Economic and Social Council

Statistical Commission
Forty-seventh session
8-11 March 2016
Item 3 (c) 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 2015/216, the Secretary-General has the honour to transmit the report of the Global Working Group on Big Data for Official Statistics. Since its first meeting, held in Beijing in October 2014, the Global Working Group has delivered advocacy materials, papers on access and partnerships for big data and scoping papers on big data classifications and methodology and the repository of projects; conducted a global survey on big data; and undertaken pilot projects involving satellite imagery data and mobile phone data. The priority list and programme of work for the short term include pilot projects on big data for the indicators of the Sustainable Development Goals; the use of the "big data sandbox" for training; the launch of the big data repository; and further progress on principles for access to data owned by the private sector. The report also includes the outcome of the second International Conference on Big Data for Official Statistics, held in Abu Dhabi in October 2015.

The Statistical Commission is invited to express its views on the progress achieved and to comment on the plans regarding data access and partnerships, capacity-building and the use of big data for the indicators of the Sustainable Development Goals. Points for discussion are contained in the last section of the report.

Please recycle

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[1] and its decision 46/101 (sect. III),[2] the global working group provides strategic vision, direction and coordination for a global programme on big data for official statistics, including indicators of the 2030 Agenda for Sustainable Development. It also promotes the practical use of big data sources, while promoting capacity-building, training and the sharing of experience. Finally, the global working group fosters communication and advocacy with respect to the use of big data for policy applications and offers advice on building public trust in the use of big data from the private sector.

2.   The global working group[3] organized its programme of work for 2014 and 2015 into eight streams, each of which is carried out by a specific task team. It established three teams on mobile phone data, satellite imagery and social media data, respectively, to draft guidance and develop practice through pilot projects. Another team was established on access to data and the building of partnerships with the private sector and other communities, which were also seen as important issues; that team drafted provisional umbrella agreements for access to data with globally operating big data providers.

3.   An additional team was established to communicate the benefits and value of big data, which included fundraising strategies to enable developing countries to actively participate in pilot projects. Given the context of the 2030 Agenda for Sustainable Development, it was also agreed that one team would be tasked specifically with keeping track of the links between the indicators needed for monitoring the Sustainable Development Goals and the big data applications. Finally, two more teams were created: one on training, skills and capacity-building, given the need for skills different from those currently available in statistical offices, and one on cross-cutting issues such as methodology, classification and quality frameworks. The previous report (E/CN.3/2015/4) listed a number of deliverables of the global working group related to the work of the eight teams. In addition, members of the group participated in a number of pilot projects on the implementation of big data for official statistics and continued to work on an inventory of big data projects.

4.   The present report starts with a brief overview of the progress made by the task teams of the global working group (sect. II) and then presents the results of the global survey on big data strategies and projects (sect. III), followed by the outcome of the second International Conference on Big Data for Official Statistics (sect. IV). Thereafter, the priorities and programme of work of the global working group are reviewed (sect. V). The report concludes with a number of points for discussion.

_____

[1]   See E/CN.3/2015/4.
[2]   See E/CN.3/2015/34.
[3]   An updated list of members of the global working group is available from http://unstats.un.org/ unsd/bigdata/.

II. Progress made by the global working group

5. As noted above, the objectives of the global working group are directing the global programme of big data for official statistics, including with a view to its use for the measurement of indicators of the 2030 Agenda for Sustainable Development, for various policy applications and for the promotion of capacity-building on big data issues. A full overview of the progress made by the group's eight task teams is provided in a background document available from the group's website.[4] Some of the highlights are set out below.

6. The global working group delivered outputs on advocacy and communication, namely, a strategy document, a brochure, videos and a newsletter. Three papers on access and partnerships[5] were prepared, including draft principles for access to big data sources, which will be further developed in close consultation with relevant stakeholders. As a consolidated effort of two of the task teams, an inventory of big data projects was compiled, in which projects were also mapped to the Sustainable Development Goals and their targets. In addition, scoping papers were drafted on the classification of big data sources, the establishment of a big data quality framework,[6] which is aimed at bringing together a bottom-up (examples) and top-down (theory) approach, and the testing of the framework, which should be adjustable to specific big data sources and processes.

