UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT

June 2018

UNCTAD Research Paper No. 23
UNCTAD/SER.RP/2018/5/Rev.1

Steve MacFeely
Statistics and Information Branch, Division on Globalization and Development Strategies, UNCTAD
steve.macfeely@unctad.org

The Big (Data) Bang: What will it mean for compiling SDG indicators?

Abstract

In March 2017 the United Nations Statistical Commission adopted a measurement framework for the UN Agenda 2030 for Sustainable Development, comprising of 232 indicators designed to measure the 17 Sustainable Development Goals (SDGs) and their respective 169 targets. The scope of this measurement framework is so ambitious it led Mogens Lykketoft, President of the seventieth session of the UN General Assembly, to describe it as an 'unprecedented statistical challenge'.

Following partial success in compiling Millennium Development Goal indicators, where after 15 years, only two-thirds of the indicators were populated, statisticians are wondering, whether in the absence of some dramatic change, will the result for the SDG framework be any better? Could big data be that dramatic change? This paper outlines the opportunities, challenges and governance issues involved with big data from the perspective of producing SDG indicators. In particular the paper will examine some of the challenges surrounding access, and for confidentiality and privacy.

Key words: Governance, Privacy, Confidentiality, National statistical offices, National statistical systems, Global monitoring framework

The findings, interpretations, and conclusions expressed herein are those of the author(s) and do not necessarily reflect the views of the United Nations or its officials Member States. The designations employed and the presentation of material on any map in this work do not imply the expression of any opinion whatsoever on the part of the United Nations concerning the legal status of any country, territory, city, or area or of its authorities, or concerning the delimitation of its frontiers and boundaries.

This paper represents the personal views of the author(s) only, not the views of the UNCTAD secretariat or member States. The author(s) accept sole responsibility for any errors. Any citation should refer to the author(s) and not the publisher. This paper has not been formally edited.

UNITED NATIONS

© 2017 United Nations

Contents

Acknowledgements

Sincere thanks to Aidan Condron and Padraig Dalton (Central Statistics Office - Ireland), Steven Vale (United Nations Economic Commission for Europe - Switzerland) and Anu Petola (United Nations Conference on Trade and Development - Switzerland) for their insightful comments and suggestions.

Introduction

In March 2017 the United Nations (UN) Statistical Commission adopted a measurement framework for the UN Agenda 2030 for Sustainable Development (United Nations, 2015), comprised of 232 indicators designed to measure the 17 Sustainable Development Goals (SDGs) and their respective 169 targets[1]. These universal goals cover all three key development pillars: economic, social and environmental. They also include indicators for enablers such as institutional coherence, policy coherence and accountability. The scope of Agenda 2030 is vast, attempting to span the full spectrum of development issues, including not only aspects of society, economy and the environment but also institutional coordination. The intricacies and ambition of this challenge led Mogens Lykketoft, President of the seventieth session of the UN General Assembly, to describe it as an 'unprecedented statistical challenge' (Lebada, 2016).

National statistical offices (NSOs) around the world and members of the Inter-Agency and Expert Group on SDG Indicators (IAEG-SDGs), the group established by the United Nations Statistical Commission to develop and implement the global indicator framework for the Goals and targets of the 2030 Agenda are faced with two questions: the first question is whether this unprecedented statistical challenge can be met? The second is what contribution, if any, might big data make? Of the 232 SDG indicators, only 93 are classified as Tier 1, meaning that the indicator is conceptually clear, has internationally established methodology and standards, and data are regularly compiled for at least 50% of countries. The remaining indicators are Tier 2 (66 indicators) meaning the indicator is conceptually clear, has an internationally established methodology and standards are available, but data are not regularly produced by countries or Tier 3 (68 indicators), meaning that no internationally established methodology or standards are yet available. Five indicators have multiple tier classifications (United Nations Statistical Division, 2017). In other words, as of December 2017, less than half (only 40%) of the SDG indicators can be populated. At the end of the Millennium Development Goals (MDGs) lifecycle in 2015, countries could populate, on average, only 68 percent of MDG indicators (United Nations Conference on Trade and Development, 2016). Compared with the 169 targets set out by the SDG programme (United Nations Economic and Social Council, 2016), the MDGs requirements were modest, both in number (21 targets and 60 indicators) and complexity (United Nations Statistics Division, 2008). If past performance is any indication of the future, then it is not unreasonable to predict, that unless something changes dramatically, the proportion of populated indicators for the SDG monitoring framework will not be significantly different to the MDGs. Could big data be that dramatic change?

Over recent years the potential of big data for government, for business, for society, for official statistics has excited much comment, debate and even evangelism. Described as the new science with all the answers (Gelsinger, 2012) and a paradigm destroying phenomena of enormous potential (Stephens-Davidowitz, 2017) big data are all the rage. Official statisticians, already with a long history of using non-survey data, must decide whether big data are something new and useful. On the one hand, some argue that 'Big Data needs to be seen as an entirely new ecosystem comprising new data, new tools and methods, and new actors motivated by their own incentives, and should stir serious strategic rethinking and rewiring on the part of the official statistical community' (Letouzé and Jütting, 2015: 5) whereas others argue to the contrary that big data are just hype and that 'Big Data, Small Data, Little Data, Fast Data, and Smart Data are all "Just Data"' (Thamm, 2017: 2). This paper will examine these issues from the perspective of compiling SDG indicators and outlines some of the opportunities, challenges and governance and privacy issues that big data present.

