Harnessing Big Data for Development[1]

Jose Ramon G. Albert, Ph.D. and Arturo Martinez, Jr., Ph.D. [2]

ABSTRACT: Governments recognize the importance of statistics for managing their economies more effectively, especially as they seek to accelerate the pace of meeting national development goals, and global commitments exemplified in the Sustainable Development Goals (SDGs). Monitoring the SDGs and national development plans require huge demands from national statistical systems, particularly more, faster, and better data. Official statistics, sourced from censuses, sample surveys, and administrative reporting systems, are now being challenged by statistics that can be generated from a tsunami of data shared and transmitted on the web and by way of various electronic means. Big Data, typically characterized by 3V's: immense volume, velocity and variety, is not only creating business opportunities, but also showing potential as an alternative source of statistics on illness, inflation, sales, people's movements, including traffic. This paper reviews issues on Big Data, including ethics, opportunities and the role of the official statistics community in ensuring that statistics generated from Big Data will complement those generated from traditional sources and ultimately matter to everyone. It discusses how the readiness of official statisticians to leverage big data for official statistics crucially depend technological as well as capacity issues.
Key Words: big data, official statistics, quality, privacy, National Statistical Systems, SDGs

1. Introduction

Governments, civil society, research and academic institutions, as well as the development community, are one in recognizing the importance of statistics for managing economies more effectively. In particular, there is interest to monitor and find ways of accelerating progress in meeting national development plans, and global commitments to reduce poverty and related goals exemplified in the Sustainable Development Goals (SDGs). The SDGs, also called the Global Goals, entail broad and interconnected social, economic, environmental and governance aspects of sustainable development. They constitute a universal call to action with 169 specific, time-bound, and quantifiable targets for 2030 to leave no one behind (United Nations, 2015). The SDGs continue work of and expand on the Millennium Development Goals, a predecessor global agenda for monitoring socioeconomic progress (but with far more limited goals and targets than the SDGs).

Monitoring the Sustainable Development Goals (SDGs)

Monitoring the 17 SDGs and the concomitant 169 targets have created new and huge demands on national statistical systems (NSSs) since the number of global indicators for monitoring the SDGs total 232, and many of these indicators require disaggregation by location, sex, gender, age, income, migratory status, ethnicity status, disability status and other relevant dimensions. What complicates matters is that statistics development is likely not going to be given considerable increases in resources from national governments given the political economy and the competition for resources in delivering social services.

[1] This work in progress is supported by ADB's Data for Development Project.

[2] Authors are senior research fellow of the Philippine Institute for Development Studies and Statistician of the Economic Research and Regional Cooperation Department of the Asian Development Bank. Views expressed are those of the authors and do not necessarily reflect those of the institutions they are a part of.

Common instruments used to fund statistical programs such as bilateral grants and multi-donor trust funds are also neither sufficient nor substantially increasing across time (PARIS21, 2017). In consequence, national statistics offices (NSOs) and other statistics producers in NSSs particularly in developing economies need to find ways of addressing these and other data demands given the level of resources they are getting.

In Asia and the Pacific alone, while disaggregation of statistics by location is available for several SDG indicators, but data granularity is sparse for some SDG indicators by sex, and it is even scarcer, if not absent, for special groups such as disabled persons, and indigenous peoples (Figure 1). This is one of the key findings of a 2017 survey of NSOs undertaken by the Asian Development Bank (ADB), in collaboration with the United Nations Economic and Social Commission for Asia and the Pacific (UNESCAP). The same survey reported that many NSOs acknowledge that the only way that they will be able to meet the disaggregated data requirements of the SDGs is for them to utilize innovative methods and data sources (such as big data, an umbrella term used to describe the constantly increasing flows of data emanating from connected persons and things), to complement conventional approaches. Results of the 2017 ADB/UNESCAP survey also suggest that more than half (56%) of reporting NSOs from ADB member countries are utilizing small area estimation (SAE)[3] methods for getting more disaggregated data especially on poverty, and in some cases, on nutrition.

