United Nations ECE/CES/2019/41

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
Sessions 1 and 2: Accessing new data sources

New data sources for official statistics – access, use and new skills

Note by United Nations Statistics Division

Summary

This document offers a few examples of access to and use of big data and outlines how international collaboration can offer opportunity and support to every national statistical office, which is interested in working with new data sources.

Big data refers to a wide variety of data sources, as shown in the list of use cases by the Big Data Public Working Group of the National Institute of Standards and Technology within the Department of Commerce of the United States and as classified by the United Nations Economic Commission for Europe Big Data Working Group. Access challenges to the various data sources are different, but some common aspects can be derived and can be translated in terms of quality factors in the discovery phase of data acquisition, such as transparency and reliability of the data provider and the completeness of the metadata. Access to proprietary big data (mostly held by private companies) requires weighing the concessions made by statistical agencies in their agreements with data owners against the needs of maintaining independence, assuring privacy and confidentiality and producing high-quality statistics.

These and other considerations were reported by the United Nations Global Working Group on Big Data for Official Statistics to the Statistical Commission in 2017 with the request of incorporating them as appropriate in the updates of the national quality assurance frameworks and the Fundamental Principles of Official Statistics. The recently developed UN Global Platform offers participating institutes access to a wide network of experts, access to big data and access to shared services and advanced methods. This means, that through the platform collaboration statistical offices of least developed countries or small island developing states can also use – for example – satellite data to compile agriculture or environment statistics and indicators.

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.

GE.19-08532(E)

1908532

Contents

I. Introduction

1. The purpose of this seminar is to discuss the opportunities and challenges that statistical offices have encountered in using new data sources; and the skills staff need today and tomorrow to use these data successfully. This paper highlights and brings together a number of initiatives, which have been looking into the issue of using big data for official statistics, of accessing big data and of partnership models with data providers. The paper offers a few examples of access and outlines how international collaboration can offer opportunity and support to every national statistical office, which is interested to access to and use of new data sources.

2. Harnessing big data for public policies and social good is still at the forefront, when it comes to a debate on timelier, more frequent and more granular data for policy making. The topic of big data entered into the agenda of the community of official statistics around 2013/2014 challenged by the recommendations contained in the Data Revolution report. Around that time three new initiatives emerged almost simultaneously at national, regional and global level: (a) the Big Data Public Working Group of the National Institute of Standards and Technology (NIST) within the Department of Commerce of the United States; (b) the United Nations Economic Commission for Europe (UNECE) Big Data Working Group; and (c) the UN Global Working Group (GWG) on Big Data for Official Statistics.

A. Independent Expert Advisory Group on a Data Revolution

3. At the end of 2014, the Independent Expert Advisory Group (IEAG) on a Data Revolution for Sustainable Development gave some key recommendations to the UN Secretary General for actions to be taken in the short term: (i) develop and adopt principles concerning legal, technical, privacy, geospatial and statistical standards to facilitate openness and exchange of information; (ii) share technology and innovation for the common good; (iii) scale investments for statistical capacity development and technology transfer; (iv) mobilize and coordinate global action through a World Data Forum, while brokering global public-private partnerships for data sharing; and (v) create a "SDGs data lab".

4. The IEAG firmly believed that timely, frequent and detailed data will be one of the fundamental elements of the accountability framework for the Sustainable Development Goals (SDGs). Achieving the SDGs means embracing the data revolution. In their report "A World that Counts", they urged the UN Member States and system organizations to dramatically speed up their work in this field.

B. Big Data Public Working Group of the National Institute of Standards and Technology

5. Since 2013, the Big Data Public Working Group[1] of the National Institute of Standards and Technology (US Department of Commerce) has worked with industry, academia and government to establish a consensus-based big data interoperability framework, which would be vendor-neutral and technology- and infrastructure-independent and would enable big data stakeholders (e.g. data scientists, researchers, etc.) to use standard interfaces with swappable architectural components. Among others, this group collected and published a large number of use cases[2] from a variety of domains, such as government, business, healthcare, or environmental sciences. Examples of use cases are

- Preservation of data at the US National Archives and Records Administration

[1] See https://bigdatawg.nist.gov/home.php

[2] See https://bigdatawg.nist.gov/usecases.php

- Fraud detection in the financial industries (banking, securities & investments, insurance)
- Persistent surveillance (object identification and tracking from high-resolution imagery or full motion video) by the US Department of Defense
- Genomic measurements
- Particle physics: analysis of Large Hadron Collider (LHC) data (Discovery of Higgs particle)
- Climate studies using the Community Earth System Model.

6. A common denominator for all these use cases is that each of them deals with the processing of terabytes of data. These use cases demonstrate that technology-wise very large datasets can be successfully exploited. It shows that it is possible to use big data.

7. Another lessons-learned from projects working with very large datasets is that you don't try to transfer those datasets, but rather perform processing at the place where the data are generated. A case in point is the processing of full motion video. The footage is processed and analysed (with specific algorithms) when it comes in and results are being transferred for further analysis. This principle of "edge computing" (or "analytics on the edge") has been explained and promoted in a recent paper[3] by Prof. Bertrand Loison and Prof. Diego Kuonen. They see this approach as the most efficient way to control data quality at the point where the data are created, instead of cleaning up data downstream (and hence centralised), which is more expensive and not scalable. As such, this approach is about *moving the analytics and the data quality frameworks to the data* and not the data to the (centralised) analytics and (centralised) data quality frameworks.

