ISSN 1727-9917

SERIES

STUDIES AND PERSPECTIVES

ECLAC SUBREGIONAL HEADQUARTERS FOR THE CARIBBEAN

An assessment of big data for official statistics in the Caribbean

Challenges and opportunities

Abdullahi Abdulkadri
Alecia Evans
Tanisha Ash

UNITED NATIONS

ECLAC

SERIES

STUDIES AND PERSPECTIVES

48

ECLAC SUBREGIONAL HEADQUARTERS FOR THE CARIBBEAN

An assessment of big data for official statistics in the Caribbean

Challenges and opportunities

Abdullahi Abdulkadri
Alecia Evans
Tanisha Ash

UNITED NATIONS

ECLAC

This document has been prepared by Abdullahi Abdulkadri, Coordinator of the Statistics and Social Development Unit, Alecia Evans, Consultant, and Tanisha Ash, an on-the-job trainee and also a member of the Statistics and Social Development Unit of the subregional headquarters for the Caribbean of the Economic Commission for Latin America and the Caribbean (ECLAC). The study also benefited from comments from Dillon Alleyne, Johann Brathwaite, Luanne Serieux-Lubin and Robert Williams, also of the subregional headquarters for the Caribbean.

United Nations publication
ISSN 1727-9917
LC/L.4133
LC/CAR/L.485

Printed at United Nations, Santiago, Chile
S.15-01378

Contents

Tables

Figures

Box

Abstract

The data revolution for sustainable development has triggered interest in the use of big data for official statistics such that theUnited Nations Economic and Social Council considers it to be almost an obligation for statistical organizations to explore big data. Big data has been promoted as a more timely and cheaper alternative to traditional sources of official data, and one that offers great potential for monitoring the sustainable development goals. However, privacy concerns, technology and capacity remain significant obstacles to the use of big data. This study makes a case for incorporating big data in official statitics in the Caribbean by highlight the opportunities that big data provides for the subregion, while suggesting ways to manage the challenges. It serves as a starting point for further discussions on the many facets of big data and provides an initial platform upon which a Caribbean big data strategy could be built.

Introduction

In September 2015, the United Nations General Assembly formally adopted a new set of internationally-agreed development goals (IADGs) as part of the 2030 Agenda for Sustainable Development. These Sustainable Development Goals (SDGs), expected to drive development for the next fifteen years, resulted from a broad consultative process, informed by the lessons learned in the implementation and monitoring of the Millennium Development Goals (MDGs). Key among these lessons was that "data are an indispensable element of the development agenda" (United Nations, 2014, p. 10). Earlier in the MDG monitoring process, it was evident that data were largely inadequate to monitor progress in the achievement of the goals. This reality galvanized efforts to strengthen the data collection networks of many countries as the MDG framework was mainstreamed in national development strategies. The result was an improvement in the monitoring and reporting capacities of countries such that, by 2014, 79 per cent of the developing countries had at least two data points on 16 out of 22 selected MDG indicators, as opposed to 2 per cent in 2003 (United Nations, 2014).

Not only has the official data production capacity increased on a global level due to the MDGs, there has been a better appreciation of the usefulness of data in evidence-based decision making. There are also increasing calls for open access to data held by public entities and technology has enabled data users to source data through alternative means.[1] Furthermore, technology has made it easier for data producers to collect and disseminate data at minimal costs. Increasingly, governments look to technology for solutions, and it will be no different for the SDGs.

The MDGs consisted of eight goals and 60 indicators. In comparison, the SDGs, which will come into effect on 1 January 2016, have 17 goals and 169 associated targets. At their last meeting in October 2015, the Inter-agency and Expert Group on Sustainable Development Goal Indicators recorded general agreement on 159 proposed SDG indicators. It is expected that many more indiactors will be added to the list by March 2016 when a global indicator framework for the SDGs will be finalised.

[1] International agencies routinely publish statistics that are not disseminated by national authorities while quoting these national authorities as the source of data.

The comprehensive nature of this framework being developed suggests that "quality, accessible, timely and reliable disaggregated data will be needed to help with the measurement of progress and to ensure that no one is left behind".[2] This, no doubt, will place a greater burden on the National Statistical Systems (NSS) than did the MDGs. Unfortunately, the Caribbean did not fare as well in the monitoring of the MDGs. Among the 24 Caribbean Development and Cooperation Committee (CDCC) member countries[3] and associate members[4] with MDG reporting data available, only four countries had at least one data point on a minimum of 30 of 60 indicators in both 2005 and 2010. As of August 2015, no Caribbean country had reported on 50 per cent or more of the indicators for 2013 using the same criterion. Seven countries improved on their percentage reporting between 2005 and 2013 while 12 regressed. In addition, all countries had a lower reporting record for 2013 when compared with 2010.[5]

The forgoing points to two issues. Despite efforts dedicated to MDG monitoring at the global and regional levels, the Caribbean has had less than an impressive performance in monitoring progress in the achievement of the MDGs at the national level. There is also a significant time lag in reporting, with half of the countries not meeting the reporting record set for 2005 in 2010, and none meeting their reporting record for 2010 in 2013.

The Caribbean is usually regarded as "data poor", not just because data are limited but also because the available data are not disseminated widely and in a timely fashion. Censuses are conducted at regular intervals and surveys such as the survey of living conditions and the labour force survey are done periodically in most countries. Only a few countries publish the results of these censuses and surveys widely, and even fewer make the anonymized survey data available for public use. For the most part, this is a result of capacity and resource limitation of the NSS. There is also an element of resistance to change where NSS decision-makers are unwilling to explore the paradigm shift, such as the incorporation of big data in official statistics, or unable to take advantage of modern technology, for example the use web technology to disseminate data, even when these offer the prospect of enhancing the statistical services that they provide.

The world is awash with information and the volume of data keeps increasing at an astronomical rate. Over 90 per cent of the data available in the world today are reported to have been created in the last two years.[6] The private sector has found value in these data but the public sector is slowly warming up to the idea of looking more closely at the benefits that these vast data sets offer. In the international arena, the United Nations has been championing the data revolution for sustainable development. This began with the launching of the Global Pulse initiative by the Secretary-General in 2009. Global Pulse was formed "to leverage innovations in digital data, rapid data collection and analysis to help decision-makers gain a real-time understanding of how crises impact vulnerable populations" (Global Pulse, 2012, p. 3). In its May 2012 report, Global Pulse alerted the world to the potentials of Big Data, not only in revealing socioeconomic trends but also in complementing official statistics. The report equally highlighted the challenges in harnessing big data for development.

The High-Level Panel of Eminent Persons on the Post-2015 Development Agenda convened by the UN Secretary-General in July 2012 was the first to officially call for a data revolution for sustainable development, including seizing the advantage that new technology and crowd sourcing provide. In responding to this call, the Secretary-General in August 2014 established the Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG). This group was charged with crafting a strategic framework for data revolution for development. The group also served to advise the Secretary-General on practical measures to implement this framework.

---

2 Finalized Text of SDG Proposal, para, 48.

3 Antigua and Barbuda, The Bahamas, Barbados, Belize, Cuba, Dominica, Dominican Republic, Grenada, Guyana, Haiti, Jamaica, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Suriname, and Trinidad and Tobago.

4 Anguilla, Aruba, British Virgin Islands, Cayman Islands, Montserrat, Puerto Rico, Turks and Caicos Islands, and United States Virgin Islands.

5 Based on data from the MDG Indicators database.

6 See http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

The IEAG made key recommendations in its November 2014 report that included developing a global consensus on data, sharing technology and innovations for the common good, creating new resources for capacity development, establishing a UN-led Global Partnership for Sustainable Development Data and establishing an ***SDGs data lab***. These recommendations are bold, specific, targeted and complement other ongoing initiatives at the United Nations. Among these initiatives is the Global Working Group (GWG) on Big Data for Official Statistics established by the Statistical Commission of the United Nations.

The GWG was created in May 2014 following the recognition by the Statistical Commission at its 45th session that "Big Data constitute a source of information that cannot be ignored and needs to be evaluated on its merits." In pursuing its mandate (see Box 1) and its broader terms of reference, the Commission emphasized that the working group should give consideration to links with the post-2015 development agenda, the data revolution initiative, and the Fundamental Principles of Official Statistics, as well as build upon work done by the United Nations regional commissions.

To this end, the Economic Commission for Latin America and the Caribbean subregional headquarters for the Caribbean embarked on this study to gain an understanding of the state of affairs of big data in the Caribbean and to examine the potential for integrating big data in official statistics in the subregion. It is aimed at initiating an informed discussion on the topic and to help guide a regional strategy for addressing big data. In pursuit of this goal, this paper provides a brief historical background of the data revolution for sustainable development as well as a non-technical exposition of the big data concept. The opportunities and challenges that big data presents are also discussed with a view to highlighting the potential benefits of big data in official statistics, while recognizing the obstacles to and risks of incorporating big data in offcial statistics. The paper also provides suggestions on how to mitigate these risks and proposes actionable steps for integrating big data into official statistics in the Caribbean.

Box 1
Mandate of the global working group on big data for official statistics

With these strategic considerations of the Fundamental Principles, the Post-2015 Agenda and the Data Revolution as a basis, and with reference to Decision 45/110, the mandate of the working group is then formulated as follows:

- Provide strategic vision, direction and coordination of a global programme on Big Data for official statistics, including for indicators of the post-2015 development agenda;
- Promote practical use of Big Data sources, including cross-border data, while building on the existing precedents and finding solutions for the many challenges:
  - Methodological issues, covering quality concerns and data analytics;
  - Legal and other issues of access to data sources;
  - Privacy issues, in particular those relevant to the use and reuse of data, data linking and reidentification;
  - Security, IT issues and management of data, including advanced ways of data dissemination, assessment of cloud computing and storage, and cost-benefit analysis;
- Promote capacity building, training and sharing of experience;
- Foster communication and advocacy of use of Big Data for policy applications, especially for monitoring of the post-2015 development agenda;
- Build public trust in the use of Big Data for official statistics.

Source: Terms of Reference and mandate of the global working group on big data for official statistics (UN ECOSOC 2014).

This study is timely, as it is essential that the Caribbean not lag behind in the 2030 Agenda for Sustainable Development even before it starts. Of ten countries that responded to a survey of National Statistical Offices (NSOs) in the Caribbean as part of this study, only four identified any role for big data in the execution of their work. This is against the backdrop that no Caribbean country has achieved a 50 per cent reporting rate on the MDGs for 2013. As the SDGs are poised to be even more data-demanding, the subregion needs to be proactive and forward-looking. NSOs in the subregion need to look at both traditional and emerging sources of data that could help in telling the sustainable development story of the Caribbean. Big data is already part of the fabric of our lives and its imperfections need not be a deterrence to exploring it. In contrast, its potential, given the subregion's data circumstances, should galvanize efforts to make the best out of it.

The strategy proposed in this paper, whilst not a panacea, is intended to serve as a framework for addressing the use of big data in official statistics in the Caribbean. This is in keeping with other international initiatives to ensure that governments are better able to provide high quality data, more frequently and in a more timely manner.

I. What is big data?

The rapid development in information and communications technology (ICT) has enabled information to be generated and shared quickly nowadays. Electronic gadgets, such as cellular phones, satellites, Global Positioning Systems (GPS), and scanning devices, and fora like social media and e-commerce create volumes of data on a daily basis, and in some instances by the second. There are over 7 billion mobile phone subscriptions and 3 billion internet users worldwide. Mobile broadband subscriptions increased from 268 million in 2007 to over 2.1 billion in 2013 (ITU, 2013). The information generated by these media constitutes *data exhaust* and is defined as "the digitally trackable or storable actions, choices, and preferences that people generate as they go about their daily lives" (Global Pulse, 2012 p. 9). IBM reported that over 2.5 quintillion[7] bytes of data are generated daily.[8] The stock of digital data rose from 150 exabytes in 2005 to 1200 exabytes in 2010 (Global Pulse, 2012). This kind of fast moving, high volume data have been dubbed ***Big Data***.

A. Defining big data

Although no uniform definition has yet been agreed upon for big data, the properties of big data are broadly accepted. Sometimes, however, big data and big data sources are synonymously defined. Irrespective of this lack of a uniform definition, it is undeniable that volumes of data exist for traffic, businesses, healthcare, telecommunications and many other spheres of life that can be used beyond their initially intended purposes. These kinds of data are unique in that they are available in greater volume, veracity, variety and at higher velocity– the four V's. The characteristics of big data are manifested in these four V's. Table 1 provides a summary of these key characteristics.

[7] One quintillion bytes is equal to one exabyte and an exabyte is defined as 1018 bytes.
[8] See http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

Table 1
The characteristics of big data

Characteristics	Description	Attributes	Drivers
Volume	The amount of data generated or data intensity that must be ingested, analysed and managed to make decisions based on complete data analysis	-Exabyte, zettabyte, yottabyte[a], etc.	-Increase in data sources -Higher resolution sensors -Scalable infrastructure
Velocity	How fast data is being produced and changed and the speed at which data is transformed into insight	-Batch -Near real-time -Real-time -Streams -Rapid feedback loop	-Improved throughput connectivity -Competitive advantage -Pre-computed information
Variety	The degree of diversity of data from sources both inside and outside an organization	-Degree of structure -Complexity	-Mobile -Social Media -Video -Genomics -M2M/IoT
Veracity	The quality and provenance of data	-Consistency -Completeness -Integrity -Ambiguity	-Cost -Need of traceability and justification

Source: TechAmerica Foundation (2012, p.11).

[a] One zettabyte equals 1000 exabytes (10^{21} bytes). One yottabyte equals 1000 zettabytes (10^{24} bytes).

Big data has gained much attention in recent times in large partdue to the fact that it is available in a timelier manner than other data and can be generated at minimal cost. The low cost to collect, store and process data can be attributed to technological advancements over the last two decades. New sources of data become available every day. This presents an opportunity, not only to the business community, but also to governments. These unstructured and complex data can prove to be invaluable input in governance and policy making and there have been developments in methodologies such as advanced computational tools to enable the exploitation of big data for these and other purposes. The international community, under the auspices of the United Nations, has identified critical roles for big data in official statistics and in particular in assisting in the monitoring of IADGs such as the SDGs.

B. Big data for development

In light of the emerging push for the use of big data in official statistics, it is important to make a distinction between big data for development purposes and big data as a modern type of data. Common characteristics of big data for development entail that data be digitally generated, passively produced, automatically collected, geographically or temporally trackable, and continuously analysed in real time (Global Pulse, 2012). The data have to be generated digitally and can be stored as binary data.[9] They should be produced by daily interactions with digital services. The automatic collection of the data suggests that relevant data are extracted and stored as they are generated by an implemented system. Based on this, mobile phone data, Twitter, Facebook, and online queries can all be classified as big data for development. According to Global Pulse (2012), the following data sources are relevant for global development: data exhaust, online information, physical sensors, and citizen reporting or crowd-sourced data.

