United Nations E/CN.3/2019/27

Economic and Social Council

Distr.: General
19 December 2018

Original: English

Statistical Commission
Fiftieth session
5–8 March 2019
Item 4 (e) of the provisional agenda*
Items for information: big data for official statistics

Report of the Global Working Group on Big Data for Official Statistics

Note by the Secretary-General

In accordance with Economic and Social Council decision 2018/227 and past practices, the Secretary-General has the honour to transmit the report of the Global Working Group on Big Data for Official Statistics. In the present report, the Global Working Group responds to the requests made by the Statistical Commission at its last session, in particular to make products and services available to the global statistical community on the United Nations Global Platform of trusted data, methods and learning for official statistics, by addressing concerns on privacy and confidentiality and by providing further details on the business model of the Global Platform. The Global Platform is a research and development environment for sharing and testing trusted methods, trusted data and trusted training materials, and offers technology infrastructure and services for official statistics working in collaboration with the private sector, academia and civil society. Furthermore, the various task teams of the Global Working Group (on earth observations, mobile phones, social media and scanner data, and privacy-preserving techniques) conducted training workshops, prepared handbooks and collaborated on innovative projects. The Commission is invited to take note of the report.

* E/CN.3/2019/1.

18-21303 (E) 070119

Report of the Global Working Group on Big Data for Official Statistics

I. Introduction

1. The Statistical Commission created the Global Working Group on Big Data for Official Statistics at its forty-fifth session, in 2014. In accordance with its terms of reference (see E/CN.3/2015/4) and decision 46/101 of the Statistical Commission (see E/2015/24-E/CN.3/2015/40), the Global Working Group provides strategic vision, direction and the coordination of a global programme on big data for official statistics, including for the compilation of the Sustainable Development Goal indicators in the 2030 Agenda for Sustainable Development.

2. In its decision 49/107 (see E/2018/24–E/CN.3/2018/37), the Statistical Commission reaffirmed that the use of big data and other new data sources was essential for the modernization of national statistical institutions so that they remain relevant in a fast-moving data landscape and highlighted the opportunity for big data to fill gaps, make statistical operations more cost effective, enable the replacement of surveys and provide more granularities in outputs. The Commission endorsed the proposal of the Global Working Group to further develop a global platform as a collaborative research and development environment for trusted data, trusted methods and trusted learning, reiterated the need to present the business case for the platform, encouraged the Global Working Group to build on the success achieved thus far by delivering practical products and services for the global statistical system to support the production of statistics and indicators, including the Sustainable Development Goal indicators, and emphasized the need to carefully address societal challenges of trust, ethics, privacy, confidentiality and security of data.

3. Section II of the present report highlights the annual meeting of the Global Working Group and its Open Day, which were held in Dubai on 21 October 2018. The Open Day offered sessions on the Global Platform of trusted data, methods and learning for official statistics and the work of the various Global Working Group task teams.[1] Section III provides some more information on the Global Platform and the achievements of those task teams, including capacity-building activities, while Section IV outlines the next steps to be taken by the Global Working Group to advance its work programme.

II. Global Working Group annual meeting and Open Day

4. Less time was allocated to the fifth annual meeting of the Global Working Group, than at previous occasions because it was held on the same day as the Open Day on the Global Platform. The annual meeting nevertheless covered the following topics: the organization of the fifth International Conference on Big Data, which is scheduled to be held in Kigali in the spring of 2019; the progress made on the Global Platform; the data management issues to be handled by the Global Working Group, especially in relation to the management of the Global Platform; a brief overview of the achievements of the Global Working Group task teams; and the preparation of the Global Working Group report to the Statistical Commission.

5. In previous years, the Global Working Group held its annual meeting as a full one-day meeting just ahead of its international conference on big data, as was the case in Beijing (2014), Abu Dhabi (2015), Dublin (2016) and, most recently, in Bogotá

[1] A report of the meeting can be found on the Global Working Group website. Available at https://unstats.un.org/bigdata/.

