UNITED NATIONS

DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS

STATISTICS DIVISION

United Nations Statistical Commission
Global Working Group on Big Data for Official Statistics
Second Meeting, 19 October 2015, Abu Dhabi, United Arab Emirates

Report

Agenda

(i) Results of the Global Survey

(ii) Progress report and evaluation of the work of the Task Teams

(iii) "How can we produce Official Statistics with Big Data and what is the role of the GWG to make this happen?"

(iv) Evaluation of the working methods

(v) Report to the Statistical Commission

Participants

Countries: Mr. Trevor Sutton (Australia), Mr. Siu-Ming Tam (Australia), Mr. Alessandro Alasia (Canada), Ms. Shu Jiang (China), Ms. Julieth Solano Villa (Colombia), Ms. Mara Bravo Osorio (Colombia), Ms. Sandra Rodriguez (Colombia), Mr. Khaled El Dib (Egypt), Mr. Edward Lambe (Ireland), Mr. Paolo Righi (Italy), Ms. Hae Ryun Kim (Korea Republic), Mr. Juan Muñoz (Mexico), Mr. Peter Struijs (Netherlands), Mr. Khalifa Abdullah Al-Barwani (Oman), Mr. Ahmed Al-Mufarji (Oman), Mr. Candido Astrologo (Philippines), Mr. Deogratius Malamsha (Tanzania), Mr. Sufyan Daghra (UAE), Mr. Brian Moyer (United States)

Organizations: Mr. Albrecht Wirthmann (Eurostat), Ms. Elham Saleh (GCC-Stat), Ms. Susan Teltscher (ITU), Mr. Paul Schreyer (OECD), Mr. Steven Vale (UNECE), Ms. Margarita Guerrero (UNESCAP), Mr. Arman Bidarbakht-Nia (UNESCAP), Mr. Marko Javorsek (UNESCAP), Mr. Robert Kirkpatrick (UN Global Pulse), Mr. Gregory Scott (UNSD), Mr. Karoly Kovacs (UNSD), Mr. Ronald Jansen (UNSD), Ms. Amparo Ballivian (World Bank)

Observers: Mr. Arnold Dekker (Earth Observation)

Minutes

(i) Results of the Global Survey

a. Overview

Ronald Jansen (UNSD) presented the Global Survey on Big Data, which was conducted in June-August 2015. The questionnaire was sent to all statistical offices with the request to consult with all stakeholders in the national statistical system. A total of 91 replies were received from 32 OECD countries, 58 non-OECD and Eurostat. The structure of the questionnaire consisted of 18 questions on Big Data management and then 24 questions per Big Data project. A total of 115 Big Data projects were reported, of which 89 by OECD and 22 by non-OECD countries plus 4 by Eurostat.

b. Results

Statistical offices see 'faster statistics', 'reduction of response burden' and 'modernization of statistical production' as the main benefits of engaging with Big Data. Meeting the new SDG demands was not seen as a driver, especially not by the OECD countries. Government institutes are the main partners in Big Data projects, followed by Satellite providers and the research communities. Most used data sources were scanner data, satellite data and web-scraping data. However, scanner and web-scraping data is used much more in OECD than in non-OECD countries. The Big Data are mostly used for Price, Population and Labour statistics. Finally, urgent need for guidance was seen for 'skills and training, 'quality framework', 'access to Big Data" and 'estimation methods'.

(ii) Progress reports of Task Teams

a. Task Team on Advocacy and Communication

Marko Javorsek (UNESCAP) presented the Advocacy and Communication Strategy paper and described the outputs of the Task Team so far: Brochure, Videos and Newsletter. Marko encouraged the other GWG members to have a close look at the Strategy paper, so that it could be adopted by the GWG as a whole. The main objective of the strategy is to effectively communicate the opportunities and challenges of Big Data for official statistics and advocate the importance of active engagement and collaboration of NSOs with key public and private stakeholders. Specific objectives are to convince the statistical community and policy makers of the benefits, raise awareness of relevant issues, such as modernization of statistical systems and develop alliances and partnerships. The following points were raised as well: statisticians are not advocacy and communication experts, and communication professionals need to be engaged; advocacy needs to be directed towards owners of big data (private sector); and for effective advocacy we need concrete examples of how Big Data are used in production of official statistics, which are not widely available at the moment.