7. Through various task teams, the global working group also initiated two pilot projects using mobile phone data and a number of other pilot projects using social media data. With respect to satellite imagery, remote sensing techniques and geospatial data, the group supports approaches for collecting representative training data of sufficient quality to develop and implement assessment methods that are fit for purpose, including for predictive modelling. As an aspirational goal, a "turn-key" statistical application is being tested for the prediction of crop production yields.

8. A full progress report on the work of the eight task teams has been provided in a background document available from the website of the global working group.

III. Results of the global survey on big data strategies and projects

9. In section III of its decision 46/101, the Statistical Commission approved the proposal that the global working group conduct a global survey on big data for official statistics. 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,
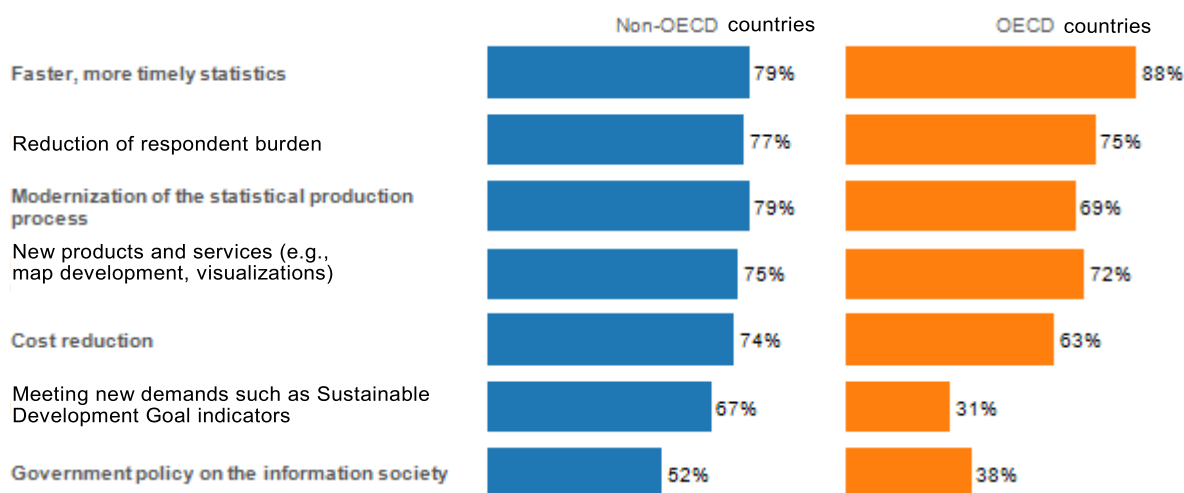
––––––––––––––––––

[4] http://unstats.un.org/unsd/bigdata/.

[5] Available from http://unstats.un.org/unsd/trade/events/2015/abudhabi/gwg-agenda.asp.

[6] For further information, see http://www1.unece.org/stat/platform/display/bigdata/2014+Project.

the questionnaire contained detailed questions on big data projects, intended for those offices that had been engaged in one or more projects.

10.   The survey was conducted from June to August 2015. A questionnaire was sent to each national statistical office with the request that it consult with the relevant stakeholders in its national statistical system. A total of 93 countries completed the questionnaire (32 member countries of the Organization for Economic Cooperation and Development (OECD) and 61 non-OECD countries). In addition, a reply was received from Eurostat. The questionnaire consisted of 18 questions on big data management and 24 questions on each reported big data project. About half of the countries reported at least one big data project, with a few reporting more than four. However, it should be noted that the other half of the countries did not report a project. Nevertheless, a total of 115 big data projects were reported: 89 by OECD countries, 22 by non-OECD countries and 4 by Eurostat. The results are shown in the figure below.