Data exhaust creates *networked sensors of human behaviour*, including passively collected data from mobile phones, web searches, purchases, or real time data collected by non-governmental organizations (NGOs) and aid organizations to assist in project monitoring.

[9] Binary data is data that takes on one of two possible states –a zero or one.

Online information is sourced from social media such as Twitter and blogs, news media, job postings, etc. Online information creates a sensor of human intent, sentiments, perceptions, and wants while physical sensors detect changes in human activity. Such changes can be indicated by traffic patterns, light emissions, and satellite imagery of changing landscapes or urban development. Crowd sourcing uses the crowd as a free source of information. Such information is provided through mobile phone-based surveys or hotlines and is used for verification and feedback.

Given its attributes, big data could play an important role in the monitoring of the SDGs in the Caribbean. Examining this potential role and devising ways to actualize it forms an essential element of the data revolution for sustainable development.

II. The data revolution

A noteworthy limitation in the implementation and monitoring of the MDGs was the lack of high quality data, especially for developing countries. Even in cases where these data are collected, they are not made available for years. Countries have, in general, not consistently reported on the MDG indicators. Even though efforts have been made to improve the availability and quality of data that are needed for monitoring and accountability, there are still challenges in accounting for some groups of people, thus hindering the effective dissemination of information to the public (IEAG, 2014).

Reporting on the progress of countries in attaining the MDGs therefore becomes difficult when there is missing information on indicators. In such cases, it canbe difficult to establish whether or not the MDG targets are being met, by which countries, and on which spheres of development. There are also problems in comparing countries when the most current data are not available or not disaggregated at the subnational level to provide useful information (Bello and Suleman, 2011; Sanga, 2011; IEAG, 2014). Another issue is that there are still groups in the world that are virtually invisible. Data may not reflect indigenous populations and persons who live in slums. A number of children are unrecorded. About 57 million infants worldwide were not registered with civil authorities in 2012 (IEAG, 2014).

A. The United Nations as a champion of the data revolution

As the world prepares to embark on a new development agenda, there will be an even greater demand for data. Since 2013, the United Nations Statistical Commission has been addressing the issue of how big data can contribute to satisfying the data needs that will arise from the implementation of the SDGs. The theme of the March 2013 United Nations Statistical Commission's Friday Seminar on Emerging Issues was "Big Data for Policy, Development and Official Statistics." The meeting concluded that big data is a potential source of information "**that cannot be ignored by official statisticians and that official statisticians must organize and take urgent action to exploit the possibilities and harness the challenges effectively"** (United Nations Economic and Social Council, 2013a, p.2). The session was motivated by the fact that with the prevalence of the internet, mobile devices and other technological gadgets, a new type of data has emerged that has an advantage over traditional data sources in its timeliness. In this era, "data are no longer centralized, highly structured and easily manageable, but are

highly distributed, loosely structured (if structured at all), and increasingly large in volume" (United Nations Economic and Social Council, 2013b, p.2).

As argued in the concept note of the Friday Seminar, the desirable qualities of big data include volume, type, and speed. There is a rapid increase in the volume of data that is made available on a daily basis via the web, mobile devices, and the Information Technology structure. The number of device users and digitization of transactions have increased the speed at which such data are created. Data are not represented in the traditional format but are generated instead in unstructured and semi-structured forms. The United Nations Economic and Social Council (2013b) justifies a careful consideration into big data analytics by illustrating how inadequate and inappropriate traditional data management tools are in analysing the volume of data that is now available. Furthermore, it would be unrealistic and expensive to attempt to tailor traditional infrastructure to process big data. These issues call for new means of processing and analysing data. Even though there are existing structures for processing big data, to fully capitalize on this new data form, there is the need to invest in Data Scientists who are able to use advanced tools to decipher the meaningful insights embedded in big data.

The Friday Seminar advocated for a place for big data in official statistics. Previously, the High-Level Group for Strategic Developments in Business Architecture in Statistics[10] raised concerns about how NSOs will approach big data. Should NSOs actively engage in capitalizing opportunities that big data now affords or should they become a third party that certifies data from private sources? There is a possible role for NSOs in collecting non-traditional data and to use them to complement traditional sources of official statistics.

In furthering the discussion on big data, the theme of the Friday Seminar on Emerging Issues at the forty-fifth session of the United Nations Statistical Commission in 2014 was "Managing the Data Revolution: Integrated Statistics and Partnerships in Data for Statistical Organizations in the Post-2015 Era." This session facilitated interactive discussions between NSOs and other stakeholders regarding formulating statistical objectives in relation to the post-2015 development agenda. These discussions were vital because the post-2015 development agenda will require partnership of traditional and non-traditional data producers – ranging from civil society to research institutions. There are institutions that are already collecting the type of information that will be relevant to this development agenda. For example, the initiatives of the World Resources Institute have made available information on food security, ecosystems and human well-being.[11] There is a need to develop plans on how such data can be harnessed from the various private and public sources.

The post-2015 development agenda creates the need for "integrated economic, social and environmental statistics".[12] The forty-fifth session of the United Nations Statistical Commission rallied NSOs to rise to the cause of strengthening their operations by introducing "innovations in their structures, systems and processes to increase their efficiency and ability to adapt to new requests."[13] Most importantly, NSOs should standardize their statistical systems to facilitate the sharing within and across statistical agencies of not only data but also processes and methods.

In his report to the forty-fifth session of the United Nations Statistical Commission, the Secretary-General outlined a clear case for an energized focus on big data. The report argued that big data could fill the gaps in traditional data and noted that, since traditional national household and business surveys were losing popularity, big data could provide timely and relevant statistics. The report highlighted the fact that the private sector had a pivotal role in providing big data yet enough was not being done in most countries to fully engage and incorporate them in the process of using big data in official statistics. The report surmised that "by incorporating big data sources into their production of official statistics, national, regional and international statistical organizations could be better positioned to obtain official

[10] The High-Level Group for Strategic Developments in Business Architecture in Statistics was created in 2010 by the Bureau of the Conference of European Statisticians. They are now called the High-Level Group for Modernization of Statistical Production and Services.

[11] http://unstats.un.org/unsd/trade/events/2014/unsc/data-revolution.asp.

[12] Ibid.

[13] Ibid.

statistics on the economy, society and the environment in terms of improved timeliness and cost-efficiency, and a lessened resource burden" (United Nations Economic and Social Council, 2013a, p.2).

Also tabled at the forty-fifth session of the Statistical Commission, the recommendations made by the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda called for a Data Revolution. This Panel was appointed in July 2012 by the UN Secretary-General to address the post-2015 development agenda. The report, dubbed "*The New Global Partnership: Eradicate Poverty and Transform Economies through Sustainable Development*", was developed by a twenty-seven person panel through consultation across various regions and sectors. Though the Panel lauded the achievements of the MDGs in reducing poverty, child mortality and death from malaria; they highlighted the MDGs' inability to reach the poorest of the poor and the most excluded persons (United Nations, 2013). They noted that the MDGs fell short in addressing issues such as the impact of global conflict on economic development, good governance, sustainable consumption and production, and most importantly sustainable development within the ambits of fully integrating the social, economic, and environmental aspects for the common goal of development.

The report from the High-Level Panel was compiled through series of meetings and consultations. The meetings initiated in New York, where the UN Secretary-General implored the Panel to develop "bold yet practical" visions for post-2015 (United Nations, 2013). The subsequent meeting in Monrovia, Liberia, centred on sustainable development and the necessary economic, political, and social buildings blocks that will facilitate such. The meeting in Bali, Indonesia, resulted in affirmations to enhance the development agenda by having developed countries satisfy their aid commitments and by reforming market activities to achieve sustainable development. It is expected that developing countries will be able to do more in terms of financing their development process as their income rises. From the Panel's perspective, the post-2015 development agenda must be universal and one of equity and sustainability. Most importantly, it must be one of "offering hope – but also responsibilities – to everyone in the world" (United Nations, 2013, p.8). The Panel also called for a rigorous monitoring system to accompany all new goals and for a data revolution for sustainable development with a new international initiative to improve the quality of statistics and information available to citizens" (United Nations, 2013, p. 9).

The IEAG (2014, p. 6) defines the data revolution as "an explosion in the volume of data, the speed with which data are produced, the number of producers of data, the dissemination of data, coming from new technologies such as mobile phones and the Internet of Things, and from other sources, such as qualitative data, citizen-generated data and perceptions data; a growing demand for data from all parts of society." The data revolution within the ambit of sustainable development focuses on fusing these new data with traditional data to create more timely and detailed information. It calls for transparency, increased levels of privacy, and minimizing inequalities in the production, use, and access to data.

The IEAG has cautioned that the whole development agenda can be compromised if there is a loss of confidence in persons who collect and use data. With increased access to data, the risk of harming others also becomes greater. Persons can be harmed in material ways if information gleaned is used in malicious ways such as hacking into bank accounts. Also, data should not be used to cause embarrassment or result in social isolation for any individual or group of people. Occurrences of this nature would lead to distrust and therefore undermine the potential of big data serving as a tool of social development.

The IEAG reiterated the need to combat the potential of increased inequalities. Data that cannot be accessed by all have limited use and impact. With poverty, lack of education, language barriers, and lack of technological infrastructure some individuals are excluded from the growing body of data and information. For example, while the price of mobile broadband subscription is less than 0.1 per cent of GDP per capita in France and Korea, it exceeds 10 per cent in Honduras, Bolivia, and Nicaragua. Therefore, for big data to work effectively in social development, "the information society should not force a choice between food and knowledge" (IEAG, 2014, p. 7).

Central to the data revolution is the improvement of national statistical agencies and the liberalization of data. National statistical agencies have to revolutionize and modernize the sources, ways, and frequency with which they collect data. Data liberalization entails making data open to all.

Pivotal to the data revolution is the acknowledgment that the new SDGs will require detailed disaggregated data that go beyond the ambit of traditional data. Data selection should be based on quality, cost, and timeliness and big data has all of these attributes. It is an innovative means of complementing traditional data sources and filling the gaps were they exist.

Some key principles that are essential for a successful data revolution for sustainable development have been identified by the IEAG (2014). These include data quality and integrity, data disaggregation, data timeliness, data transparency and openness, data usability and curation, data protection and privacy, data governance and independence, data resources and capacity, and data rights. The use of poor data can result in incorrect and misleading policy prescriptions and guarding against this necessitates a universal framework to ensure the quality of the data is maintained, especially for data to be utilized for official statistics. With the new dimensions of SDGs, data should be disaggregated as much as possible to allow better comparisons. The quality and value of data will be strengthened by a steady flow of timely data from national and international sources and also from the sources that generate big data. The data revolution will enhance sustainable development if data on public matters are open by default. The development process will continually be hampered if data are not made available in a standardized machine readable form (technically open) and can be used and reused without restrictions for both commercial and non-commercial use (legally open). Not only should these data be open and public, they must also be available in a form that is usable. In promoting openness, however, every effort must be made to protect privacy A successful data revolution will guarantee that policies and legal frameworks are in place to minimize the risk of violating individuals' privacy. Importantly, where proper procedures are adhered to, data producers and users should have protection from recrimination.

In order to address these issues identified by the IEAG, the United Nations Statistical Commission in 2014 supported the formation of a Global Working Group on Big Data for Official Statistics. The objective of this group is to "make an inventory of ongoing activities and examples regarding the use of big data, address concerns related to methodology, human resources, quality and confidentiality, and develop guidelines on classifying various types of big data sources" (United Nations Economic and Social Council, 2015b, p. 2). As pointed out in the introduction to this paper, the mandate of the group incorporated efforts made by international statistical agencies and other regional commissions and zeroed in on the needs of developing countries in terms of the legal frameworks and disadvantages that may exist regarding information technology structure. As the United Nations Economic and Social Council indicated, big data is supported by the Fundamental Principles of Official Statistics. "**Based on these principles, it is not only a possibility but almost an obligation for statistical organizations to investigate and pursue the use of big data sources for statistical purposes, as long as the rights of privacy and confidentiality are strictly observed**" (United Nations Economic and Social Council, 2015b, p. 6).

The first meeting of the working group was held on October 31, 2014 in Beijing following the International Conference on Big Data for Official Statistics. That conference focused on three big data sources: mobile phones, GPS, and other tracking devices; satellite imagery; and Twitter and other social media. Discussions at the conference surrounded the types of data sources that exist and the necessary means of exploiting them as well as the benefits and challenges of these sources. In all, a case was made for big data. Information on projects presented at the conference indicated that big data can improve official statistics and that "these innovative tools can be added to the portfolio of a statistical system so as to complement existing applications or provide more flexible, short-term solutions to highly relevant policy questions" (United Nations Economic and Social Council, 2015b, p. 3).

In 2014, The United Nations Statistics Division (UNSD) and the United Nations Economic Commission for Europe (UNECE) conducted a global survey on existing big data projects in official statistics. The results of the survey revealed interesting findings. Fifty seven projects were identified, 54 from countries and three from international organizations. Most countries were yet to define how big data will be incorporated into official statistics and only a few countries had a long term vision for using big data in official capacities. Some had been making strides in carrying out pilot projects to test whether or not big data can complement national statistics.

However, many projects were done through international collaborations to minimize the risk of failure. Such partnerships included the United Nations Economic Commission for Europe Sandbox project and the Global Working Group on Big Data for Official Statistics and the Eurostat Task Force (United Nations Economic and Social Council, 2015b; UNSD/UNECE, 2014). Most countries had no specific privacy framework to deal with big data. Some chose to only transfer data aggregates to statistical offices and examine the data at the source. From the survey, the limitations identified centered on methodology, analytical skills, information technology and access to data sets. The survey showed that access to data sets may require collaboration with the private sector which owns such data. It has been acknowledged that access to big data sources will determine the success or failure of these projects and accessing international data may even be more challenging.

From the global survey, big data was identified to be of potential use in financial, social and demographic, and price statistics. Interestingly, most of the projects were being conducted through private-public partnerships. These partnerships were generally with data providers for ease of access to big data. According to Landefeld (2014), official statistical agencies can benefit from partnerships with firms such as Google, Baidu and Yahoo. Such firms can assist statistical agencies in understanding and accessing official statistics through broader internet dissemination and branding. This association can also help search engines identify the relevant data that match their customer's statistical needs. Successful collaboration in joint projects include the IBM and the US Bureau of Economic Analysis' development of hedonic computer prices, Google and the development of regional search engines and Chrysler's partnership with other similar manufacturers to develop new auto and truck pricing data (Landefeld, 2014).