(2017). Reports of those meetings, as well as reports and documents of the meetings of the Bureau of the Global Working Group, can now be found on the website of the Global Working Group.[2] Originally, the fifth annual meeting of the Global Working Group was scheduled to precede the fifth International Conference on Big Data, which, according to the traditional schedule, would have been held in the fall of 2018. Once the second United Nations World Data Forum was announced to be held in Dubai, United Arab Emirates, at the end of October 2018, however, the Bureau decided that the timing of International Conference would move to the spring, beginning with Kigali in 2019. At the same time, however, the annual meeting of the Global Working Group still needed to be held in order to avoid having a 12-month period without a physical meeting of the full membership and was therefore exceptionally organized the day before the start of the United Nations World Data Forum.

6. The two main points of attention for the Global Working Group are the Global Platform and the corresponding data management issues. The Global Platform has evolved from a concept of the Global Working Group into a reality, with delivery of data, methods and learning. In this regard, the Global Working Group needs to more precisely define and agree on the concepts of its four basic pillars: trusted data, trusted methods, trusted partners and trusted learning. This implies agreement on the ownership of and access to the various large data sets on the Global Platform, whether data and algorithms need to be "open" and how software, services and tools will be "Platform independent". These questions have a direct bearing on the business model for the Global Platform. The Global Working Group decided that a paper on the business model and on data management would be developed in close cooperation with, respectively, the Office for the National Statistics of the United Kingdom of Great Britain and Northern Ireland and Statistics Canada.

7. The Open Day on the Global Platform was organized by the Global Working Group and hosted by the Federal Competitiveness and Statistics Authority of the United Arab Emirates. The programme of the Open Day consisted of a demonstration of the Global Platform, followed by sessions on agricultural crop and land cover statistics (using satellite data), measuring human mobility (using mobile phone data), measuring price fluctuations (using scanner data) and privacy-preserving techniques.[3]

8. The Global Platform is open for collaboration on data projects of the global statistical community. It holds an increasing number of data sets, such as Landsat and Sentinel data, trial satellite data from Planet.com, AIS ship positioning data and ADS-B aircraft positioning data. It offers services such as various cloud servers, geospatial analytics services and Jupyter Notebook. The Data Science Campus of the Office for the National Statistics, made use of the Global Platform in studies such as the urban vegetation index in 112 cities in the United Kingdom (using 17 million images from Google Street View) or estimating the proportion of the rural population that lives within 2 km of an all-season road (indicator 9.1.1 of the Sustainable Development Goals) in Northern Ireland (using Open Street Map and population data).

9. The subsequent sessions at the Open Day demonstrated the use of new data sources and new technologies for official statistics, some of which have been using the Global Platform for executing the projects. Over time, the Global Working Group wants more projects to actively use the Global Platform and store the tested methods and data for shared use by others. The statistical office of Colombia estimated, in a pilot project, the yield of cereal crops by applying a deterministic model that incorporates the results of satellite image processing and other data sources. Among

[2] See https://unstats.un.org/bigdata/bureau/.

[3] Details can be found at https://unstats.un.org/unsd/bigdata/conferences/2018/open-day/default.asp under "Agenda".

other things, the office uses algorithms in Google Earth for pre-processing and processing satellite imagery to obtain the Vegetation Condition Index and the Temperature Condition Index, which are needed to estimate crop yield.

10. Statistics Canada improves its crop yield models using satellite data. It tested (for the month of September) the yield estimates for 19 crops using satellite data. For 15 of those crops, the estimates were evaluated to be of sufficiently high quality to be published as official statistics. This implies that crop yield surveys could be gradually replaced by yield estimates based on satellite data. Overall, Statistics Canada concluded that there was a need to accelerate learning, provide an environment in which people could experiment and assess quality. The benefits of the Global Platform might be in facilitating collaboration using trusted methods, data and partnerships.

11. The United Nations Environment Programme (UNEP) demonstrated the Global Surface Water Explorer application, which provides free and open access to national, basin and subbasin aggregated data on water extent and which supports the measurement of indicator 6.6.1 of the Sustainable Development Goals. This can encourage more engagement in places where other sources of data do not exist. Moreover, the global application ensures comparability across time and places and provides a window back in time. This application leverages the best expertise in Earth observation algorithms from around the world.

12. Positium supports the statistical office of Indonesia with improving the quality of its tourism statistics. Using mobile positioning data, gaps were filled for cross-border statistics of tourists coming from neighbouring countries. In addition, the number and destination of trips regarding domestic tourism could be estimated more accurately using mobile positioning data, while increasing frequency and reducing cost.