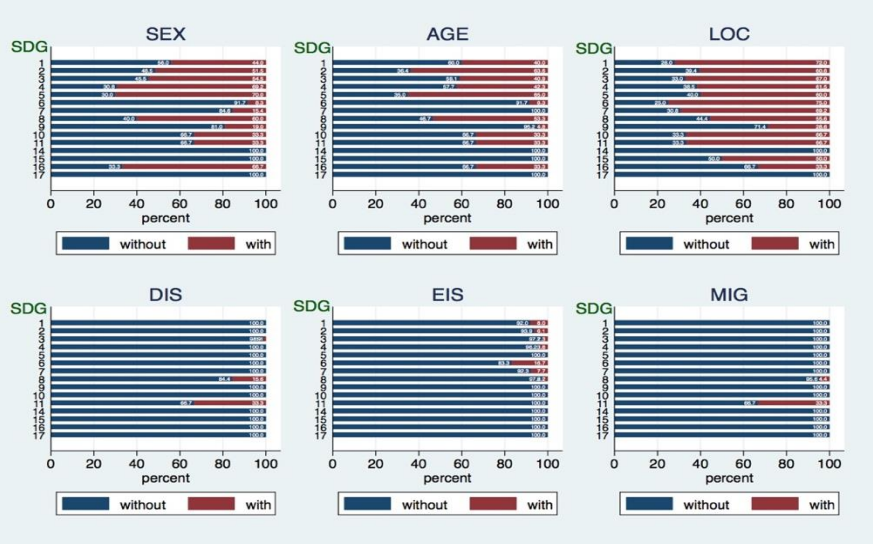

Figure 1. Distribution of SDG indicators by disaggregation

Source: 2017 ADB-UN ESCAP Survey

Notes: (i) SEX = Sex and gender; AGE = Age; LOC = Location or spatial disaggregation (e.g. by metropolitan areas, urban/ rural, or districts); DIS = Disability status; EIS = Ethnicity and indigenous status ; and MIG = Migration status. (ii) Percentages in red for each goal represent average of "with disaggregation" responses to the total responses at the indicator level.

[3] SAE methods strengthen direct survey estimates for small areas (or small sub-populations) with auxiliary information (such as census records).

Traditional Data Sources in Official Statistics

The NSOs across the world live up to a code of practice summarized by the United Nations Fundamental Principles of Official Statistics (FPOS). The first two principles of the FPOS summarizes the relevance of official statistics, the impartiality required in the production process, as well as the professional standards and ethics for ensuring the credibility of official statistics. Implicit in the first principle is a definition of official statistics, i.e., "data about the economic, demographic, social and environmental situation", and a mandate of national statistical systems to be auditors of a country's socio-economic performance. Maintaining independence and adherence to a professional conduct in the production and release of official statistics help guarantee the credibility of official statistics in the public arena.

Data sources in official statistics have used "tried and tested" mechanisms for ensuring the credibility of official statistics. National income accounts, data on prices, among others typically follow a Statistical Quality Assessment Framework to ensure integrity and credibility of resulting figures. Fellegi (1996) suggests that credibility is fundamental in official statistics: "Credibility plays a basic role in determining the value to users of the special commodity called statistical information. Indeed, few users can validate directly the data released by statistical offices. They must rely on the reputation of the provider of the information. Since information that is not believed is useless, it follows that the intrinsic value and usability of information depends directly on the credibility of the statistical system. That credibility could be challenged at any time on two primary grounds; because the statistics are based on inappropriate methodology, or because the office is suspected of political biases.

While the quality of statistics involves various criteria, there has undoubtedly been more focus by NSOs in the production of official statistics on managing precision and accuracy over timeliness and other quality issues. Official statistics are currently almost exclusively based on surveys and censuses, as well as administrative data reporting systems from government programs, often resulting from legislative mandates provided to a NSS.

Blending Traditional and New Data Sources in the Data Revolution

Across the world, organizations in both public and private sectors collect data, sometimes as a by-product of an administrative function, and in other cases, designed specifically for generating statistics to inform decision makers. In the private sector, firms either directly collect data and produce statistics for internal use, or subcontract market research organizations to conduct data collection activities and summarize business insights from data collected.

With the dawn of the information and communications technologies (ICT) age, and the increasing use of ICT, a data revolution has arisen: more data being captured, produced, stored, accessed, analyzed, archived, and re-analyzed, and at an exponential pace. Further,

the resulting hyper-connectivity that connects person to person, people to machines, machines to machines has led to a deluge of digital data (known as "big data") characterized by three V's, volume, velocity and variety (Hilbert, 2013). While databases from traditional sources of official statistics have a big "volume," the resulting data can hardly be called "Big Data" unless data collection for these data sources is more frequent, i.e., hourly, daily or weekly instead of the usual monthly, semi-annually, or annually. An administrative reporting system such as a country's civil registration system, for instance, may somewhat qualify in terms of large volume, and even have a large velocity if the current population in a country is fairly large and growing by the minute, but the records from civil registrations systems are fairly structured.

Big data can be categorized largely into three main sources: human-sourced information (e.g., social networks), process-mediated data (e.g., search engines, commercial transactions), machine-generated data (e.g., mobile phone location). All of these voluminous, fast-paced, and complex data, however, are often by-products of transactions from hyper-connectivity, and as such, do not necessarily involve a target population, much unlike traditional data sources of official statistics.