For public and private collaborations in big data to be mutually beneficial, both parties must recognize and acknowledge the possibilities of such gains. In some countries, statistical agencies may need to initiate these collaborations and indicate the benefits to the private sector. Wherever such collaborations exist, transparency must be emphasized on how data are collected and estimated. Though important, this can be difficult. Private firms have a need to protect their data, considered to be intellectual property, to maintain competitiveness. Without trade secrets, products can be easily replicated and owners prevented from recouping their investments in research and development. Due to this, privacy and confidentially laws must be established and enforced. Landefeld (2014) highlights that this is especially important considering that there may be distrust of the government in light of cases where the government has accessed private big data for national security purposes without proper legal authorization. There are also cases where official statistical agencies have inappropriately used confidential data (Landefeld, 2014). Caution is necessary to prevent repeated occurrences of such incidents. When dealing with business data, care must be taken to protect details of business practices and personal information of customers. The public should also feel reassured that, in going about their daily activities, their personal data are secured and protected from unlawful access and disclosure. Also, data that could have discriminatory outcomes in terms of employment, loans or government programmes must be protected. Importantly, the government, as a result of gaining access to big data, should not exploit the citizenry by using the information so gathered in unauthorized or unethical ways.

B. A global agenda for big data

Having established a case for the use of big data, discussions on big data at the forty-sixth session of the United Nations Statistical Commission progressed to how big data could be used to monitor the SDGs. The theme of the Friday session was "The Development of an Indicator Framework for the Post-2015 Development Agenda: Towards a nationally owned monitoring system for the SDGs" and discussions focused on the monitoring requirements for the SDGs at the national level.[14] The session reiterated that with the new goals and the requirements for disaggregated data, there will be increased data requirements placed on NSOs. The session facilitated discussions by stakeholders on how to complement traditional statistics to meet the data needs.

[14] http://unstats.un.org/unsd/statcom/statcom_2015/seminars/post-2015/default.htm.

Undoubtedly, this will require investments to improve the data collection capacity, particularly of small island developing states, fragile states and landlocked countries (United Nations Economic and Social Council, 2015c).

The report of the forty-sixth session reiterated the importance of a data revolution and the role of NSOs in coordinating non-official sources of data and validating such to ensure that all is in keeping with the Fundamental Principles for Official Statistics (United Nations Economic and Social Council, 2015c). Data quality is paramount to the success of incorporating unofficial data sources in official statistics and the report designated the United Nations Statistics Division as the entity to provide guidance for the development and implementation of such. Notwithstanding, it is well-recorgnized that countries have different priorities and as such will participate in the data revolution within the ambits of their national priorities.

The Statistical Commission recommended that the Global Working Group on Big Data proceed by focusing on the following (United Nations Economic and Social Council, 2015c):

- Sharing of experience of lessons learned from big data projects;
- Quality concerns related to the use of big data for official statistics;
- Coordination of the work of the Global Working Group with the work undertaken at the regional level, especially in Europe, in order to build on achievements and repurpose and reuse outputs;
- Existing technology gap of developing countries and need for funding for developing countries to be able to take part in big data projects;
- Legal frameworks for access to big data sources, especially from the private sector, while addressing privacy concerns;
- Skills, human resources and capacity-building needed to address big data adequately, especially in relation to data science and information technology;
- Demonstration and explanation of the limitations of big data for official statistics;
- Building of partnerships, which is necessary to exploit and harness big data, especially with the private sector, research institutes and academia.

Also, other data sources such as e-commerce transactions should be explored in addition to the mobile positioning data, satellite imagery and social media data. These data sources must be seen as secondary and complementary to administrative data.

III. Examining big data

A. Opportunities presented by big data

The benefits of big data are still emerging but there are already clear indications that big data could enhance timeliness in the production and dissemination of information while being more cost-effective than traditional means. Whereas conducting a new survey could cost over US$20 million in the United States of America (USA), using big data can produce similar results at one-fifth of this cost. For instance, a study on US foreign investment matched firm level data to plant level data and expanded the available data from 100 to over 600 industries covering all 50 states at a cost of US$3 million (Landefeld, 2014). EUROSTAT did a similar application by matching trade data and data from a business register to compile a detailed database of importing and exporting firms. It has already been identified that big data can be used for public health purposes such as tracking disease outbreaks. Other public sector applications include "geographic planning for infrastructure and the provision for services; for understanding behavioural responses and characteristics of the population for designing health and unemployment insurance, for tax and other policies; …and for evaluating the effectiveness of government programs" (Landefeld, 2014 p. 13).

There have been studies on the economic value of big data. In the United Kingdom (UK), Deloitte produced a report that estimated the economic value of data held by the public to be in the realms of £5 billion per year (IEAG, 2014). Of this amount, £400 million was attributed to saving the lives of cardiac patients and between £15-58 million was ascribed to using live data from the *Transport for London* app. Another study conducted by McKinsey Global Institute estimated the global value of big data to be around $3 trillion yearly (IEAG, 2014). In Uganda, UNICEF established a social monitoring platform that controlled the spread of an infectious banana disease which could have otherwise cost the country over $360 million annually. The exploitation of big data can also result in increased productivity in businesses. A study mapped a relationship between productivity and employee interaction using call data records. Such results guided changes in the scheduling of Bank of America workers' coffee break. This seemingly minor adjustment in operational rules has resulted in an increase in productivity to the tune of $15 million per year. Instead of one person going on coffee break, the bank allowed workers to take breaks simultaneously to increase interaction amongst employees which ultimately resulted in higher productivity.

Historically, official statistics such as National Accounts have incorporated public and private data that were collected for purposes other than for national statistics (Landefeld, 2014). Such data are often extrapolated to provide early estimates of the state of the economy. Big data facilities this and similar processes when data are automatically collected and analysed in near real-time.

Numerous cases have illustrated a place for mobile phone records, satellite imagery and Twitter and other social media in international development and particularly for developing and emerging economies. Global Pulse (2012) believes that since decolonization and the Green Revolution, the increased access to mobile phones may be the most significant change in the developing country. In 2010, there were more than 5 billion mobiles phones of which 80 per cent were in operation in developing countries. Mobile phone usage in Sub-Saharan Africa is particularly of significant note where it is used to complement weak financial and banking systems, telecommunications, and transport infrastructure. The director of mobile banking strategy of GSM Association, Gavin Krugel, notes that there are 18,000 new mobile banking users per day in Uganda and 15,000 and 11,000 in Tanzania and Kenya, respectively. In the Caribbean, mobile service penetration rate was in excess of 100 per cent in 2014 for 13 countries.[15]

Big data does not proffer solutions to development problems but, if used properly, has significant value to the statistical community. Big data analysis is useful in identifying correlations and stylized facts in large datasets (Global Pulse, 2012). Even though correlations by no means imply causation, the presence of correlations can yield interesting insights. There are situations where new data sources can provide cheap and fast proxies for official statistics. The consumer price index (CPI) is reported on a monthly basis. A study done by researchers at the Massachusetts Institute of Technology (MIT) accurately estimated inflation rates by examining daily prices of goods sold on the web. The method "may help detect inflation spikes sooner than traditional methods, or offer new insights into the transmission of price fluctuations to various goods and areas" (Global Pulse, 2012, p. 37).

Mobile positioning is another useful source of big data. Mobile positioning data provide very accurate and precise geographical information. Therefore, they can be useful in providing mobility statistics and in conducting spatial behavioural analyses. Mobile positioning data collection is relatively easy to set up and is not demanding of the respondents. The drawback to using mobile positioning data lies in the limits to the requests that can be made via mobile network operator's network infrastructure and the possible small sample size that could result from the need to seek permission from subscribers (Tiru, 2014).

In developing countries, the potential use of big data goes beyond supplementation of official statistics. It can also be used to drive government programmes and services such as in providing "health, education, financial services and agriculture" to the poor (Word Economic Forum, 2012 p.3). Mobile data can be predictive of the people's needs and can be used to target the relevant population. With the prevalence of mobile phones in developing countries, real time data can be provided at very low cost on development issues. At the International Conference on Big Data, various statistical applications and national experiences were presented surrounding the use of mobile data in poverty mapping and in estimating population and mobility patterns in cases of disease outbreaks (United Nations Economic and Social Council, 2015b). The World Economic Forum (2012) also highlighted that mobile money services could indicate spending habits and assist in generating credit histories.

Health information collected through similar means by health workers or provided by the individuals themselves can be useful in predicting and stopping disease outbreaks. Mobile data have been used successfully in tracking persons in Haiti after the 2010 earthquake using the movement of two million SIM cards. The study conducted by researchers at the Karolinska Institute and Columbia University successfully tracked individuals displaced from Port-au-Prince. Mobile data were also used to track persons from affected areas in a cholera outbreak subsequent to the earthquake. This information was useful for government and humanitarian organizations in responding to the crises (World Economic Forum, 2012).

[15] ITU Database.

Geospatial applications for smart phones were similarly applied in Trinidad and Tobago to contain an outbreak of the chinkingunya virus in 2014 where the Ministry of Health used these applications to identify the location of infected persons (United Nations, 2015). In times of natural disasters and afterwards, it is difficult to collect longitudinal data on population movements through conventional surveys. With the prevalence of mobile phones in low and middle income countries, mobile positioning data can serve as an effective option. In such studies mobile data have "the advantage of high resolution in time and space, being instantaneously available with no interview bias, and they provide longitudinal data for very large number of persons" (Lu et al., 2012 p. 11576).

The poor are the most vulnerable and susceptible to economic crisis (such as food and oil crises) and environmental disasters and they often have minimal coping mechanisms to recover from these shocks. When evidence of the plight of the poor becomes apparent to policy makers, it is often too late or expensive to respond (Global Pulse, 2012). In some cases, policy makers may even know the main triggers of distress to the poor such as drought, war, climate change, and oil and food crises, but it is still difficult to identify and locate vulnerable and affected groups. Even though Early Warning Systems exist, traditional data sources are rife with limitations such as being time-consuming and expensive to collect, verify and publish. Big data offers a less expensive and more timely alternative to addressing these challenges.

Toursim is a major economic sector for a number of Caribbean countries. Mobile data can complement existing official tourism statistics. Domestic (D), inbound (I) and outbound (O) call data mirror the movements of persons. Such mobile data can also be explored in compiling statistics for balance of payments, monitoring tourism investments, marketing for tourism attractions and for internationals travels. According to Tiru (2014), mobile data can provide the following statistical indicators in the tourism domain:

- Number of trips (I,O,D);
- Number of unique travellers (I,O,D);
- Duration of the visit in a destination country (I,O,D)/ in smaller sub-regions (I,D);
- Breakdown by the home administrative unit within the country (O,D);
- Temporal breakdown: day/week/month (I,O,D);
- Overall duration of the trips in spent nights, hours, days present (O,D);
- Geographic accuracy: country (I,O,D), lower level administrative units (I,D);
- Trajectory of tourism trip (I-only inland, O-only country level, D);
- Repeating visits (I,O,D);
- Destination, secondary destinations, transits (I,D);
- Destination country, transit countries (O).

The daily movements of persons can be tracked using the data provided by mobile devices. It respresents a "more reliable, accurate, timely, and less expensive alternative to traditional methods" to establish "travel patterns between home and work-time locations" (Tiru, 2014 p. 14). For example, statistical data that include daily commuting patterns, average trips per day per person, travel times, distance travelled, and spatial accuracy up to $100m^2$ grid can be collected from mobile devices. This type of data are useful for transportation planning, traffic density data, and new business and outdoor media poster location planning. In between censuses, mobile data provide a source of population statistics. Also, in addition to its use in market, tourism, and long term migration research, mobile data can be used in the areas of spatial segregation, emergency and public safety, social networking analysis and mobility, and climate change.

Other sources of big data are equally important. Satellite imagery has great potential, particularly for the agricultural sector.

The possibility exists for replacement or complementation of surveys on agricultural crop production when satellite imageries become available once every two weeks (United Nations Economic and Social Council, 2015b). A wealth of high frequency, low cost data can be generated from social media such as Twitter and Facebook. United Nations Economic and Social Council (2015b) indicates how the Netherlands has been using Facebook and Twitter data to derive estimates on consumer sentiments. China and Italy have also estimated job vacancy rates via web scrapping. Research conducted by Global Pulse in collaboration with Crimson Hexagon examined over 14 million Tweets to get an understanding of individuals' perception of food, fuel and housing prices in both the USA and Indonesia. The official statistics reporting actual food price inflation were very closely related to changes in the volume of Tweets mentioning the price of rice. Another study by the SAS Institute showed some correlation between social media discussions on unemployment and the actual unemployment rate in the US and Ireland. It found that there was a spike in discussions about unemployment on blogs and online forums up to three months before an official announcement of increased unemployment in Ireland and in the USA. Talks about loss of housing showed up two months before increases in the unemployment rate (Global Pulse, 2012).

B. Challenges of big data

As much as big data presents new opportunities for sustainable development, it comes with many challenges as well. These are mainly related to privacy concerns, access, technology, methodology, and capacity.

In the statistical community, there are genuine concerns about the validity of information scoured from the web. In the case of user-created content, the data could be false or misleading. Also, there are concerns about the wrong perceptions that data collected from the web may convey. There are challenges in understanding the data architecture, interpreting the data, and defining and detecting anomalies (Global Pulse, 2012). Consider Google Flu Trends, which was able to predict nonspecific respiratory illnesses well, but was not as accurate in predicting the actual flu. Persons could have flu-like symptoms and not have the virus, while others could be afflicted with flu and not manifest the usual flu-like symptoms. Hence, care must be taken in analysing and interpreting big data. When adapting big data for use in official statistics, the critical question is whether the sample is representative of the entire population. Many factors contribute to traditional data sets being less biased than big data. Mobile phone users may not be representative of the entire population. For example, we would expect less use of mobile devices by the very young and the very old. Or, if data are only sourced from one mobile service provider, this may not be representative of the entire population of mobile device users. In such cases where the sample is non-representative of the population, generalizations cannot be made beyond the sample. The nature of the data may also encourage judgement based on correlations, some of which may be spurious correlations at best.