13. Eurostat is in the process of developing a unified methodological view for the processing of mobile phone data for official statistics. The so-called reference methodological framework will facilitate the interworking of mobile network operators with statisticians at the technical and organizational levels and ensure the consistency, reproducibility and portability of processing methods. Such cooperation will also provide a concrete basis to clarify legal aspects, such as the General Data Protection Regulation, and enable multi-mobile network operator processing and analysis (i.e., fusion of data from different operators).

14. Eurostat proposes to work in three layers: a mobile network operator data layer in which the processing of raw data takes place; a statistics layer in which the statistical methods are applied and processed; and an in-between convergence layer in which the confidential mobile phone data meet the statistics methods. In the convergence layer, statisticians can design algorithms and the mobile network operator can execute those algorithms using secure multi-party computations.

15. International migration is at the forefront of the political debate, which results in an unprecedented demand for migration statistics. The Statistics Division of the Department of Economic and Social Affairs of the Secretariat supports countries in improving capacity in the collection and dissemination of migrant stock and flow statistics and works with a number of project countries, including Georgia, where a State commission on migration issues was established specifically to build a comprehensive national migration data infrastructure, using censuses, surveys and administrative data initially. This work supports all Sustainable Development Goal indicators with respect to breakdowns by migratory status.

16. In addition to those more traditional data, the project has begun to look into using new data sources and technologies and has partnered with the national mobile

network regulator in Georgia, which has access to and processes mobile phone data. Together with international partners, such as Eurostat, the International Telecommunication Union, the International Organization for Migration and Positium, the Division attempts to improve the measurement of human mobility using mobile positioning data and social media data, identifying migration, tourism and commuter statistics.

17. To date, the modernization of the measurement of price indices has focused on the use of scanner data from retailers aiming to increase the effective use of such data in official statistics. The Global Working Group task team in this area wants to deliver a tool hosted on the Global Platform for analysis, monitoring and index estimation using historic scanner data from the information, data and measurement company Nielsen, with accompanying training and instructional material on the use of the tool, methodological guidance material and a catalogue of good practice. Among other things, scanner data provides an opportunity to change the expenditure weights over time. On the Global Platform, national statistical offices now have access to tested index method code and can practise calculating the indices using different methodologies on some training data.

18. Nielsen partners with the Global Working Group. For Nielsen, data are its business, and it collects, enriches and delivers data about what people watch, listen to and buy. Nielsen has also begun to collect e-commerce data globally, with a 76 per cent coverage worldwide by 2018. Nielsen and the Global Working Group have been seeking common ground over the past year and are developing win-win scenarios in their cooperation.

19. Since the beginning of 2018, the Global Working Group has a privacy-preserving techniques task team led by the Office for the National Statistics, which is expected to draft the encryption section of the data policy framework for governance and information management. This task team will develop and propose principles, policies and open standards for encryption within the Global Platform. This will cover the ethical use of data, taking full account of data privacy, confidentiality and security issues when designing methods and procedures for the collection, processing, storage and presentation of data. Among other things, this should reduce the risks associated with handling proprietary and sensitive information.

20. Cybernetica is an active member of the Global Working Group task team on privacy-preserving techniques and demonstrated them for the use of scanner and mobile phone data. Cryptography is used to make the reidentification of data subjects unfeasible and reduce the risk of insider attacks without reducing the accuracy of results. Examples are secure multi-party computation and homomorphic encryption, which Cybernetica successfully applied in cases for the Government of Estonia. Anonymization is another technique that adds noise and makes the reidentification of data subjects harder, but also reduces the accuracy of results. Examples are differential privacy and k-anonymization.

