UNITED NATIONS E

ECONOMIC AND SOCIAL COUNCIL

Distr.
LIMITED
E/ESCWA/SD/2015/IG.1/5
13 January 2015
ORIGINAL: ENGLISH

Economic and Social Commission for Western Asia (ESCWA)

Statistical Committee
Eleventh session
Amman, 4-5 February 2015

Item 6 of the provisional agenda

OFFICIAL STATISTICS AND EMERGING SOURCES OF DATA: IMPLICATIONS FOR ESCWA STATISTICAL ACTIVITIES

BIG DATA

Summary

This document on big data was prepared as a background paper for the eleventh session of the Statistical Committee of the Economic and Social Commission for Western Asia (ESCWA), which has proposed as its leading theme "Data revolution for support and monitoring of the post-2015 development agenda". It is intended to be informative and non-technical, aiming to raise the awareness of participants by bringing deserved attention to big data and its potential role in improving official statistics, and measuring and assessing the post-2015 development agenda.

This work could be further developed and expanded to include best practices, strategies and guidance for selecting, deploying and managing big data analytics, including data virtualization and text mining.

CONTENTS

I. INTRODUCTION

1. The development of Internet and mobile phone services nearly two decades ago instigated an information and communications technology (ICT) revolution, which made smartphones, tablets and connected devices widely available. A digital universe was created where people and devices are connected in a seamless network, dubbed the Internet of Things (IoT), allowing them to constantly produce and share information, thus generating a massive amount of digital data at an unprecedented rate.

2. In 2013, the IoT universe encompassed around 5 billion connected devices, generating 4.4 zettabytes of data (or 4.4 trillion gigabytes). This number is forecasted to reach 200 billion devices in 2020, accounting for 44 zettabytes of data that would double every two years.[1]

3. Data generated in the IoT universe has been labelled as big data; it is mostly derived from connected embedded systems, mobile and smartphone usage, web content, social media, credit card and financial transactions, sensors, cameras, etc. This data revolution, if information is stored, tagged and analysed properly, presents a tremendous opportunity; it provides snapshots of the well-being of populations at high frequency and high degrees of granularity, and from a wide range of angles, thus narrowing both time and knowledge gaps.[2] In 2013, the United Nations, through its Global Pulse initiative, spearheaded world efforts to harness big data for humanitarian purposes and global development. Furthermore, the United Nations considers that big data represents a new renewable natural resource with the potential to revolutionize sustainable development and humanitarian practice.[3]

II. DEFINITION

4. The concept of big data was coined in 2001[4] to describe large and complex sets of data that are difficult to process using traditional data processing methods. The main reason driving data sets to grow in size is that data are increasingly being gathered by information-based sensors and ICT devices. This definition was updated in 2012, when big data was described as data sets characterized by high volume, high velocity and high variety (known as the 3 Vs), requiring new forms of processing to enable enhanced decision-making, insight discovery and process optimization.[5]

5. Further work on data science has recently introduced a new dimension to Gartner's definition by adding a fourth V for "veracity"; it is an indication of "accuracy", "fidelity" or "truthfulness" of data and its ability to be used to make crucial decisions.[6]

III. SOURCES AND TYPES OF BIG DATA

6. While sources of traditional data collected by national statistical organizations (NSOs) have mainly focused on sample surveys and administrative data sources (including registers), big data is derived from

[1] Vernon Turner and others, "The digital universe of opportunities: rich data and the increasing value of the Internet of Things", International Data Corporation (IDC) White Paper, No. 1672 (Framingham, MA, 2014). Available from http://idcdocserv.com/1678.

[2] United Nations Global Pulse, "Big data for development: challenges and opportunities", May 2012, p. 6. Available from www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf.

[3] See www.unglobalpulse.org/about-new.

[4] Doug Laney, "3D data management: controlling data volume, velocity and variety", 6 February 2011. Available from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. This article was published by Meta Group, a provider of IT research, advisory services and strategic consulting, acquired by Gartner in 2005.