Big data use case template

8. To organize the use cases, the working group developed an extensive template[4], which is shown in abbreviated form in annex II. The organizing elements of the template are:

- Overall project description
- Big data characteristics
- Big data science
- General security and privacy
- Classify use cases with tags
- Overall big data issues
- Workflow processes
- Detailed security and privacy.

9. The big data characteristics contain questions about 4 of the 5 Vs, namely Volume, Velocity, Variety and Variability, whereas big data science covers the Veracity, as well as Data quality, Data types and Metadata. Overall, the template is a very rich and comprehensive tool to describe big data projects, which could also be used to describe projects within the community of official statistics.

C. UNECE Big Data Working Group

10. Under the umbrella of the Conference of European Statisticians the UNECE Big Data Working Group developed guidance on several issues, including a Big Data Quality

[3] See http://www.dgins2018.ro/wp-content/uploads/2018/10/15-Are-current-frameworks-enough-Paper_DGINS_Version_v2.pdf

[4] See https://bigdatawg.nist.gov/_uploadfiles/M0621_v2_7345181325.pdf

Framework and a Classification[5] for Types of Big Data. The UNECE task team on big data proposed the following taxonomy to classify big data:

(a) **Social networks** (human-sourced information): This information is the record of human experiences. Human-sourced information is now almost entirely digitized and stored everywhere from personal computers to social networks. Data are loosely structured and often ungoverned.

- Social Networks: Facebook, Twitter, Tumblr, etc.
- Blogs and comments
- Pictures: Instagram, Flickr, Picasa etc.
- Videos: Youtube, etc.
- Internet searches
- User-generated maps.

(b) **Traditional business systems** (process-mediated data): These processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected is highly structured and includes transactions, reference tables and relationships, as well as the metadata that sets its context. Usually structured and stored in relational database systems. (Some sources belonging to this class may fall into the category of "Administrative data").

- Data produced by Public Agencies
    - Medical records
- Data produced by businesses
    - Commercial transactions
    - Banking/stock records
    - E-commerce
    - Credit cards.

(c) **Internet of Things** (machine-generated data): Derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world. The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed are beyond traditional approaches.

- Data from sensors
    - Fixed sensors
        - Home automation
        - Weather/pollution sensors
        - Traffic sensors/webcam
        - Scientific sensors
        - Security/surveillance videos/images
    - Mobile sensors (tracking)
        - Mobile phone location
        - Cars

[5] See https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data

- Satellite images.

- Data from computer systems
  - Logs
  - Web logs.

1. Big Data Quality Framework

11. The Big Data Quality framework[6] developed by the UNECE Big Data Working Group provides a structured view of quality at three phases of the business process:

- Input – acquisition, or pre-acquisition analysis of the data
- Throughput –transformation, manipulation and analysis of the data
- Output – the reporting of quality with statistical outputs derived from big data sources.

12. The framework is using a hierarchical structure composed of three hyper-dimensions with quality dimensions nested within each hyper-dimension. The three hyper-dimensions are the *source*, the *metadata* and the *data*. The concept of hyper-dimensions has been borrowed from the administrative data quality framework developed by Statistics Netherlands. The hyper-dimension *source* relates to factors associated with the type of data, the characteristics of the entity from which the data is obtained, and the governance under which it is administered and regulated. The hyper-dimension *metadata* refers to information available to describe the concepts, the contents of the file data set, and the processes applied to it. The hyper-dimension *data* relates to the quality of the data itself.

13. At the input phase of the business process, a statistical institute should engage in a detailed quality evaluation of a big data source both before acquiring the data (this known as the 'discovery' component of the input phase), and after (this is the 'acquisition' component). Details of the Big Data Quality framework are shown in annex I.

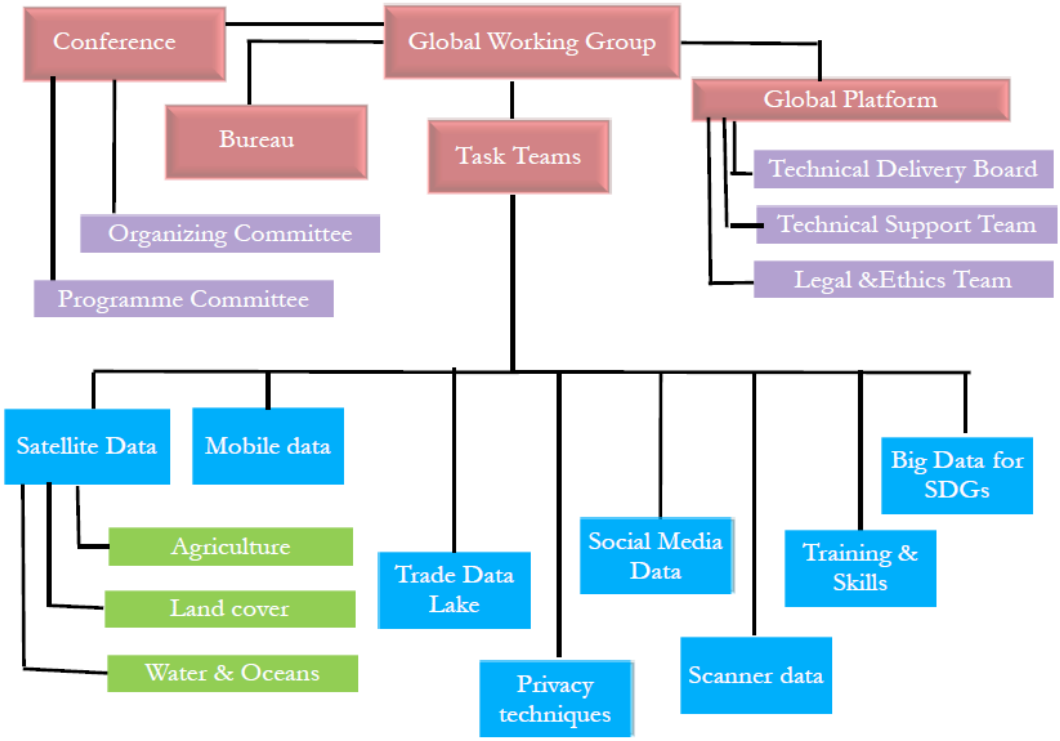
D. UN Global Working Group on Big Data for Official Statistics