While the private sector has been primarily engaged in harnessing big data, statistics that describe socio-economic development sourced from big data are emerging and complementing official statistics from traditional data sources. For instance, official statistics on flu incidence released by the US Center for Disease Control (CDC) have been found to correlate strongly with the number of google searches from the US on the term "flu" (Ginsburg *et al*., 2009). Twitter conversations in Jakarta City on rice prices have also been reported to be a reasonable means of monitoring actual prices of rice in the Indonesian capital (Letouzé, 2012). Through the Open Transport Partnership, near real-time traffic data and statistics, including speeds, flows, and delays at intersections, that are sourced from anonymized GPS data of ridesharing drivers[4] are getting used to examine critical areas in traffic management in the Philippines and other developing countries (Krambeck *et al*., 2015). More granular data on poverty have been sourced from anonymized call detail records (CDRs) and other information on behavior of mobile users (Smith *et al*., 2013). Digital traces of mobile phone usage have also been used to track population movements (see e.g. http://vimeo.com/37238326 for a visualization in Geneva) and examine people's behavior during disaster events. (Liang, *et al*., 2014). Nighttime luminosity maps with high-resolution daytime satellite images have been used to yield estimates of household consumption and assets (Jean et al., 2016).
In the Philippines, much work on making use of big data for development has been in the area of improving disaster preparedness. there is growing recognition that climate disasters (including storms and floods) that batter the country every year with their increasing

[4] These big data have the same use as travel time surveys, the traditional method of collecting traffic congestion data, but there is much more data gathered from ridesharing drivers in near real-time at nearly no cost. Data, however, from travel time surveys are still required to obtain analysis of different vehicle types, such as motorcycles and trucks, but traffic congestion information from big data here validate statistics from traditional sources.

intensity are becoming a very serious threat to the country's growth and development. As reported by Thomas et al. (2012), disaster data from the Centre for Research on the Epidemiology of Disasters suggests that within Asia and the Pacific, the Philippines experienced the fourth highest frequency (98) of intense hydrological disasters during 1971–2010, topped only by Indonesia (124), India (167), and the PRC (172), all of which have much larger land areas, and the Philippines experienced the highest frequency (218) of intense meteorological disasters in the region during the span of four decades.

The Philippine government's weather bureau suggests that from 1951 to 2010, the annual average frequency of tropical cyclones affecting the country has remain unchanged at around 19 to 20 cyclones per year. However, an examination of the typical paths of tropical cyclones per decade indicates that cyclones have been shifting southward toward central and southern Philippines. In addition, there is evidence that the amount of precipitation with cyclones appear to be increasing. To manage the risks associated with these hazards of nature, the Philippine government, through its Department of Science and Technology (DOST) started a flagship project called Nationwide Operational Assessment of Hazards (NOAH) in June 2012. Project NOAH involved the development of hydromet sensors (e.g. automatic rain gauges, water level sensors, stream gauges) as well as high resolution geo-hazard maps. The latter can provide national and local chief executives lead time early warning (i.e., targeted 6 hours or less lead time) to minimize the costs to lives, property and livelihood from these hazards of nature. Project NOAH used topographic maps generated by light-detection and ranging (LiDAR) for flood modeling but currently the maps generated are limited to selected locations around the country's major rivers basins. These maps and other weather information are shared publicly through the NOAH website http://noah.up.edu.ph/#/ and some mirror sites. These high velocity data has led national and local governments to become more disaster prepared. In Cagayan de Oro city alone, there is evidence of how information has brought about improved disaster readiness. In 2011, typhoon Sendong led to 676 deaths in Cagayan de Oro. A year later, a typhoon with a similar strength (Pablo) only had one associated death reported. The deaths during Typhoon Haiyan, whose direction was predicted accurately by Project NOAH, unfortunately, also suggests that information has no power to lessen disaster costs If local chief executives have lack of capacity to understand data on disaster risks that are provided to them. Thus, the communication of information is equally important as the information itself. Project NOAH ended at DOST, but has subsequently been resurrected at the University of the Philippines.

2. Readiness of Official Statistics Community to be part of Big Data ecosystem

There is undoubtedly enthusiasm about the emerging data revolution, and the possibilities of making use of Big Data for measuring and monitoring progress in societies. Official statisticians are taking note of this alternative data source, but with some degree of caution, as bigger data need not always mean better data. There is some tension between official statistics and big data, as the latter were not tailor made for statistical purposes and its use provides the risk of yielding figures that are far from reality. Big data is largely unstructured, unfiltered data exhaust from digital products, such as electronic and online transactions, social media, sensors (GPS,

climate sensors), and consequently, analytics can be poor, unlike traditional data sources utilized for official statistics that are well-structured with good analytics, but with a fairly high cost (for data collection), and typically infrequent conduct with time lags that stakeholders find unreasonable in an age of fast data.