[5] Doug Laney and Mark Beyer, *The Importance of 'Big Data': A Definition* (Gartner, 2012).

[6] See www.ibmbigdatahub.com/infographic/four-vs-big-data.

various public and private sources. The data in its raw format is loosely structured and is often incomplete and inaccessible.

7. The following is not an exhaustive list of big data types and sources; however, it gives the reader a better idea about the diversity of data sources. These sources may include structured data (databases, sensors and location data) and unstructured data (emails, social media and images).

(a) Administrative data: this includes public and private sector data, such as government open data, medical and health related records, insurance records and bank records;

(b) Commercial or transactional data: this includes credit card transactions, online transactions (e-commerce), financial transactions and stock market data;

(c) Sensor data: this includes satellite imaging, road and car sensors, traffic cameras, office buildings, appliances and climate sensors;

(d) Data from the web (a main source for detecting behaviour through browsing and online searches): this includes search engines, websites, online content and emails;

(e) Social media data (a main source of public opinion and sentiment): this includes social media applications such as Facebook, Twitter, Instagram, blogs, Flikr, Pinterest, LinkedIn and Tumblr;

(f) Multimedia data: this includes images, video files, audio files, flash animations, live streaming and podcasts;

(g) Data storage: this includes relational and other forms of databases, such as SQL, NoSQL, Hadoop,[7] document repositories and file archives;

(h) Machine generated data logs (resulting from the execution of a process or application, without human intervention): this includes event logs, call logs, mobile telephone locations, mobile application usage, tracking devices and the Global Positioning System (GPS).

IV. BIG DATA AND OFFICIAL STATISTICS

8. Apart from the interest which big data sparked in the private sector given its numerous commercial opportunities, it caught the attention of the statistical world as a new possible source of data, coupled with the potential to complement or improve the production of official statistics.

9. Traditionally, official statistics have been derived mainly from data collected through surveys and administrative government data gathered by NSOs at regular intervals. These data are then processed, stored and managed by NSOs in a very structured manner.

10. In contrast, big data is mostly unstructured – having no pre-defined data model – which renders its storage incompatible with conventional relational databases. Moreover, the private sector is the proprietor of most available big data today, placing them at an advantage to produce more relevant and timely statistics than NSOs.

11. Nevertheless, NSOs have started to explore the possibility of using big data for official statistics. Apart from combining big data with official statistics, other areas of experimentation included replacing official statistics with big data and filling new data gaps to address emerging phenomena, for which traditional approaches are not effective.

12. Initial results were promising, depending on the area, given that NSOs, with their extensive statistical know–how, are better positioned than the private sector to measure the accuracy of big data, ensure the

[7] Hadoop is an open-source software framework by Apache, used for distributed storage and distributed processing of big data on clusters of hardware.

consistency of official statistics, and provide sound interpretations while constantly working on relevance and timeliness. Thus, the role and importance of official statistics in this hybrid model will be safeguarded.

13. Combining official statistics with other forms of data is not novel; NSOs have resorted to such practice in the past when combining administrative data with official statistics. What is probably different in the case of big data is resorting to extensive statistical modelling for combining the two forms of data; a practice that may lead to enhanced estimates of good quality as a result of near real time measurements obtained from big data.

V. CASE STUDIES: APPLICATIONS OF BIG DATA IN OFFICIAL STATISTICS

14. Successful uses and applications of big data have harnessed emerging big data sources, such as mobile phone data, GPS and other tracking device data. These applications have proven to have a great value added given that the widespread use of mobile phones provides great potential in supplying real-time low cost information. This type of data is being utilized to produce and complement data on tourism and daytime mobility statistics, estimate population census data, map poverty and track mobility patterns in case of disease outbreaks.