14. The UN Statistical Commission created the UN Global Working Group (GWG) on Big Data for Official Statistics in 2014. The GWG provides strategic vision, direction and the coordination of a global programme on big data for official statistics, including for compilation of the SDG indicators for the 2030 Agenda for Sustainable Development. The use of big data is essential for the modernisation of national statistical institutions so that they remain relevant in a fast-moving data landscape. In 2018, the GWG put into place the so-called UN Global Platform, which is a collaborative research and development environment for the global statistical community and all its stakeholder groups. Its platform organization is based on networking and marketplace principles, which facilitates the exchange, development and sharing of data, methods, tools and expertise, and accelerates data innovation. The platform will also be used as an environment for capacity building activities, which includes development of skills sets related to Data Science, since processing of big data requires use of advance technology and machine learning techniques.

15. During the five years since its creation, the GWG has organized its activities through several task teams (such as task teams on the use of satellite imagery data, mobile phone data, social media data, and scanner data and a task team on training, skills and capacity building), a committee for the UN Global Platform, committees for the organization of the international conferences, and associated workstreams and projects. Each of the task teams and committees arranges regular meetings to advance its work. An overview of this elaborate organizational structure is given in figure 1.

[6] See https://tinyurl.com/y5qd8r4m

Figure 1
Overview of organizational structure



II. Examples of access to and use of big data sources

16. In 2015, GWG conducted a big data survey[7] among all statistical institutes. The objective of the survey was to assess the situation regarding the steps undertaken thus far by statistical agencies in relation to big data. It enquired about the strategic vision of national statistical offices and their practical experience with big data. The questionnaire contained questions on the management of big data, advocacy and communication, linking big data with the Sustainable Development Goals, access, privacy and confidentiality, skills and training, and on the most urgent needs of statistical offices regarding the use of big data. In addition, the questionnaire contained detailed questions on big data projects, intended for those offices that had been engaged in one or more projects.

17. Survey results showed that the most used sources are scanner data, satellite imagery and web-scraping data. In particular, it is worth noting that more than 80 percent of the OECD countries have used or considered using web-scraping data and scanner data. Social media and mobile phone data are for the moment much less used, due to a number of factors, in particular issues related to privacy and confidentiality. Non-OECD countries appear to be focusing on trying to make use of satellite imagery and mobile phone data.

18. In 2015, the statistical community expressed the need for guidance in the areas of skills and training, quality frameworks for big data and access to big data. There was also demand for guidance on the usage of specific data sources such as web-scarping and mobile phone data.

19. In the following sections, a few examples are given of access to and use of satellite data, mobile phone data and scanner data.

[7] See https://tinyurl.com/y3vbbu8b

A. Access to and use of satellite data

20. The GWG task team on satellite data completed its Handbook[8] in December 2017, which contains information about sources of Earth observation data, methodologies for producing crop and other statistics using satellite imagery data, outlines of the task team's pilot projects and general guidance for the statistical community exploring the use of earth observation data for statistical purposes.

21. The task team uploaded satellite data, crop survey data, agro-climatic data, crop yield source code and supporting documentation on the UN Global Platform for the purpose of sharing and testing. This can be done relatively easily because some of the satellite data (notably Landsat and Sentinel) are nowadays in public domain. The mentioned datasets and corresponding code and documentation were provided for the purpose of learning.

22. For the purpose of statistics production, we need analysis-ready, timely, more frequent and higher resolution satellite data. To get access to that level of data, we need to establish partnerships with satellite data providers (like Planet, European Space Agency or NASA), as well as with some service providers for the preparation and processing of the satellite data, and with additional service providers for the analysis and visualization of the resulting statistics and indicators. Given that satellite data providers are global data providers, the GWG can make collective service level agreements for access to satellite data on the UN Global Platform.