Main reasons for and business benefits of using big data

	Non-OECD countries	OECD countries
Faster, more timely statistics	79%	88%
Reduction of respondent burden	77%	75%
Modernization of the statistical production process	79%	69%
New products and services (e.g., map development, visualizations)	75%	72%
Cost reduction	74%	63%
Meeting new demands such as Sustainable Development Goal indicators	67%	31%
Government policy on the information society	52%	38%

11.   As shown in the figure above, statistical offices consider "faster, more timely statistics", "reduction of response burden" and "modernization of the statistical production process" to be the main benefits of using big data, followed by "new products and services" and "cost reduction". Interestingly, whereas two thirds of the developing countries consider meeting new demands such as the indicators of the Sustainable Development Goals to be a benefit, only one third of the OECD countries share that view.

12.   Many offices have already used or considered using big data sources, mainly scanner data, web-scraping data, mobile phone data and satellite imagery data. This result corresponds more or less with the actual reported big data projects; most used were mobile phone data (42 projects), followed by web-scraping data (31 projects) and scanner data (23 projects). However, scanner and web-scraping data were used much more in OECD countries than in non-OECD countries. These results are also reflected in the statistical domains for which big data was most used: price statistics (based on scanner data), followed by tourism (mobile phone data), population (mobile phone data) and transport and labour statistics (web-scraping data).

13.	Given that the use of big data in general involves working together with other offices to gain access to data and acquire the necessary technical skills, the survey asked what partnerships respondents had established for dig data projects. Most often, respondents had partnered with government institutes, followed closely by the research and academic communities. They had established far fewer partnerships with data providers and information technology (IT) companies.

14.	Furthermore, the surveys determined that most reported big data projects were in the exploration and research stage. However, if a project was intended to go into production, it complemented an existing data source. In addition, the survey revealed an interesting finding: both developing and developed countries continued to rely on traditional statistical methods in the processing and analysis of big data sources. That finding may be interpreted in at least two ways. First, big data projects can apparently be processed and analysed using traditional methods, which should significantly reduce the threshold for undertaking a big data project. Second, national statistical offices may not yet be very familiar with advanced big data tools for processing and analysis, which may hamper the full exploitation of big data sources.

15.	In response to the question "Which skills are important for your office to acquire in order to better deal with big data?", respondents indicated "methodologist on big data issues", "data scientist" and "mathematical modelling specialist" as the three most-needed skills, with the more IT-oriented skills of "IT architecture specialist", "data visualization specialist" and "cybersecurity specialist" given less priority. Whereas statistical offices clearly see a need for these new advanced big data skills, they have not yet begun to hire candidates who possess them, nor have they sent their staffs to receive training on them.

16.	The survey explicitly asked about the urgent needs for guidance that national statistical offices might have. The respondents indicated "skills and training for big data", "quality framework for big data", and "access to big data" as the three top guidance priorities, followed by "estimation methods", "use of web-scraping data" and "use of mobile phone data". That result is consistent with the needed skills indicated in the areas of big data methodologies, estimation methods and data science.

17.	The following conclusions can be drawn from these results: (a) training and capacity-building on big data topics (methodologies, estimation and a quality framework) is necessary; and (b) more pilot projects are needed, especially those with participation from developing countries. In addition, easier access to big data will lower the threshold for undertaking a big data project. Consequently, training, big data methodology and quality frameworks, and data access have been chosen as the three top priorities of the programme of work of the global working group and are included in the themes for the next international conference. The full report on the survey has been provided to the Statistical Commission in a background document.