Big data are the by-product of a technological revolution. In simplistic terms, one can think of big data as the collective noun for all of the new digital data arising from our digital activities. Our increasing day-to-day dependence on technology is leaving 'digital footprints' everywhere. Those digital footprints or digital exhaust offer official statisticians rich and tantalizing opportunities to augment or supplant existing data sources or

[1] These indicators were adopted by the UN Statistical Commission in March 2017 (see UNSC 48 – E/CN.3/2017/35) and were subsequently endorsed by the United Nations Economic and Social Council (ECOSOC) in June 2017 and by the United Nations General Assembly on 6 July 2017 (see A/RES/71/313).

generate completely new statistics. With the computing power available today these digital data can be shared, cross-referenced, and repurposed as never before, opening up a myriad of new statistical possibilities. Big data also present enormous statistical and governance challenges and potential pitfalls: legal; ethical; technical; and reputational. Big data also present a significant expectations management challenge, as it seems many hold the misplaced belief that accessing big data is straight-forward and that their use will automatically and dramatically reduce the costs of producing statistical information. As yet the jury is out on whether big data will offer the statisticians tasked with compiling indicators for the SDG monitoring framework anything especially useful. Beyond the hype of big data, and hype it may well be[2], statisticians understand that big data are not always better data and that more data doesn't automatically mean more insight. In fact more data may simply mean more noise. As Boyd and Crawford (2012: 668) eloquently counsel 'Increasing the size of the haystack does not make the needle easier to find.'

Organisation of the paper

To understand the opportunities, challenges and governance issues involved with big data from the perspective of producers of SDG indicators, it is useful to first define what we mean by big data, administrative data and official statistics. Thereafter the paper will look at sources of big data before examining the opportunities and some of the challenges presented by big data, in particular problems with access, and confidentiality and privacy. Before concluding, the paper will briefly outline some of the governance structures that National Statistical Offices (NSOs) and International Organisations (IOs) may need to consider putting in place if they intend to harvest big data for the purposes of compiling SDG indicators.

1. Definitions

This section briefly outlines what we mean by big data, administrative data and official statistics.

1.1 Big Data

What are big data? While some, such as, Stephens-Davidowitz argue that big data are 'an inherently vague concept' (2017: 15) it is nevertheless important to try and define it. This is important, if only, to explain to readers that big data are not simply 'lots of data' and that despite the name 'big data', size is not the defining feature. So if not size, what makes big data big? One of the challenges in trying to answer this question is that 'there is no rigorous definition of "big data"' (Mayer-Schonberger and Cukier 2013: 6).

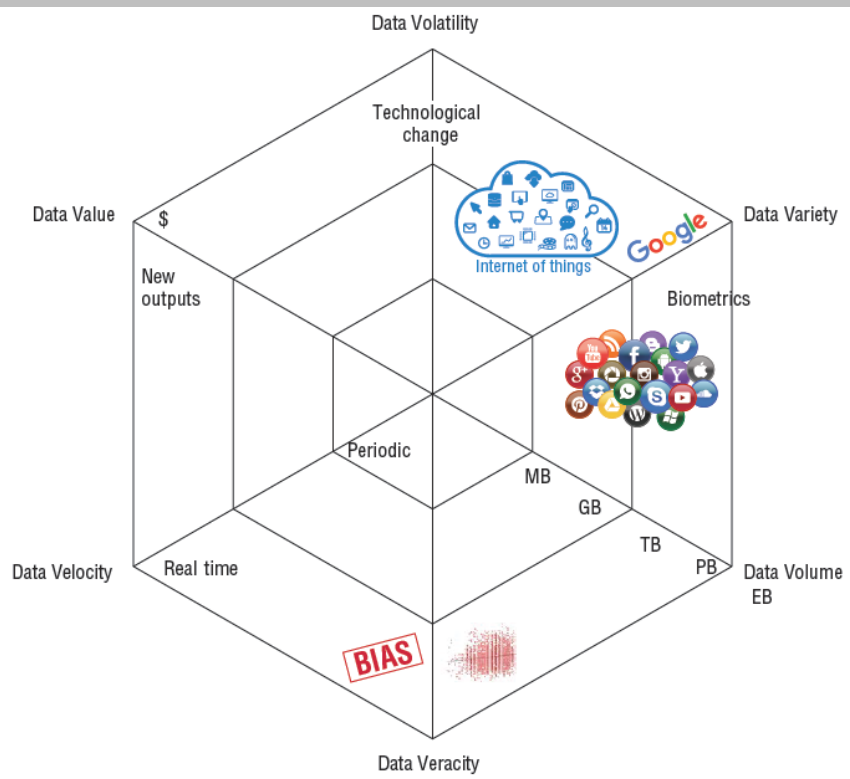
Gartner analyst Doug Laney (2001) provided what has become known as the three 'Vs' definition. He described big data as being high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. In other words, big data should be huge in terms of volume (i.e. at least terabytes), have high velocity (i.e. be created in or near real-time), and be varied in type (i.e. contain structured and unstructured data and span temporal and geographic planes). The European Commission (2014) definition of big data; 'large amounts of data produced very quickly by a high number of diverse sources' is essentially a summary of the 3V's definition.

It seemed that the 3Vs definition was generally accepted, within official statistics circles at least, with the United Nations Statistical Commission (2014: 2) adopting a very similar definition - 'data sources that can be described as; high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making.' But Tam and Clarke (2015: 3) have used a more general definition, simply describing big data as 'statistical data sources comprising both the traditional sources and new sources that are becoming available from the "web of everything".'

[2] Buytendijk (2014) argued that big data had passed the top of the 'Hype Cycle' and was moving towards the 'Trough of Disillusionment' and that expectations regarding the use of big data would now become more realistic.

Over the intervening years, the '3Vs' have swollen to '10Vs.[3]' Perhaps more usefully, in 2017 Hammer et al. selected a 5V definition (the original 3Vs plus an additional two V's - volatility and veracity). Veracity refers to noise and bias in the data and volatility refers to the 'changing technology or business environments in which big data are produced, which could lead to invalid analyses and results, as well as to fragility in big data as a data source' (2017: 8). At first glance, the additional Vs may seem odd as they are not per-se defining characteristics of the data or intrinsic to it. Nevertheless, volatility and veracity are extremely important additions, in particular for understanding the contribution that big data might make to compiling official statistics. Certainly the '5Vs' definition is more balanced and useful from an analytical perspective than the 3Vs as it flags some of the downside risks that prompted Borgman (2015: 129) to note that using big data is 'a path with trap doors, land mines, misdirection, and false clues.' But arguably a '6V' definition that includes 'value', where value means that something useful is derived from the data offers a better definition - introducing the notion of cost-benefit and yet striking a balance between parsimony and utility. The introduction of value is extremely important, as the costs of investing in big data must be carefully weighed up against what they might deliver in practical terms. See Figure 1.1.