III. United Nations Global Platform and the Global Working Group task teams

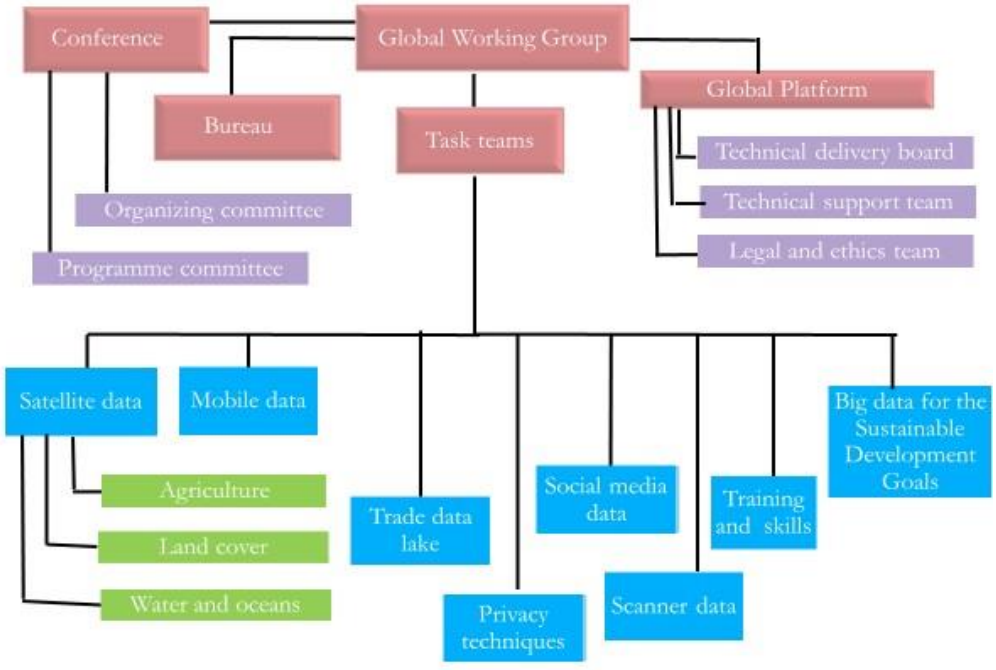
21. In the four years since its creation, the Global Working Group has organized its activities through several task teams, a committee for the Global Platform, committees for the organization of the international conferences and associated work streams and projects. The Bureau of the Global Working Group meets every two weeks to manage and guide the work of the task teams and committees. Each task team and committee has a leading institute, which itself arranges regular meetings

and steers the programme of work. An overview of this elaborate organizational structure is given in the figure below.

A. United Nations Global Platform

22. The Bogotá Declaration (see E/CN.3/2018/8) outlined the details and drivers of the Global Platform as the facilitator for sharing, exchanging and developing trusted data and metadata, trusted methods, trusted partners and trusted learning. The Global Platform will be part of an interconnected and federated network of platforms, be based on the best practices of private and public big data initiatives and offer technology infrastructure for data innovation throughout the community of official statistics. The Global Platform is expected to support capacity-building through a library of trusted training materials, methods and software applications and by conducting workshops on the modernization of official statistics, the use of alternative data sources (e.g., big data) and the application of new tools, services and analytical techniques. The development and maintenance of the Global Platform will be undertaken under the auspices and guidance of the Statistical Commission, in support of the national statistical systems of developed and developing countries.

Figure
Organizational structure of the Global Working Group



23. The Global Platform currently contains several alpha services such as access to Alibaba Cloud, Amazon Web Services, Google Cloud Platform and Microsoft's Azure cloud, combined with a number of other services for code collaboration, methods publishing and Earth observation and location data analysis. Users of the Global Platform can search, build, deploy and consume algorithms and statistical methods and can further develop methods using the main programming languages used by the community (R, Python, Java and Scala). The Global Platform can also host machine-learning models and publish API endpoints to these. Partners on the Global Platform

from around the world can make use of the algorithms from their own environments by calling the APIs. They will also have access to several global data sets, such as the ADS-B flight data dating back to July 2016, AIS shipping data and high-resolution commercial satellite imagery.

B. Global Working Group task teams

24. After delivering its handbook a year ago, the task team on satellite imagery and geospatial data[4] successfully conducted a full five-day training workshop on the use of Earth observation data for statistical purposes and methodologies for estimating crop yield and related statistics using satellite imagery data. Because of the addition of other topics beyond agriculture crop statistics, the task team created three work streams, namely, the estimation of agricultural crop production, land cover and land use statistics and water-related ecosystem statistics.