15. Another emerging technology that has proven useful in official statistics is satellite imagery. For example, the Australian Bureau of Statistics (ABS) has been working on complementing and partly replacing surveys for measuring agricultural crop production with data retrieved from satellite imagery. Although ABS is still testing the accuracy of its estimation methods when utilizing this data source, the ongoing work has indicated that, with this technology, disaggregated agricultural statistics can be provided more frequently, at a lower cost and in a timely manner. Other countries, such as China, Colombia and Mexico, have also been examining and testing the application of this technology in ecosystem accounting.

16. Satellite imagery is also used by the United Nations Office for Drugs and Crime for statistical monitoring of poppy production in some sensitive areas where poppies are used to produce opium and derive illicit drugs.

17. Social media provides another large source of big data that can be utilized in official statistics. The applications of such data in terms of human behaviour can vary among the areas of social science. The Dutch initiative for assessing consumer sentiment estimates from Facebook and Twitter data is a promising application, which may allow for the production of sentiment indices at a lower cost and higher frequency through social media data. Another example addresses the labour market, as China and Italy have employed web scraping tools to estimate job vacancy rates that can complement labour statistics data by improving monthly predictions and territorial estimates.

18. Experimenting with big data applications has proven fruitful when it comes to complementing official statistics. Researchers at the Massachusetts Institute of Technology, for example, have been estimating inflation by collecting and analysing the prices of goods advertised or sold online. Another initiative has been undertaken by the United States Geological Survey; it developed a system that monitors Twitter for significant spikes in the volume of messages on earthquakes, along with location information, so that seismologists can verify the occurrence of an earthquake and quantify its magnitude more rapidly. Moreover, researchers at Harvard University conducted a retrospective analysis of the 2010 cholera outbreak in Haiti and demonstrated that mining Twitter and online news reports could have provided health officials with a highly accurate indication of the actual spread of the disease with two weeks lead time.

VI. CHALLENGES OF USING BIG DATA

19. In spite of success stories and promising results, the use of big data in official statistics faces several challenges,[8] which can be categorized as follows:

[8] See United Nations Economic Commission for Europe, "What does 'big data' mean for official statistics?", 10 March 2013. Available from http://www1.unece.org/stat/platform/download/attachments/77170614/Big%20Data%20HLG%20Final%20Published%20Version.docx?version=1&modificationDate=1370507520046&api=v2.

(a) Methodological: representativity of big data is a main challenge for statisticians who follow a very structured process to identify a target/survey population, then build a survey frame to reach this population, draw a sample and collect the data. In contrast, with big data, data come first. Another challenge of equal importance is that using traditional statistical methods, developed for analysing small samples, fails to accommodate big data, thus requiring new statistical methods and tools;

(b) Management: using big data for official statistics requires new management policies and directives to deal with the protection and management of additional data. Another challenge is the lack of qualified data scientists[9] with expertise in national statistics; academia and the private sector could assist in this regard;

(c) Legislative: not all countries have similar legislation on data access. Some countries guarantee the right to access data from both government and non-government sources, while others may restrict the right of access to public sources only;

(d) Privacy: privacy concerns the right of individuals to allow or restrict the disclosure of information related to them. Private sector companies use privacy measures to protect consumers and their competitiveness. In a world of big data, users of certain services are most likely unaware that the data they are generating could be used for other purposes;

(e) Financial: the acquisition of big data by NSOs is most likely to be associated with financial costs, especially when big data is held by the private sector. This might create significant financial burdens for NSOs, but the potential benefits might outweigh the costs. NSOs should therefore assess those benefits against the cost of acquiring big data and make their decisions accordingly.

20. In addition to the above, the use, management, processing and storage of big data requires cutting-edge technical solutions and skills that are less likely to be found in NSOs, especially in developing counties. This additional challenge should be addressed by government programmes when planning and budgeting for the public sector.

