United Nations ECE/CES/2014/7

Economic and Social Council

Distr.: General
24 January 2014

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-second plenary session
Paris, 9-11 April 2014
Item 7(b) of the provisional agenda
Outcomes of the in-depth reviews carried out by the Conference of European Statisticians Bureau

In-depth review of big data

Prepared by the temporary Task Team on Big Data and the Secretariat

Summary

The present note is an updated version of the in-depth review paper on "Big Data". The Bureau of the Conference of European Statisticians conducted the in-depth review at its meeting in October 2013. The purpose of the reviews carried out by the Bureau is to improve coordination of statistical activities in the region of the United Nations Economic Commission for Europe, identify gaps or duplication of work and address emerging issues.

The note has been revised based on the discussion held at the review. The note summarises the international statistical activities related to "Big Data", identifies issues and challenges, and makes recommendations how the international statistical community could tackle the issues.

The outcome of the review is provided in document ECE/CES/2014/7/Add.1.

I. Executive Summary

1. This in-depth review considers the response of the official statistics community to the “Big Data” phenomenon. It defines Big Data and summarizes the activities in this area of various national and international statistical organizations. It outlines the main issues and challenges identified so far, and concludes that it would be more efficient for the international statistical community to tackle these issues in a collaborative way, rather than each organization seeking its own solution. The review concludes with the following three recommendations:

(a) Specify the key priority areas relating to Big Data to be tackled as a collaborative activity by the international statistical community. The project proposal in the Annex provides the thinking so far in this area at the expert level. The views of the CES on whether this proposal covers the right activities would be very welcome;

(b) As knowledge and experience of using Big Data are gained, a mechanism for sharing that information is needed. The inventory of Big Data activities started by the informal Task Team should be consolidated and expanded as a resource for the whole statistical community;

(c) The two activities above should be overseen by the High-Level Group on the Modernisation of Statistical Production and Services, to ensure a sufficiently strategic focus.

II. Introduction

2. The Bureau of the Conference of European Statisticians (CES) regularly reviews selected statistical areas in depth. The aim of the reviews is to improve coordination of statistical activities in the UNECE region, identify gaps or duplication of work, and address emerging issues. The review focuses on strategic issues and highlights concerns of statistical offices of both a conceptual and a coordinating nature. The current paper provides the basis for the review by summarising the international statistical activities in the area of Big Data, identifying issues and problems, and making recommendations on possible follow-up actions.

3. In our modern world more and more data are generated on the web and produced by sensors in the ever growing number of electronic devices surrounding us. The amount of data and the frequency at which they are produced have led to the concept of “Big Data”. The term “Big Data” is used to describe data sets of increasing volume, velocity and variety; the three V's. Sources described as ‘Big Data’ are often largely unstructured, meaning that they have no pre-defined data model and/or do not fit well into conventional relational databases. Apart from generating new commercial opportunities in the private sector, Big Data is also potentially very interesting as an input for official statistics; either for use on its own, or in combination with more traditional data sources such as sample surveys and administrative registers. However, harvesting the information from Big Data and incorporating it into a statistical production process is not easy.

4. Big Data has the potential to produce more relevant and timely statistics than traditional sources of official statistics. Official statistics have been based almost exclusively on survey data collections and acquisition of administrative data from government programmes. But this is not the case with Big Data where most data are readily available or with private companies. As a result, the private sector may take advantage of the Big Data era and produce more and more statistics that attempt to beat official statistics on timeliness and relevance. It is unlikely that statistical organizations will lose the "official statistics" trademark but they could slowly lose their reputation and relevance unless they

get on board. One big advantage that statistical organizations have is the existence of infrastructures to address the accuracy, consistency and interpretability of the statistics produced. By incorporating relevant Big Data sources into their official statistics process, statistical organizations are best positioned to measure their accuracy, ensure the consistency of the whole systems of official statistics and provide interpretation while constantly working on relevance and timeliness. The role and importance of official statistics will thus be protected.

5. However, the topic of Big Data is still rather new for many statistical organizations, and there is uncertainty about what it really means for official statistics, and how best to react.

