United Nations

E/CN.3/2019/30

Economic and Social Council

Distr.: General
5 December 2018

Original: English

Statistical Commission
Fiftieth session
5 - 8 March 2019
Item 4 (f) of the provisional agenda*
Items for discussion and decision: 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 2017/220 and past practices, the Secretary-General has the honour to transmit the report of the Global Working Group (GWG) on Big Data for Official Statistics. In this report the GWG responds to the requests made by the Commission at its last session, in particular by making products and services available to the global statistical community on the UN Global Platform, by addressing concerns on privacy and confidentiality and by providing further details on the business model of the platform. This 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. Further, the various task teams of the GWG (on earth observations, mobile phone, social media and scanner data, and on 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.

I. Introduction

1. The Statistical Commission created the Global Working Group (GWG) on Big Data for Official Statistics at its 45th session in 2014. In accordance with its terms of reference[1] and Decision 46/101 (iii)[2], 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.

2. In Decision 49/107 at its last session in 2018, the Commission reaffirmed that the use of Big Data and other new data sources is essential for the modernisation of national statistical institutions so that they remain relevant in a fast-moving data landscape, and that it creates the opportunity to fill gaps, make statistical operations more cost effective, enable the replacement of surveys, and provide more granularities in outputs. The Commission requested that the GWG further develops the UN Global Platform as a collaborative research and development environment for trusted data, methods and learning, that the business case for sustainability of the platform will be made by 2020, that practical products and services will be delivered for the global statistical system to support the production of statistics and indicators, including the SDG indicators, and that societal challenges of trust, ethics, privacy, confidentiality and security of data will be addressed.

3. Section II highlights the annual meeting of the GWG and its Open Day, which were held in Dubai on 21 October 2018. The Open Day offered sessions on the UN Global Platform and the work of the various GWG task teams[3]. Section III provides some more information on the UN Global Platform and the achievements of the GWG task teams, including capacity building activities, while Section IV describes the next steps to be taken by the GWG to advance its work programme.

II. The GWG annual meeting and Open Day

4. For the fifth annual meeting of the GWG less time was allocated than at previous occasions, because it took place on the same day and just ahead of the Open Day on the UN Global Platform. The annual meeting nevertheless covered 6 topics, namely the organization of the 5th International Conference on Big Data, which is scheduled for the spring of 2019 in Kigali; the progress made on UN Global Platform; the data management issues to be handled by the GWG, especially in relation to the management of the platform; a brief overview of the achievements of the GWG Task Teams; and the preparation of the GWG report to the Statistical Commission.

[1] See E/CN.3/2015/7

[2] See E/CN.3/2015/34

[3] A report of the meeting can be found on the GWG website, see https://unstats.un.org/bigdata/

5. In previous years the GWG 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 Bogota (2017). Reports of those meetings as well as reports and documents of the meetings of the GWG bureau can now be found on the website of the GWG[4]. Originally, the fifth annual meeting of the GWG was planned to precede the 5th international conference on Big Data, which according to the traditional schedule would take place in the fall of 2018. However, once the second UN World Data Forum was announced to be held at the end of October 2018 in Dubai, the GWG bureau decided that the timing of the Big Data conferences would move to the spring, starting with the spring of 2019 in Kigali, Rwanda. At the same time, however, the annual meeting of the GWG still needed to be held – 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 World Data Forum.

6. The two main points of attention for the GWG are the UN Global Platform and the corresponding data management issues. The platform has evolved from a concept of the GWG into a reality with delivery of data, methods and learning. In this regard, the GWG 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 and access to the various large datasets on the 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 platform. The GWG decided that two papers on the business model and on data management will be developed in close cooperation, by ONS, UK, and Statistics Canada, respectively.

7. The Open Day on the Global Platform was organized by the GWG and hosted by the Federal Competitiveness and Statistics Authority of the United Arab Emirates. The program of the Open Day consisted of a demonstration of the UN Global Platform followed by sessions on agricultural crop and land cover statistics (using satellite data), on measuring human mobility (using mobile phone data), on measuring price fluctuations (using scanner data) and on privacy preserving techniques[5].