Figure 1.1
The 6Vs of Big Data for Official Statistics



Derived from Hammer et al. (2017)

In understanding big data from a statistical perspective, it is important to understand that like administrative data, big data is conceptually quite different to traditional survey or census data. 'Big Data sets are made available to us, rather than designed by us' (Dass et al., 2015: 256). They are a collection of by-product data rather than data designed by statisticians for a specific purpose. In other words the derivation of statistics is a secondary purpose. This difference is perhaps obvious but profoundly important. We find ourselves today in a situation reminiscent of the storyline in 'The Hitchhikers Guide to the Galaxy' (Adams, 1979), where we now have the answers but we are still struggling to define the question. As big data is a by-product of our

[3] Volume, Velocity, Variety, Variability, Veracity, Validity, Vulnerability, Volatility, Visualisation and Value.

interactions with technologies that are evolving quickly, we must also accept that as a consequence big data are not a stable platform but a very dynamic one, and so any definition is likely to require further refinement over time.

1.2 Administrative Data

Many NSOs already make extensive use of administrative data. Blackwell (1985: 78) defined administrative or public sector data as 'information which is collected as a matter of routine in the day-to-day management or supervision of a scheme or service or revenue collecting system.' Similarly, United Nations Economic Commission for Europe (2011) defined administrative data 'as collections of data held by other units or entities of government, collected and used for the purposes of administering taxes, benefits or services.' In other words across public services, a huge volume of administrative records are collected, maintained and updated on a regular basis. These data pertain to the wide range of administrative functions in which the state is involved, ranging from individual and enterprise tax payments to social welfare claims or education or farming grants. Typically these administrative records are collected and maintained at the lowest level of aggregation, i.e. at individual transactions level. The interactions of individual taxpayers, applicants and recipients make these data very rich from an analytical perspective (MacFeely & Dunne, 2014). Brackstone (1987) identified four key distinguishing characteristics of administrative data: (1) the data are collected by an agency other than an NSO; (2) the methodology and processing are controlled by an agency other than an NSO; (3) the data were collected for non-statistical purposes; (4) the data have complete coverage of the target population.

Although administrative datasets are often very large (high volume), they are not typically considered big data, in the sense that they are not updated in real time (low velocity), they tend to be relatively stable (low volatility) and they are typically structured (low variety). It is worth noting that in 1997, Eurostat, proposed a narrow and a broad definition of administrative data. The narrow view saw administrative data comprising of only public sector non-statistical sources whereas the wider definition includes private sector sources (presumably including big data). The Conference of European Statisticians adopted a definition of administrative data consistent with this wider concept in 2000 (United Nations Economic Commission for Europe, 2000). Notwithstanding their differences, administrative and big data do share some important characteristics: both are secondary data; neither is originally compiled for statistical purposes; and both may suffer from problems with veracity. But there are some important differences too: administrative data are typically national whereas big data may be supra-national or global[4]; big data are inherently more unstable and volatile than administrative data; and big data will in all likelihood compromise a greater variety of sources and types of data.

1.3 Official Statistics

It is also useful to define official statistics, as it is official statisticians who have been tasked with populating the SDG monitoring framework (United Nations, 2017). The purpose of official statistics is to provide 'an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honor citizens' entitlement to public information' (United Nations, 2014: Principle 1). This is a challenging and ever more complicated role, as in addition to measuring the traditional economic, social and environment dimensions, demands for new indicators, on topics as varied as peace, security and wellbeing have emerged.

[4] This is an important difference as a NSO may be able to influence the content or quality of national administrative data, whereas it is highly unlikely to be able to exert any influence over global datasets.

Official statistics can be national or international (or both). Official National Statistics are all statistics produced by the NSO in accordance with the *Fundamental Principles of Official Statistics* (United Nations, 2014), other than those explicitly stated by the NSO not to be official; and all statistics produced by the National Statistical System (NSS) i.e. by other national organisations that have been mandated by national government or certified by the head of the NSS to compile statistics for that specific domain. So in practice, if a NSO produces statistics they are, de facto, official unless stated otherwise. If another organisation within a NSS produces statistics, then they are typically also considered official.

Official International Statistics are statistics, indicators or aggregates produced by a UN agency or other international organisation in accordance with the *Principles Governing International Statistical Activities* formulated by the Committee for the Coordination of Statistical Activities (2014). It is often necessary for a UN agency, or other international organisation, to modify official national statistics that have been provided by an NSO or another organisation of a NSS, in order to harmonise statistics across countries or to correct evidently erroneous values. Furthermore, in the absence of an official national statistic, a UN agency or other international organisation may compile estimates. Thus, it is insufficient to define official international statistics as simply the reproduction of official national statistics.