III. Scope/Definition of the statistical area covered

6. The topic of Big Data cuts across all statistical activities, and could be relevant for all statistical domains. In terms of the Classification of Statistical Activities (Rev. 1, October 2009[1]), Big Data probably fits best in activity 4.3 (Data Sources), but it does not readily fit into any of the more specific subject areas under this heading, unless a very broad definition of “administrative sources” is used, such that Big Data could be included in subject area 4.3.5 (Other administrative sources).

7. This review focuses on Big Data as a source, rather than an output of statistical organizations, as statistical outputs do not (yet) meet the criteria of volume, velocity and variety to be truly considered as Big Data.

IV. Overview of international statistical activities in the area

A. United Nations Economic Commission for Europe (UNECE)

8. UNECE, together with Rosstat, the Russian Federal State Statistics Service, organized a High-level Seminar on Modernization of Statistical Production and Services, in St. Petersburg in October 2012[2]. Following some discussion of Big Data, one of the conclusions of the Seminar was:

"Big Data is an increasing challenge. The official statistical community needs to better understand the issues, and develop new methods, tools and ideas to make effective use of Big Data sources. This includes closer integration with geographical data and standards."

9. As a follow up activity, it was proposed that the High-Level Group for the Modernization of Statistical Production and Services (HLG) should provide "a document explaining the issues surrounding the use of Big Data in the official statistics community". HLG convened a group of leading international experts, facilitated by the UNECE, to prepare a paper to address this requirement. The resulting paper, “What does Big Data mean for official statistics?”, was published by HLG in March 2013[3], and presented to the Conference of European Statisticians three months later at their 2013 Plenary Session.

10. This paper provoked further discussion on the topic of Big Data in official statistics, including at the joint UNECE / Eurostat / OECD / UN-ESCAP meeting on the Management

[1] http://www1.unece.org/stat/platform/display/disaarchive
[2] http://www.unece.org/stats/documents/2012.10.hls.html
[3] http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614

of Statistical Information Systems (Paris and Bangkok, April 2013)[4]. This meeting decided that Big Data is a key issue for official statistics, and noted the following key points:

- It is an ideal time to start collaborating on Big Data, as organizations typically did not have systems in place yet, and could develop them collaboratively;
- Each organization faces common issues in relation to using Big Data, so it could be more efficient to work together to find common solutions. This should be a priority for HLG;
- Statistical organizations have traditionally focused on producing consistent time series, but increasingly there is also a need for short-lived measures that address a phenomenon in a country when it happens;
- Experience gained in the use of administrative sources may be helpful for Big Data;
- It is important to take a multidisciplinary approach to Big Data, currently different groups are all looking at this issue from their own perspectives;
- Agreeing a common classification of the different types of Big Data should be an early priority;
- A concrete project to produce specific statistics from Big Data, and to find real solutions would be useful;
- A virtual task team should be set up to define the issues and formulate a clear project proposal, which would be passed to HLG.

11. In May 2013, a temporary task team[5] was set up to identify the key issues with using Big Data for official statistics, determine priority actions and formulate a project proposal. The team worked virtually, by Wiki and web conferencing, during May and June, and produced the following outputs:

- A draft project proposal consisting of three major work strands: an exploration of strategic and methodological issues; analysis of Big Data sources and international replication of outputs using a shared computing environment; and training and dissemination activities (see Annex);
- A draft classification of types of Big Data[6];
- A specification for an inventory of Big Data sources and projects, based on the above classification. The inventory has subsequently been launched, and populated with information about projects in several countries[7], and will be developed further in the context of UNECE work on the modernisation of statistics.

12. Following the presentation of the annual report of the HLG at the 2013 Plenary Session of the CES, delegates were asked to identify priorities for HLG projects for 2014. Most speakers recognised the need for international collaboration activities in the area of facilitating the use of Big Data for official statistics.

13. HLG discussed and decided the key priorities for 2014 with representatives of expert groups at the annual Workshop on the Modernisation of Statistical Production and Services, held in Geneva on 25-27 November 2013. Big Data was chosen as the subject of a major international collaboration project for 2014, overseen by HLG.

[4] http://www1.unece.org/stat/platform/display/msis/MSIS+2013
[5] http://www1.unece.org/stat/platform/display/msis/Members+of+the+task+team
[6] http://www1.unece.org/stat/platform/display/msis/Classification+of+Types+of+Big+Data
[7] http://www1.unece.org/stat/platform/display/msis/Big+Data+Inventory

B. Eurostat

14. Eurostat is investigating the potential use of Big Data for official statistics in areas such as price statistics (using Internet price data) and information and communication technology (ICT) usage statistics.

