

15. There is likely to be a cost to the NSOs to acquire Big data, especially Big data held by the private sector and especially if legislation is silent on the financial modalities surrounding acquisition of external data. As a result, the right choices have to be made by NSOs, balancing quality (which encompasses relevance, timeliness, accuracy, coherence, accessibility and interpretability) against costs and reduction in response burden. Costs may even be significant for NSOs but the potential benefits far outweigh the costs, with Big data potentially providing information that could increase the efficiency of government programs (e.g. health system). Rules around procurement in the government may come into play as well.

16. One of the findings in the report prepared by TechAmerica Foundation's Federal Big data Commission in the United States[5] was that the success of transformation to Big data lies in:

> *Understanding a specific agency's critical business imperatives and requirements, developing the right questions to ask and understanding the art of the possible, and taking initial steps focused on serving a set of clearly defined use cases.*

17. This approach can certainly be transposed in an NSO environment.

[4] http://essnet.admindata.eu/WorkPackage?objectId=4251 (p41)
[5] http://www.techamerica.org/Docs/fileManager.cfm?f=techamerica-bigdatareport-final.pdf

Management

18. Big data for official statistics means more information coming to NSOs that is subject to policies and directives on the management and protection of the information to which NSOs must adhere.

19. Another management challenge is one related to human resources. The data science[6] associated with Big data that is emerging in the private sector does not seem to have connected yet with the official statistics community. The NSOs may have to perform in-house and national scans (academic, public and private sectors communities) to identify where data scientists are and connect them to the area of official statistics.

Methodological

20. Administrative data presents issues but representativity is the fundamental issue with Big data. The difficulty in defining the target population, survey population and survey frame jeopardizes the traditional way in which official statisticians think and do statistical inference about the target (and finite) population. With a traditional survey, statisticians identify a target/survey population, build a survey frame to reach this population, draw a sample, collect the data etc. They will build a box and fill it with data in a very structured way. With Big data, data comes first and the reflex of official statisticians would be to build a box! This raises the question *is this the only way to produce a coherent and integrated national system of official statistics?* Is it time to think outside of the box?

21. Another issue is both IT and methodological in nature. When more and more data is being analysed, traditional statistical methods, developed for the very thorough analysis of small samples, run into trouble; in the most simple case they are just not fast enough. There comes the need for new methods and tools:

a. Methods to quickly uncover information from massive amounts of data available, such as visualisation methods and data, text and stream mining techniques, that are able to 'make Big data small'. Increasing computer power is a way to assist with this step at first.
b. Methods capable of integrating the information uncovered in the statistical process, such as linking at massive scale, macro/meso-integration, and statistical methods specifically suited for large datasets. Methods need to be developed that rapidly produce reliable results when applied to very large datasets.

22. The use of Big data for official statistics definitely triggers a need for new techniques. Methodological issues that these techniques need to address are:

- Measures of quality of outputs produced from hard-to-manage external data supply. The dependence on external sources limits the range of measures that can be reported when compared with outputs from targeted information-gathering techniques.
- Limited application and value of externally-sourced data.

[6] Wikipedia defines data science as a science that incorporates varying elements and builds on techniques and theories from many fields, including mathematics, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modelling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products.

- Difficulty of integrating information from different sources to produce high-value products.
- Difficulty of identifying a value proposition in the absence of the closed loop feedback seen in commercial organisations.

Technological

23. As discussed in paragraph 9, improving data velocity of accessing administrative data means to also use intensively standard Application Programme Interfaces (APIs) or (sometimes) streaming APIs to access data. In this way it is possible to connect applications for data capturing and data processing directly with administrative data. Collecting data in real time or near real time maximize in fact the potential of data, opening new opportunities for combining administrative data with high-velocity data coming from other different sources, such as:

- Commercial data (credit card transactions, on line transactions, sales, etc.).
- Tracking devices (cellular phones, GPS, surveillance cameras, apps) and physical sensors (traffic, meteorological, pollution, energy, etc.).
- Social media (twitter, facebook, etc.) and search engines (online searches, online page view)
- Community data (Citizen Reporting or Crowd-sourced data).

24. In an era of Big data this change of paradigm for data collection presents the possibility to collect and integrate many types of data from many different sources. Combining data sources to produce new information is an additional interesting challenge in the near future. Combining "traditional" data sources, such as surveys and administrative data, with new data sources as well as new data sources with each other provide opportunities to describe behaviours of "smart" communities. It is yet an unexplored field that can open new opportunities.