25. Within the workstream on agriculture statistics, Statistics Canada uploaded to the Global Platform, for the purpose of sharing and testing, its satellite data, crop survey data, agro-climatic data, part of the crop yield source code of Canada and supporting documentation. This collaborative approach of sharing and testing leads to establishing trusted data, methods and learning. This work stream will help to inform indicator 2.4.1 of the Sustainable Development Goals, on the proportion of agricultural land area under productive and sustainable agriculture. A few targeted projects using satellite data to identify specific crops and determine their yield in Canada, Kenya and Rwanda are also foreseen.

26. The work stream on land cover and land use statistics has specific interest in indicator 11.3.1 of the Sustainable Development Goals, on the ratio of land consumption rate to population growth rate, and 15.1.1, on forest area as a proportion of total land area. These indicators include measurements of land cover and land use change and condition over time. The methods used in the measurement are validated through the Committee of Experts on Economic Environmental Accounting. This work stream is looking into executing a project on changes in peatland over time.

27. A third work stream in the task team on satellite imagery and geospatial data focuses on changes in the extent of water-related ecosystems over time (indicator 6.6.1 of the Sustainable Development Goals). As mentioned, a global application is available for the use of satellite data in the estimation of fresh water extent. At the national level, the global application can be adapted slightly for local circumstances, such as was done in Canada owing to frozen water surfaces for large parts of the country in the winter months. This work stream is considering a project on specific water basins in one or more developing countries.

28. With regard to the other Global Working Group task teams, the task team on the use of mobile phone data has finalized the first full draft of its handbook, which details data sources, methods, partnerships models and applications primarily for estimating tourism statistics. The team will soon begin work on a second volume of the handbook, which will include additional applications for the measurement of human mobility and cover privacy by design, such as the framework of Eurostat, described earlier. High priority is given by the team to the execution of a project on measuring migration, tourism and related statistics in Georgia. There are plans to test the methods and algorithms developed during this project in Colombia, Indonesia and Italy.

29. The task team on scanner data in the calculation of consumer price indices is testing statistical methods and software code, using mostly open source applications

[4] See https://unstats.un.org/bigdata/taskteams/satellite/.

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

30. Lastly, as presented during the Global Working Group Open Day, a new task team was created on privacy preserving techniques, which develops methods and procedures for securely processing and exchanging proprietary and sensitive information on the Global Working Group platform among the trusted partners. The team will develop and propose principles and policies for encryption, using open standards and open source algorithms. A handbook is being prepared and expected to become available during 2019.

IV. Next steps

31. The Global Working Group needs to further develop a sustainable business model for the Global Platform as a collaborative research and development environment for the global statistical community. Projects and trusted learning are proofs of concept for the relevance and sustainability of the Global Platform. The Global Working Group will organize and participate in a series of events in 2019 to demonstrate the progress and readiness of the Global Platform.

A. Business model for the United Nations Global Platform

32. The main aspects of the business model for the Global Platform are given below. A background document to this report will provide further details.

Legal entity

33. A legal entity is needed as a vehicle for the overall operation of the Global Platform. The structure and ownership of this entity needs to be worked out. It is expected that the entity will raise and control funding and will be able to underwrite the risks associated with operating the Global Platform. A model, such as that of the UNEP World Conservation Monitoring Centre, might provide a good solution.

Operational entity

34. It is envisioned that the trusted partners of the Global Platform will be both providers and users of products and services. Different parts of the same participating institutes may be both providers and users, in which institutes of official statistics and their partners have a special position. The rules of engagement for different types of partners need to be established. The Global Platform will initially focus on open data sources allowing differential access to more sensitive data over time. The Global Platform is intended to have global 24/7 support. The model may extend to regional hubs for co-development and capacity-building activity.

Funding

35. With substantial funding, the Global Platform could grow very quickly, in which funding is expected to come from development and philanthropic sources. Typical investors may include development funding sources, foundations or philanthropic funding from large technology providers.