## IV.	The outcome of the second International Conference on Big Data for Official Statistics

18.	The second International Conference on Big Data for Official Statistics was held in Abu Dhabi from 20 to 22 October 2015 on the theme "Moving from examples to guidance" and was attended by approximately 250 statisticians from

around the world. The Conference was opened by the Minister of Culture, Youth and Community Development of the United Arab Emirates, who placed the relevance of big data in the context of the 2030 Agenda for Sustainable Development and made reference to the theme of World Statistics Day, "Better data, better lives". The Conference showcased advances made by big data projects using mobile phone data, social media data and satellite data for a variety of statistical applications. Progress was also evidenced in the areas of capacity-building, data access and partnerships, quality and methodology and how to communicate the value of big data more effectively. All presentations are available from the website of the Conference.[7] Some of the conclusions reached in the panel discussions at the Conference are set out below.

19. Big data represents a great opportunity, but it is not a "magic bullet" and requires investment in time, money and capability. Pilots are a good way to establish the feasibility of approaches. Blended approaches to big data are likely to be the dominant and successful models in the future, which means using data from a range of sources and producing estimates that are fit for purpose and whose level of quality and uncertainty are established and made transparent. At the end of the day it is all about performance: is the new estimate better than the existing one?

20. The use of big data fits within the agenda of modernizing statistical production and involves significant cultural change for many national statistical offices. In the communication strategy, it should be emphasized that the use of big data by national statistical offices must be understood as part of a gradual process of change in the statistical production process. In the not-too-distant past, many national statistical offices introduced business surveys alongside business censuses, and the business surveys progressively became the instrument of choice for the collection of business statistics. Thereafter, many national statistical offices started to blend administrative (tax) data with business survey data. Whereas administrative data were initially seen as unfit for statistical purposes, today statisticians are almost required to exhaust the possibility of available administrative data before being allowed to introduce a new survey. This may also happen over time with the use of certain big data sources in statistical production.

Big data methodology and estimation

21. Another conclusion reached in the panel discussions was that big data needs official statistics as much as official statistics need big data. This is not only because the production of official statistics is anchored in internationally agreed quality frameworks and methodologies and based on principles of professional independence and trust; it is the official statistics using traditional source data that allow methods and techniques for generating statistics from big data sources to be calibrated, "trained" and, ultimately, validated. Other findings were that statistical methodology can turn big data into small data, for example, through sampling, and that the transfer of data is not always necessary, as the method or algorithm can be applied at the location of the data source.

22. Furthermore, the potential of satellite imagery and geospatial data was established for statistics with respect to monitoring the environment at high temporal and spatial resolutions. In this regard, it is important to note that during the Conference, the

_____

[7] http://unstats.un.org/unsd/trade/events/2015/abudhabi/default.asp.

global intergovernmental geospatial and Earth observation mechanisms — the Committee of Experts on Global Geospatial Information Management and the Group on Earth Observations — offered to support the efforts of the global working group to explore the plethora of possibilities provided by satellite imagery and geospatial data, including how to make use of existing methods for estimating official statistics and indicators for the Sustainable Development Goals. The global working group can also benefit from the experimentation and testing taking place in ecosystem accounting when it compiles official statistics regarding measures of ecosystem extent, condition and services undertaken under the aegis of the United Nations Committee of Experts on Environmental-Economic Accounting.

Training and capacity-building

23.   It was very encouraging to note that a number of the developing countries are exploring the opportunities provided by big data (especially satellite imagery) owing to the limited availability of data in more traditional areas of statistics. The big data "sandbox" in Dublin represents an excellent low-cost opportunity for countries to begin to increase their familiarity with big data tools, methods and capabilities in multidisciplinary teams. Other opportunities should also be explored, including leveraging the skills and capabilities of the private sector and not-for-profit organizations in utilizing their human capital and infrastructure, such as Global Pulse, and, where appropriate, forming long-term partnerships.

Data access and partnerships

24.   Given that big data projects, almost by definition, involve players from the private sector and research institutes, it is important that compelling business cases be made as to why these partnerships are win-win situations for all involved. Long-term partnerships, particularly those with private sector entities, will require that there be strong incentives on both sides. It is also important that the statistical community position itself in a unified way when seeking partnerships, especially with globally operating data providers.