Big data analytics are the "tools and methodologies that aim to transform massive quantities of raw data into data about the data – for analytical purposes" (Global Pulse, 2012, p. 13). This entails algorithms or advanced visualization techniques called sense-making tools. The techniques are used to identify patterns, correlations and trends in the data. Patterns and interactions which would otherwise remain undetected can be highlighted with reality mining. Big data for development uses the three means of reality mining. Continuous data analysis over streaming data is concerned with scraping the Web to monitor and analyse online data. An example of this method is gathering online product prices systematically for real time analysis. The other method attempts to understand persons' needs and wants and pinpoint topical issues, in a way representing an online digestion data that are semi-structured or unstructured. Finally, there is correlation of streaming data (fast stream) with slowly accessible historical data in real time. This method is about correlating and integrating real-time data with historical records. Visualization is another way of analysing big data for development. For example, word clouds are useful

in providing new perspectives on difficult findings. Expertise in these tools and methodologies is scarce and this constitutes a barrier to the full exploitation of the benefits of big data.

When new big data are utilized as extrapolators, certain steps should be followed to ensure its credibility. Big data should be analysed for performance, coverage, and definition against an established benchmark (Landefeld, 2014). After establishing new ways of use, big data should then be evaluated for accuracy against existing extrapolators and benchmarks. Following this is the development of seasonal adjustments factors. Landefeld (2014) cautioned against using complex econometric techniques that are hard to understand and interpret while adding little accuracy to the performance of extrapolators. He noted further that these complex models are inadequate for large scale forecasting and concluded that "if the models are complex or produce only aggregated results, there will be a loss of drill down capacity and links to key indicators" (Landefeld, 2014, p.8).

Therefore, before big data can be introduced into national statistics or be used in any of the promising sectors that have been discussed, issues relating to methodology, data access and quality, technology, legislation, privacy, management and finance must be addressed. Additionally, the cost-benefit of big data integration must be established (United Nations Economic and Social Council, 2015b; Global Pulse, 2012). Privacy is particularly of grave concern in the acquisition, storage and use of big data and must be maintained for individuals and for institutions. In some cases, individuals may not even be aware that they are providing data or have no knowledge about how the data will be used. It is questionable if blogs and social media users actually authorize the use of their data harvested by web crawlers. Violation of privacy could be more worrying since it has been revealed that anonymized data can be de-anonymized when combined with other databases. For example, even if call data records are stripped of all personal identification, when combined with other data such as geo-located tweets or check-ins, it may be possible to reveal individuals' identities. The use of communication metadata can also even reveal patterns between various interacting entities (Global Pulse, 2012; ITU, 2013).

The use of mobile positioning data is not without administrative, technological and methodological challenges. As already emphasized, issues of privacy must be considered when accessing and using data gathered from mobile devices. Such issues can greatly affect the accessibility of data from mobile network operators since they generate, process, and store such data. Operators use data for legal purposes such as in law enforcement and in limited official statistics, but also for commercial purposes such as in geo-marketing, and for internal native technical business purposes such as customer billing. Therefore, "the first step for getting the access to the data is laying down the appropriate legislation that ensures the justification and legality of the data access and processing" (Tiru, 2014, p. 17). There are many players involved in the collection, processing, and dissemination of big data thus care must be taken in clearly identifying and assigning the roles that each entity would play to ensure that the opportunites that big data presents are fully capitalized upon. This calls for a business model that clearly outlines each entity's responsibilities, the beneficiaries of the result as well as the financial arrangements in the implementation of any big data project. There may even be opportunities for data brokerage companies that process and aggregate the data and then pass this on to the end users (Tiru, 2014).

As it relates to technology, it is necessary to clearly identify the available data sources and whether or not they are sufficient to satisfy the expected results. The end application of the data will greatly influence the quality of the data that is required. Whereas simple call detail records (CDRs) are appropriate for tourism statistics, clearly a more dense and accurate data set is needed for a transportation model (Tiru, 2014). There are also issues surrounding technical ability and resources to process mobile data. The resources and maintenance costs will escalate as the delivery window gets smaller. Near- real time results require more resources to generate than the processing of historical data. Resource needs and maintenance costs can also rise due to domain-specific processing requirements. Core processing may be similar across domains but domain-specific processing demands individual methodology thus adding to the resource needs and maintenance costs.

There are certain types of legislation that govern the use of data. These include the privacy protection legislation, the telecommunications data processing legislation and the Statistics Act.

Some data, for example data on apps, contain highly sensitive data about the mobile device owner. The collection and processing of such data must be governed by strict privacy and confidentially agreements that protect the subscribers. Also, consent must be sought from the subscriber or the legal entity.

The alternative is to transform the data in anonymized data so as to protect the subscribers' identity. In using mobile data for official statistics, special care must be exercised to maintain anonymity.

As with any data set, big data could contain errors. Such errors in mobile positioning data may arise from accidental roaming subscribers, missing values due to system down-time when no data are recorded or incorrect data are being recorded, whether from duplications in recorded data or incorrect time values, or incorrect geographical coordinates from antennae. There may even be recording of non-human related data from machine-to-machine devices. If such errors are recognized there needs to be a methodology to deal with them to reduce the possibility of biased results and decreased quality. Spatio-temporal algorithms may be useful in identifying undesirable characteristics within the data.

There are domain-specific methodological issues. The domain in which the data are used will greatly influence many aspects of big data. Definitions and calculations will vary depending on the end use. For example, Tiru (2014) highlights that the definition of environment will differ based on whether the mobile data are being used in a tourism statistics domain or for a short term migration statistics domain. There are other problems that can also arise. In terms of tourism statistics, these data can be under-representative if tourists decide to use local SIM cards instead of incurring roaming charges.

For statistical purposes, the sample that is compiled from mobile data is used to represent a general population. The use of reference data, whether from population census results, mobile network operators' costumer profile, or national transportation is pivotal to accurately extrapolating to an entire population. There are various selection biases that the use of reference data can assist in eliminating. In some cases individuals may have multiple devices from more than one network provider. Relatedly, there is the question of how to deal with mobile data from more than one mobile network operator. Will the data be merged and analysed or will they be examined separately and merged at the end?

A general disadvantage of the use of big data is the loss of control or independence of statisticians. The nature of big data implies some reliance on external sources. If for some reason companies decide on halt their data collection or change the scope of the data collected, official statisticians will be handicapped as they have no control over the matter (Landefeld, 2014). Closely linked to this is the inaccessibility of valuable data. The UNSD/UNECE (2015) survey indicates that the biggest challenge faced by most big data projects is limited access to datasets. Corporations may prefer not to share their data for privacy and other reasons. In some cases, the data may be stored in a form or a location that makes it difficult to access or transfer. Global Pulse (2012) highlights that even with the UN system, it can be difficult to get agencies to share their programme data. To fully realize the benefits of big data, it will be necessary to promote partnerships between the private and public sectors to guarantee reliable data streams and access to backup data. Global Pulse advocates data philanthropy, where data owners anonymize their data and make it available for whoever wants to convert this into meaningful data to analyse people's behaviour. Data should be made open by default by government, international research institutions and the public sector and they should also advocate for other agencies to allow open access to their data. This means that data should be readily made available to anyone who wants to use it, reuse it or to distribute it. The G8 countries have recognized the economic potential of open data and therefore expect governments to publish these data by default.

IV. Application of big data

A. Big data in the Caribbean

In as much as big data is still evolving, there are already numerous examples of big data applications, although only a few are from the Caribbean. The most cited examples of big data use in the Caribbean region originate from Haiti, where mobile phone data collected from 2.9 million Digicel subscribers were used to examine the movement of people after the 2010 earthquake (Lu et al., 2012). This study spanned 42 days before the disaster to 341 days after the event. The relevance of the study lies within the insights it yields in understanding the movement of persons during and after natural disasters to better inform and organize disaster relief by the relevant agencies. The results suggest that population movement after and during a disaster is relatively predictable using big data. Immediately after the earthquake, around 19 days after, the Haitian population had decreased by 23 per cent. The study highlighted a strong correlation between the destinations of people after the earthquake and their destination during normal times, particularly to areas where individuals have strong social ties.

Chunara et al. (2012) documented a study, also done in Haiti, that sought to understand disease epidemiology. A mixture of official and unofficial statistics were used to study the cholera outbreak in Haiti after the 2010 earthquake. Unofficial data sources included Twitter and HealthMap while official data were collected from the Ministere de la Sante Publique et de la Population (MSPP). The study examined the correlation among tweets, government reported cases of cholera, and HealthMap news media reports in the first 100 days of the outbreak. The study concluded that alternative data can provide a timely and effective means of informing intervention during a disaster and that big data served as a source of early data of unfolding events before official data were reported. In the case of the cholera outbreak such data were available up to two weeks ahead of the official reports.

The United Nations (2015), in its 2015 MDG Report, also cited the use of geospatial information in supporting health care and designing social interventions in response to the chikungunya virus outbreak in the Caribbean in 2014.

Apart from these examples, there is a dearth of documented big data application for development in the region.

There are not many examples of governments in the region complementing official statistics with big data sources. If such applications exist in other realms beyond official uses, they are not documented.

B. Big data in other regions

There are numerous examples of how big data has been used in an official capacity in other regions. Eurostat launched three big data projects in 2013. These projects included using the internet for the collection of information society and other statistics, the use of mobile positioning data for tourism statistics and collection of price information via the internet. Karlberg and Skaliotis (2013, p.2) pointed out that these three early applications of big data highlighted the potential issues of "privacy, methodology, trust, access to third party data, skills development, [and] technology."

Since 2009, Estonia has been using mobile positioning data in tourism statistics (Karlberg and Skaliotis, 2013). The initiative started when border surveys fell victim to financial cutbacks. These tourism statistics are used in the calculation of national balance of payment. A feasibility study commissioned by Eurostat emphasized the benefits of using mobile data to document tourist movements. The method is cheap and offers better time-space insights in comparison to traditional data collection methods. Also, such data are highly quantitative and thus a myriad of methodologies exist to process and analyse them. Mobile data can accurately describe people's behaviour within a destination and even document trips to locations where it would have been difficult to reach tourists with a questionnaire. The method however lacks qualitative rigour. It may be difficult to decipher information such as purpose of visit or mode of transportation.

In Europe, big data is being used in the compilation of the consumer price index. With ecommerce rising from 15 per cent in 2005 to 35 per cent in 2012, online shopping is quickly replacing the physical market place. Now, virtually all products that households use along with product information are available online. This means the internet can be used to collect information on prices. The Netherlands pioneered a study in 2013 to use the internet to collect prices on real estate, movie tickets, clothes, and airline tickets. The project utilized internet robots to collect prices from various websites. This initiative was motivated by the need to "reduce costs of data collection, reduce burden on respondents and to produce results much faster as compared to traditional prices collections (Karlberg and Skaliotis, 2013, p. 6)". This along with other studies suggests that there is a potential for the use of internet to collect prices to construct price indices.

Karlberg and Saliotis (2013) also outlined a study conducted by The European Commission's Directorate-General for Communication Networks, Content and Technology to examine how Internet as a Data Source (IaD) can be incorporated in traditional statistic sources. Such methods include user-centric, network-centric and site-centric measurements. Whereas user-centric looks at client side behaviour of individual users and network-centric measures properties of the network, site-centric captures data from webservers. Of the three, the study utilized user-centric and site-centric measurements. After examining household and business surveys, the conclusion reached was that digital footprints cannot measure offline items nor items that elicit opinions. The study findings indicate that it is not possible to collect all data over the internet. While internet use of individuals can be measured and e-commerce and e-skills are feasible, activities such as use of computer and access to Information and Computer Technology (ICT) cannot be measured. With some redefinition, the use of e-government can also be measured.

The quality and efficiency of health care can be improved through big data applications. There is a growing body of medical data such as mammograms, 3D MRIs, CT scans and x-rays that are collected routinely that have the potential to improve the way in which health care is administered resulting in cost reduction (ITU, 2013). Nowadays, technology even permits the remote monitoring of patients through machine-to-machine communications. For example, the glucose level of persons with diabetes can be monitored and tracked remotely in near real time from the information generated by their glucometer. From this data, it is possible to not only track patients, but also to inform treatment decisions.

This means that "exploiting remote patient monitoring systems for chronically ill patients can reduce physician appointments, emergency department visits and in hospital bed days; improving the targeting of care and reducing long-term health complications (ITU, 2013, p. 8)". Fitness products and sleep monitors serve similar purpose in tracking and analysing individuals' movements and performances. These are generally referred to as quantified self.

The large volume of health-related data that are produced can also prevent the over and under treatment of patients and ultimately reduce the cost of delivering health care services. This is avoided by finding the most cost effective methods to administer care by analysing the differences in patients, healthcare costs, and outcomes. Such research can have a global impact on how health care is administered. A disease such as pneumonia which results in the death of over 1.2 million children below the age of five is preventable with low cost low tech medication and care. However, the disease is becoming more resistant to conventional antibiotics. The analysis of volumes of healthcare data can provide a broader scope under which to examine individual diseases. By examining trends in global diseases, medical and pharmaceutical experts can make informed R&D decisions by modelling future costs and demand for their products.

ITU (2013) illustrates how big data can improve the movement of people. Data are gathered daily from GPS, traffic cameras, and passengers swiping public transport passes. Such data assist in understanding, predicting, and solving traffic-related problems such as traffic jams and routing emergency vehicles more effectively. Two California-based companies, Drivewise.ly and Zendrive aim at making drivers safer and improving performance while being more eco-friendly. This is all premised on the fact that data collected from smartphones, GPS etc. can reveal patterns and habits in commuting resulting in improved efficiencies in transportation and improved energy efficiencies through reduced fuel consumption and emissions. Cellular phone records can also assist in transportation modelling. A study was conducted in Cote d'Ivoire using 2.5 billion call and text records from more than five billion users over a five month period. The research entailed selecting calls made from residential areas in the evenings and then tracking the calls made from the same phone throughout the next day. This process generated a mobility map to indicate the number of people that commute and also established where and when. A similar study was done in Seoul to improve the public bus services offered at nights. Previously, the night service was based on the day service schedule. However, this was not efficient since the patterns of movement of people differ during the day and at night. By examining over 300 million Call Detail Records provided by Korea Telecom, the efficiency of the bus system was improved by establishing the optimal number of locations for the night bus to stop. The results informed decisions to use routes at nights that would have been congested in the days, avoid certain stops such as the Seoul Art Center, and include popular stops such as the Konkuk University. In total, seven new night bus routes were added resulting in reduced transportation cost for citizens.

Some social science and policy applications were highlighted by Global Pulse (2012). Research conducted at both Northeastern University and MIT indicated that mobile positioning data could be used to detect changes in human activities. The Northeastern University experiment was able to predict with 93 per cent accuracy an individual's location based on past movements indicated by their mobile device. The MIT research concluded that the data that mobile devices generate can quantify human movements. Big data has assisted in public health studies. Online data can be used for syndromic surveillance (infodemiology). The Google Flu Trends launched in 2008 showed that it is possible to estimate level of flu activities in regions of the USA using big data. There is also a Google Dengue Trends that has yielded similar results. Johns Hopkins University modelled the flu rate, using more than 1.6 million health related tweets and found a 0.958 correlation with the official flu rate. The same study also showed the possibilities of using Facebook updates to identify college students drinking problems.