B. Proof of concept: projects

36. Almost every task team has begun work on one or more projects that are to be executed on the Global Platform as proof of concept. The following projects, which are in various stages of development, can be pointed out:

(a) Estimation of crop yield using satellite data. A good example of this project is the work done by Statistics Canada, mentioned above, in which it successfully used satellite data for the estimation of the yield of 15 crops. Similar projects are scheduled for African countries;

(b) Measuring changes in land cover and land use. A project is planned that focuses on measuring peatlands;

(c) Measuring the extent of water-related ecosystems. A project is under way for the estimation of fresh water extent in Canada, with possible additional applications in specific delta areas (e.g., the Mekong Delta);

(d) Measuring human mobility. A project has commenced in Georgia to measure migration, tourism, seasonal workers and day-time/night-time population using mobile phone data;

(e) Estimating the consumer price index using Nielsen data. Initial testing of Nielsen data for price index calculations in Canada has been done. Further testing with a larger basket of products and more countries is scheduled for 2019;

(f) Exploratory data analysis using trade and transport data. A data lake with trade statistics, AIS shipping data and ADS-B flight data has been set up on the Global Platform, and testing will take place in 2019.

C. Proof of concept: trusted learning

37. Four two-day training workshops were organized in Bogotá in November 2017, which demonstrated that the Global Working Group was able to offer a course curriculum on the use of new data sources for the compilation of official statistics. A five-day regional training workshop on the use of satellite imagery data for crop and related statistics was organized in Bangkok in June 2018 for the Asia and the Pacific region. A similar offering of training workshops is scheduled for 2019, with shorter training workshops just before the fifth International Conference on Big Data and a five-day regional training workshop on the use of mobile phone data for official statistics, to be held in Jakarta in the first half of June 2019.

38. The task team on training, skills and capacity development has begun a new phase with new deliverables, focusing more on skills development in a changing data environment. Statistics Poland has taken the lead on this. In Europe, it is a long-term goal to have a large pool of statistics graduates with data science skills available throughout the European Statistical System. By 2020, data science skills should be an integral part of official statistics education. The programme on skills development of the Global Working Group will link closely with existing programmes around the world, notably the European Statistical Training Programme, which has been offering courses on the use of big data sources, such as text analytics regarding social media and web searches or use of mobile phone data for official statistics.

D. Events

39. In 2019, the Global Working Group will have several opportunities to showcase the progress of the Global Platform. The Global Working Group will organize the following events:

(a) Side event at the fiftieth session of the Statistical Commission, to be held in New York in March 2019. As mentioned, the Global Working Group will submit a background document to this report laying out the options for the sustainable business model for the Global Platform. During the side event, these options will be explained and open for discussion among the wider statistical community;

(b) The fifth International Conference on Big Data. After Asia and the Pacific, the Middle East, Europe and South America, Africa will host this global conference on big data, giving the countries of this continent the opportunity to demonstrate the use of new data sources and technologies in their compilation of official statistics. As indicated, on the margins of the International Conference, training workshops will be organized on a variety of big data-related topics;

(c) Satellite event on big data and new technologies, to be held in Kuala Lumpur from 15 to 17 August 2019 at the sixty-second World Statistics Congress of the International Statistical Institute. In cooperation with the organizing committee of the Congress, the Global Working Group plans to offer hands-on exercises on the use of satellite data, mobile positioning (training) data, social media (training) data and scanner (training) data on the Global Platform. Statisticians and data scientists of statistical offices will be the target audience.

V. Action required by the Statistical Commission

40. The Statistical Commission is invited to take note of this report.

Annex

Membership of the Global Working Group on Big Data for Official Statistics

Countries

Australia

Bangladesh

Brazil

Cameroon

Canada

China

Colombia

Denmark

Egypt

Georgia

Germany

Indonesia

Ireland

Italy

Mexico

Morocco

Netherlands

Oman

Pakistan

Philippines

Poland

Republic of Korea

Saudi Arabia

Switzerland

United Arab Emirates

United Kingdom of Great Britain and Northern Ireland

United Republic of Tanzania

United States of America

Organizations

African Development Bank

Caribbean Community

Economic and Social Commission for Asia and the Pacific

Economic Commission for Africa

Economic Commission for Europe

Eurostat

Food and Agriculture Organization of the United Nations

International Monetary Fund

International Telecommunication Union

Organization for Economic Cooperation and Development

Statistical Centre for the Cooperation Council for the Arab States of the Gulf

Statistical Institute for Asia and the Pacific

Statistics Division

United Nations Global Pulse

Universal Postal Union

World Bank