2. Sources of Big Data

In a world where our day-to-day use of technology and applications are leaving significant 'digital footprints', it seems that just about everything we think or do is now potentially a source of data. Big data are being generated from a bewildering array of activities and transactions. Our spending and travel patterns, our online search queries, our reading habits, our television and movies choices, our social media posts - everything it seems now leaves a trail of data. Some examples – big data were generated by the 227.1 billion global credit/debit card purchase transactions made in 2015 (Nilson, 2018). The 7.7 billion mobile telephone subscribers in 2017 around the world (International Telecommunications Union, 2017) also unwittingly created big data every time they used their phone. In fact even when they didn't use their phones, they were still generating data - according to Goodman (2015) mobile phones generate 600 billion unique data events every day. Every day we send 500 million tweets (Krikorian 2013), 8 billion snapchats (Aslam 2015) and conduct 3.5 billion google searches. Everyday approximately 1.25 million trades are made on the London Stock Exchange alone (Statistica, 2018a). Each one of these transactions leaves several digital footprints, from which new types of statistics can be compiled. In fact as Stephens-Davidowitz (2017: 103) explains, today 'Everything is data.' The torrent of by-product data being generated by our digital interactions is now so huge it has been described variously as a data deluge; data smog; info-glut or an information overload. This deluge is also the result of an important behavioral change, where people now record and load content for free. Every day upload 1.8 billion images (Meeker, 2017), approximately 1.7 million TripAdvisor reviews (Statistica, 2018b) and every minute of every day we upload 400 hours of video to YouTube (Taplin, 2017). Weigand (2009) described this phenomenon where people actively share or supply data directly to various social networks and product reviews and which led to the evolution of the wiki model as a 'social data revolution'. Every day upload 1.8 billion images (Meeker, 2017)

The United Nations Economic Commission for Europe (2013) have devised a useful classification of big data types to try and simplify matters. They identified three broad sources or types: Social Networks or human-sourced information - these would include data sourced from social networks, blogs, pictures, videos, internet searches and so forth; Traditional Business systems or process-mediated data - includes data sourced from public agencies, medical records, commercial and e-commerce transaction etc.; and the 'Internet of Things' or machine-generated data - includes data sourced from from sensors, surveillance and computer systems and logs.

Not only have the sources changed, the very concept of data itself has changed - 'the days of structured, clean, simple, survey-based data are over. In this new age, the messy traces we leave as we go through life are becoming the primary source of data' (Stephens-Davidowitz, 2017: 97). Now data includes text, sound and images, not just neat columns and rows of numbers. Begging the question, in this digital age, how much data now exist. Definitional differences again make this a difficult question to answer, and consequently there are various estimates to choose from. Hilbert and Lopez (2012) estimated that 300 exabytes (or slightly less than one third of a zettabyte[5]) of data were stored in 2007. According to their 2017 Big Data factsheet, Waterford Technologies (2017)[6] estimated that 2.7 zettabytes of digital data exist. Goodbody (2018) states that 16 zettabytes of data are produced globally every year and that by 2025 it is predicted that that estimate will have risen to 160 zettabytes annually. IBM (2017) now estimates we create an additional 2.5 quintillion bytes[7] of data every day.

Although definitions and consequent estimates differ, it is clear that, a massive volume of digital data now exists. But as Harkness (2017: 17) wisely counsels, the 'proliferation of data is deceptive; we're just recording the same things in more detail'. Nor are all of these data necessarily accessible or of good quality. As Borgman (2015: 131) warns, big data must be treated with caution. A threat 'to the validity of tweets as indicators of social activity is the evolution in how online services are being used. A growing proportion of twitter accounts consists of social robots used to influence public communication….As few as 35 percent of twitter followers may be real people, and as much as 10 percent of activity is social networks may be generated by robotic accounts.' Furthermore, Goodman (2015) states that 25% of reviews posted on Yelp are bogus. Facebook themselves have admitted that 3% of accounts are fake and an additional 6% are clones or duplicates, the equivalent of 270 million accounts (Kulp, 2017). Taplin (2017) also states that 11% of display ads, almost 25% of video ads, and 50% of publisher traffic are viewed by bots[8] not people - 'fake clicks.'

There are also issues of coverage, as sizeable digital divides exist. For example, the International Telecommunication Union (2017) estimates that global Internet penetration is only 48% and global mobile broadband subscriptions 56%, although they are as high as 97% in the developed world. Although global coverage is improving rapidly, it still means that in 2017 almost half of the world's population did not use the web. The digital divide - limited access and connectivity to the web or mobile phones is creating a data divide. To quote William Gibson (2003) 'The future is already here - it's just not very evenly distributed.' Anyone excluded from web access or mobile phones will not have a digital footprint or at best, a rather limited one. Even within countries, digital divides exist arising from a range of access barriers: social; gender; geographic; or economic. This may lead to important cohorts being excluded, with obvious bias implications for statistics (see Struijs and Daas 2014). Representativity is extremely important in the context of Agenda 2030 where one of the underlying ambitions is to ensure that 'no one gets left behind' (United Nations, 2015) or as Mogens Lykketoft, President of the seventieth session of the UN General Assembly, translated this; 'leaving no one uncounted' (Lebada, 2016). The question being asked by NSOs is whether these data can be safely accessed and whether they are representative and stable enough to be used to compile official statistics.

3. Opportunities for compiling SDGs

There will almost certainly be opportunities in the future to compile SDG indicators in new and exciting ways. Assuming access problems can be overcome then big data offer the potential to contribute to the measurement of SDGs in a number of ways. According to the *Big Data Project Inventory*[9] compiled by the UN Global Working Group on Big Data, 34 NSSs from around the world have registered 109 separate big data projects (see Table 3.1). NSOs and agencies are attempting to use satellite imagery, aerial imagery, mobile phone data, data scraped from websites, smart meters, road sensors, ships identification data, public

[5] A zettabyte is 10^{21} bytes (i.e. 1,000,000,000,000,000,000,000 bytes) or 1,000 exabytes or 1,000,000 petabytes

[6] It is not clear whether these estimates include data on the 'Deep Web' or 'Dark Net'. Goodman (2015) estimates that the Deep Web is 500 times larger than the google-able 'Surface Web'.

[7] A quintillion bytes is 10^{18} bytes or 1 exabyte.

[8] Goodman (2015) refers to these bots as WMD - Weapons of Mass Disruption.