23. The most likely scenario for access to satellite data on the UN Global Platform is as follows: the data provider (for example Planet) prepares analysis-ready satellite data for the whole world on a weekly basis; these data (or particular parts of these data, for instance satellite data for East Africa) can be accessed by GWG members on the Cloud infrastructure of Planet through services available on the UN Global Platform. Support staff of the UN Global Platform will regularly liaise with technical staff of Planet to check and test data and services. On the global platform itself domain experts and data scientists will work collaboratively on algorithms and APIs to develop and test appropriate methods, and derive agriculture or environment statistics and indicators from the satellite imageries. Given the network of experts on the global platform, also relatively small statistical offices can still engage in the work on using satellite data for official statistics.

B. Access to and use of mobile phone data

24. This task team has also finalized its Handbook[9] on the use of mobile phone data for official statistics, which describes in detail applications, data sources and methods. The Handbook also includes partnership business models between national statistical offices and mobile operators for access to the mobile phone data and concludes with two country cases from France and Indonesia. The team will shortly start on a second handbook, which will include additional applications (on measuring migration, tourism and related statistics) for the measurement of human mobility and will cover the privacy by design, such as the framework of Eurostat.

25. Eurostat presented, in October 2018 in Romania, a paper[10] on public-private partnerships between statistical institutes and mobile network operators. Such cooperation can be designed to prevent potential conflicts between the public and private interests, e.g. by the provision of adequate protection for business confidentiality, methodological quality and process transparency. Synergies between the production of official statistics and commercial analytic products can be positively leveraged within the framework of a well-

[8] See https://unstats.un.org/bigdata/taskteams/satellite/UNGWG_Satellite_Task_Team_Report_WhiteCover.pdf

[9] See https://tinyurl.com/y62zeu3x

[10] See http://www.dgins2018.ro/wp-content/uploads/2018/10/25-DGINS_paper_TSS_architecture_V20_final-1.pdf

designed partnership model. By doing so, such partnership does not represent a risk to the business interests of the mobile network operator, nor a diminution of the role and independency of statistical institute, but rather represents a win-win opportunity for both sides.

26. The approach promoted by Eurostat shows that technological solutions exist that allow eliciting only the desired output information, without requiring the disclosure of input data across different parties. One prominent solution to this class of problems is provided by Secure Multi-Party Computation (SMPC) methods. In a nutshell, SMPC allows processing confidential input data across different institutes (for example, the mobile network operators and the statistical institute) without disclosing the input data or leaking any related information other than the desired output.

27. Such technology has considerably matured in the last decade, making its way out of academic laboratories into commercial products. Pilot projects and early deployments based on SMPC technology are being carried out. Besides the technical aspects, such pioneering activities are contributing to clarify the legal aspects around the use of this technology for the processing of personal data.

28. Two examples of access to mobile phone data: (1) Statistics Indonesia (BPS Indonesia) and (2) Geostat in Georgia.

29. In Indonesia, the national statistical office (BPS Indonesia) has reached an agreement with the largest mobile phone company on use of the mobile phone data for the estimation of domestic tourism statistics and some population statistics. With a user base of about 150 million subscribers the daily number of phone records is very large (about 7.5 billion mobile phone signalling records each day). These records stay on the server of the mobile network operator (MNO) and algorithms are run on that server as well. Two data scientists of BPS Indonesia work inside the MNO for the data processing. The analysis of the results and subsequent refinement of algorithms can be done by a larger team at BPS Indonesia.

30. In Georgia, the national statistical office (Geostat) collaborates with the national mobile network regulators, which receives mobile phone records from the three largest MNOs in Georgia. Experts of the GWG task team on mobile phone data are supporting Geostat in this pilot project. These experts have reached an agreement with the regulator to assist in the pre-processing of the mobile phone data, which remain at the site of the regulator. The regulator will subsequently run the algorithms to derive statistics from the mobile phone records. The GWG has offered to use the UN Global Platform to execute the processing programs, reducing the compute cost for the regulator. The GWG task team is looking into using SMPC to incorporate the mobile phone data in secure form into the platform Cloud.

C. Access to and use of scanner data

31. This is a more recent task team, which was established in April 2017. It hit the ground running by making available the software code and statistical methods developed by Australian Bureau of Statistics, Statistics New Zealand and Statistics Netherlands for the use of scanner data in production of consumer price indices (CPI). These statistical methods and software code were sharable and well-documented. The methods have been reviewed and released as open-source applications using R. The task team made an explicit point of keeping the Inter-Secretariat Working Group on Price Statistics informed about the progress of its work.

32. The task team on scanner data in the calculation of CPI is testing statistical methods and software code, using mostly open-source applications and has documented this also in a handbook. This will allow other statistical offices to experiment and test scanner data for potential use in their statistical production process, along with web scraped and survey data. Scanner data could be obtained nationally from large department stores or supermarket chains.

33. A promising development is also the collaboration with Nielsen, which has made some of its price data available to the global statistical community. Nielsen data constitute a global, standardized and therefore comparable data source, which would allow sharing of trusted methods for the CPI calculation.

34. Nielsen could become a key partner and key data provider for the statistical community. Initially, two commodities derived from scanner data were provided as a proof of concept. As of March 2019, an enduring data set covering five commodities is available on the UN Global Platform. It covers transaction data spanning three years from the Canadian market, broken down by region. This data is available to all staff from GWG members to use for training and testing purposes. Moreover, the FEWS method has been implemented using open source code and is available for use. Other methods (Geary-Khamis, GEKS-T and chained Jevons) will be uploaded shortly. Finally, an instructional guide has been available on the Global Platform since March 2019. It has been peer reviewed by experts from our participating countries.