8. The UN Global Platform is open for collaboration on data projects of the global statistical community. It holds an increasing number of datasets, such as Landsat and Sentinel data, trial satellite data from Planet, AIS ship positioning data, and ADS-B aircraft positioning data. It offers services such as various Cloud servers, geospatial analytics services, Jupyter Notebook and many more. The Data Science Campus of ONS, UK, made use of the UN Global Platform in studies such as the urban vegetation index in

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

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

112 cities in the UK (using 17 million images from Google Street View) or estimating the proportion of rural population who live within 2 kilometers of an all-season road (SDG 9.1.1) in Northern Ireland (using Open Street Map and population data).

9. The subsequent sessions at the Open Day demonstrated use of new data sources and new technologies for official statistics, some of which has been using the platform for executing the projects. Over time, the GWG wants more projects to actively use the platform and store the tested methods and data for shared use by others. The statistical office of Colombia, DANE, estimated in a pilot project the yield of cereal crops, by applying a deterministic model, which incorporates the results of satellite image processing and other data sources. Among others, DANE uses algorithms in Google Earth for preprocessing and processing satellite imagery to obtain the Vegetation Condition Index (VCI) and the Temperature Condition Index (TCI), which are needed to estimate crop yield.

10. Statistics Canada further improves its crop yield models using satellite data. It tested (for the month of September) the yield estimates for 19 crops using satellite data. For15 of those crops the estimates were evaluated as 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 is a need to accelerate learning, provide an environment where people can experiment and assess quality. The benefits of the UN Global Platform might be in facilitating collaboration using trusted methods, data and partnerships.

11. UNEP demonstrated the Global Surface Water application, which provides free and open access to national, basin and sub-basin aggregated data on water extent and which supports the measurement of SDG indicator 6.6.1. 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 observations algorithms from around the world.

12. Positium[6] supports the statistical office of Indonesia with improving the quality of its Tourism statistics. Using mobile positioning data (MPD) gaps were filled for cross-border statistics of tourists coming from neighbouring countries. Further, for domestic tourism statistics the number and destination of trips could be estimated more accurately using MPD while increasing frequency and reducing cost.

13. Eurostat is in the process of developing a unified methodological view for processing of mobile phone data for official statistics. This so-called Reference Methodological Framework will facilitate interworking of the mobile network operators

[6] See https://www.positium.com/

(MNOs) with statisticians at technical and organizational level and ensure consistency, reproducibility and portability of processing methods. Such cooperation will also provide a concrete basis to clarify legal aspects, like GDPR, and enable multi-MNO processing and analysis (fusion of data from different MNO).

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

15. International migration is at the forefront of the political debate, which results in an unprecedented demand for migration statistics. UNSD supports countries to improve capacity in collection and dissemination of migrant stock and flow statistics; and works with a number of project countries, including Georgia. In Georgia, a state commission on migration issues was established specifically to build a comprehensive national migration data infrastructure, using initially census, surveys and administrative data. This works supports all SDGs with respect to breakdowns by migratory status.

16. In addition to those more traditional data, the project has started looking into using new data sources and technologies and partnered with the national mobile network regulator, which has access to and processes mobile phone data. Together with international partners, such as Eurostat, International Telecommunication Union, the International Organization of Migration and Positium, UNSD attempts to improve the measurement of human mobility using MPD and social media data, identifying migration, tourism and commuter statistics.

17. The modernization of the measurement of price indices focused so far on the use of scanner data from retailers aiming to increase the effective use of scanner data in official statistics. The GWG task team in this area wants to deliver a tool hosted on the UN Global Platform for analysis, monitoring and index estimation using historic scanner data from Nielsen with accompanying training and instructional material on the use of the tool, methodological guidance material and cataloguing good practice. Among others, scanner data gives an opportunity to change the expenditure weights over time. On the platform, national statistical offices now have access to tested index method code, and can practice calculating the indices using different methodologies on some training data.