Opportunities however to use big data for monitoring sustainable development are growing. One major concern of official statisticians on big data use is related to privacy, security, intellectual property and related issues (UNECE, 2013). Much of Big Data being generated includes personal information. Precise, geo-location-based information certainly pushes the boundary of privacy/confidentiality. We are all well aware that Amazon, Visa and Mastercard are watching our shopping preferences; Google is examining our browsing habits; Twitter is looking into our minds from our tweets; Facebook is inspecting various information about us, including our social relationships and what we share; and Mobile providers are collecting data on whom we talk with or SMS.

Privacy not only has legal issues but also technological and ethical ones. While users of technology routinely tick a box to routinely consent to the collection and use of web-generated data and may decide to have some information put on public view, it is unclear whether they actually consent to having their data being analyzed, especially if it can be to their disadvantage (see, e.g, Stopczynski, *et al*., 2014). Can users give "informed consent" to an unknown use? For instance, when Google Flu Trends was developed in 2008, did Google have to contact all its users for approval to use old search queries for this project? Even if that were possible, the time and cost for doing that would have been enormous for Google. So, should users be asked to agree to any possible future use of their data?

Although various mechanisms to protect privacy are in place, including asking people to opt out of studying the information they give, and anonymization methods, such as differential privacy and "space time boxes", these methods are not fool proof. That is, when we anonymize, there is potential to re-identify. While official statisticians have established protocols on data confidentiality, this is not the case for big data which are usually in the hands of firms in the private sector that do not necessarily make this tsunami of data publicly accessible. When big data is publicly available, it may only be a minute portion of the actual data. For instance, only a very small subsample of twitter data is available publicly for free (while the entire data has to be purchased for use), and there are questions on whether the actual sample made available by twitter is representative (Fan and Bifet, 2012). Digital traces can be incomplete. While big data can enlighten, it can also obscure information, especially if limitations of such data are poorly understood and if the data are examined inadequately, with bias and with malice. Complementing traditional with innovative data sources to portray socio-economic conditions should thus be undertaken with much care and preparation to ensure that resulting statistics are reliable and accurate.

Technology

Retrieving and examining big data streams require adequate technological infrastructure, both hardware and software. Many data mining tools are neither suitable nor efficiently used for large datasets in a sequential computer. NSOs that intend to routinely use big data will thus need better ICT infrastructure to download these big data sources (bandwidth), as well as to catalogue, organize and process the complex collage of data in a sufficiently timely manner. Recent practice in big data analytics is to utilize a cluster of physical computers running a framework tool such as Hadoop-MapReduce [5], and/or use cloud computing/processing. The availability of interfaces by some statistical packages such as open sourced R to Hadoop or MapReduce for most used statistical platforms has, however, significantly contributed to the use of big data analytics. Further, the cloud has also emerged as an ideal computing environment for big data (Agrawal *et al*., 2011). On the infrastructure side, cloud computing, through "infrastructure as a service" in a public cloud or "platform as a service" in a private cloud, provides options for accessing and managing very large data sets as well as for supporting powerful infrastructure elements at a relatively low cost. Further, an increasing number of "software as a service" in a hybrid cloud are also capable of performing the processing and data integration tasks.

A related technological issue is the curation of big data. The big data sources result in a messy collage of data points. There are those who think that there are big gains in velocity and cost over sacrificing precision and accuracy, i.e. Big Data may not be completely accurate, but it is "good enough" and in near real-time. But how good is "good enough?" Some work, for instance, (D. Butler, Nature, Feb., 2013) has noticed the over-estimation of Google Virus Trends of flu levels (11% in the US public for the flu season in 2013, almost double the CDC's estimate of about 6%). A study of Twitter and Foursquare data before, during and in aftermath of Hurricane Sandy (Grinberg, et al., 2013) revealed interesting results: (i) grocery shopping peaks the night before the storm; (ii) nightlife picked up the day after; (iii) greatest number of tweets about Hurricane Sandy came from Manhattan. The latter creates the illusion that Manhattan was the most hit in the US by Hurricane Sandy, and it certainly wasn't; it merely had the most twitter users.