In additon to the documented application of big data presented in this chapter, a synopsis of cases in which big data was used to complement or replace traditional data is presented in Annex 1 of this paper. The list is by no means exhaustive, but provides an indication of the various arenas in which big data has been or is currently being applied.

In the UN System, Global Pulse and the UNSD have been involved in projects promoting big data and studies documenting the use of big data, respectively. Annex 2 presents a synopsis of selected big data projects that have been executed as part of the Global Pulse initiative[16] while Annex 3 reports more specifically on big data projects implemented by NSS across countries and organizations (UNSD/UNECE, 2015).

[16] See http://www.unglobalpulse.org/projects.

V. Big Data: the Caribbean reality

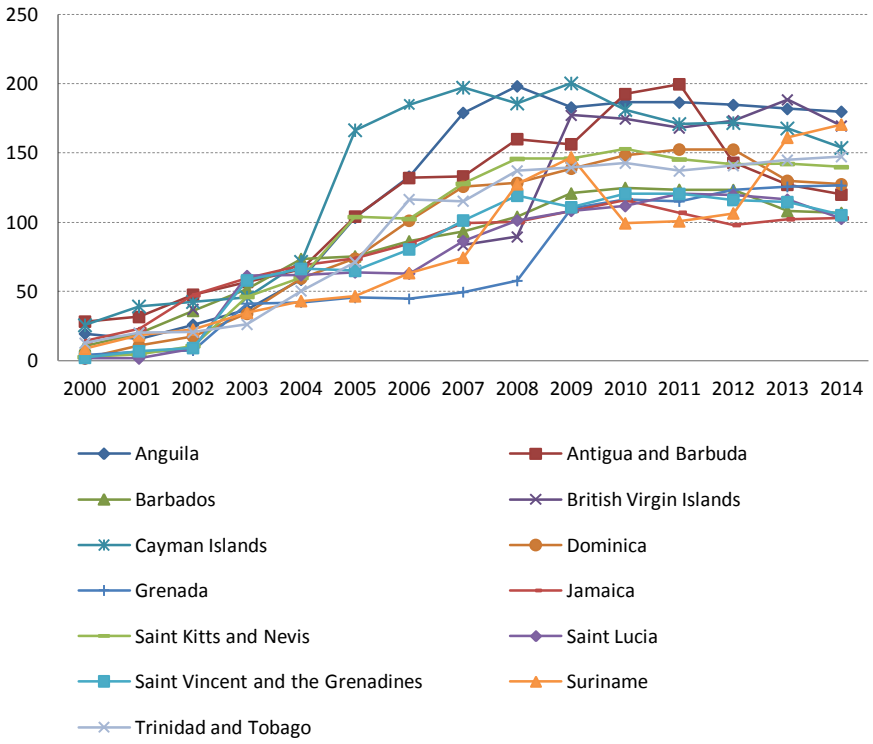
A. A situation analysis

1. Prevailing environment

The Caribbean is one of the subregions with the highest mobile connectivity. Mobile service is very prevalent and internet use is on the rise. In 2014, 13 Caribbean countries had mobile service penetration rate in excess of 100 per cent[17] with most countries reaching this feat as early as 2009 (see Figure 1). , On average, the penetration rate for the 13 countries was 11 per cent in 2000, increasing to 83 per cent in 2005 and reaching in excess of 140 per cent in 2010. The rate declined slightly to 135 per cent in 2014 (Figure 2). While mobile service subscription in the subregion may be stabilizing, the individual internet subscription is growing (see Figure 3), moving from an average of 7 per cent of thethe population in 2000 to 57 per cent in 2014 (Figure 4). Therefore, it is safe to assume that digital information is being created in the Caribbean, as in the rest of the world, and perhaps at a higher rate when compared to developing countries in other regions. Meanwhile, the region has been slow in reporting on progress on the MDGs. It is ironic that the region performed better as a whole in reporting on the MDG indicators in 2005 than it did in 2010 or 2013. Figure 5 shows the Caribbean subregion's progress in reporting on the MDGs. As the figure shows, no country has reported on up to 50 per cent of the 60 MDG indicators from among 24 CDCC countries whereas four countries did so for 2005 and 2013. It is legitimate to ask if it is that those countries that recorded 50 per cent or greater reporting in 2010 no longer have the data to report on the same indicators three years later. The critical issue sometimes is not the availability of data, but the timeliness of the data. In such instances, big data could provide countries with a more timely option for monitoring progress in the attainment of their development goals.

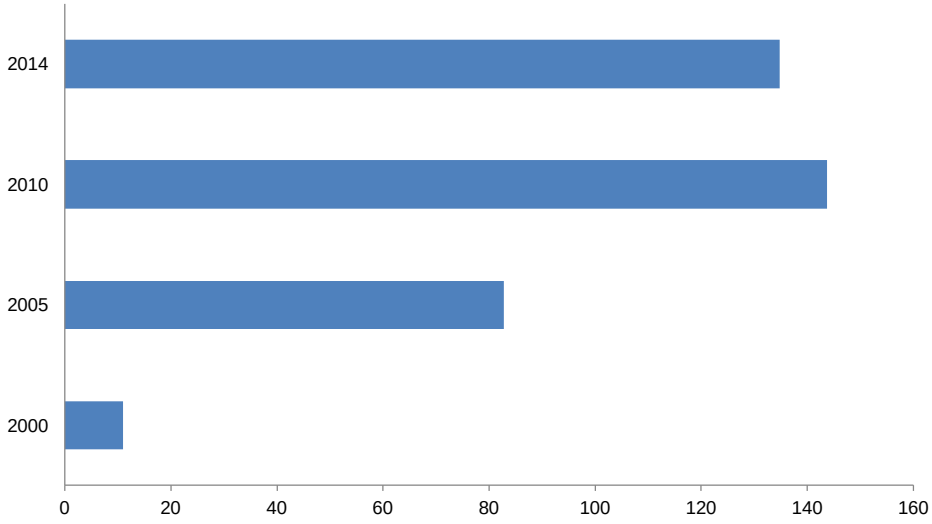
[17] A penetration rate of 100% does not translate to universal mobile service for all citizens. It is common in the Caribbean for mobile service users to have more than one subscription.

Figure 1
Mobile-cellular telephone subscriptions per 100 inhabitants
(Subscription rate percentage)



Source: Based on ITU data.

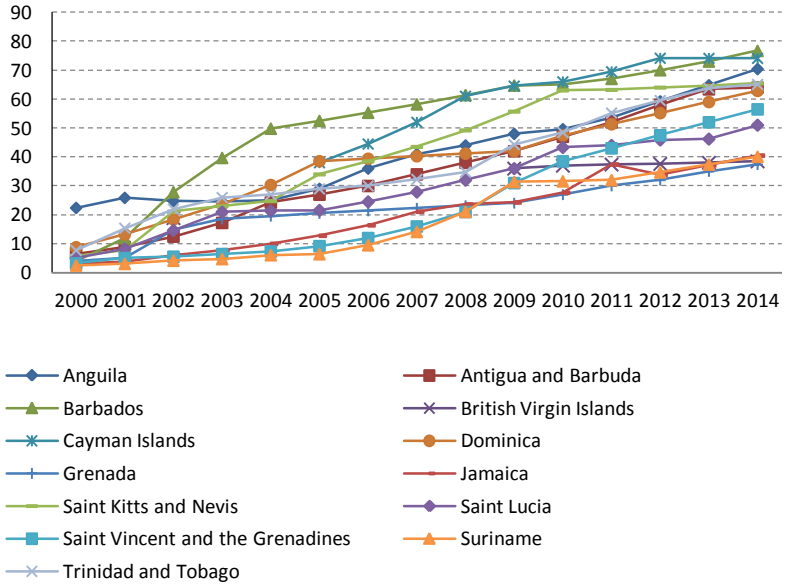
Figure 2
Average mobile-cellular telephone subscription rate for selected Caribbean countries[a]
(Subscription rate percentage)



Source: Based on ITU data.

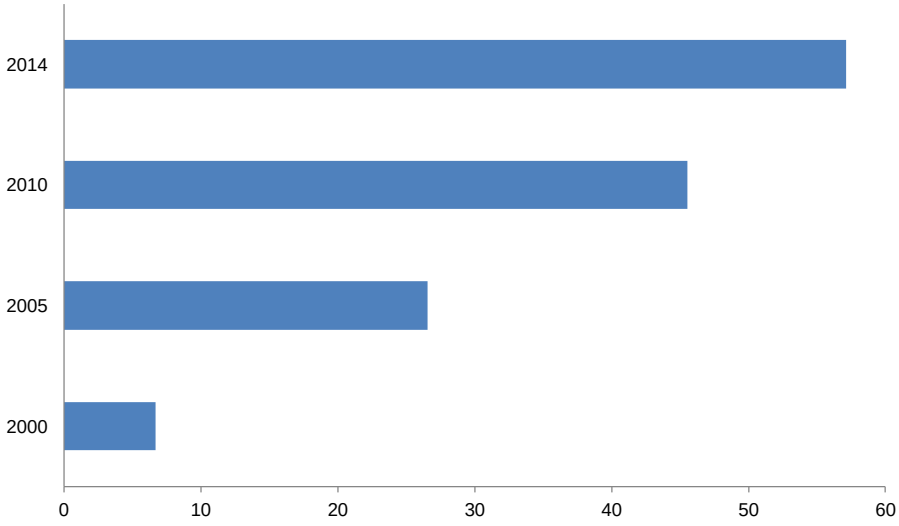
[a] The countries represented are those shown in Figure 1.

Figure 3
Individuals using the Internet
(Proportion of population percentage)



Source: Based on ITU data.

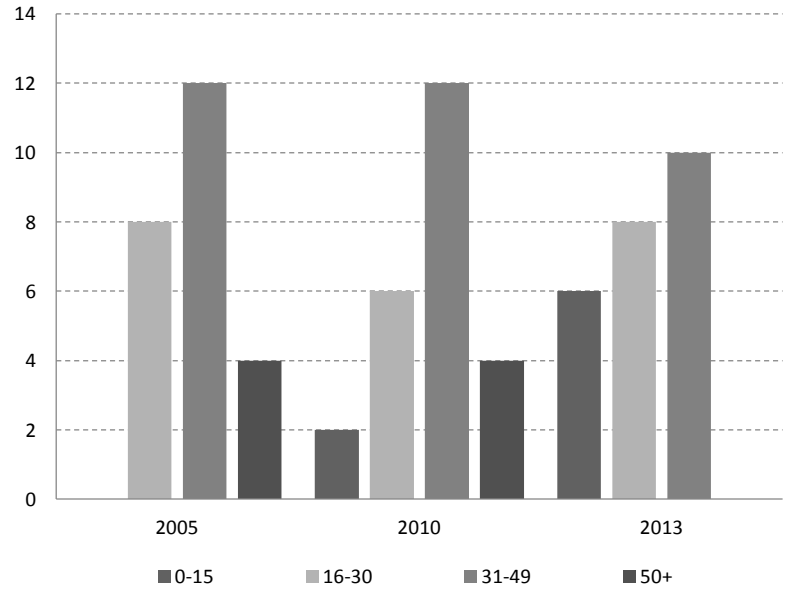
Figure 4
Average of persons using the internet in selected Caribbean countries[a]
(Percentage of population)



Source: Based on ITU data.

[a] The countries represented are those shown in Figure 3.

Figure 5
Proportion of MDG indicators on which CDCC countries have at least one data point
(Percentage)



Source: Based on MDG Indicators data.

2. Issues to address

In the Caribbean context, three main issues emerge from the analysis of opportunities and challenges of big data presented in the earlier sections of this paper and they pertain to classification, benchmarking, and legislation.

Big data is unstructured or at best semi-structured. In contrast, data from censuses and surveys as well as administrative data from which official statistics are traditionally derived are highly structured. Using big data for official statistics will require that big data sources be categorized and the data types classified according to their potential use in official statistics and the area of application. Such classification will indicate the type of big data with multiple utility that can provide the most benefits to governments and organizations in the region.

For instance, Global Pulse reported a study that assessed the potential use of mobile phone data as a proxy for food security indicators. The study, done in East Africa, revealed high correlations between purchases of airtime credit and the consumption of different food items. In the Caribbean, there may be use for such big data not only for food security purposes but also for controlling non-communicable diseases (NCDs) (of which diet is a major risk factor).

Big data can offer no value if the information contained in it is not discernible. Worse still, it could do a huge disservice if the information is erroneously interpreted or applied. It is thus essential that appropriate studies be done to understand the information contained in big data to enable data users to benchmark big data to traditional data sources, especially when it comes to official statistics. Many studies have reported high correlation between big data and sociodemographic variables. While correlation is desirable, it is not sufficient. We need to understand for example, if a 10 per cent increase in the rate of Facebook users reporting the fear of job loss translates to a 1 per cent rise in unemployment in three months or a 5 per cent rise in unemployment in one year. This can only be established after thorough research is done to establish such a benchmark. Given the high frequency of big data, such benchmarks will have to be dynamic and continually refined to ensure validity.

In many countries, the Statistical Act predates developments in ICT and there are legal obstacles to the use of non-traditional data sources for official statistics as well as to the wider dissemination of data. In such situations, any consideration for big data in official statistics will meet a stiff barrier. Therefore, making practical steps towards incorporating big data in official statistics wll require change in legislation in most countries.

Issues of privacy and ownership of big data also require legalization. Adequate legal protection must be accorded data producers to protect against identity disclosure or unlawful access to their private information. At the same time, with careful consideration of the usefulness of big data in the interest of the public, legal authority must be provided to the NSOs to access big data from data producers when such a case is established.

3. Perception of big data by the NSOs

As the UN Statistical Commission emphasized, the fundamental principles of official statistics place an obligation on the NSS to explore big data. To gauge how the region is performing in this respect, ECLAC conducted a survey of NSOs in the region during July-August 2015. The study was similar to that conducted by UNSD and UNECE[18]. For the current study, a brief questionnaire (See Annex 4) was sent to 23 NSOs in the Caribbean by email. Each NSO was given the option to complete the survey online or return the completed questionnaire by email. The questionnaire consisted of two sections. The first section sought information on the NSO and the respondent. The second section contained questions about big data. Of the ten countries that completed the questionnaire, six did so online. Although the response rate was low (43 per cent), the respondents represented diverse characteristics of the Caribbean in terms of population size, geographical location, independence status and the autonomy of NSOs within the public sector of their respective countries.