V.   Priorities of the programme of work

25.   As indicated in section II above, the global working group has made some progress. Documents have been prepared on advocacy, principles for access, good practices with regard to data access and partnerships, templates for memorandums of understanding, classifications, quality and methodology and the repository of big data projects. Moreover, the project repository is an important deliverable of the global working group that should serve the statistical community as well as a wider audience. Pilot projects are ongoing using data from satellite imagery, mobile phones and social media. However, more needs to be done in order to move from research to production and from examples to guidance.

26.   The respondents to the global survey expressed an urgent need for guidance in the areas of skills and training, quality frameworks for big data, access to big data and estimation methods. The conclusions reached at the Conference also reflected this need for training, methodological guidance and data access. Therefore, those topics will be given priority in the future programme of work of the global working group.

27. Regarding access to data, the global working group will continue to work on the scope and the formulation of the principles for data access, as well as their format and the way in which they will be packaged. The scope of the principles is seen as "access to proprietary data", including data held by private sector entities, government institutes or other organizations. It is expected that the work will move forward in close consultation with stakeholders, in particular those outside the statistical community, with the outcome to be presented at the next big data conference, to be held in Dublin.

28. "Learning by doing" is one way of building the capacity to use big data for official statistics. The global working group will continue to be actively engaged in a number of pilot projects, some of which will have application to the measurement of the indicators of the Sustainable Development Goals. Developing countries in particular should take the opportunity to be involved in the pilot projects. Possibilities for active participation are offered through the "sandbox" in Dublin. In addition, the use of the Global Pulse labs in Jakarta and Kampala will be further explored.

29. Given the expressed need for a big data methodology and quality assurance framework, the global working group agreed to organize its work more along the lines of the big data quality framework, by grouping elements of its programme of work under "input quality" (data access), "throughput quality" (big data methodology and estimation methods, the repository of big data projects and statistical production processes) and "output quality" (application of big data to official statistics, including the Sustainable Development Goal indicators). This refocusing and slight rearrangement of the work of the task teams could also improve the sharing of the findings of the pilot project and their integration into the work of the task teams on data access, methodology and training.

30. In anticipation of the third International Conference on Big Data for Official Statistics, which is expected to be held from 28 to 30 June 2016 in Dublin, the provisional agenda is organized along the lines of the big data quality framework, as follows: theme 1, "Data access and partnerships: the win-win scenarios"; theme 2, "Statistical production: the role of big data and the skills and capacity required"; and theme 3, "Big data for official statistics and Sustainable Development Goal indicators".

31. To further emphasize the importance that the global working group attaches to training and capacity-building, it was agreed that one or more workshops would be held on the margins of the conference in Dublin. It was suggested that a workshop be held on big data methodologies, in line with the needs expressed in response to the global survey. Another workshop could be held on "Leading and managing big data projects", which would cover issues related to problem formulation, developing business cases, the acquisition of data sources, project management and statistical procedures. Finally, it was suggested that a workshop also be held on "Contemporary and emerging IT tools and technologies for big data".

VI. Points for discussion

32. **The Commission is invited to express its views on:**

    (a) **The progress of the work of the global working group on big data for official statistics;**

    (b) **The results of the global survey on big data strategies and projects;**

(c)   **The outcome of the second International Conference on Big Data for Official Statistics;**

(d)   **The priorities of the programme of work of the global working group:**

(i)   **Pilot projects in the use of big data for official statistics, including those on the measurement of selected indicators of the Sustainable Development Goals;**

(ii)   **Training and skills in the use of big data involving more developing countries, especially through the use of the "sandbox" in Dublin and the Global Pulse labs for training purposes;**

(iii)   **Enhancing data access by further developing guiding principles for access to proprietary data;**

(iv)   **Guidance on big data methodologies and estimation methods;**

(v)   **A quality framework for big data for official statistics.**

_____