[9] https://unstats.un.org/bigdata/inventory/ [examined on 27 April, 2018]. These numbers are a best estimate. Projects are not always well defined or explained on the inventory. Some projects seem to incorporate several projects or big data sources.

transport usage data, social media, scanner data, health records, patent data, criminal record data, Google alerts, and credit card data as sources to compile a wide range of official statistical indicators. These include, improving registers, compiling mobility, transport and tourism statistics, road safety indicators, price indices, indicators on corruption and crime, energy consumption, population density, nutrition, land use, wellbeing and measures of remoteness, labour market and job vacancies. The big data inventory is not of course an exhaustive catelogue of all big data activity, but it nevertheless provides a good overall picture of the types of activities that are underway. From Table 3.1, it is clear that NSSs are targeting web scraping, scanner and mobile phone data in particular - these three sources account for half of the big data projects underway. Although it should be noted that several projects are speculative or aspirational, where the big data source has not yet been identified or where access to the data (particularly mobile phone/CDR) has not yet been secured. Improving price indices using scanner data or prices scraped from the web are by far and away the most popular projects. This is not surprising as these approaches have been in development for many years[10] and typically have fewer data access problems.

Table 3.1 Big Data sources and project topics registered by National and International Organisations on the UN Big Data Project Inventory

Data Source	National	International	Project topic	National	International
Web scraping	22	4	Prices	22	4
Scanner	20	1	Population/migration	10	4
Mobile phone/CDR	14	18	Transport/mobility	9	11
Social media	8	23	Geographical/spatial	8	7
Satellite imagery	6	7	Labour market	7	2
Smart meter	5	1	Agriculture/Land use	6	4
Credit card	3	1	Tourism	5	1
Road sensor	5	-	Health/disease	4	7
Health records	5	2	Energy/Enviroment	4	6
Ship identification	2	-	Crime/Corruption	2	4
Criminal records	1	2	Poverty/inequality	1	9
			Disaster risk reduction	-	8
Other	20	31	Other	31	24
Total	111	90	Total	109	91

Source: Authors own calculations derived from UN Big Data Project Inventory[11]

IOs, most particularly the World Bank and the United Nations Global Pulse, are also investigating big data - they have logged 91 projects on the *Big Data Project Inventory.* Here too a wide variety of big data sources are being explored - mobile phone CDR records, Wikipedia, Google Trends, scanner data, web scraping, road sensor data, satellite imagery, credit card transactions, ATM withdrawls data, online purchases, aerial imagery, financial transaction data, taxi GPS data, freight data, medical insurance records, crime records, building certification data, OpenStreetMap, Twitter, publically available Facebook data, social media, aid data, bus fleet AVL data and electricity data. These big data may be used in conjunction with or as a replacement for traditional data sources to improve, enhance and complement existing statistics - Table 3.1 suggests that IOs are focusing on social media and mobile phone records to try and address issues regarding, in particular, transport, poverty and disaster mitigation. In 2017, the United Nations Global Pulse also listed 20 big data projects in their annual report[12] - these projects are using similar big data sources to those listed in Table 3.1 and have similar stated objectives. Figure 3.1 summarises the SDG goals towards which big data projects were focused in 2017 - goals 3, 8, 11 appear to be the most targeted, being

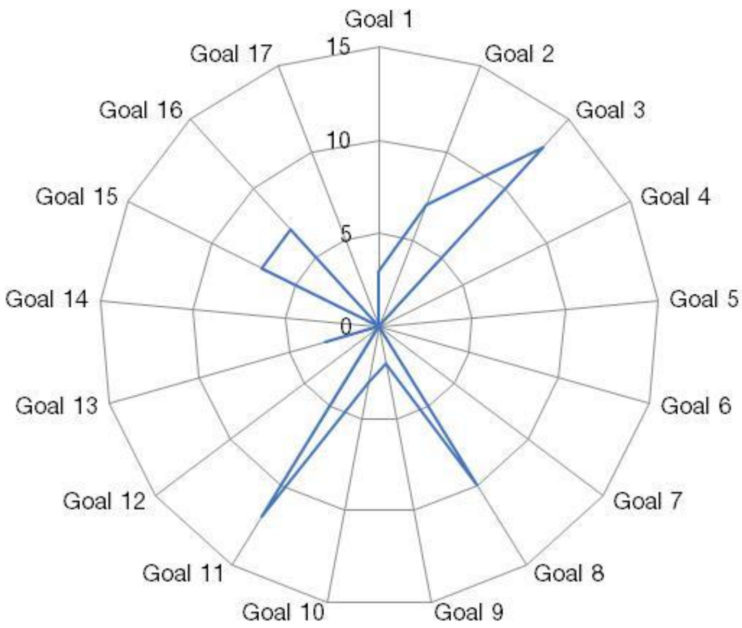
[10] See Guerreiro et al. (2018); Van Loon & Roels (2018); or Nyborg Hov (2018) for some recent examples.

[11] Readers will note that the totals for the data sources and projects do not match. This apparent mis-match arises as some projects use several sources, whereas in other cases a single source can be used on several projects. The data presented in Table 3.1 is a best estimate based on the text available in the project plans. Several projects are not well defined or don't appear to have any clear objective - hence the 'other' categories are quite large.

[12] These projects appear to be separate to the projects registered with the UN Big Data Project Inventory but given the detail available it is impossible to be certain.

included in at least 10 projects each. Goals 2, 15 and 16 enjoying somewhat less attention, included in 7 projects each.

Figure 3.1
SDG Goals being assisted by Big Data as reported by UN Global Pulse in 2017



Source: Derived from UN Global Pulse 2017 Annual Report[13]

Big data may offer new cost-effective or efficient ways of compiling indicators, improve timeliness or offer some relief to survey fatigue and burden. Big data also offers the tantalizing potential of being able to generate more granular or disaggregated statistics, allowing for more segmented and bespoke analyses, or the possibility of generating completely new statistics. Again, from an SDG perspective, this could be very important, not only in terms of realizing the lofty ambitions of leaving no one behind, but also from the perspective of realizing the general aim of target 17.18[14].