The Chief Statistician or Director, or someone acting in that capacity completed the survey for six of the eight countries. For the remaining two countries, a statistician completed the survey. Hence, the responses represent opinion at the highest level of authority within the NSOs.

The responses conveyed different levels of awareness of big data. It also conveyed an indication that big data may be misconstrued as a large dataset from traditional surveys. This perception was explicitly conveyed by one of the NSOs. For this reason, the responses from that NSO were excluded from further analysis as they were based on the understanding that big data represented datasets from large surveys.

For the remaining seven countries, all the NSOs indicated that they were not aware of any big data initiative or projects in their country. Only one out of the seven NSOs indicated that they had a big data strategy in their organization; that NSO also indicated foreseeing a role for big data in the work of the organization. Among the six NSOs that did not yet have a big data strategy, only two indicated foreseeing a role for big data in the work that they did. In general, some of the roles highlighted for big data include: supplementation of other data sources; serving as an additional source of data; a means to reduce cost; and a means to produce more timely output. When asked for existing or potential sources of big data in their country, the responding NSOs cited telecommunications companies, financial institutions and online merchants as big data sources.

There was a general consensus among the NSOs that big data could be used to supplement official statistics. Five out of six responding NSOs agreed that big data has the potential to alleviate data challenges in the Caribbean. Along this line, they indicated that it could offer a more cost-effective way to supplement other official data sources as well as provide trend data in a more timely fashion. In listing the major risks to the use of big data in official statistics, lack of standardization, inability to verify data quality, lack of experience in big data analytics, and non-representativeness of big data were the main risks itemised.

[18] Findings of that study is presented in Annex 3.

In terms of the specific obstacles to the use of big data in their respective organizations, the respondents indicated technological constraints, access to big data, privacy barriers, lack of expertise, lack of resources and challenges with adherence to statistical standards as major obstacles.

The results of the survey are very informative. They show that more than half of the NSOs do not foresee a role for big data in their organization. This raises the question of whether this is due to a philosophical adherence to tradition or is a result of the obstacles to big data use in their organization that they have identified. It also raises the possibility that countries with small populations may not see benefits in big data.

Nonetheless, the concerns raised about the risks of big data and the obstacles to its use are similar to those that have been identified in other regions. Given these risks and obstacles, it is important that the NSOs, or more genrally the NSS, of each country devise a big data strategy where one does not currently exist to guide the country's approach to big data.

4. How data producers view big data

In an attempt to seek the views of big data producers in the subregion, especially those in the telecommunications sector, a focused interview was conducted with the Caribbean Telecommunications Union (CTU)[19], an intergovernmental organisation dedicated to facilitating the development of the regional information and communications (ICT) sector. CTU's membership includes governments, the private sector and civil society, including the main telecommunications companies in the Caribbean.

CTU identified govenments, ICT service providers, financial services institutions, and energy, tourism and cultural sector players as existing or potential sources of big data in the Caribbean. The CTU, however, was unaware of any big data intiative or project being implemted by its members. The CTU was also unaware of any overarching strategic plan for big data at the subregional level.

However, the CTU supported the notion that big data has the potential to alleviate data challenges in the Caribbean but expressed reservations about the capacity of the subregion to capitalise on the potential of big data, indicating that the requisite infrastructure and skills were lacking. To address these challenges, the union considered parnerships and collaborations among stakeholders, including government, research institutions and data producers as the most efficient means to make big data accessible to researchers and policy makers. Implemnting such a framework, the CTU suggested, would require gaining a clear understanding of the needs that big data can fulfil.

The CTU indicated its readiness to support big data projects that the governments in the region may propose through public education, project implementation and management, and by serving as a repository of big data. In that light, the union was supportive of the call to creat a center of excellence for big data in the subregion and emphasized that its mandate includes serving as a repositoty of big data for ICT information.

As much as the CTU acknowledged that a role existed for big data in the development of the Caribbean, the union considered lack of finance, limited infrastructure, unintegrated data sources, lack of skills and capacity, and limited cooperations among stakeholders as major risks to the use of big data in official capacity.

[19] The interview was conducted with the Secretary General of the Caribbean Telecommunications Union on 4 October 2015.

VI. Towards a big data strategy for the Caribbean

If NSOs in the Caribbean are to effectively respond to the call to incorporate big data in official statistics, then the issues surrounding standards, benchmarking and legislation have to be addressed strategically and any strategy put in place must address capacity, funding and access to big data.

Further to the call for NSOs to be independent statutory bodies of the government, Statistical Acts across the subregion will have to be updated to provide for such in countries where it does not exist. Revision of the act should also reflect the realities of today's world. In addition to promoting open data in official quarters, such amendments should remove obstacles to the use of big data in official statistics. While this may not be achievable in the near term, awareness needs to be raised from now in order to enlighten the statistical community, policy makers, and politicians alike about big data and its benefits.

The statistical community must take an active role in debating the pros, and not just the cons, of big data. Such a debate must include a broad range of stakeholders in an effort to define standards that will assist in classifying big data. Discussions along this line should begin soon so that the Caribbean perspective could inform any global consensus on data that may be developed as recommended by the IEAG.

It is even more important that benchmarks be developed at a regional level for big data to be more impactful. The size of many economies in the Caribbean may preclude the development of a national benchmark, although national circumstances may suggest that national benchmarks are the most useful criteria. Benchmarking will require rigorous research and investment in data analytics capacity. Creating a Centre of Excellence in Big Data Analytics in one of the regional institutions would be an efficient way to concentrate available skills and help develop new expertise in this critical area. Such a centre must be developed through public-private partnership where big data producers are integrally involved in research to derive value from the data that they generate. The Centre, once fully established, would serve as a training facility for statisticians in the NSOs and other organizations in the subregion.

Funding for the Centre of Excellence would come mainly from the private sector, philanthropies, and the international development partners. In particular, as part of the Funding for Development agenda, the region needs to make a case for capacity strengthening in big data as a means of implementation of the SDGs. However, under the current circumstances in which most countries do not have a big data strategy, it is difficult to evaluate the big data capacity needs and the financial implications of incorporating big data in official statistics.

As a starting point therefore, every NSO needs to critically appraise the relevance of big data in the environment in which they operate in an effort to develop a strategy for big data. International agencies have a role to play here to assist countries in doing the background study to inform such strategies.

Having the capacity does not automatically resolve the challenges to big data. There is also the issue of access to big data. As most big data producers operate across national boundaries, a regional approach also needs to be taken in addressing the issue of access. Similar to the proposal that a centre of excellence in big data analytics be created, a clearing house for big data in the region should also be established. A regional organization, institution, or business with the requisite technology to handle big data should be designated as the Caribbean Big Data Clearing House. The clearing house would then be better placed to negotiate access to big data from regional and international big data producers on behalf of all Caribbean nations and territories. This would reduce overhead and transaction costs and avoid duplication of efforts in an emerging area where expertise is limited.

The big data initiative proposed in this paper is to serve as a guide in engineering a more robust discussion on the way forward for big data in the Caribbean that would involve all stakeholders, including civil society. It is by no means prescriptive; rather it is indicative of what the essential elements of a regional approach to big data should entail.

VII. Conclusions

In this paper, we have presented big data as a viable option for computing official statistics in a world in which the demand for high quality data is ever increasing. Big data differs from traditional data in that it is available in greater volume, veracity, and variety and occurs with higher velocity. Big data sources include mobile phone data; social media such as Twitter, Facebook, and blogs; and websites such as those used for e-commerce, e-business, job postings, and online searches. The major benefits of incorporating big data in official statistics rest in its timeliness and low cost of production. Big data can be generated, processed, and stored at lower costs than traditional surveys. Its application spans various sectors. It can be used in the provision of health services, in the understanding of population movements, and in the design of social programmes.

The use of big data to advance the international development agenda offers promising prospects. Developing countries, particularly in Sub-Saharan Africa, are already using ICT devices such as mobile phones to complement weak financial and banking systems. Big data can also positively impact on development by serving as an early warning system to facilitate swift responses in crises by providing real time awareness and real time feedback on whether or not policies and programmes are working. In terms of statistical reporting, it is possible to use big data to identify correlation and stylized facts in large datasets, thereby providing cheap and fast proxies for official statistics. For example, inflation rates can be accurately estimated by examining the prices of goods sold on the internet. Social media is also a good source of assessing individuals' sentiments about economic conditions and can indicate trends in inflation and unemployment before official statistics are produced.

Big data can be instrumental in studying the movements within a population. Mobile phone data provide very accurate and precise geographical information. Therefore, it can be used to provide mobility statistics and similar applications within the tourism sector. The use of mobile positioning data in healthcare application is promising. Mobile data have successfully tracked displaced individuals after an epidemic, for example the 2010 earthquake and the resulting cholera outbreak in Haiti. In the case of the Haitian cholera outbreak, mobile data provided insights into the magnitude of the epidemic up to two weeks before official reports were made. In other cases, the internet has been used as a tool for passive surveillance for health information. The Web has the potential to predict outbreaks and population-based events that are related to health. Overall, big data can serve as an early warning system to indicate the plight of the poor and vulnerable before traditional data sources can.

Despite the potential of big data, the Caribbean region has not warmed up to the idea of using big data in an official capacity. As reflected in a survey of NSOs in the subregion, more than half of the respondents did not foresee any role for big data in the implementation of their work. This is of concern considering that no Caribbean nation has recorded a reporting rate of 50 per cent on the MDG indicators for 2013. With the SDGs due to begin in January 2016, there will be greater demand on countries to provide timely reports on their progress. The international community, under the aegis of the United Nations, has identified critical roles for big data in the attainment of the goals and is supporting several big data initiatives. The Caribbean has not undertaken any project under these initiatives and the region does not appear to be making serious enough efforts to explore big data.

However, to successfully incorporate big data in official statistics, the issues relating to methodology, quality, technology, data access, legislation, privacy, management, and finance have to be addressed. Privacy must be maintained when acquiring, storing, and using big data. There are technological issues surrounding identifying the possible data sources and determining how applicable they are in achieving the desired end result. Although various methods exist to process big data, there may be a need for a core methodology to transform raw data into a model from which domain specific methodologies can be applied. Other issues include errors in the data, lack of representativeness, the loss of independence of statisticians, and limited access to key data sets. Private entities may refuse to share their data for privacy reasons or the form in which the data are stored may make it difficult to be accessed or transferred.

For big data use in official statistics to be successful, it is essential to drive legislative changes that would provide a legal framework for accessing, processing and utilizing big data in official capacities. There is also the need for a business model that outlines roles and responsibilities, beneficiaries, and funding arrangements for a successful big data initiative. In this paper, we have made some suggestions on developing a regional strategy for big data in the Caribbean. Key among those suggestions is the creation of a Centre of Excellence in Big Data Analytics and a Caribbean Big Data Clearing House. We have also indicated that public-private partnership will be required to finance these proposals. In as much as big data is a new and evolving phenomenon in the subregion, we should caution that these suggestions are only to guide more elaborate discussions that should involve a broader group of stakeholders if the region is to realize the full benefits of big data in official statistics.

Bibliography

Baker, L, Wagner, T.H., Singer, S. and Bundorf, M.K. (2003), "Use of the Internet and email for health care information: results for a national survey" JAMA 289: 2400-6.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F. and Pentland, A. (2014), "Once Upon a Crime: Towards crime Prediction from Demographics and Mobile Data" [online] [date of reference 20 May 2015] <http://arxiv.org/pdf/1409.2983.pdf>

Chunara, R., Andrews, J. and Brownstein, J. (2012) "Social and News Media Enable Estimation of Epidemiological Patters Early in the 2010 Haitian Cholera Outbreak" American Journal of Tropical Medicine and Hygiene, 86:39-45.

Cooper, C.P., Mallon, K.P., Leadbetter, S., Pollack, L.A. and Peipins, L.A. (2005), "Cancer Internet Search Activity on a Major Search Engine" J Med Internet Res 2005; 7:e36 United States 2001-2003.

Culotta, A. (2010), "Detecting Influenza outbreaks by Analysing Twitter Messages", [online] [date of reference 15 May 2015]< http://arxiv.org/pdf/1007.4748v1.pdf>

Ettredge, M., Gredes, J. and Karuga, G. (2005), "Using Web-based search data to predict Macroeconomic Statistics" Commun ACM, 48(11): 87-92

Frias-Martinz, V., Ruiz-Soguero, C. and Josephidou, M. (2013), "Forecasting Socioeconomic Trends with Cell Phone Records" [online] [date of reference 15 May 2015] http://www.vanessafriasmartinez.org/uploads/dev13.pdf

Evans, A. (2015), Annotated Bibliography on a Review of Developments with Big Data and the Data Revolution. Report submitted to ECLAC, May 2015.