18. Nielsen[7] partners with the GWG. For Nielsen data is its business. Nielsen collects, enriches and delivers data about what people watch, listen to and buy. Nielsen has also started collecting e-commerce data globally with a 76% coverage worldwide by 2018.

[7] See https://www.nielsen.com/us/en.html

Nielsen and the GWG have been seeking common ground over the last year and are developing win-win scenarios in their cooperation.

19. Since the beginning of 2018, the GWG has a Privacy Preserving Techniques task team led by ONS, UK, 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 others, this should reduce the risks associated with handling proprietary and sensitive information.

20. Cybernetica[8] is an active member of the GWG task team on privacy preserving techniques and demonstrated these techniques for the use of scanner and mobile phone data. Cryptography is used to make re-identification of data subjects infeasible and reduce the risks 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 Estonian government. Anonymization is another technique which adds noise makes re-identification of data subjects harder, but also reduces the accuracy of results. Examples are differential privacy and k-anonymization.

III. The UN Global Platform and the GWG task teams

21. During the 4 years since its creation, the GWG has organized its activities through several task teams, a committee for the UN Global Platform, committees for the organization of the international conferences, and associated workstreams and projects. The GWG bureau meets every two weeks to manage and guide the work of the task teams and committees. Each of the task teams and committees has a leading institute, which itself arranges regular meetings and steers the program of work. An overview of this elaborate organizational structure is given in Figure 1.

A. UN Global Platform for official statistics

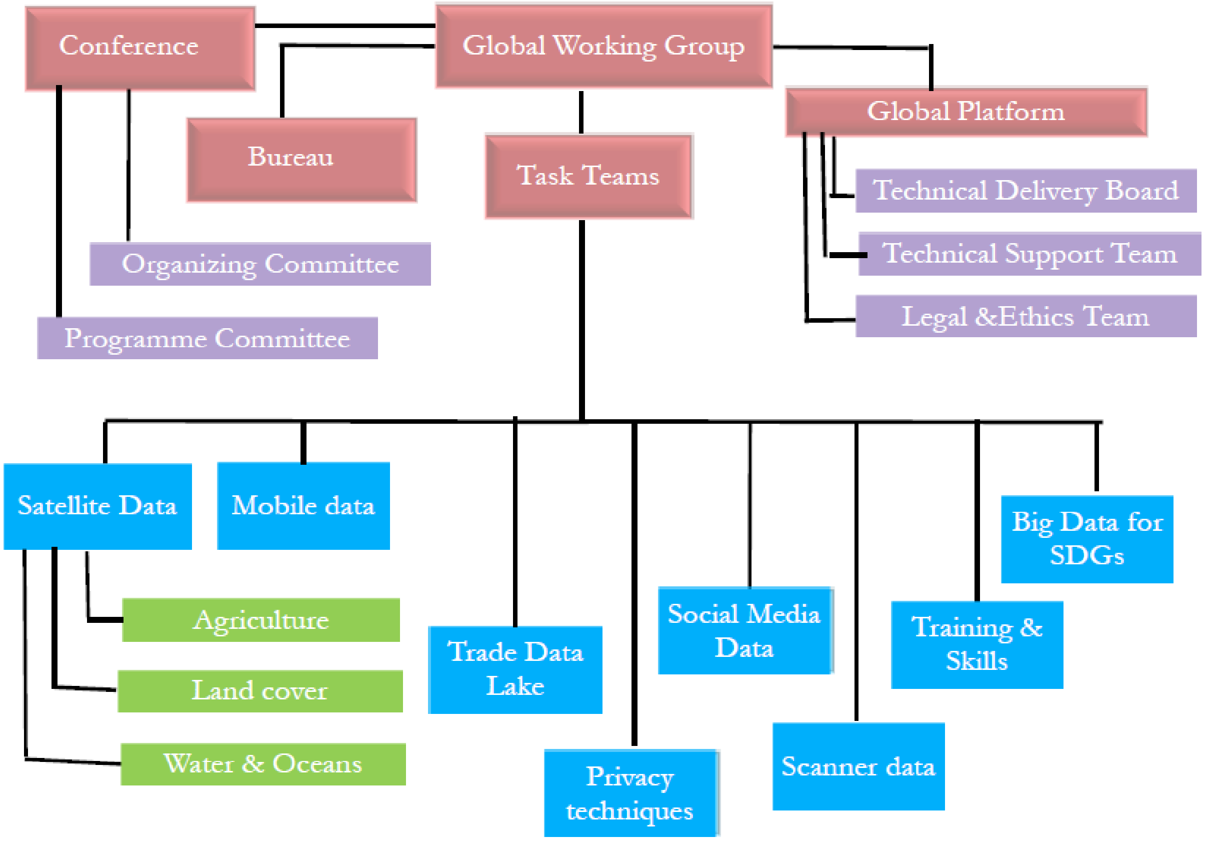
22. The Bogota Declaration[9] outlined the details and drivers of the UN Global Platform as facilitator for sharing, exchanging and developing trusted data and metadata, trusted methods, trusted partners and trusted learning. The platform will be part of an interconnected and federated network of platforms, will be based on the best practices of private and public Big Data initiatives and will offer technology infrastructure for data