VII. ROLE OF THE UNITED NATIONS

21. Recognizing the pivotal role that big data could play in socioeconomic development, the United Nations launched Global Pulse in 2013; a flagship initiative intended to accelerate the discovery, development and scaled adoption of big data for sustainable development and humanitarian action.

22. In July 2012, the United Nations Secretary-General formed the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda.[10] A year later, the Panel issued a landmark report[11] calling for a data revolution to improve accountability and decision-making, and to tackle the challenges of measuring sustainable development progress. It also called for collaboration with international agencies, NSOs and the private sector to draw on existing and new sources of data to fully integrate statistics into decision-making, promote open access to and use of data, and ensure increased support for statistical systems.

23. Furthermore, in August 2014, the Secretary-General formed an Independent Expert Advisory Group to make concrete recommendations on bringing about a data revolution in sustainable development.[12] The aim is to provide an achievable vision for a future development agenda beyond 2015, replacing the Millennium

[9] Data scientists investigate complex problems in the fields of mathematics, statistics and computer science, which are broad knowledge areas. A data scientist can probably be an expert in only one or two of these areas at the most, and merely proficient in the others.

[10] See www.un.org/sg/management/beyond2015.shtml.

[11] United Nations, *A New Global Partnership: Eradicate Poverty and Transform Economies Through Sustainable Development* (New York, 2013). Available from www.un.org/sg/management/pdf/HLP_P2015_Report.pdf.

[12] See www.undatarevolution.org.

Development Goals. The Group released a flagship report in November 2014[13] that called for global action, led by the United Nations, to mobilize the data revolution for sustainable development. According to the report this endeavour could be achieved through the following:

(a) Experimenting with how traditional and new data sources, including big data, can be brought together for better and faster data on sustainable development;

(b) Developing new infrastructures for data development and sharing and supporting innovations that improve the quality and reduce the costs of producing public data;

(c) Reducing the data gap between developed and developing countries and between data-poor and data-rich people;

(d) Improving cooperation between old and new data producers, ensuring the engagement of data users, and developing statistical standards to improve data quality and protect people from abuses in a rapidly changing data ecosystem.

VIII. THE WAY FORWARD

24. The present paper has described big data as the process of extracting actionable intelligence from disparate and non-traditional data sources, which might include structured and unstructured data. The importance of big data is paramount in official statistics; a practice commended and supported by the United Nations for its potential in revolutionizing sustainable development and humanitarian practice.

25. However, big data derived from multiple sources at high velocity, volume and variety requires unconventional processing power, analytics capabilities and skills. This poses numerous challenges to NSOs, especially when big data is harnessed to improve the production of official statistics.

26. Against this backdrop, the ESCWA Statistics Division proposes that NSOs in the region undertake the following:

(a) Consider adapting big data analytics tools and systems to official statistics by identify and carrying out pilot projects that can serve as a proof of concept, with the participation, collaboration and support of the private sector and the international community;

(b) Formally address big data requirements in their work programmes by conducting research in related areas and allocating appropriate financial and human resources for that purpose;

(c) Build or further develop their analytical capacities and capabilities through specialized training on the subject.

27. The ESCWA Statistics Division will continue to provide countries in the Arab region with the support (expert group meetings, workshops and guidelines) and advisory services required to ensure synergy with global statistical developments and trends. In addition, the Statistics Division is committed to strengthening the capacities of NSOs to improve the production of official statistics while facing the challenges of measuring and assessing the post-2015 development agenda.

IX. ACTION REQUIRED OF THE STATISTICAL COMMITTEE

28. The Committee is invited:

(a) To consider how big data impact official statistics at present and in the future;

(b) To make recommendations on how the work programme of ESCWA should be adjusted to absorb the issues related to big data, taking into account the points in paragraphs 26 and 27 above, in particular.

---

[13] United Nations, *A World That Counts: Mobilising the Data Revolution for Sustainable Development* (New York, 2014). Available from http://www.undatarevolution.org/report.