35. The access model for scanner data is almost always a mixed-mode approach, meaning that only part of the basket of goods and services for CPI calculation will be based on scanner data. These scanner data could be acquired nationally through agreements with major department stores or major supermarket chains, or could be acquired – in the near future – through global agreements with globally operating companies like Nielsen (which is in the business of price data) or Amazon (which has internal logs of price data from about every country in the world). An additional advantage of working with global companies on the UN Global Platform would be the global harmonization of methods in the calculation of CPI, when using scanner data.

III. How does access to big data affect the quality of statistics?

36. As indicated earlier, the statistical process can be viewed from the input phase, the throughput phase and the output phase. Quality issues can occur at each of these phases. Most of the distinct quality issues for using big data are located in the input phase.

37. Within the input phase – as shown in annex I – the *source* and *metadata* dimensions deal with aspects of the data source that may be discoverable prior to actually obtaining the data. For this reason, *source* and *metadata* hyper-dimensions provide the opportunity for assessment before the data is obtained. Such an assessment is sometimes referred to as the ‘discovery’ phase, and can be undertaken, for example, to decide the fitness of the data for its intended use, determine what uses the data might be put to, or how much effort should be expended in acquiring it.

38. Factors to consider regarding the data provider are the expected sustainability of the data provider, its reliability status and its transparency. This will also involve the existing legislation to deal with such data providers and their social perception.

39. Regarding the metadata of the big data source, factors to consider are technical constraints; whether the data is structured or unstructured; whether the metadata is available, interpretable and complete; whether additional resources are required to use the data; the timeliness, periodicity, and changes over time of the big data source; and the transparency and soundness of available methods and algorithms to process the big data.

40. Quality dimensions in the *data* hyper-dimension, on the other hand, can only be assessed once the data is actually acquired. In some cases, the requirements of the national statistical office and the intended use of the data will be known prior to the start of the data quality evaluation. In other cases, potential use of the data will be discovered as the data is explored further. For both cases, at the onset, or as the evaluation progresses, the intended use should be clearly documented. The new data may be useful for new statistics, or improvement of accuracy and relevance of existing statistics.

A. Access to proprietary data – 2017 Global Working Group report to the Statistical Commission

41. In 2017, the GWG reported[11] to the Statistical Commission in some detail on the issue of access to proprietary data, such as mobile phone data, scanner data or social media data. Access to these kinds of big data means in many cases access to data held by private data owners with commercial business interests. That fact raises concerns regarding the agreements that statistical agencies are allowed to make with those data owners, all while maintaining independence, assuring privacy and confidentiality and producing high-quality statistics. The issues to be considered include the equal distribution of the burden across data owners, the cost and effort required to provide data, the role of data in the value propositions of businesses, the achieving of a fair balance between public and business interests and the operational arrangements between statistical agencies and data providers. Such agreements may – as mentioned in the earlier part of this section – need to cover transparency of data provision, provide clear descriptions of metadata and include a public disclaimer on the use of the data to ensure both the quality of official statistics and the public trust.

42. In other words, access to proprietary data not only concerns data quality, but also concerns data privacy, protection and security, and the partnership arrangements made with data providers and other parties involved in the compilation of those data.

43. Whether obtained directly or obtained through a contract with a third-party data provider, data should be collected, analyzed or otherwise used through lawful, legitimate and fair means. In particular, data access (or collection, where applicable), analysis or other use should be in compliance with applicable laws, including data privacy and data protection laws, as well as the highest standards of confidentiality, moral, and ethical conduct. Any data use must be compatible or otherwise relevant, and not excessive in relation to the purposes for which it was obtained.

44. Obtaining access to proprietary data does not necessarily mean obtaining a full data set with full access control. In many practical cases, the statistical office does not acquire the big data in full, but either has the data provider as a partner in the processing of the data, or uses an intermediary to do the pre-processing of the data. Moreover, the hosting, processing and otherwise treatment of the data may require strategic partnerships with technology companies, academia or research institutes. These partnership models should be well documented as they are a crucial part of the quality assurance framework. The following types of partnerships[12] are among those relevant for statistical institutes in engaging with third parties, such as private sector and academia, in relation to the use of big data for official statistics namely (1) cooperative research and development agreements, (2) cooperative agreements, and (3) joint venture partnerships. This last type is a public-private partnership in which a government agency and its business partner jointly plan, invest resources in, and carry out a project to meet an agency mission need and share any revenue generated from the project.

IV. Accessing and using big data require new skill sets

45. A multifaceted transformation of national statistical systems is needed to meet the new data innovation challenges and to reap benefits from using big data. Existing capacity-building programmes should be broadened and possibly be focused on transforming the technology architecture and the workforce, exploiting more big data sources and redirecting products and services. The transformation of the technology architecture should facilitate

[11] See https://unstats.un.org/bigdata/bureau/documents/reports/GWG%20report%20-%202017-7-BigData-E.pdf

[12] See an article by Alexander Kostura and Daniel Castro (of the Center for Data Innovation) titled *"Three Types of Public-Private Partnerships That Enable Data Innovation"*, which can be found at https://www.datainnovation.org/2016/08/three-types-of-public-private-partnerships-that-enable-data-innovation/

the shift from physical information technology equipment on-site towards the introduction of a cloud-computing environment along with the adoption of common services and application architecture for data collection, registers, metadata and data management, analysis and dissemination.