Various tools have to be used to assess the veracity of big data. Bias does not necessarily disappear in voluminous big data. The gains in velocity (and cost) in yielding statistics from big data sources, as well as the complexity and the sheer size of big data however requires a

[5] Hadoop is an open-source software project, managed by the Apache Software Foundation, targeted at supporting the execution of data-oriented application on clusters of generic hardware. The Hadoop project is comprised of four modules: Hadoop Distributed File System (HDFS), the MapReduce model, and Hadoop Common and YARN. The first two modules are critical: HDFS allows data to be stored in an easily accessible format, across a large number of linked storage devices, while MapReduce carries out two basic operations - reading data from the database and putting it into a format suitable for analysis (map); and performing mathematical operations (reduce). Hadoop Common provides the tools (in Java) needed for s user's computer systems (e.g., Windows, Unix, MacOS, *etc.*) to read data stored under the Hadoop file system, while YARN manages resources of the systems storing the data and running the analysis. (White, 2012).

different type of data processing and analytic tools from those used for “small data” to yield statistics that are fit for use.

Capacity

Analytics on big data requires new skill sets. While NSOs have had experience in curating data from traditional data sources, they often have no data scientists that are strong in both data and computational focus. Related to analytics, big data can also be burdened with methodological challenges regarding its veracity. In January 2013, the Google Virus Trends estimate (11%) of flu levels in the United States was nearly double the official estimate (6%) from the CDC (Butler, 2013). The reliability of pictures portrayed by big data crucially depends on whether digital traces can represent information on an entire population (Cox et al., 2018).

Unlike traditional data sources of official statistics that are designed for producing precise and accurate statistics that estimate parameters of populations, many types of big data do not have clear target populations. Big data sources are usually produced as a by-product in the course of some other activity (such as making a call on a mobile or taking a photograph and sharing it).

Statistics production blending traditional and innovative data sources requires a new skill set for all NSO staff, from managers to methodologists to IT staff. Managers and leaders in NSOs will also need to have paradigm shifts in statistical production and will require soft skills in building partnerships in big data ecosystem.

Ecosystem

Big data is not just about data sources, sets or streams, but is also about a complex ecosystem (Letouzé, 2015). Disruptive technologies have the effect of ‘disintermediating’ producers of official statistics and developing country citizens and the private sector and other organizations (who supply data traditionally to NSSs), since emerging technologies empower citizens to collect and publish their own data.
Technological solutions of using big data in official statistics require strengthening institutions and developing proper skills, a process that requires building trust, which takes time, perseverance, as well as soft skills. New business models must be developed by NSOs to leverage data resources, human talent, and decision-making capacity. Institutional frameworks and arrangements, such as public private partnerships and linkages with various institutions engaged in data science need to be developed and enhanced to further promote official statistics as a public good, whose quality (including timeliness, disaggregation and meaningfulness) requires constant improvement. Countries will require guidance, such as, statistical standards and knowledge materials. The latter should include not only what works but what does not. Legal protocols will also be required by NSOs to access big data holdings for development purposes (without infringing on data privacy), as well as to prevent misuse of big data.

3. Ways Forward

Statistics tell a story. In the case of Big Data, the accuracy of the story may be suspect as each piece of information compiled is not given weights. While it is clear that Big Data is here to stay, it should also be equally clear that the data revolution and the use of Big Data do not mean the end of official statistics as we know it, but a reinvention of ways of doing things.

Recognizing the many barriers and bottlenecks in meeting the data demands for the SDGs, most NSOs from ADB member countries that participated in the 2017 ADB/UNESCAP Survey consider big data as a promising means of addressing data gaps for SDGs. The use of big data as an additional source for official statistics to be integrated alongside with traditional data sources, however, is important especially because of the costs associated to traditional data collection activities. Further, the increasing levels of non-response due to the burden associated with primary data collection, even if proposed with advanced modalities (such as web surveys) potentially yield losses in quality. Although a few NSOs in the 2017 ADB/UNESCAP Survey reported having access to aerial photos/ satellite imagery, mobile data, web-scraped online price data, and social media data, only a limited number mentioned having current big data projects. Those that have big data projects are currently working on using satellite imagery and geo-spatial data and social media data to improve the granularity of statistics on poverty and welfare. Global, regional and national initiatives have been undertaken or are underway on making use of big data.

Big Data Work of Development Partners

Recognizing the need for a typology of development data that will be required to comprehensively measure all of the elements identified in the SDG agenda, development partners have begun to support countries effectively in big data use. Some development organizations These big data though have often been ad hoc and the result of individual initiatives instead of a coordinated institutional approach among development partners. Further, several of these projects have been at the piloting or incubation phase.