V. How Big data can be used in Official Statistics Communities

25. Several examples of Big data studies being conducted or planned are discussed in this section. The first two are from the Netherlands.

26. **Traffic and transport statistics:** In the Netherlands, approximately 80 million traffic loop detection records are generated a day. These data can be used as a source of information for traffic and transport statistics and potentially also for statistics on other economic phenomena. The data are provided at a very detailed level. More specifically, for more than 10,000 detection loops on Dutch roads, the number of passing cars in various length classes is available on a minute-by-minute basis. Length classes enable the differentiation between cars and trucks. The downside of this source is that it seriously suffers from under coverage and selectivity. The number of vehicles detected is not available for every minute and not all (important) Dutch roads have detection loops yet. At the most detailed level, that of individual loops, the number of vehicles detected demonstrates (highly) volatile behaviour, indicating the need for a more statistical approach. Harvesting the vast amount of information from the data is a major challenge for statistics. Making full use of this information would result in speedier and more robust statistics on traffic and more detailed information of the traffic of large vehicles which are indicative of changes in economic development.

27. **Social Media statistics:** Around 1 million public social media messages are produced on a daily basis in the Netherlands. These messages are available to anyone with Internet access. Social media has the potential of being a data source as people voluntarily share information, discuss topics of interest, and contact family and friends. To respond to whether social media is an interesting data source for statistics, Dutch social media messages were studied by Statistics Netherland from two perspectives: content and sentiment. Studies of the content of Dutch Twitter messages (the predominant public social media message in the Netherlands at the time of the study) revealed that nearly 50% of messages were composed of 'pointless babble'. The remainder predominantly discussed spare time activities (10%), work (7%), media (TV & radio; 5%) and politics (3%). Use of these, more serious, messages was hampered by the less serious 'babble' messages. The latter also negatively affected text mining approaches. Determination of the sentiment in social media messages revealed a very interesting potential use of this data source for statistics. The sentiment in Dutch social media messages was found to be highly correlated with Dutch consumer confidence; in particular with the sentiment towards the economic situation. The latter relation was stable on a monthly and on a weekly basis. Daily figures, however, displayed highly volatile behaviour. This highlights that it is possible to produce weekly indicators for consumer confidence and could be produced on the first working day following the week studied, demonstrating the ability to deliver quick results.

28. The following are a list of planned studies under Eurostat's programme of work and include a number of feasibility studies which aim at exploring the potential of Big data for official statistics.

29. **Price Statistics**: The use and analysis of prices collected on the Internet This is a 24-month project starting in January 2013, which will develop an open source advanced scraping software that assists the Consumer Price Index (CPI) specialists in the automated Internet collection of prices. It has similarities to the Massachusetts Institute of Technology (MIT) Billion Prices Project and it will be tested in five countries in order to address the technological and methodological issues. The software which will be developed within the scope of the project will be made available 'as is' to other statistical organisations on request under an open source license (EUPL).

30. **Tourism Statistics**: Feasibility study on the use of mobile positioning data for tourism statistics. A 15-month project is expected to start in January 2013. The study will explore the usefulness of using mobile positioning data for tourism statistics (and related domains) and will assess the strengths and weaknesses. Issues to be studied include access (and continuity of access), trust (of producers and users of statistics), costs, concepts (in translating the existing tourism statistics concept to a new data source) and other methodological topics (e.g. representativeness, sampling within a very large number of observations). The ability of handling large data files held by mobile operators is considered a critical obstacle to overcome if the project is to be successful. The inclusion of this project in the work programme is, among other reasons, based on promising research results in a number of countries.

31. **Information and Communications Technology (ICT) usage Statistics**: Feasibility study on exploiting internet traffic flows for collecting statistics on the Information Society. Under this project Eurostat is aiming at piloting and assessing the feasibility of 'user-centric' and 'web-centric' measurement approaches under a multi-dimensional perspective which

include technical, methodological, cost, legal and socio-political issues. Amongst the other deliverables of this project, which may be valuable for national and international statistical agencies, it is worth mentioning three planned reports: (i) How to develop an accreditation procedure; (ii) Methodological and process implementation Handbook for NSOs and (iii) Testing the concept of 'federated open data'. This concept of 'federated open data' is the counterpart (or supplement) of the so called 'open data' of governments. It refers to a shared sub-set of Big data from private sector entities which will be 'open' for use by NSOs.