15. Eurostat has contributed actively to the work of the Task Team facilitated by the UNECE, and the development of the resulting project proposal. Eurostat staff have also prepared several papers on this topic for international conferences.

16. One session of the annual DGINS (Director Generals of national statistical organizations) meeting, held in The Hague, Netherlands, in September 2013[8], was devoted to the topic of Big Data. This included presentations from national statistical organizations and private companies. It resulted in the Scheveningen Memorandum on Big Data and Official Statistics, which encourages members of the European Statistical System to develop a Big Data strategy, share experiences, and collaborate at the level of the European Statistical System and beyond. An action plan and roadmap should be adopted by mid-2014, and integrated into the Eurostat work programme.

17. Eurostat held a workshop on the topic of Big Data, in cooperation with the UNECE, in Rome in April 2014.

C. Organisation for Economic Cooperation and Development (OECD)

18. OECD is currently investigating the use of Big Data in the areas of innovation indicators, quality of Internet connections, and well-being / better life indicators.

19. OECD has contributed actively to the work of the Task Team facilitated by the UNECE, and the development of the resulting project proposal. OECD staff have also prepared papers on this topic for international conferences. The OECD has also released a policy paper "Exploring data-driven innovation as a new source of growth: Mapping the Policy Issues Raised by Big Data"[9]

D. United Nations Statistical Division

20. The United Nations Statistical Division organized a one-day side-event to the 2013 meeting of the Statistical Commission, with the title "Big Data for Policy, Development and Official Statistics"[10]. This included presentations from national statistical agencies, international organizations and private companies.

21. The 2014 meeting of the Statistical Commission also included an item on "Big Data and modernization of statistical systems"[11], which proposed the "creation of a global working group on the use of big data for official statistics whose activities would complement the work carried out by the regional commissions and manage the globally relevant issues".

[8] http://www.cbs-events.nl/dgins2013/

[9] http://www.oecd-ilibrary.org/science-and-technology/exploring-data-driven-innovation-as-a-new-source-of-growth_5k47zw3fcp43-en

[10] http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/default.html

[11] http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf

E. World Bank

22. The World Bank has organized several events on the topic of Big Data, including an event "Turning Big Data into Big Impact"[12] in October 2012 and a live webcast "What happens when Big Data meets official statistics?"[13] in December 2012.

F. Other

23. There are many other events and discussions on the topic of Big Data outside the area of official statistics. The numbers of Big Data consultants, and software tools specifically designed for handling Big Data are growing rapidly. It is clear that in developing a response to the emergence of Big Data, the international official statistics community should actively follow external developments and determine how they might apply to our activities.

V. Country practices

24. The Task Team on Big Data identified a wide range of activities relating to the use of Big Data in participating countries. Most of these activities are at the planning or experimental stage, aiming to determine the feasibility of using Big Data sources for statistical production. However, a few countries are starting to take the next step, and move towards regular data production using Big Data.

25. One issue identified was the lack of a mechanism for sharing information on current and planned activities. This resulted in the proposal to develop an inventory of Big Data projects and resources.

VI. Impact of crisis on the statistical area

26. The financial crisis starting in 2009 has strongly encouraged statistical organizations to look for ways to increase efficiency and cut data costs. Traditionally data collection has been one of the most cost-intensive parts of the statistical production process, so the interest in alternative data sources, including Big Data, is growing.

VII. Issues and challenges

27. The following issues and challenges were identified in the HLG paper "What does Big Data mean for official statistics?"

A. Legislative

28. Legislation in some countries may provide the right to access data from both government and non-government sources while in other countries, legislation may provide the right to access data from public authorities only. This can result in limitations for accessing certain types of Big Data.

[12] http://www.linkedin.com/groups/World-Bank-Event-Turning-Big-137043.S.177763767
[13] http://live.worldbank.org/what-happens-when-big-data-meets-official-statistics-live-webcast

29. Even if legislation has provision to access all types of data, the statistical purpose for accessing the data might need to be demonstrated to an extent that may be different from country to country.