Eysenbach, G. (2006), "Infodemiology: Tracking Flu-related searches on the Web for Syndromic Surveillance" AMIA Annual Symposium Proceedings 2006: 244-248

Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S. and Brilliant, L. (2009), "Detecting Influenza Epidemics using search engine query data" Nature Volume 457

Global Pulse (2012), "Big Data for Development: Challenges & Opportunities" [online] [date of reference 14 May 2015] http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf

Hulth, A. and Rydevlk, G. (2011), "Web query-based Surveillance in Sweden during the Influenza A(H1N1) 2009 Pandemic" April 2009 to February 2010. Euro Surveli. 201; 16 (18) [online] [date of reference 14 May 2015]< http://www.eurosurveillance.org/images/dynamic/ee/v16n18/art19856.pdf>

ITU (2013), "Big Data: Big Today, Normal Tomorrow" [online] [date of reference 20 May 2015] <http://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000220001PDFE.pdf>

Independent Expert Advisory Group (2014), "A World that Counts: Mobilizing the Data Revolution for Sustainable Development" [online] [date of reference 18 May 2015] < http://www.undatarevolution.org/report/>

Johnson, H.A., Wagner, M.M., Hogan, W.R., Chapmen, W., Olszewski, R.T., Dowling, J. and Barnas, G. (2004), "Analysis of Web Access Logs for Surveillance of Influenza" [online][date of reference 14 May 2015]< http://rods.health.pitt.edu/Technical%20Reports/2004%20IMIA%20-%20Johnson.pdf>

Karlberg, M. and Skaliotis, M., (2013), "Big Data for Official Statistics – Strategies and Some Initial European Applications", Seminar on Statistical Data Collection [online] [date of reference 8 May 2015]< http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2013/mgt1/WP30.pdf>

Landefeld, S. (2014), "Uses of Big Data for Official Statistic: Privacy, Incentives, Statistical Challenges and Other Issues" Discussion Paper, International Conference on Big Data for Official Statistics, Beijing China, 28 – 30 October [online] [date of reference 8 May 2015] < http://unstats.un.org/unsd/trade/events/2014/beijing/Steve%20Landefeld%20-%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf>

Lu, X., Bengstsson, L. and Holme, P. (2012), "Predictability of Population Displacement after the 2010 Haiti Earthquake" PNAS, Vol. 109, No. 29, 11576-11581

Moreno, M.A., Christakis, D., Egan, K.G., Brockman, L.N. and Becker, T. (2011), "Associations Between Displayed Alcohol Reference on Facebook and Problem Drinking Among College Students" Archives of Paediatrics and Adolescent Medicine [online] [retrieved 14 May 2015] <http://archpedi.jamanetwork.com/article.aspx?articleid=1107693>

Mykhalovskiy, E. and Weir, L. (2006), "The Global Public Health Network and Early Warning Outbreak Detection: A Canadian contribution to Global Health" Can J Public Health, 97(1):42-4

O'Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A. (2010), "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series" International AAAI Conference on Weblogs and Social Media, Washington, D.C. 2010

Oliver, N. (2014), "Big Data for Social Good: Opportunities and Challenges" 12th World Telecommunication/ICT Indicators Symposium (WTIS-14) [online] [date of reference 20 May 2015] <http://www.itu.int/en/ITU-D/Statistics/Documents/events/wtis2014/003INF-E.pdf>

Pastor-Escuredo, D. and others (2014), "Flooding through the Lens of Mobile Phone Activity" [online] [date of reference 20 May 2015] <http://arxiv.org/ftp/arxiv/papers/1411/1411.6574.pdf>

Paul, M.J. and Dredze, M. (2011), "You are what you Tweet: Analysing Twitter for Public Health" Rep. Centre for Language and Speech Processing at John Hopkins University [online] [date of reference 15 May 2015] < http://www.cs.jhu.edu/~mpaul/files/2011.icwsm.twitter_health.pdf>

Statistics Netherlands (2014), "Innovation of Tourism Statistics Through the use of New Big Sources" 12th World Telecommunication/ICT Indicators Symposium (WTIS-14) [online] [date of reference 20 May 2015] <http://www.itu.int/en/ITU-D/Statistics/Documents/events/wtis2014/002INF-E.pdf>

TechAmerica Foundation (2012), "Demystifying Big Data: A Practical Guide to Transforming the Business of Government [online] [date of reference 01 June 2015] <http://www.techamerica.org/Docs/fileManager.cfm?f=techamerica-bigdatareport-final.pdf>

Tiru, M. (2014), "Overview of the Sources and Challenges of Mobile Positioning Data for Statistics", International Conference on Big Data for Official Statistics, Beijing [online] [date of reference 6 May 2015] < http://unstats.un.org/unsd/trade/events/2014/beijing/Margus%20Tiru%20-%20Mobile%20Positioning%20Data%20Paper.pdf>

United Nations (2013), "A New Global Partnership: Eradicate Poverty and Transform Economies through Sustainable Development: The Report of the High-Level of Eminent Persons on the Post-2015 Development Agenda" [online] [date of reference 17 July 2015] <http://www.un.org/sg/management/pdf/HLP_P2015_Report.pdf>

United Nations (2014), "The Millennium Development Goals Report 2014. New York: The United Nations.

United Nations Economic and Social Council (2013a), "Big Data and Modernization of Statistical Systems" Report of the Secretary-General, E/CN.3/2014/11 [online] [date of reference 18 May 2015] < http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf>

United Nations Economic and Social Council (2013b), "Concept Note: Friday Seminar on Emerging Issues, Big Data for Policy, Development and Official Statistcs, [online] [date of reference 15 July 2015] < http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/concept_note.pdf>

United Nations Economic and Social Council (2015a), "Emerging Issue: The Data Revolution" Report of the Secretary-General, E/CN.3/2015/3 [online] [date of reference 6 May 2015]< http://unstats.un.org/unsd/statcom/doc15/2015-3-EmergingIssues-E.pdf>

United Nations Economic and Social Council (2015b), "Report of the Global Working Group on Big Data for Official Statistics, Report of the Global Working Group on Big Data for Official Statistics", E/CN.3/2015/4 [online] [date of reference 6 May 2015] <http://unstats.un.org/unsd/statcom/doc15/2015-4-BigData.pdf>

United Nations Economic and Social Council (2015c), "Statistical Commission: Report on the Forty-sixth Session, E/2015/24-E/CN.3/2015/40 [online] [date of reference 19 July 2015] <http://unstats.un.org/unsd/statcom/doc15/Report-E.pdf>

UN Global Pulse "Unemployment Through the Lens of Social Media, UN Global Pulse" [online] [date of reference 10 May 2015]< http://www.unglobalpulse.org/projects/can-social-media-mining-add-depth-unemployment-statistics>

UNSD/UNECE (2015), "Results of the UNSD/UNECE Survey on Organizational Context and Individual projects on Big Data" [online] [date of reference 14 May 2015] < http://unstats.un.org/unsd/statcom/doc15/BG-BigData.pdf>

Wasshausen D. and Mouton, B.R., (2006). The Role of Hedonic Methods in Measuring Real GDP in the United States [online] [date of reference 13 July 2015] < http://www.bea.gov/papers/pdf/hedonicGDP.pdf>

Annexes

Annex 1
Annotated listing of studies documenting the use of big data

Author	Summary of Study
Bogomolov, A. (2014)	This research illustrated the fact that human behavioural data can have some success in predicting and understanding social issues. The paper outlines a study that used data on mobile network activity and basic demographic information to predict if a particular area in London will be a hotspot for crime. The paper reported a 70 per cent prediction accuracy for this approach when compared with real crime data.
Cooper, C.P. et al. (2005)	This study used the Yahoo! Buzz Index to establish a correlation between cancer incidence, estimated cancer mortality and the volume of cancer news coverage. The conclusion was that media coverage influenced online search for cancer information. Also, internet search activities wereregarded as an innovative tool for passive surveillance for health information.
Culotta, A. (2010)	This analyzed 500 million tweets over eight months to predict future influenza outbreaks. The results suggested that document classification can limit the impacts of false alarms but more research would be needed for the large cases of spurious messages.
Ettredge, M. (2005)	This article documents an example of how big data can be used to complement economic statistics. It concluded that common search terms on the internet can produce accurate statistics about the unemployment rate. It examined the relationship between rates of employment-related internet searches and the unemployment levels reported by the US in their monthly reports.
Eysenbach, G. (2006)	The study explored the possibilities of using internet search for flu-related reports (instead of data from clinical outcomes) to predict influenza epidemic for a 33 weeks period in the 2004/05 flu season in Canada. The study concluded that The “Google and sentinel method” that was used can be more cost effectively and more accurately predicted influenza-like illnesses than the traditional means. Thus tracking web search has the potential to predict outbreaks and other population based events related to health. It however cautioned that such studies can be confounded by “epidemics of fear”.
Frias-Martinz, V. et al. (2013)	This study proposed the use of human behavioural patterns to predict future values of socioeconomic indicators. The study focused on the following socioeconomic indicators collected by the National Statistical Institute of an emerging economy: total assets, total number of employed citizens, total number of workers employed by the private sector, public sector employment, number of subcontracted civil servants, and number of subtconracted workers. The research focused on whether or not consumption and mobility data derived from call logs can forecast these socioeconomic indicators indicated. The study used both multivariate regression models and multivariate time-series forecasting models. The multivariate time-series yielded good results whereas the multivariate regression models did not provide good forecasts.
Ginsberg et al. (2009)	This article presented a methodological approach to analysing the large amount of influenza related google queries in the United States using search queries between 2003 and 2008. It identified a strong correlation between the percentage of doctor visits of patients showing influenza-type symptoms and the relative frequency of certain queries and was therefore able to accurately predict weekly flu activity in each region of the US with about a one day lag. The paper concluded that flu epidemics can be detected by using search queries if the population of Web users is large.
Hulth, A. & Rydevlk, G. (2011)	This study conducted in Sweden used queries submitted to the Swedish medical website about influenza and influenza symptoms to test statistical models for predicting the H1N1 2009 pandemic. The results indicated that these queries may be a bit more precise method than the Google Flu Trend in estimating flu activity.
Independent Expert Advisory Group Secretariat (2014)	IEAG reported that data on forest fires that are publicly available have been used by Indonesian’s two largest pulp and paper companies to respond to fires on land managed by the companies. The data are provided by the World Resources Institute's Global Forest Watch website which also provides ultra-high resolution imagery that has been used by Indonesia and Singapore to eliminate illegal burning by companies. This system along with the Indonesian Government's Karhulta Monitoring system has allowed firefighters to reduce their response time by 32 hours, from 36 to four hours.

Independent Expert Advisory Group Secretariat (2014)	Reported that Uganda has been making strides in its effort to deal with malaria by using SMS surveys done by health workers. The Mtrac programme uses SMS to alert public officials about malaria outbreaks and about medicine available at health facilities. The programme successfully collected and analysed data from thousands of health facilities within 48 hours at a cost of less than US$150 per poll. On average, more than 350 actionable reports were made per month and about 70 per cent of these were followed up within two weeks. Also, the number of facilities that run out of the key medicine to treat malaria fell from 80 per cent to 15 per cent.
Independent Expert Advisory Group Secretariat (2014)	As reported by the IEAG, the University of Minnesota School of Public Health generated epidemiological models using information from the Orange mobile network and information from WHO on the spread of malaria. The results of these models are useful in creating rapid response systems to notify the public before epidemics.
Johnson, H.A et al. (2004)	The focus of this article was on the timing of the signal of data gleaned from websites over traditional influenza surveillance data. It used one year of Web data from Healthlink to determine the correlation between influenza search activities and influenza itself in order to establish if the frequency of influenza search precede increases in the occurrence of the virus. The research found moderately strong correlation between search and the occurrences of influenza. However the result on timeliness was inconclusive.
Moreno, M.A. et al. (2012)	This study looked at the association between alcohol use and intoxication/problem drinking and references to such on Facebook. It measured drinking problems using the Alcohol Use Disorders Identification Test (AUDIT) scale. It concluded that Facebook alcohol references are positively related to being at risk of having a drinking problem.
Mykhalovskiy, E. & Weir, L. (2006)	This article ipresented a global example of using the internet as an early warning system to detect infectious disease outbreak. It illustrated a reporting technique that transcends national and geographic boundaries and goes beyond official reporting systems. The Global Public Health Intelligence Network combines human expertise and technological development (automated translation and electronic text search) to create a new methodology of monitoring global health. It provides an effective mechanism to identify and respond to infectious disease outbreaks and makes it difficult for countries to hide disease outbreaks of international significance.
O'Connor, B. et al. (2010)	Resulst form this study suggest that text streams are a potential replacement and supplement to traditional expensive polling. The study examined the correlation between tweets and consumer confidence and political opinion during 2008 and 2009.
Pastor-Escuredo, D. (2014)	The results from this study add to the literature that suggests that Call Detail Records (CDR) can improve warning and emergency management mechanisms relating to natural disasters. The study examined the effects of CDR on the 2009 Mexican floods in Tabasco. The results showed that for emergency operations, even though the data from cell towers were anonymized and aggregated, observed changes can be useful. The study also suggests that attention should be paid to how people respond to disasters and that such information should be included in future emergency management strategy. In general, it was concluded that public-private partnership can improve humanitarian response in disasters.
Paul, M.J. & Dredze, M. (2011)	This was the first study to attempt to use social media to examine a broad range of public health research. Previous studies only focused on the flu epidemic. It found correlations between the Ailment Topic Aspect Model and tweets about various ailments including flu, allergies, insomnia and obesity.
Polgreen, P.M. et al. (2008)	This study used yahoo search to find out about influenza originating in the USA for March 2004 and May 2008 to examine the relationship between flu search and the occurrence of flu. The results indicated that positive influenza cultures can be predicted 1-3 weeks in advance.
PriceStats (n.d.)[a]	This research entailed daily scraping of the Web for online prices and converting these unstructured data into a structured data set. The research concluded that online prices can be used to accurately predict inflation, PPP, and other economic indicators.

Statistics Netherlands (2014)	Statistics Netherlands in 2012 – 2013 carried out three innovative research studies on how big data can assist in tourism statistics. The studies were based on internet data, data from an app installed on mobile devices, and aggregated data provided by Vodafone and Mezuro. The studies revealed the importance of big data as a new source of data, especially for statistics that cannot be produced through traditional sources. The research highlighted the fact that the greatest challenges are in the areas of developing appropriate methodology, privacy protection, knowledge of big data, big data resource constraints and current culture of statistics production.
UN Global Pulse	The study compared online qualitative data with unemployment figures. It analysed social media data, blogs, forums, and news sites from US and Ireland for references to unemployment and how persons were coping. The study concluded that online conversations can complement official unemployment statistics.

Source: Evans, A. (2015)

[a] http://www.pricestats.com/

Annex 2
Selected projects implemented under the United Nations Global Pulse

Project Title	Project Location	Project Description/ Summary
Mining Citizens Feedback Data for Enhanced Local Government Decision Making	Indonesia	Program Area: Economic Well- Being Data from the complaint systems and other indifferent feedback from social media sites were essential parts in this research in an attempt to influence the local government's policy decisions. The outcome shows data can be utilized to grasp different aspects of citizens' feedback on policy issues and generate timely responses.
Feasibility Study: Supporting Forest and Peat Fire Management using Social Media	Indonesia	Programme Area: Climate and Resilience In support of the Emergency Response Management in Indonesia, the Pulse Lab conducted a feasibility study to understand the relationship between social media signals (real time information or communication trends) and current events. They found that ndonesians 'tweet' more about these hazards during and after the events. The study recommended that more research be conducted to better utilize the data from social media and other digital data to generate a better insight on the impact of disaster on human well-being. These types of data will also provide a more timely response to emergencies.
Using Mobile Phone Data and Airtime Credit Purchases to estimate Food Security	East Africa	Programme Area: Food Security & Agriculture This study considered the use of mobile phone data as a data source for food security and poverty indicators. Data from airtime credit purchases and mobile phone activity were compared to the results of a survey conducted by the UN World Food Program (WFP). The results showed high correlations between the two. The study indicated that proxies derived from mobile phone data could provide valuable indices of food security, thereby filling the data gaps that existed in povety assessment.
Using Mobile Phone Activity for Disaster Management during Floods	Mexican state of Tabasco	Programme Area: Humanitarian Action In this study, mobile phone activity data were combined with remote sensing data in an attempt to explore ways to improve disaster response. The results show patterns of mobile activities in areas affected by disaster which could be used as indicators for flooding impacts and other disasters.
Analysing Seasonal Mobility Patterns using Mobile Phone Data	Senegal	Programme Area: Food Security & Agriculture This study showed how mobile phone data could be used to analyse seasonal mobility patterns by extracting and visualising movement patterns. The results showed that for vulnerable populations, changes in mobility patterns could indicate changes in livelihood or coping strategies or exposure to new shocks. Furthermore, constant monitoring can lead to timely response thus allowing informed decisionsto be made in a timely manner.
Analysing Social Media Conversations to understand Public Perceptions of Sanitations	New York	Programme Area: Public Health This research was done in support of the UN Deputy Secretary General's Call for Action on Sanitation. The United Nations Millennium Campaign along with the Water Supply and Sanitation Collaborative Council analysed social media in an attempt to acquire a better understanding of the public perceptions of sanitation. Thirty three per cent of the relevant tweets focused on cholera. The research further stated that apart from the cholera cases, tweets were mostly in the context of human rights, health, and policy and governance.
Mapping the Risk Utility Landscape of Mobile Data for Sustainable Development and Humanitarian Action		Programme Area: Data Privacy and Protection The aim of this project was to analyse how mobile data can be used to maximum effects in support of policy planning and crisis response, and with limited risk to privacy. The project results indicated that the relationship between the risk to privacy and utility is dependent on the context and purpose of use.