46. This approach should be accompanied by capacity-building programmes that support the progressive diversification of the new skill sets of the staffs of the national statistical systems, ranging from data scientists and data engineers using new multisource data and modern technology, to lawyers strengthening the legal environment, to managers leading the change in corporate culture with a continuously improving quality standard. Those new capabilities should allow for the adoption of a standardized corporate business architecture that is flexible and adaptable to emerging demands and should be process-based rather than product-based, with an increasing use of administrative and big data sources for multiple statistical outputs. In addition, our dissemination and communication strategy should be upgraded and made adaptable to target different segments of users by applying a diverse set of data dissemination techniques, including mobile device applications and data visualization of key findings.

A. Global Working Group task team on training, competencies and capacity development

47. This GWG task team[13] focuses on new competencies and new skill sets, which are needed for the staff of statistical institutes to work with the new kinds of data sources, like big data. The main objectives of this task team are to develop methods and tools for needs identification and assessment of "big data" competencies in national statistical systems (NSSs); to build a competency framework for big data acquisition and processing in the current data landscape; and to identify the existing supply of training of data scientists in academic and research centers. The members of the task team include national and regional statistical offices, development banks and some institutes with specific focus on Data Science, notably the Data Science Campus of the Office of National Statistics (ONS) of the United Kingdom.

48. The goals of ONS's Data Science Campus[14] are to investigate the use of new data sources, including administrative data and big data for public good and to help build data science capability for the benefit of the United Kingdom. A new generation of tools and technologies is being used to exploit the growth and availability of these new data sources and innovative methods to provide rich informed measurement and analyses on the economy, the global environment and wider society. The mission of the ONS Data Science Campus is to work at the frontier of data science and artificial intelligence – building skills and applying tools, methods and practices – to create new understanding and improve decision-making for public good. It defines data science as "applying the tools, methods and practices of the digital and data age to create new understanding and improve decision-making".

49. Data science is a concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyse actual phenomena. Figure 2 illustrates how data science can be seen to sit at the intersection of mathematics, statistics, computer science, machine learning, traditional research and domain expertise. This Venn diagram is adapted from original work of Drew Conway[15].

50. Tom Smith, Director of ONS Data Science Campus, presented recently[16] about "data science as a team sport" at the 5th International Conference on Big Data for official statistics in Kigali, Rwanda. He emphasized that partnerships and knowledge exchange are

[13] See https://unstats.un.org/bigdata/taskteams/training/

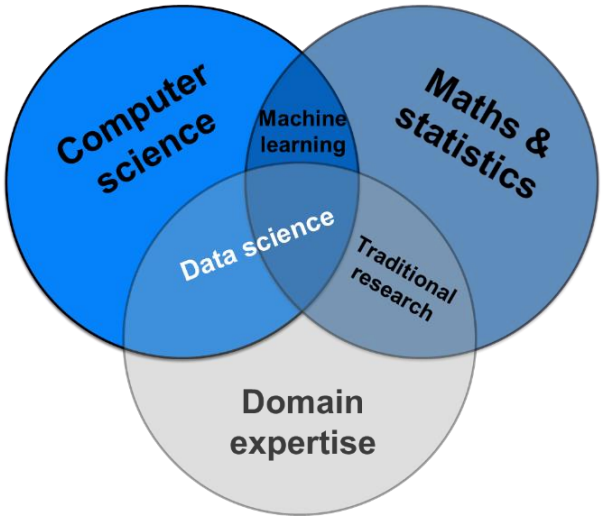
[14] See https://datasciencecampus.ons.gov.uk/about-us/

[15] See for example http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

[16] See Conference agenda, Day 2, session at "Data Science And Capacity Development In Official Statistics", https://unstats.un.org/unsd/bigdata/conferences/2019/default.asp

crucial elements for a successful application of data science. In the United Kingdom, a Government Data Science partnership was created to solve complex business problems using a combination of domain expertise, coding knowledge, machine learning and statistics skills on large and varied datasets.

Figure 2
Data science at the intersection of multiple disciplines



51. Several statistical institutes have started creating data innovation centers to experiment with data science and big data sources. Besides the Data Science Campus of ONS in the United Kingdom, Statistics Netherlands has created a big data center[17] for data innovation in official statistics. China, Republic of Korea and Rwanda have concrete plans to establish such data innovation centers as well. The reasons for wanting to have a data science facility in conjunction with a national statistical office are (1) to harness and exploit large digital datasets and data streams, (2) to develop and test algorithms, which lead to statistics and insights, and (3) to develop new skills in the task force of the statistical office, as well as attracts partner communities to work with the statistical office. Private sector, academia, research institutes and civil society can and are willing to support the statistical community, especially in the new domain of data science for official statistics. Together, significant progress can be made to fulfil the promise of timely, more frequent and more granular data to inform and achieve the sustainable development goals and targets.

V. Conclusions

52. Access to big data is a necessary condition to use big data. As obvious as this sounds, access has been a major challenge for many big data projects, as was one of the main findings in the 2015 global big data survey, conducted by the GWG.