b. Task Team on Access and Partnerships

Peter Struijs (Netherlands) presented the deliverables of this Task Team, namely Good practices for data access and partnerships, a Template for MOUs with global data providers

and the draft Principles of access to Big Data sources. This "Good practices" document is based on analyses of practices in data access and partnership provided by statistical organizations and study of literature. It provides a list of goods practices (not exhaustive) that are relevant when establishing partnerships in relation to Big Data project including considerations when acquiring data access from Big Data providers. Where possible, recommendations are formulated. These include recommendations on building partnerships aimed at developing tools and skills that can facilitate data access. The "Template for MOUs" document is based on analyses of current partnership agreements and templates provided by statistical organizations. It provides a list of items that are relevant when engaging into a partnership agreement with data providers and a template of a partnership agreement which can be adjusted to the specific purpose of a partnership between a data provider and a statistical organization. Finally, Peter mentioned that the draft Principles for access to Big Data sources had been posted on the Conference website and would be presented in detail at the session on Access and Partnerships.

c. Task Team on Big data and SDGs

Amparo Ballivian (World Bank) gave a presentation on the work done by her Task Team. She presented the consolidated inventory of Big Data projects, which includes projects from UNECE, the Sandbox, the World Bank, Global Pulse and Positium. The Task Team has started mapping each project to one or more of the SDG targets. Going forward, the Task Team agreed to join forces with the Task Team on Cross-cutting issues to put the consolidated inventory in a database format, with searchable capabilities and visualization tools. The Task Team also did a survey of NSOs and academic institutions to uncover the use of Big Data for estimating SDG targets, the advantages and the constraints and Amparo presented the results.

d. Task Team on Cross-cutting issues

Paolo Righi (Italy) gave a presentation of the three documents provided by the members of the Task Team, namely a paper on "Revision and further development of the Classification of Big Data", a paper on "Further developing work on quality and methodological frameworks and analytical tools" and a note on "Maintenance and update of the repository of Big Data projects". The classification paper is a scoping paper, which proposes that a new classification of Big Data sources be constructed, building on and replacing the 2013 UNECE Classification of Big Data for Official Statistics. The construction of the classification is to be based on the intended uses. Several uses of the UNECE classification were mentioned, and other needs were also identified. Future uses may include informing the discussion on the use of Big Data for compiling SDG indicators. The "Quality" paper defines the objectives for the work to be done in this area. The overall deliverable is planned to summarize the convergence of a bottom-up (examples) and top-down (theory) approach. The potential output should test and strengthen the work already done on the Quality framework and organise a Methodological framework according to each specific Big Data source. The latter output seems to be innovative in the Big Data field and for such a reason further feedback based on practical applications will contribute to improve the deliverable with further versions. Finally, the note on the repository sets the scope

e. Task Team on Training, Skills and Capacity-building

Marge Guerrero (UNESCAP) gave an overview of the terms of reference and methods of work of the Task Team. The presentation focused on the results of the Global Survey pertaining to Skills, Training and Capacity-building and related results from the 2014 UNECE Survey as these, together with a follow-up focused survey on skills, would be the bases for the future work of the Team. The Survey showed that all respondents pointed out that *guidance on Skills and training for Big Data* is a priority for their offices. Respondents emphasized skills relating to methods rather than the tools or technologies required, namely: methodologist, data scientist and mathematical modelling specialist. However, in the UNECE survey results, skills relating to creative problem solving were mentioned. Even though new tools for working with big data are widely available, the majority of respondents used traditional technologies and tools such as R or relational databases. This can be seen as a practical approach since mastering Big Data tools may require time and investment that national statistical institutions do not have. Moreover, partnerships with data providers or research institutes often result in arrangements where processing of raw data is handled by the partners. The derived secondary and often smaller data sets can be analysed by statistical offices using traditional tools and technologies.