The UN Global Pulse[6], an information initiative launched by the Executive Office of the United Nations Secretary-General, has upscaled its “now-casting” of twitter data for monitoring rice prices to those of other commodities (beef, chicken, onion and chilli) in Indonesia. It has also analyzed GPS-stamped tweets in Jakarta City and anonymized data from GPS navigation smartphone apps (such as Waze) to investigate commuting patterns in relation to near-real-time traffic conditions in the Indonesian capital. Global Pulse has also examined anonymized mobile phone data (particularly, call details records and airtime credit purchases) to produce a set of proxies for education and household characteristics in Vanuatu. It has also examined anonymized financial records from four financial service providers in Cambodia to determine factors affecting savings and loans mobilization, with a

[6] https://medium.com/pulse-lab-jakarta/tracking-the-sdgs-using-big-data-dad0ad351f2e

focus on gender disaggregation. Through its VAMPIRE (Vulnerability Analysis Monitoring Platform for Impact of Regional Events) platform, Global Pulse has also been examining satellite imagery to map locations with climate and rainfall anomalies and to provide climate data visualizations and early warning alerts to policymakers and the general population.

Recognizing the opportunities for using big data to accelerate development outcomes as well as to potentially close data gaps in monitoring sustainable development, UNESCAP has also been exploring linking various datasets (geospatial data especially from satellite imagery, census and household survey data, and administrative records) focusing on the need for granular data on multi-dimensional data on poverty, population dynamics (including movements, urbanization and migratory status), and disaster risk reduction.

Country Efforts on Big Data

The UN Statistics Division (UNSD) has also developed and maintained an inventory[7] of big data projects. The inventory includes work in several developing countries in Asia-Pacific, such as past and on-going undertakings on making use of scanner data from supermarket chains and other retailers, as well as online prices obtained from webscraping to generate price indices in China, Japan and Korea. mobile phone data with secondary data (such as land use and transportation networks) and primary data (from surveys) to yield information on population movement with high granularity and high frequency in Bangladesh and soon in Sri Lanka. Table 1 provides an overview of some of these initiatives in the region.

Table 1. List of Select Big Data-Related Initiatives in Asia and the Pacific

Economy	Institute or Department	Big Data Project
Australia	UN - Global Pulse	Estimating migration flows using online search data
Bangladesh	World Bank Group	Predicting vulnerability to flooding and enhancing resilience using big data
China, People's Rep. of	National Bureau of Statistics	Using web scraping price data for price index of e-commerce
		Crop survey by farmland: using satellite and aerial remote sensing to help estimate agricultural statistics
		Comparison of data of interbank transactions with retail sales: credit card data for use in verifying retail sales
		Application of big data for highway and waterway transport statistics
		Online price changes of means of production
		Big data enterprise statistical indicator
	World Bank Group	Using big data analytics to discover patterns of medical insurance utilization for medical cost monitoring in the People's Republic of China
	UNDP and Baidu	Using big data to support e-waste management in the People's Republic of
Japan	Ministry of Internal Affairs and Communications	Web scraping and scanner data for price statistics
Korea, Republic of	Statistics Korea	Online price index
		Daily migration of population: using mobile call detail record data for daily migration data
India	World Bank Group	Tracking light from the sky version 2.0 or monitoring rural electrification from space
		Real-time forecasting of skills demand and supply: analytics of big data from Babajob in India

[7] https://unstats.un.org/bigdata/inventory/

Economy	Institute or Department	Big Data Project
	UN - Global Pulse	Understanding immunization awareness and sentiment through analysis of social media and news content
Indonesia	World Bank Group	Big data for freight transport and logistics policy making
		Using mobile phone data for national, subnational, and geo-coded average prices
		Using big data to predict student achievement in low-income school settings
	UN - Global Pulse	Understanding public perceptions of immunization using social media
		Mining citizen feedback data for enhanced local government decision making
	ILO and UN Global Pulse Lab Jakarta	Using social media to track workplace discrimination against women in Indonesia
Pakistan	World Bank Group	Using high-resolution satellite imagery and detection algorithms to better track poverty in Pakistan
Philippines	World Bank Group	OpenRoads Philippines: improved real-time decision making of infrastructure investments for the Philippines by linking geospatial road network data with rich geo-tagged social data collected through mobile phones
Singapore	Department of Statistics	Integrated environment system (IES): using environmental sensing systems and data analytics for real-time environmental information
		Population estimates: using administrative data from many sources for population estimates
Sri Lanka	World Bank Group LIRNEasia	Enabling up-to-date and accurate authoritative country mapping with crowdsourced geospatial data
		Potential of mobile network big data as a tool in Colombo's transportation and urban planning
Viet Nam	World Bank Group	Using big data to predict student achievement in low-income school settings

Notes: ILO = International Labour Organization, UN = United Nations, UNDP = United Nations Development Programme.