3.1 Linked data

Big data also offer the potential to compile datasets that are linkable, offering enormous potential to undertake more cross-cutting and dynamic analyses that may help us to better understand causation, offering the potential for more policy-relevant, outcome-based statistics. One of the short comings of many existing development indicators is that each indicator is derived from an official statistic which was compiled discretely, most likely from a sample. While this bespoke approach offers many advantages regarding bias, accuracy and precision, it has the disadvantage that as discrete data, those data cannot be easily connected or linked (other than at aggregate level) with other data i.e. the data cannot talk to each other. It is not always possible subsequently to construct a comprehensive analyses or narrative for many complex phenomena. As big data sets are more likely to have full or universal coverage, then provided there are common identifiers, the potential to match those data with other datasets increases enormously, significantly increasing the

[13] There is an element of double counting in Figure 3.1 as the report bundles several SDG goals against each project.

[14] 17.18 - By 2020, enhance capacity-building support to develop countries, including for least developed countries and small island developing States, to increase significantly the availability of high-quality, timely and reliable data, disaggregated by income, gender, age, race, ethnicity, migratory status, disability, geographic location and other characteristics relevant in national contexts.

analytical power of the data. For example, being able to link individual spatial pattern data with social data will massively improve our ability to model and understand how pandemics spread[15].

3.2 Improved timeliness

The possibility of improving timeliness by utilizing big data is enormously attractive. Policy makers require not only long-term structural information but they also require up-to-date, real time information - particularly during emergencies such as natural disasters or economic crises. Official statistics (and consequently MDG and SDG indicators) have generally been very good at providing the former but rather poor at the latter. This has been a long standing criticism of development indicators. In the words of the United Nations Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development (2014: p.22) 'Data delayed is data denied…The data cycle must match the decision cycle.' This presupposes, of course, that the public policy cycle has the capacity to absorb and analyse more volumous and timely statistics - it is not always clear that this is the case. Nevertheless, big data may offer the possibility of publishing very current indicators, using what Choi and Varian (2011: p.1) describe as 'contemporaneous forecasting' or 'nowcasting.' This may allow the identification of turning points much faster, which, from a public policy perspective could be very useful to making better decisions. This will be of critical importance for containing, not only pandemics, but also financial crises[16] and reacting quickly to natural disasters[17]. It should be noted that the SDG indicators are essentially performance metrics and, as such, are only reported annually. However more timely data may be of much greater importance for policy formulation and intervention stages required to implement Agenda 2030.

3.3 International production

Many digital data are supra-national or global in scope. This globalized aspect of big data offers exciting, although strategically sensitive, opportunities to reconsider the national production models currently employed by NSOs and NSSs all around the world. Switching from a national to a collaborative international production model might make sense from a pure efficiency or international comparability perspective, but it would be a dramatic change in approach, and possibly a bridge too far for many NSOs and governments. The sensitivities surrounding this topic are evident from the document 'Guidelines on Data Flows and Global Data Reporting for Sustainable Development Goals' prepared by the IEAG-SDG (United Nations Statistical Commission, 2018) where strong emphasis is placed on using nationally produced statistics as inputs into the global indicators. Nevertheless, in the case of global digital data, the most logical and efficient approach might be to centralize statistical production in a single center rather than replicating production many times over in individual countries. Obviously, this would not work for all domains, but for some indicators that could conceivably be derived from globalized big data sets it would offer the chance of real international comparability. Some examples of this might be land use, maritime and fishery statistics[18] derived from satellite imagery. Such an approach poses some difficult questions, not least legal. Globalized data present a particular challenge as they defy national sovereignty, putting the owners and the data themselves beyond the reach of national legal systems. Governments cannot always enforce national laws or ensure their citizens are protected.

[15] This is of particular relevance to target 3.d - Strengthen the capacity of all countries, in particular developing countries, for early warning, risk reduction and management of national and global health risks.

[16] For example, see target 17.13 - Enhance global macroeconomic stability, including through policy coordination and policy coherence. The Office for National Statistics in the United Kingdom are examining big data sources to see whether it can be used to developed what they call 'superfast indicators' of GDP growth (Mahajan, 2018).

[17] Targets 1.5, 2.4. 11.5, 11.b, 13.1.

[18] Targets 14.2, 14.3, 15.1, 15.2, 15.3, 15.4

3.4 Jump ahead

Across the world there exists not only a digital divide but also a significant data divide. For many developing countries, the provision of basic statistical information remains a real challenge. The Global Partnership for Sustainable Development Data (2016) notes that much of the data that does exist is 'incomplete, inaccessible, or simply inaccurate.' Jerven (2015) too has also been very critical of the quality of statistics available for many developing countries. In 2015, at the end of the fifteen year MDG life cycle, developing countries could populate, on average, only two thirds of the MDG indicators (United Nations Conference on Trade and Development, 2016). If this is a barometer for data availability in developing countries, then it is clear, that despite significant progress, serious problems with data availability persist. Some, Long & Brindley (2013); Korte (2014); and Ismail (2016) have argued, that owing to the falling costs associated with technology, big data may offer developing countries opportunities to skip ahead, and compile next-generation statistics today. Examples, such as, the massive growth of M-Pesa mobile money services in countries like Kenya, where almost half of the population use it, lend some credence to this argument (Donkin, 2017). Nevertheless others (Mutuku, 2016; United Nations Conference for Trade and Development, 2016; MacFeely & Barnat, 2017; Runde, 2017) have cautioned that in order to do so, there will need to be improved access to computers and internet, significant development in numeric and statistical literacy, and in basic data infrastructure. For example, the problems experienced with the UN Development Programme 'one laptop per child' scheme are an illustration of what can happen when ambitions run ahead of supporting infrastructure and connectivity (Keating, 2009). There are also concerns that as statistical legislation and data protection are often weak in many parts of the developing world, focusing on big data before addressing these fundamental issues might do more harm than good in the long term.