[8] See https://cyber.ee/

[9] See https://unstats.un.org/unsd/statcom/49th-session/documents/2018-8-BigData-S.pdf

innovation across the community of official statistics. The platform is expected to support capacity building via a library of trusted training materials, methods and software applications and via conducting workshops on modernization of official statistics, the use of alternative data sources (such as Big Data) and the application of new tools, services and analytical techniques. The development and maintenance of the platform will be undertaken under the auspices and guidance of the Commission, in support of national statistical systems of developed and developing countries.

Figure 1: Organizational structure of the GWG



23. The UN Global Platform currently contains a number of alpha services such as access to the 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, earth observation and location data analysis. Users of the 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 platform can also host machine learning models and publish API endpoints to these. Partners on the platform from around the

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

B. The GWG Task Teams

24. After delivering its Handbook a year ago, the task team[10] on satellite imagery and geospatial data successfully conducted a full 5-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 streams of work, namely 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 on the UN Global Platform – for the purpose of sharing and testing – its satellite data, crop survey data, agro-climatic data, part of Canada's crop yield source code and supporting documentation. This collaborative approach of sharing and testing leads to establishing trusted data, methods and learning. This workstream will help inform the SDG indicator 2.4.1 "Proportion of agricultural land area under productive and sustainable agriculture". It is also foreseen to execute a few targeted projects using satellite data to identify specific crops and determine their yield in Canada, Rwanda and Kenya.

26. The workstream on land cover and land use statistics has specific interest in SDG indicators "11.3.1 Ratio of land consumption rate to population growth rate", and "15.1.1 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 UN Committee of Experts on Economic Environmental Accounting. This team looks into executing a project on changes in peatland over time.

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

28. Of the other GWG task teams, the task team on the use of mobile phone data has finalized the first full draft of its Handbook, which describes in detail data sources,

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

methods, partnerships models and applications mostly for estimating Tourism statistics. The team will shortly start on a second Handbook, which will include additional applications for the measurement of human mobility and will cover the 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. The methods and algorithms developed during this project are planned to be tested in Colombia, Indonesia and Italy.

29. The task team on scanner data in the calculation of consumer price indices (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. A promising development is also 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 sharing of trusted methods for the CPI calculation.

30. Finally, as presented during the GWG 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 GWG 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 is expected to become available during 2019.

IV. Next steps

31. The GWG needs to further develop a sustainable business model for the UN 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 platform. The GWG will organize and participate in a series of events during 2019 to demonstrate progress and readiness of the UN Global Platform.

A. Business model for the UN Global Platform

32. The main aspects of the business model for the UN Global Platform are given hereafter. A background document to this report is giving further details.

- *Legal Entity*

A legal entity is needed as a vehicle for overall operation of the platform. The structure and ownership of this entity needs to be worked out. It is expected that this entity will raise and control funding and will be able to underwrite the risks associated with operating the platform. A model like that of the UNEP-WCMC[11] might provide a good solution.