53. Big data refers to a wide variety of data sources, as shown in the list of use cases by the NIST Big Data Working Group and as classified by the UNECE Big Data Working Group. Access challenges to the various data sources are different, but some common aspects can be derived and can be translated in terms of quality factors in the discovery phase of data acquisition, such as transparency and reliability of the data provider and the completeness of the metadata.

54. Access to proprietary big data (mostly held by private companies) requires weighing the concessions made by statistical agencies in their agreements with data owners against the needs of maintaining independence, assuring privacy and confidentiality and producing high-quality statistics. These and other considerations were reported by the GWG to the Statistical Commission in 2017 with the request of incorporating them as appropriate in the

[17] See https://www.cbs.nl/en-gb/our-services/innovation/big-data

updates of the national quality assurance frameworks and the Fundamental Principles of Official Statistics.

55. A few examples of access to and use of big data sources were shown in this paper:

- Satellite data – these data usually remain on the cloud of the data provider; the data need to be prepared to be analysis ready; and additional services are needed to be able to use the data. Frequent and high-resolution satellite data is expensive, and collective bargaining for such global dataset is an advantage; the UN Global Platform can do the bargaining for the global community of official statistics
- Scanner data – for the calculation of CPI only part of the basket of goods and services will consist of scanner data. Many national statistical offices have agreements in place with large department stores and supermarket chains to regularly obtain scanner data; the UN Global Platform can help with expertise on testing price calculation methods; and can help negotiate deals with global companies, like Nielsen and Amazon, on making price data available in a cost-effective manner
- Mobile phone data – in general, mobile phone data will remain with the Mobile Network Operator (as in the case of Indonesia) or with the Mobile Network Regulator (as in the case of Georgia). Processing of mobile phone data is an example of "edge computing". You bring your algorithms to the data, instead of bringing the data to the algorithms.

56. The UN Global Platform, as the collaboration environment for the global statistical community, can help statistical office in getting access to global data sets, to shared services, to tested methods and to a network of experts. The platform can empower relatively small statistical offices to engage in working with big data, new technologies and new methods. The platform organization offers the participating institutes access to a wide network of experts in data science, in handling of satellite data and its services and in advanced methods. This means, that through the platform collaboration statistical offices of least developed countries or small island developing states can also use – for example – satellite data to compile agriculture or environment statistics and indicators.

57. Finally, promising additions to the toolset of the platform are the privacy preserving techniques. These techniques are important to maintain public trust in the handling of sensitive data, such as mobile phone or social media data, and they can be a vehicle to obtain cooperation and support from the national data protection authorities in the efforts of the statistical community to gain access to big data.

Annex I

Big data quality dimensions

Table 1
Big data quality dimension and factions

Hyper-dimension	*Quality Dimension*	*Factors to consider*
Source	Institutional/Business Environment	Sustainability of the entity-data provider
		Reliability status
		Transparency and interpretability
	Privacy and Security	Legislation
		Data Keeper vs. Data provider
		Restrictions
		Perception
Metadata	Complexity	Technical constraints
		Whether structured or unstructured
		Readability
		Presence of hierarchies and nesting
	Completeness	Whether the metadata is available, interpretable and complete
	Usability	Resources required to import and analyse
		Risk analysis
	Time-related factors	Timeliness
		Periodicity
		Changes through time
	Linkability	Presence and quality of linking variables
		Linking level
	Coherence - consistency	Standardisation
		Metadata available for key variables (classification variables, construct being measured)
	Validity	Transparency of methods and processes
		Soundness of methods and processes
Data	Accuracy and selectivity	Total survey error approach
		Reference datasets
		Selectivity
	Linkability	Quality of linking variables
	Coherence - consistency	Coherence between metadata description and observed data values
	Validity	Coherence between processes and methods and observed data values

Annex II

Big data use case template

A. Big Data Public Working Group of the National Institute of Standards and Technology

This template was designed by the NIST Big Data Public Working Group (NBD-PWG) to gather big data use cases. The use case information you provide in this template will greatly help the NBD-PWG in the next phase of developing the NIST Big Data Interoperability Framework. We sincerely appreciate your effort and realize it is nontrivial.

The template can also be completed in the Google Form for Use Case Template 2: http://bit.ly/1ff7iM9.

More information about the NBD-PWG and the NIST Big Data Interoperability Framework can be found at http://bigdatawg.nist.gov.

B. Template outline

1. Overall Project Description

- Use Case Title *

Please limit to one line. A description field is provided below for a longer description.

- Use Case Description *

Summarize all aspects of use case focusing on application issues (later questions will highlight technology).

- Use Case Contacts *

Add names, phone number, and email of key people associated with this use case. Please designate who is authorized to edit this use case.

- Domain (“Vertical”) *

What application area applies? There is no fixed ontology. Examples: Health Care, Social Networking, Financial, Energy, etc.

- Application *

Summarize the use case applications.

- Current Data Analysis Approach *

Describe the analytics, software, hardware approach used today. This section can be qualitative with details given in Section 3.6.

- Future of Application and Approach *

Describe the analytics, software, hardware, and application future plans, with possible increase in data sizes/velocity.

- Actors / Stakeholders

Please describe the players and their roles in the use case. Identify relevant stakeholder roles and responsibilities. Note: Security and privacy roles are discussed in a separate part of this template.