The core elements of the 2016-2017 work programme for the Task Team, as presented, included: raising awareness by sharing experiences, pilot projects and knowledge among countries; development of web-based knowledgebase and document sharing platform; sharing knowledge among countries; further identifying skills, both methods and tools, including the project management skills; conduct of a follow-up focused training and capacity building survey; and stepping-up development of training curricula and courses to address skills gap.

f. Task Team on Mobile Phone Data

An overview was presented on the mandates, deliverables and progress of this Task Team, as well as the session at the Conference and the results of the Global survey. The Task Teams set out to clarify which kind of mobile phone data can be collected, how it can be collected, how it can be analyzed and processed into statistics, useful for policy purposes. The deliverables are one or more pilot projects on the use of mobile phone data for official statistics, documenting and publishing the lessons learned from those projects, maintaining a knowledge base on the use of these data, developing tools and applications, as well as templates of project plans.

Pilot projects are ongoing in Oman and at Eurostat, and a knowledge base can be built from the gathered lessons learned from projects on Tourism, Population and Population mobility statistics and on Poverty estimates. Most of those topics were put on the programme of the Big Data Conference. The results of the Global Survey showed 17 Mobile Phone projects, mostly on Tourism and Population statistics and mostly for exploration. About half of those projects claimed that the mobile phone data covered the whole country and most of the market. Data visualization and Hadoop were the most used tools to process and analyse these data.

To speed up the pace of delivering by this Task Team the leadership was handed over to Oman and Eurostat. As mentioned these two institutes have ongoing pilot projects using mobile phone data.

g. Task Team on Social Media Data

In a very similar manner to the overview provided for the mobile phone Task Team, the overview of the Task Team on Social Media data covered mandates, deliverables and progress of this Task Team, as well as the session at the Conference and the results of the Global survey. The objectives and deliverables of this Task Team are identical to the mobile phone Task Team, except that the focus is on project using Social media data.

Pilot projects for this Task Team are ongoing in Mexico, in the Sandbox in Ireland and with UN Global Pulse, and a knowledge base is being prepared starting from an inventory of ongoing activities (covering more than the mentioned pilot projects). The programme of the Big Data Conference covered presentations by Statistics Netherlands, Mexico, Global Pulse and the Sandbox (done by Eurostat). The results of the Global Survey showed only a few (6) Social Media data projects, but many (25) web-scraping projects. Whereas none of the Social Media projects were intended to go towards production, about half of the web-scraping projects claimed that the intention was to use the results in the production of the official statistics. Hadoop was by far the preferred tools to process Social media data, whereas a large variety of mostly traditional tools were used to process and analyse web-scraping data.

To speed up the pace of delivering by this Task Team the leadership was handed over to Mexico with close support from UNSD.

h. Task Team on Satellite imagery data

Siu-Ming Tam (Australia) provided an overview on the work of the Task Team on Satellite Imagery, Geospatial Data and Remote Sensing. The overall objective is to utilize satellite imagery and geo-spatial data for official statistics and indicators for post-2015 SDGs by identifying the most reliable and accurate statistical methods for estimating quantities of interest, suggesting approaches for collecting representative training data of sufficient quality, researching, developing and implementing assessment methods for predictive models including measures of accuracy and goodness of fit, and by establishing strategies to reuse and adapting algorithms across topics and to build implementations for large volumes of data. Some of the challenges, which the Task Team is working on, are pre-processing of satellite imagery data, gathering ground truth data, image processing and setting of priorities and resourcing. The Task Team will complete and document some pilot projects in 2015 and will report on this to the UN Statistical Commission in 2016. As an aspirational goal it has set a "turnkey" statistical system for predicting crop classification and crop yields.

(iii) Discussion on Role of the GWG

Peter Struijs and Ronald Jansen introduced the discussion note on *"How can we produce Official Statistics with Big Data and what is the role of the GWG to make this happen?"*

The Global Working Group has its work divided over 8 task teams. Progress has been made, but a general and fair conclusion would be that we are still very much in an initial phase. Guidelines for the three specific categories of Satellite imagery, Mobile Phone data and Social Media data are at a beginning stage, initial inventories of projects have been made by the Big Data and SDGs team and the Cross-cutting team, some documents have been prepared on advocacy, principles for access, good practices for data access and partnerships, MOU templates, classifications, quality and methodology and project repository. So work is ongoing, but how much more needs to be done to move from pilot to production, and from examples to guidance. What are the factors determining our speed of progress, and what time schedule can we realistically follow? This discussion note is giving some of those factors as input to the discussion.