- *Operational entity*

It is envisioned that the Trusted Partners of the platform will be both 'providers' and 'users' of products and services. Different parts of the same participating institutes may be both providers and users, where 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 platform will initially focus on Open Data sources allowing differential access to more sensitive data over time. The platform is intended to have global 24/7 support. The model may extend to regional hubs, for co-development and capacity building activity.

- *Funding*

With substantial funding the platform could scale very quickly, where 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

33. About every task team has started work on one or more projects, which are to be executed on the UN Global Platform as proof of concept. The following projects – which are in various stages of development – can be pointed out.

- *Estimation of crop yield using satellite data*. A good example of this project is the mentioned work done by Statistics Canada, in which it successfully used satellite data for the estimation of yield of 15 different crops. Similar projects are scheduled for African countries.
- *Measuring changes in land cover and land use*. A project is planned which focuses on measuring Peatlands.
- *Measuring extent of water-related ecosystems*. A project is underway for the estimation of fresh water extent in Canada with possible additional applications in specific delta areas (such as the Mekong delta).
- *Measuring human mobility*. A project has started in Georgia to measure migration, tourism, seasonal workers and day-time / night-time population using mobile phone data.

[11] See https://www.unep-wcmc.org/

- *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 and more countries is scheduled for 2019.
- *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 platform and testing will take place during 2019.

C. Proof of concept – Trusted learning

34. Four 2-day training workshops were organized in Bogota in November 2017, which demonstrated that the GWG is able to offer a course curriculum on the use of new data source for the compilation of official statistics. A 5-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 Big Data conference in Kigali, Rwanda and a 5-day regional training workshop on the use of mobile phone data for official statistics to take place in the first half of June 2019 in Jakarta.

35. The task team on Training, Skills and Capacity Development has started 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 across the European Statistical System. By 2020, data science skills should be an integral part of official statistics education. The program on skills development of the GWG will link closely with existing programs around the world, notably the European Statistical Training Program, which has been offering courses in the use of Big Data sources, such as text analytics of social media and web searches or use of mobile phone data for official statistics.

D. Events

36. The GWG will have a number of opportunities in 2019 to showcase the progress of the UN Global Platform. The GWG will organize the following events.

(i) Side-event at the 50th session of the Statistical Commission in New York in March 2019. As mentioned, the GWG will submit a background document to this report, which lays out the options of the sustainable business model for the platform. During the side-event the options of the business model will be explained and are opened for discussion with the larger statistical community.

(ii) 5th international conference on Big Data for official statistics from 29 April to 2 May 2019 in Kigali, Rwanda. After East Asia, the Middle East, Europe and South America, Africa will be the place where this global conference on Big Data will be taking place, giving the countries of this continent the opportunity to demonstrate use of new data sources and technologies in their compilation of official statistics. As indicated, in the margins of the conference training workshops will be organized on a variety of Big Data related topics.

(iii) Satellite event on Big Data and new technologies on 15-17 August 2019 at the 62nd World Statistics Congress of the ISI in Kuala Lumpur. In cooperation with the organizing committee of the 62nd WSC, the GWG plans to offer hands-on exercises in the use of satellite data, mobile positioning (training) data, social media (training) data, and scanner (training) data directly on the UN Global Platform. Statisticians and data scientists of statistical offices will be the target audience.

37. The Commission is invited to take note of the report.

ANNEX I –

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
United Republic of Tanzania
United States of America

Organisations

African Development Bank
CARICOM
Eurostat
FAO
GCC-STAT
International Telecommunication Union
International Monetary Fund
Organization for Economic Cooperation and Development
United Nations Economic and Social Commission for Asia and the Pacific
United Nations Statistical Institute for Asia and the Pacific
United Nations Economic Commission for Europe
United Nations Economic Commission for Africa
United Nations Statistics Division
United Nations Global Pulse
Universal Postal Union
World Bank