B. Privacy

30. Privacy is generally defined as the right of individuals to control or influence what information related to them may be disclosed. Privacy is a pillar of democracy. The problem with Big Data is that the users of services and devices generating the data are most likely unaware that they are doing so, and/or what it can be used for. The data would become even bigger if they are pooled, as would the privacy concerns.

C. Financial

31. There is likely to be a cost to statistical organizations to acquire Big Data, especially from the private sector, particularly if legislation is silent on the financial modalities surrounding acquisition of external data. As a result, statistical organizations have to balance quality (which encompasses relevance, timeliness, accuracy, coherence, accessibility and interpretability) against costs and reduction in response burden. Costs may be significant, but the potential benefits may far outweigh the costs, with Big Data potentially providing information that could increase the efficiency of government programmes (e.g. health systems). Rules around procurement in the government may come

D. Management

32. Big Data for official statistics may mean more information coming to statistical organizations that is subject to policies and directives on management and protection of information.

33. Another management challenge relates to human resources. The data science associated with Big Data that is emerging in the private sector does not seem to have connected yet with the official statistics community. Statistical organizations may have to perform in-house and national scans (academic, public and private sector communities) to identify where data scientists are and connect them to the area of official statistics.

E. Methodological

34. Representativeness is a fundamental issue with Big Data. The difficulty in defining the target population, survey population and survey frame jeopardizes the traditional way in which official statisticians think and do statistical inference about the target (and finite) population. With a traditional survey, statisticians identify a target/survey population, build a survey frame to reach this population, draw a sample, collect the data etc. They will build a box and fill it with data in a very structured way. With Big Data, the data come first and the reflex of official statisticians would be to build a box! This raises the question is this the only way to produce a coherent and integrated national system of official statistics? Is it time to think outside of the box?

35. Another issue is both technological and methodological in nature. When more and more data are analysed, traditional statistical methods, developed for the very thorough analysis of small samples, run into trouble. In the most simple case they are just not fast enough. New methods and tools are needed, for example:

(a) Methods to quickly uncover information from massive amounts of data available, such as visualisation methods and data, text and stream mining techniques, that are able to ‘make Big Data small’. Increasing computer power is a way to assist with this step at first;

(b) Methods capable of integrating the information uncovered in the statistical process, such as linking at massive scale, data integration, and statistical methods specifically suited for large datasets. Methods need to be developed that rapidly produce reliable results when applied to very large datasets.

36. The use of Big data for official statistics triggers a need for new techniques. Methodological issues that these techniques need to address are:

(a) Measures of quality of outputs produced from hard-to-manage external data supply. The dependence on external sources limits the range of measures that can be reported when compared with outputs from targeted information-gathering techniques;

(b) Limited application and value of externally-sourced data;

(c) Difficulty of integrating information from different sources to produce high-value products;

(d) Difficulty of identifying a value proposition in the absence of the closed loop feedback seen in commercial organizations.

F. Technological

37. New tools are needed to connect applications for data capturing and data processing directly with data sources. Collecting data in real time or near real time can maximize the potential of data, opening new opportunities for using data from high-velocity sources, such as:

(a) Commercial data (credit card transactions, on line transactions, sales, etc.);

(b) Tracking devices (cellular phones, global positioning systems, surveillance cameras, ‘apps’) and physical sensors (traffic, meteorological, pollution, energy, etc.);

(c) Social media (Twitter, Facebook, etc.) and search engines (online searches, online page views);

(d) Community data (Citizen Reporting or Crowd-sourced data).

38. In the era of Big Data this change of paradigm for data collection presents the possibility to collect and integrate many types of data from many different sources. Combining traditional data sources, such as surveys and administrative data, with Big Data could provide new challenges and opportunities.

VIII. Conclusions and recommendations

39. It is clear that the official statistical community is just starting to explore the potential issues and benefits of Big Data. If each organization does this on its own, this will lead to inefficiency within the statistical system at the global level. The main recommendations of this in-depth review are therefore:

(a) Specify the key priority areas relating to Big Data to be tackled as a collaborative activity by the international statistical community. The HLG Big Data project is addressing this;

(b) As knowledge and experience of using Big Data are gained, a mechanism for sharing that information is needed. The inventory of Big Data activities started by the informal Task Team is being consolidated and expanded as a resource for the whole statistical community;

(c) The two activities above should be overseen by the High-Level Group on the Modernisation of Statistical Production and Services to ensure a sufficiently strategic focus.