32. While NSOs are at an early stage of exploring the potential of Big data for purposes of official statistics, initial evidence suggests that there are three broad areas for experimentation:

- Combining Big data with official statistics.
- Replacing official statistics by Big data.
- Filling new data gaps, i.e. developing new 'Big data - based' measurements to address emerging phenomena (not known in advance or for which traditional approaches are not feasible).

33. The combination potential of Big data with official statistics represents some analogies with what has been done during the last several decades in respect of using administrative data with official statistics. What is likely to be slightly different though, and potentially attractive, is the possibility of applying, more extensively, statistical modelling for combining the two. In doing so, estimates may be obtained that maintain the strong quality properties of official statistics and enhance them with the power of near real time measurements obtained from Big data.

VI. Conclusions/Recommendations

34. This section discusses the conclusions drawn and proposes recommendations as possible next steps. **It is clear that during the next two years there is a need to identify a few pilot projects that will serve as proof of concept** (similar to those discussed under section 5) with the participation of a few countries collaborating. The results could then be presented to the HLG.

35. Big data represents a number of critical challenges and responsibilities for national and international statistical organisations, which relate primarily to methodological, technological, management, legal, and skills issues. **Statistical organisations are, therefore, encouraged to address formally Big data issues in their annual and multi-annual work programmes by undertaking research and pilot projects in selected areas and by allocating appropriate resources for that purpose**.

36. Using enormous amounts of data is not an easy task. Solely through their size alone, getting insight from Big data and ensuring quality can be difficult where the data exploration phase would take considerable more time for Big data compared to other, often more structured, sources of high volume data. As a result, **'new' exploration and analysis methods are required**. The term new is placed in quotes here because many of the methods exist and are already used but are new in the area for official statistics. Three were found particularly fruitful, namely: *Visualization methods*, *Text mining*, and *High Performance Computing.*

37. While a limited number of NSOs are actively engaged with technological aspects of Big data, it is mainly the private sector which leads the work on Big data analytics tools and solutions. Adapting Big data analytics tools and systems to official statistics will inevitably require the involvement of NSOs. **Successful use cases should be brought to the attention of the international statistical community**.

38. Synergies between NSOs and the private sector are not limited to technological issues only. **Collaboration of NSOs with private data source owners is of critical importance** and it touches upon sensitive issues such as privacy, trust and corporate competitiveness, as well as the legislation framework of the NSOs. **National examples in this field, addressing some of the issues of granting NSOs privileged access to private sourced Big data should be part of the priority actions**.

39. To use Big data, **statisticians are needed with a different mind-set and new skills**. The processing of more and more data for official statistics requires statistically aware people with an analytical mind-set, an affinity for IT (e.g. programming skills) and a determination to extract valuable 'knowledge' from data. These so-called "data scientists" can be derived from various scientific disciplines. The Netherlands has good experiences with (PhD) researchers from areas such as Mathematics, Physics, (Bio) Chemistry and Economics/Econometrics together with an affinity for IT.

40. While in the long-run the skills issues for Big data statisticians and 'data scientists' more generally will be addressed through adaptations of university curricula (some universities have already started to offer relevant courses), in the short to medium terms, **NSOs should develop the necessary internal analytical capability through specialised training. International collaboration in this regard would be very beneficial for the official statistical community**.

41. **The HLG should also reflect in the medium term on the drafting of guidelines / principles for the effective use of Big data for purposes of official statistics**. A similar reflection should take place with regard to **a new role that NSOs could play in the future i.e. in terms of *labelling and certification* of statistics derived from Big data** and used for public policy. With the proliferation of Big data into many aspects of life, such a need of a trusted third party is becoming increasingly important; is it NSOs that should assume such a responsibility alone or as members of an independent multi-disciplinary authority?

42. There are a number of different initiatives within the official statistics community relating to Big data. Some examples include a proposal at the 44th Session of the United Nations Statistical Commission for a global conference in autumn 2013, sessions at the World Statistics Congress of the International Statistical Institute in August 2013, a discussion planned for the autumn 2013 "DGINS" meeting for heads of the organisations comprising the European Statistical System, and a proposed Eurostat / UNECE workshop on practical applications of Big data in late 2013 or early 2014. Several events are also planned on the use and integration of geo-spatial data (which are often considered to be Big data) with official statistics. **The HLG should ensure that the outputs of these and similar activities are effectively coordinated and communicated at the strategic level**.

****