Sources: United Nations Global Working Group on Big Data Project Inventory and United Nations Economic and Social Commission for Asia and the Pacific, as cited in ADB Key Indicators for Asia and the Pacific 2016

The appreciation for the value of incorporating big data in the work programs of data producing agencies is becoming more apparent in some developing countries. The National Statistical Office of Mongolia has developed a geospatial statistical framework that enables the tracking of highly mobile units of enumeration (i.e. the herders) prior to the conduct of censuses and surveys (Chimeddamba, 2017). The same framework, which makes use of satellite imagery, has also been used in the conduct of NSO Mongolia's by-Census of Agriculture to aid in the identification of crop types and estimation of production. The Department of Statistics Malaysia's (DOSM's) Statistics Big Data Analytics (STATSBDA) project[8] works on constructing a big data infrastructure that will implement the following key project components: 1) integration of business registry with trade database for identification of characteristics of enterprises engaged in international market; 2) adoption of webscraping techniques to improve the quality of Consumer Price Index; and 3) assessment of public feedback on the quality of official statistics produced through social media.

Big data also continues to be a valuable resource in the strategic formulation and implementation of government projects and programs.

In the Philippines, the Metro Manila Development Authority has partnered with data science consultancy firm Thinking Machines to develop solutions in easing the worsening

[8] http://statsbda.stats.gov.my/statsbda-info/

traffic situation in Metro Manila, through analysis of data from GPS navigation software Waze. Disaster mitigation and response has vastly improved with the use of big data in tracking and planning activities. The Government of Indonesia has worked with UN Global Pulse for the development of a crisis analysis tool Haze Gazer[9] which provides real-time information on fire and haze hotspots as well as the locations of vulnerable members of the population. The tool, which utilizes satellite data, baseline population information and citizen-generated data published in social media and national complaint system LAPOR!, can enhance the disaster management capacity of the Government of Indonesia by enabling the formulation and implementation of well-informed response strategies. A slightly similar approach is also being undertaken in China. The Ministry of Environmental Protection has been providing a platform to compile different big data sources in tackling the country's severe air pollution problem. Satellite data, drones and data from citizen reporting are used for prompt prediction and measurement of air quality, and the resulting information can be used to trigger early warning systems for potentially severe smog incidences (Zhang and Hughes, 2017).

The challenge for NSSs is to be more forward looking and open to making use of non-traditional data sources, such as Big Data. Clearly, there will also be a need to identify legal protocols and institutional arrangements so that a NSS can get access to Big Data. There will be a need for Public-Private Partnerships, perhaps in the form of bilateral arrangements of the NSS with those that own Big Data holdings. But there will also be a need to addressing privacy issues on Big Data, in order to prevent misuse of Big Data.

By integrating relevant big data sources into official statistics process, NSOs are best positioned to measure their accuracy, ensure the consistency of the whole systems of official statistics and providing interpretation while constantly working on relevance and timeliness.

Cognizant of the changing landscape in data and the need for granular data for the SDGs, the Asian Development Bank has started a 'Data for Development' knowledge and support technical assistance (KSTA) Project that aims to strengthen the capacity of NSOs in augmenting the limitations of conventional data sources with innovative data sources in official statistics production. These efforts are in support of the principle of the SDGs to leave no one behind, and its concomitant granular data requirements. ADB's project aims to keep tabs of relevant initiatives on using big data within the Asia and the Pacific region so that countries and their stakeholders can have a more nuanced understanding on the scalability of such initiatives.

Ten developing member countries of ADB participated in the project's inception meeting and workshop held in May 2018. During the meeting, half of the participating countries reported specific work currently undergoing or in the pipeline to use mobile data (PHI, BGD and MDV), webscraping and remote sensing (VTN), social media and scanner data (THA).

[9] http://hazegazer.org/

ADB's KSTA project intends to develop two country case studies that explore using satellite data (and mobile data, subject to availability) as extra auxiliary information aside from census data that will be meshed with household survey data to yield small area estimates of population, poverty, and employment indicators. Doing country-specific case studies is important in examining what resources are needed for NSOs to be able to integrate relevant big data sources into official statistics process, measure their accuracy, ensure the consistency of the whole systems of official statistics and provide interpretation while constantly working on relevance and timeliness.

Online course modules are to be also developed for the project to scale knowledge sharing. Potential synergies and opportunities for collaboration are being explored by ADB with other development partners such as UNESCAP and PARIS21 to strengthen the ecosystem on big data for development, particularly through capacity development of NSOs to meet the disaggregated data requirements of the SDGs for ensuring that as the world takes its sustainable development path, no one gets left behind.

References

Agrawal, D. , S. Das, and A. El Abbadi. 2011. Big Data and cloud computing: current state and future opportunities in Proceedings of the 14th International Conference on Extending Database Technology. ACM. https://openproceedings.org/2011/conf/edbt/AgrawalDA11.pdf (accessed June 3, 2018).

Butler, D. 2013. When Google got flu wrong. Nature. https://www.nature.com/news/when-google-got-flu-wrong-1.12413 (accessed June 3, 2018).