3.5 Better data

Big data may, in some cases, be better data than survey data. Seth Stephens-Davidowitz makes a compelling argument in his 2017 book *Everybody Lies* that the content of social media posts, social media likes and dating profiles is no more (or less) accurate than what respondents report in social surveys. However, big data has other types of data available that are of much superior quality. He explains 'the trails we leave as we seek knowledge on the internet are tremendously revealing. In other words, people's search for information is, in itself, information' (2017: 4). He describes data generated from searches, views and clicks as 'digital truth.' Thus, big data may be able to provide more honest data with greater veracity than could be achieved from survey data. Hand (2015) makes a similar argument, noting that as big data are transaction data they are closer to social reality than traditional survey and census data that are based on opinions and statements or rely on recall.

3.6 Data Broker

Finally, big data may offer the United Nations an opportunity to exercise some leadership and regain some control over an increasingly congested and rapidly fragmenting information space. Two opportunities spring to mind. Firstly, NSOs and IOs may find opportunities in rethinking and repositioning their role within the new data ecosystem. In the next section the challenges of accessing proprietary big data are discussed. This is a challenge, not only for NSOs, but for all sorts of institutions hoping to use big data. The United Nations Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development (2014) argued there is a role for someone - presumably the UN - to act as a data broker, to facilitate the safe sharing of data. The 2017 *Bogota Declaration on Big Data for Official Statistics* (United Nations Statistics Division, 2017b) hinted at a similar possibility, recommending a 'marketplace' for sharing data as a public good. At the global level, the UN would seem to be the most appropriate body. But perhaps at national level, there is a role also for NSOs to act as a trusted 3rd party intermediary, or honest broker, where big datasets could be housed, curated, anonymised and disseminated under strict and controlled conditions. This would be similar to the approach many NSOs already take for the release of anonymised microdata. If such a mechanism were available, it might encourage big data owners to release at least a sample of their data for statistical purposes.

3.7 Accreditation

Secondly, 'Statistical agencies could consider new tasks, such as the accreditation or certification of data sets created by third parties or public or private sectors. By widening its mandate, it would help keep control of quality and limit the risk of private big data producers and users fabricating data sets that fail the test of transparency, proper quality, and sound methodology' (Hammer et al, 2017: p.19). Similar proposals have been floated in the past (see Cervera et al. 2014; Landefeld, 2014; Kitchin, 2015; MacFeely, 2016). As noted in the introduction, the task of populating the SDG monitoring framework will be nothing short of herculean. Past experience suggests that adopting a 'business as usual' approach will bring only partial success. Instead, the United Nations could adopt a new proactive approach and introduce an accreditation system (with uniform standards) that would allow un-official compilers of statistical indicators to be accredited as 'official' for the purposes of populating the SDG monitoring framework. While UN Pulse[19] has already pioneered collaboration and partnership in the big data space, encouraging the sharing of big data sets, tools and expertise, what is envisaged here is a step further, offering accreditation or certification of indicators. Accreditation might take a number of forms. One could envisage an agreed, recognized and mandated body (for example, the IAEG-SDG), with the authority and competence to certify statistics as 'fit for purpose'[20], reviewing unofficial statistics to see whether they can be certified as 'official'[21] for the purposes of populating the SDG global monitoring framework. Statistics certified 'fit for purpose' could be accredited and used as official statistics. Without going into detail, this approach would only be used when particular conditions apply. For example, it might be used for Tier 3 or Tier 2 indicators that remain unpopulated by 2020 or 2025. Compilers of unofficial national indicators would need to demonstrate adherence to the *UN Fundamental Principles of Official Statistics* (United Nations, 2014). To secure global accreditation adherence to the *Principles Governing International Statistical Activities* (Committee for the Coordination of Statistical Activities, 2014) would be required. Indicators would also be required to meet a pre-defined set of quality and metadata standards, such as those set out in the *UN Statistical Quality Assurance Framework* (Committee of the Chief Statisticians of the United Nations System, 2018) and the *Common Metadata Framework* (United Nations Economic Commission for Europe, 2013b) respectively. Finally, prospective compilers of official SDG indicators would need to be able to guarantee that they can supply those indicators for the lifetime of Agenda 2030. In practical terms, this means being able to supply, at a minimum, the statistic on an annual basis for the years 2010 - 2030. Would this create sufficient incentive for big data holders to open up and reveal their metadata and help to make the idea of a multi stakeholder data ecosystem a reality? Such a move would not be without risks: legal; reputational; or equity. Landefeld (2014) also points out, that such a move might also face its share of resistance, based on ideological grounds, challenging the right of government to impose more regulation[22].

3.8 New or more granular indicators

Big data also potentially offer a range of other benefits and opportunities. The variety offered by big data provides not only new data sources but the promise of new types of data. These alternative or substitute data sources may offer mechanisms to relieve survey fatigue and burden to households and businesses. Given the exhaustive nature or massive volume of big data, they also offer opportunities to improve existing registers (or develop completely new ones) that could improve sample selection and weighting for traditional statistical

---

[19] https://www.unglobalpulse.org/about-new

[20] For the purposes of this discussion 'Fit for purpose' means that an indicator or statistic meets pre-defined quality and metadata standards and has been compiled in an impartial and independent manner. The quality and metadata standards must be clearly defined, open and transparent. The term quality can be interpreted in the broadest sense, encompassing all aspects of how well statistical processes and outputs fulfil expectations as a SDG indicator. In more concrete terms, 'fit for purpose' would mean that any statistic must be relevant, accurate, reliable, coherent, timely, accessible, and interpretable. The statistic must be produced using sound methodologies, concepts and reliable systems. The statistic must also be compiled within an institutional environment that recognises the need for objectivity, impartiality and transparency. This last point is important. In order for a statistic to be designated official, neither the input data nor the methodologies can be proprietary but must be available to all and open to scrutiny.

[21] Official statistics can be national or international (or both).