a. Factors determining our speed of progress

A number of external factors were mentioned that we need to be aware of while trying to move forward on using Big Data for official statistics, such as support from top level management, having institutional arrangements for Big Data and new techniques, enough budget, support from external stakeholders, IT environment or engagement in strategic partnerships. Factors determining the progress of the Task Teams themselves are more of a practical nature, and having a full time Programme Manager would be essential to move the work forward.

b. GWG deliverables for which a time schedule may be needed

So what are the deliverables of the GWG and when do we need to deliver? Again, quite a number of deliverables were listed in the discussion note, such as Guidelines for Satellite imagery, Guidelines for Mobile Phone data, Guidelines for Social Media data, Principles for Big Data access, Classification of Big Data sources, Use of Big Data for SDG indicators and Big Data training modules. Also a number of tools and applications were mentioned, such as the Sandbox, workshops and other capacity building activities, e-learning modules, Big Data projects repository, Big Data Quality framework and Big Data methodology.

c. Conclusions on role of GWG

A rich discussion took place, which could be summarized into the following conclusions:

- The GWG needs to be engaged in pilot projects. It was mentioned to be engaged in about 5 pilot projects, of which 2 or 3 with application to the SDG indicators.
- Big Data projects are almost by definition based on multi-disciplinary teams with involvement from private sector, research institutes and possibly others. The GWG needs to make compelling business cases, why these partnerships are win-win situations for all involved. Further work on data access needs to be done in that context.
- The repository / inventory of Big Data projects is one of the main deliverables of the GWG, which should serve the statistical community but also a wider audience. This repository should become available in the short-term.
- The GWG should have special focus on the SDG indicators and should have special focus on developing countries. In this sense the opportunities offered through the

Sandbox in Ireland and the Global Pulse labs in Jakarta and Uganda should be fully take advantage off. Pilot projects (involving SDG indicators) can be hosted in those environments and developing countries can participate. Learning by doing is the slogan within this context.

- Looking forward to the 3rd Conference in Ireland, the main themes would be (i) Big Data and SDG indictors, (ii) Access and partnerships: what are the win-win scenarios, and (iii) modernization of statistical production: how does Big Data fit in and how do we build skills and capacity to use it. It was further suggested to conduct one or more workshops in the margins of the Conference on for instance (a) Big Data project management and (b) use of IT tools and technologies.

(iv) Evaluation of Working Methods

The final discussion of the GWG meeting was on the working methods: what are the lessons learned of the last year? What do we need to do to be (even) more effective? A number of ideas and concrete proposal were provided, namely

- We need to be better informed about what is going on the other Task Teams. This can be done via Trello, but is also a responsibility of the Coordination team to discuss the cross-cutting issues at its meetings and circulate the outcomes among all GWG members.
- More cooperation an integration of work among the Task Teams is necessary to increase the speed of progress and for improving our outputs.
- In a similar sense, we need cross-fertilization of the lessons learned in the pilot projects. To make this effective we need a (small) pilot coordination group, which compares notes.
- Task Teams may want to update (if necessary) their TOR. UNSD will collect these and circulate them to have full transparency on what each team is doing.
- ABS offered a full-time Programme Manager for the GWG. This person can reinforce the sharing of information among the Task Teams.
- The GWG could organize its work more along the lines of the Big Data Quality framework, namely regarding Input quality (Data Access), Throughput quality (Integration of TTs, Repository and Statistical production) and Output quality (application to SDG indicators). It is still a point of discussion when the GWG would introduce a rearrangement of Task Teams.
- The Mobile Phone task Team will have Oman and Eurostat as new leaders, and the Social Media task team will be led by Mexico with support from UNSD.

(v) Report to the Statistical Commission

The Global Working Group on the use of Big Data for Official Statistics is required to submit a report to the Statistical Commission at its forty-seventh session in 2016 on the work done. The report should be brief and to the point, and should reflect the main achievements and proposals for next steps. The report may re-emphasize the mandate of the GWG.

Sections (ii) (iii) and (iv) of these minutes serve as input to the Report.

The report will be drafted by UNSD and ABS as chair of the GWG and is expected to be shared with the GWG members around 20 November.