Chimeddamba, L. 2017. Mongolia's experience on integrating geospatial information with administrative records to produce official statistics. Presented in Workshop on Integrating non-traditional data sources in the production of the SDG indicators 28-30 November 2017. https://unstats.un.org/unsd/capacity-building/meetings/da10-workshop-2017/4b.2-Mongolisa-EN.pdf (accessed July 13, 2018).

Cox, D.R., C. Kartsonaki, and R. H.Keogh. 2018. Big data: Some statistical issues. Statistics & Probability Letters Volume 136, May 2018, Pages 111-115.
https://www.sciencedirect.com/science/article/pii/S0167715218300609#sec3 (accessed June 3, 2018).

Fan, W., and A. Bifet. 2012. Mining big data: Current status, and forecast to the future 2012.
http://kdd.org/exploration_files/V14-02-01-Fan.pdf (accessed June 3, 2018).

Fellegi, I. 1996. Characteristics of an Effective Statistical System, International Statistical Review, Vol 64, pp165-197.

Ginsburg, J., M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. 2009. Detecting influenza epidemics using search engine query data. Nature 457:1012–14.
https://www.nature.com/articles/nature07634 (accessed June 3, 2018).

Grinberg, N., Naaman, M., Shaw, B., and Lotan, G. Extracting Diurnal Patterns of Real World Activity from Social Media. Available on the Internet: http://sm.rutgers.edu/pubs/Grinberg-SMPatterns-ICWSM2013.pdf (accessed June 3, 2018).

Jean, N., M. Burke, M. Xie, W. M. Davis., D. B., Lobell, S. Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. Science. Vol. 353, Issue 6301, pp. 790-794 DOI: 10.1126/science.aaf7894
http://science.sciencemag.org/content/353/6301/790 (accessed June 3, 2018).

Krambeck, H., N. Beschorner, V. N. Dato, B. Sanchez, A. Nuno, L. Qu, Y. C. Lu, and K. Webb. 2015. OPEN TRAFFIC Big Data Challenge Project Completion Report No: ACS15491. World Bank.
http://pubdocs.worldbank.org/en/513661445369530688/Open-Traffic-Completion-Report-1.pdf (accessed June 3, 2018).

Letouzé, E. 2015. Big Data & development: An Overview. Data-Pop Alliance White Paper Series. Data-Pop Alliance, World Bank Group, Harvard Humanitarian Initiative.

http://datapopalliance.org/wp-content/uploads/2015/12/Big-Data-Dev-Overview.pdf
(accessed June 3, 2018).
Letouzé, E. 2012. Big Data for development: Challenges and opportunities. UN Global Pulse.
http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf
(accessed June 3, 2018).
Liang, G., C. Song, Z. Gao, A.-L. Barabasi, J. P. Bagrow, and D. Wang. 2014. Quantifying Information Flow During Emergencies. Scientific Reports. 4 : 3997. DOI: 10.1038/srep03997.
https://www.nature.com/articles/srep03997.pdf
(accessed June 3, 2018).
Secretariat of the Partnership in Statistics for Development in the 21st Century (PARIS21). 2017 Partner Report on Support to Statistics PRESS 2017
http://www.paris21.org/sites/default/files/2017-10/PRESS2017_web2.pdf
(accessed June 3, 2018).
Smith C., A. Mashadi, L. Capra. 2013. Ubiquitous sensing for mapping poverty in developing countries, Proceedings of the Third Conference on the Analysis of Mobile Phone Datasets.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.408.9095&rep=rep1&type=pdf (accessed June 3, 2018).
Thomas, V., Albert, J.R., Perez, R. 2013. Climate-Related Disasters in Asia and the Pacific. ADB Economics Working Paper Series No. 358 , July 2013. http://www.adb.org/sites/default/files/pub/2013/ewp-358.pdf
(accessed June 3, 2018).
United Nations (UN). 2015. A/RES/70/1 - Transforming our world: the 2030 Agenda for Sustainable Development. Resolution adopted by the General Assembly on 25 September 2015.
https://undocs.org/A/RES/71/313
(accessed June 3, 2018).
United Nations Economic Commission for Europe (UNECE). 2013. What Does “Big Data” Mean for Official Statistics? https://statswiki.unece.org/pages/viewpage.action?pageId=77170614 (accessed June 3, 2018).
White, T. 2012. Hadoop: the definitive guide. O'Reilly. ISBN 978-1-449-31152-0.
Zhang, B. and Hughes, R. 2017. China deploys big data to clear smog. Nature 542:31. DOI: 10.1038/542031a.
https://www.nature.com/articles/542031a (accessed July 13, 2018).