- Project Goals or Objectives

Please describe the objectives of the use case.

2. Big Data Characteristics

Big Data Characteristics describe the properties of the (raw) data including the four major 'V's' of big data described in NIST Big Data Interoperability Framework: Volume 1, Big Data Definition.

- Data Source

Describe the origin of data, which could be from instruments, Internet of Things, Web, Surveys, Commercial activity, or from simulations. The source(s) can be distributed, centralized, local, or remote.

- Data Destination

If the data is transformed in the use case, describe where the final results end up. This has similar characteristics to data source.

- Volume
    - Size: Quantitative volume of data handled in the use case
    - Units: What is measured such as "Tweets per year", Total LHC data in petabytes, etc.?
    - Time Period: Time corresponding to specified size.
    - Proviso: The criterion (e.g. data gathered by a particular organization) used to get size with units in time period in three fields above
- Velocity

Enter if real-time or streaming data is important. Be quantitative: this number qualified by 3 fields below: units, time period, proviso. Refers to the rate of flow at which the data is created, stored, analyzed, and visualized. For example, big velocity means that a large quantity of data is being processed in a short amount of time.

- Unit of Measure: Units of Velocity size given above. What is measured such as "New Tweets gathered per second", etc.?
- Time Period: Time described and interval such as September 2015; items per minute
- Proviso: The criterion (e.g., data gathered by a particular organization) used to get Velocity measure with units in time period in three fields above

- Variety

Variety refers to data from multiple repositories, domains, or types. Please indicate if the data is from multiple datasets, mashups, etc.

- Variability

Variability refers to changes in rate and nature of data gathered by use case. It captures a broader range of changes than Velocity which is just change in size. Please describe the use case data variability.

3. Big Data Science

- Veracity and Data Quality

This covers the completeness and accuracy of the data with respect to semantic content as well as syntactical quality of data (e.g., presence of missing fields or incorrect values).

- Visualization

Describe the way the data is viewed by an analyst making decisions based on the data. Typically, visualization is the final stage of a technical data analysis pipeline and follows the data analytics stage.

- Data Types

Refers to the style of data, such as structured, unstructured, images (e.g., pixels), text (e.g., characters), gene sequences, and numerical.

- Metadata

Please comment on quality and richness of metadata.

- Curation and Governance

Note that we have a separate section for security and privacy. Comment on process to ensure good data quality and who is responsible.

- Data Analytics

In the context of these use cases, analytics refers broadly to tools and algorithms used in processing the data at any stage including the data to information or knowledge to wisdom stages, as well as the information to knowledge stage. This section should be reasonably precise so quantitative comparisons with other use cases can be made. Section 1.6 is qualitative discussion of this feature.

4. General Security and Privacy

The following questions are intended to cover general security and privacy topics. Security and privacy topics are explored in more detail in Section 8. For the questions with checkboxes, please select the item(s) that apply to the use case.

- Classified Data, Code or Protocols
- Does the System Maintain Personally Identifiable Information (PII)? *
- Open publisher; traditional publisher; white paper; working paper
- Data governance refers to the overall management of the availability, usability, integrity, and security of the data employed in an enterprise.
- Transparency and data sharing initiatives can release into public use datasets that can be used to undermine privacy (and, indirectly, security.)
- Current audit needs *
- Under what conditions do you give people access to your data?
- Under what conditions do you give people access to your software?

5. Classify Use Cases with Tags

The questions below will generate tags that can be used to classify submitted use cases. See http://dsc.soic.indiana.edu/publications/OgrePaperv11.pdf (Towards an Understanding of Facets and Exemplars of Big Data Applications) for an example of how tags were used in the initial 51 use cases. Check any number of items from each of the questions.

- DATA: Application Style and Data sharing and acquisition
- DATA: Management and Storage
- DATA: Describe Other Data Acquisition/ Access/ Sharing/ Management/ Storage Issues
- ANALYTICS: Data Format and Nature of Algorithm used in Analytics
- ANALYTICS: Describe Other Data Analytics Used
- PROGRAMMING MODEL
- Please Estimate Ratio I/O Bytes/Flops
- Describe Memory Size or Access issues
- User Interface and Mobile Access Issues
- List Key Features and Related Use Cases

6. Workflow Processes

- Workflow details for each stage *
- Nature of Data: What items are in the data?
- Software Used: List software packages used
- Data Analytics: List algorithms and analytics libraries/packages used
- Infrastructure: Compute, Network and Storage used. Note sizes infrastructure -- especially if "big".
- Percentage of Use Case Effort: Explain units. Could be clock time elapsed or fraction of compute cycles
- Other Comments: Include comments here on items like veracity and variety present in upper level but omitted in summary.

7. Detailed Security and Privacy

- Roles
- Personally Identifiable Information
- Covenants and Liability
- Ownership, Distribution, Publication
- Risk Mitigation
- Audit and Traceability
- Data Life Cycle
- Dependencies
- Framework provider S&P
- Application Provider S&P
- Information Assurance | System Health
- Permitted Use Cases
- Institutional S/P duties
- Curation
- Classified Data, Code or Protocols
- Multiple Investigators | Project Leads *
- Personally Identifiable Information (PII)
- Additional Formal or Informal Protections for PII
- Algorithmic / Statistical Segmentation of Human Populations
- Covenants, Liability, Etc.