Using Twitter to Measure Global Engagement on Climate Change		Programme Area: Climate & Resilience Global Pulse explored online media communications about climate change by analysing tweets on a daily basis using a social media monitor. The result showed increased discussions on climate change around the Climate Summit. The study concluded that by providing a tool to compare interest levels, the monitor could be used to support policy making and encourage public engagement.
Using Twitter Data to Analyse Public Sentiment on iEl Salvador	El Salvador	Programme Area: Economic Wellbeing The World Bank and Global Pulse utilised tweets to better understand public opinions about the fuel subsidy reform made in El-Salvador in 2011. The result showed a decline in negative comments and an increase in positive feedback about the reform. Therefore, the research suggested that tweets could be used for policy analysis.
Estimation Migration Flows Using Online Search Data	Australia	Programme Area: Economic Wellbeing Using Australia as a case study, this study evaluated how online search data can be used to understand migration flows. This was done by comparing Google search data with official migration statistics. The result showed a correlation between the two suggesting that online data could be used to estimate migration flows. Ths study made a case for more exploration of online and other digital sources as alternate sources of for migration statistics.
Data Visualisation and Interactive Mapping to support response to disease outbreak	Uganda	Programme Area: Public Health This project used a series of data visualisation to support the early response to the typhoid outbreak which occurred during January –May 2015. The project showed that visualisation data could be used as decision-making tool in public health.
Putting People's voices at the Centre of Development using big Data Analytics	Uganda	Programme Area: Post-2015 This research was conducted by the Pulse Lab to generate a better understanding of the Ugandan youth's perceptions of the Post- 2015 development topics. The study analysed 3.1 million messages from UNICEF's U- report platform to arrive at the conclusion that better health care, good education, and better job opportunities were the top priorities of Ugandan youths.
Feasibility Study: Analysing Large- scale News Media content for early warning of Conflict	Tunisia	Programme Area: Humanitarian Action This study showed how data mining of large scale online news data can be used to prevent conflicts. The study involved analysing news media archives from periods before and after the government transition in January 2011 and concluded that data could provide an insight to emerging conflict.
Analysing Attitudes Towards Contraception & Teenage Pregnancy using Social Data	Uganda	Programme Area: Public Health The Pulse Lab used public facebook posts to analyse the perceptions of teenage pregnancy and the different types of contraceptives. This was done using an interactive dashboard and data from UNICEP's U-report for keywords relating to contraceptives and teenage pregnancy.
Nowcasting Food Prices in Indonesia Using Social Media Signals		Programme Area: Food Security & Agriculture This research explained how how social media analytics can provide alternative metrics for daily food price statistics. Using Twitter data, the study forecastfood prices and concluded that the approach offer a faster, more affordable, and efficient way of collecting real-time food prices.
Understanding Public Perception of Immunisation Using Social Media		Programme Area: Public Health To establish that social media can be used to understand public perceptions of immunization, the Pulse Lab Jakarta filtered tweets from relevant conversations about vaccines and immunisation and showed trends in perceptions, religious issues, disease outbreak, and the launch of a new vaccine.
Online Signals for Risk Factors of Non-Communicable Diseases (NCDs)		Programme Area: Public Health In this study, the World Health Organisation examined how online signals can be used to explore risk factors for non- communicable diseases. Using key words, indices for different risk factors were built and tracked over time on social media and other internet search sites.

Mining Indonesian Tweets to understand food price crises	Indonesia	Programme Area: Food and Agriculture This study showed that Twitter feeds can be used to detect early warnings of inflation and other crises. The study involved using keywords and phrases to extract relevant tweets which were then analysed to show that around the same time when real-world events transpired, the conversations relating to food prices also spiked.
Advocacy Monitoring through Social Data: Women's and Children's health		Programme Area: Gender This project analyzed tweets in relation to women's and children's health using the Crimson Hexogen's analytical tool- Forsight. The results showed specific spikes in conversations around World Aids Day, Mother's Day, International Women's Day, MDG Summits and Women-centered conferences.
Using Twitter to understand the Post-2015 Global Conversation		Programme Area: Post-2015 Global Pulse filtered tweets for comments relevant to development topics to identify the most talked about topics related to the Post-2015 Development Agenda. An online dashboard providing this information at the country level and which automatically updates was then developed.
Unemployment through the Lens of Social Media	US & Ireland	Programme Area: Climate & Resilience This project centered ona analysing social media for moods and current topics on unemployment based on changing job conditions in US and Ireland. The results provided a qualitative picture of how respondents coped with their current unemployment status and gave an indication of how online conversations can be used to supplement official unemployment statistics.
Monitoring Perceptions of Crisis-Related stress using Social Media Data	US & Indonesia	Programme Area: Economic Well-being This study showed how big data can be used to analyse people's concerns around food, fuel, finance, and housing. Spikes and drops in the number of tweets about a particular topic were analysed to identify changes in trends that occurred overtime.
Rapid Mobile Phone Survey on Economic Conditions	Uganda, India, Mexico, Ukraine and Iraq	Programme Area: Real-time Evaluation Global Pulse conducted a mobile phone survey to understand how populations in different parts of the world perceive economic conditions over the last year, their ability to meet their household needs, changes they have made in their way of life, and their attitudes about the future. The survey was also used as a test case for rapid mobile data collection.

Source: Global Pulse (http://www.unglobalpulse.org/projects).

Annex 3
Big data projects across countries and organizations

Country/Institution	Application
China	The Chinese government signed agreements relating to big data with 17 commercial enterprises in a bid to collect internet and ecommerce transaction data for use in official statistics. The intention of the National Bureau of Statistics of China is to be a leader in big data research.
Denmark	Denmark has taken an interest in incorporating big data into official statistics. At the time of the UNSD/UNECE survey, there was one big data project related to scanner data for Price statistics. Also, Statistics Demark reported on the development of their new strategy in which big data will have a role to play.
Finland	Statistics Finland has an internal big data working group. Statistics Finland is also a member of the European Statistical System Taskforce on Big Data and Official Statistics. Even though there is no defined business process for integrating big data sources in official statistics, there are examples of big data application in Finland. The traffic sensor data of the Finnish Transport Agency is one such example. It is used in research and development to provide better drive time estimations.
Italy	In Italy, there are plans to use big data in statistical production. Istat is exploring partnerships with big data providers such as telecommunication companies and supermarket chains.
Japan	Although Japan has not discussed its plans for big data, the country has recognized the need to consider it in the future.
Portugal	At the time of the survey, Portugal had no defined business process or partnerships with regards to big data. However, the country intended to explore the possibilities of using big data as an alternative to traditional survey method.
Romania	Romania reported investigating possible big data sources to complement official statistics. In its 2014-2020 NSI Strategy, it indicated plans to consider the use of big data to modernize official statistics.
Serbia	The use of big data in official statistics existed only as an idea in Serbia. There were plans to develop and implement a big data strategy.
Sweden	Sweden uses scanner data to compute CPI. Additionally, there were plans to incorporate additional big data sources in the future.
UK	The Office for National Statistics (ONS) long term plan is to reduce the use of survey and census data sources and complement traditional statistics sources with big data. There is a 15 month Big Data Project that was implemented to explore the benefits and challenges in using big data in official statistics. The procedure involved using practical applications to assess methodological and technical issues.
Eurostat	They are concerned with maximizing the potential benefits of implementing the use of Big Data in Official Statistics. The Task Force is mandated to carry out these actions by leading and coordinating developments within the European Statistics System and the Commission. The action plans of Eurostat falls in the general framework of the ESS Vision 2020 of formulating a response to the data revolution
ESCAP	UNESCAP has recognized the limitations that less developed National Statistical Systems may face in not being able to develop their own big data solutions and not being able to adopt the ones developed elsewhere due to the differences in data needs. For example, developing countries may be more concerned about measuring poverty and inequality. In light of this, discussions have been initiated in the region to devise a regional strategy to use big data to enhance economic and social development in Asia and Pacific

Source: UNSD/UNECE (2015).

Economic Commission for Latin America and the Caribbean
subregional headquarters for the Caribbean

STUDY ON BIG DATA

Survey of National Statistical Offices

Introduction

ECLAC is currently undertaking a study on the potential of Big Data in the Caribbean and we would like you to participate in a survey of NSOs in the region. The survey is intended to gauge the perception of Big Data among NSOs in the region and to gain information on any Big Data initiatives being implemented or planned. Your individual responses to the survey questions will not be identified but will be part of summary results that will be presented in the study report. We value your opinion on the issues surrounding Big Data and we look forward to your participation. Should you have comments, suggestions or concerns about this survey, please contact Abdullahi Abdulkadri, Coordinator, Statistics and Social Development Unit, ECLAC, Port of Spain. Email: abdullahi.abdulkadri@eclac.org; Phone: (868) 224-8021. Thank You.

Instructions: Please provide your answer in the space provided or choose your response by placing an **X** against your choice.

Background Information

1. Country: ……
2. Organization: ……
3. Position/Designation of Respondent: ……

Information on Big Data

4. How would you define Big Data?

……

……

5. Are you aware of any Big Data initiative or project in your country?
   a. Yes: ____ b. No:____.
6. If Yes to Question 5, are these initiatives in the public or private sector?
   a. Public: ____ b. Private: ____ c. Both ___.
7. If yes to Question 5, please briefly describe these Big Data initiatives.

   ……

   ……

   ……

8. As the National Statistical Office for your country, does your organization have any Big Data strategy?
    a. Yes:__ b. No:___.

9. Does your organization foresee any role for Big Data in the execution of your work?
    a. Yes: ___ b. No: ___.

10. If Yes to Question 9, what are these roles?
    a. ……
    b. ……
    c. ……

11. Are there existing or potential Big Data sources in your country?
    a. Yes: ____ b. No: ____.

12. If Yes to Question 11, please list some examples:
    a. ……
    b. ……
    c. ……

13. In your opinion, does Big Data have the potential to alleviate data challenges in the Caribbean?
    a. Yes:____ b. No: ____.

14. If Yes to Question 13, in what ways do you see Big Data being used in the region for this purpose?

    ……

    ……

    ……

    ……

15. Do you agree with the notion that Big Data can be used to supplement official statistics?
    a. Yes:____ b. No: ____.

16. What do you consider to be the major risks to the use of Big Data in official statistics?
    a. ……
    b. ……
    c. ……

17. What are the major obstacles to the use of Big Data in official statistics in your organization?
    a. ……
    b. ……
    c. ……
18. Please provide any general comments you have about Big Data:

……
……
……

Thank You!

UNITED NATIONS

ECLAC **Studies and Perspectives Series – The Caribbean**

Issues published

A complete list as well as pdf files are available at

www.eclac.org/publicaciones

48. An assessment of big data for official statistics in the Caribbean – Challenges and opportunities, LC/L.4133, LC/CAR/L.485, 2016.
47. Regional approaches to e-government initiatives in the Caribbean, LC/L.4132, LC/CAR/L.483, 2016.
46. Opportunities and risks associated with the advent of digital currency in the Caribbean, LC/L.4131, LC/CAR/L.482, 2016.
45. Ageing in the Caribbean and the human rights of older persons – Twin imperative for actions, LC/L.4130, LC/CAR/L.481, 2016.
44. Towards a demand model for maritime passenger transportation in the Caribbean – A regional study of passenger ferry services, LC/L.4122, LC/CAR/L.477, 2015.
43. The Caribbean and the post-2015 development agenda, LC/L.4098, LC/CAR/L.472, 2015.
42. Caribbean synthesis review and appraisal report on the implementation of the Beijing Declaration and Platform for Action, LC/L.4087, LC/CAR/L.470, 2015.
41. An assessment of performance of CARICOM extraregional trade agreements – An initial scoping exercise, LC/L.3944/Rev.1, LC/CAR/L.455/Rev.1, 2015.
40. Caribbean Development Report – Exploring strategies for sustainable growth and development in Caribbean small island States, LC/L.3918, LC/CAR/L.451, 2014.
39. Economic Survey of the Caribbean 2014 – Reduced downside risks and better prospects for recovery, LC/L.3917, LC/CAR/L.450, 2014.
38. An assessment of mechanism to improve energy efficiency in the transport sector in Grenada, Saint Lucia, and Saint Vincent and the Grenadines, LC/L.3915, LC/CAR/L.449, 2014.
37. Regional integration in the Caribbean – The role of trade agreements and structural transformation, LC/L.3916, LC/CAR/L.448, 2014.
36. Strategies to overcome barriers to the implementation of the Barbados Programme of Action and the Mauritius Strategy in the Caribbean, LC/L. 3831, LC/CAR/L.441, 2014.
35. Foreign direct investment in the Caribbean – Trends, determinants and policies, LC/CAR/L.433, 2014.
34. Situation of unpaid work and gender in the Caribbean: The measurement of unpaid work through time use studies, LC/L.3763, LC/CAR/L.432, 2014.
33. Progress in implementation of the Mauritius Strategy: Caribbean Regional Synthesis Report, LC/L.3762, LC/CAR/L.431, 2014.

48

STUDIES AND PERSPECTIVES

ECLAC Series

ECONOMIC COMMISSION FOR LATIN AMERICA AND THE CARIBBEAN
COMISIÓN ECONÓMICA PARA AMÉRICA LATINA Y EL CARIBE
www.eclac.org