[22] Ironically, Hayak (1944), the grandfather of neo-liberal doctrine understood the role of government in regulating weights and measures.

instruments. The sheer volume of data may also allow greater disaggregation of some statistics allowing greater segmentation or granular analyses (including better geographic or spatial disaggregation). Big data may also be able to contribute to improving the quality of a number of existing statistics (for example, tourism statistics[23]) while also offering new approaches to measuring difficult concepts like wellbeing[24].

3.9 Summary of opportunities

In summary, big data offer a wide range of potential opportunities: cost savings; improved timeliness; burden reduction; greater granularity; linkability and scalability; greater accuracy; improved international comparability; greater variety of indicators; and new dynamic indicators. Big data may offer solutions to data deficits in the developing world where traditional approaches have so far failed. Big data may also offer opportunities to rethink the role of official statistics and re-position the United Nations and NSOs vis-a-vis the wider data ecosystem. But of course, big data also presents risks and challenges for compiling SDG indicators. These are examined in the next section.

4. Accessing and using Big Data

One of the biggest barriers to using big data is lack of access. Many big data are proprietary i.e. data that are commercially or privately-owned and not publically available. For example, data generated from the use of credit cards, search engines, social media, mobile phones and store loyalty cards are all proprietary and may not be available for use in official statistics. Even if these data were publically accessible, sensitivities around their repurposing to compile official statistics must be carefully considered. 'Even if there are no legal impediments, public perception is a factor that must be taken into account' warn Daas et al. (2015: 257). The current proprietary status of some data may change in the future as people around the world realise that their data are being used and traded. But for the moment many datasets are not currently accessible by NSOs, either because costs are prohibitive, data protection legislation prevents it or proprietary ownership makes it impossible. Changes to statistical legislation may be required to give NSOs or NSSs access to big data sources necessary for statistics, as recommended by United Nations Economic Commission for Europe (2018). MacFeely and Barnat (2017) made similar recommendations, arguing that in order to future-proof statistical legislation, consideration should be given to mandatory access to all appropriate secondary data, including some important, commercially held data.

NSOs must be extremely careful not to damage their reputation and the public trust they enjoy. To do so, a NSO must ensure it does not break the law or stray too far outside the culturally acceptable boundaries or norms of their country. An NSO must decide whether it is legally permissible, ethical or culturally acceptable to access and use big data. These are not always easy questions to answer. When it comes to accessing new data sources, the legal, ethical and cultural boundaries are not always clear-cut. In some cases NSOs may be forced to confront issues well before the law is clear or cultural norms have been established. Furthermore, given the speed with which the digital data world is changing, any legislation will be out of date before the ink is even dry (Rudder, 2014). This poses a challenge as public trust and reputation is fragile; hard won but easily lost. NSOs depend on the public to supply information to countless surveys and enquiries. If an NSO breaks that trust, they risk biting the hand that feeds them. Yet a progressive NSO must to some extent lead public opinion, meaning they must maintain a delicate balance, innovating and publishing new statistics that deal with sensitive public issues but without moving too far ahead of public opinion. For example, from a technical, statistical perspective the most logical and cost-effective method of deriving international travel and tourism statistics might be to use mobile phone data, but from a data protection and public opinion perspective using these type of data may not be acceptable.

[23] Target 8.9 and 12.b
[24] Target 17.19

This tension or trade-off does not appear to be well understood and is certainly not well reflected in many national and international policy documents. MacFeely (2017) notes that when it comes to big data there is a growing but discernible mismatch between potential and actual. The United Nations Economic Commission for Europe (2016) too, reflecting on their experiences, note 'High initial expectations about the opportunities of Big Data had to face the complexity of reality. The fact that data are produced in large amounts does not mean they are immediately and easily available for producing statistics. Data from mobile phones represent a notable example in this sense[25]. It has been proved that such data can be exploited for a wide range of purposes, but they are still largely outside the reach of the majority of statistical organizations, due to the high sensitivity of the data.'

5. Other challenges for compiling SDGs

5.1 Rapid change and instability

Technology, the source of many big data, continues to rapidly evolve. This continuing and rapid evolution raises questions regarding the long-term stability or maturity of big data and their practicality as a data source for the compilation of SDG indicators. As Daas et al. (2015: 258) note 'The big data sources encountered so far seem subject to frequent modifications'. For example, social media platforms may tweak their services to test alternative layouts, colours and design, which in turn may mutate the underlying data. Kitchin (2015: 9) warns 'the data created by such systems are therefore inconsistent across users and/or time'. The United Nations Economic Commission for Europe (2016) caution that statisticians using big data will need to accept a general volatility or instability in the data. They note 'Wikipedia access statistics show a general drop in the overall number of accesses from the time the mobile version of Wikipedia was released. Similarly, Twitter had a significantly lower number of geo-located tweets after Apple changed the default options for its products.' This has obvious implications for time series consistency, which in turn raises questions regarding the use of big data, as the central purpose of the SDG monitoring framework is to produce a consistent time series for the 2010 - 2030 period.[26] Hence the importance of incorporating the concept of volatility in to the definition of big data. Instability of some big data sources introduces risks to continuity of the data supply itself. NSOs must decide whether access to big data are sufficiently stable and those data are adequately mature to justify making the investment. Will this 'exhaust pipe' data be available over the twenty year period on a consistent basis? If not, then this will pose a challenge for SDG indicators. It is often said that data are the new oil. But data (just like crude oil) must be refined in order to produce useable statistics. And just like oil, if the quality and consistency of the raw input data (crude oil) keeps changing, it will be very costly and difficult to refine.

5.2 Data ownership

Ownership of source data is another issue of concern. As an NSO moves away from survey based data and becomes more reliant on administrative or other secondary data, such as big data, it surrenders control of its production system. The main input commodity, the source data, is dependent on external factors, exposing the NSO to the risk of exogenous shocks. Partnerships with third party data suppliers means, not only losing control of data generation, but perhaps also sampling and data processing (perhaps as a solution to overcome data protection concerns). Furthermore NSOs will have limited ability to shape the input data they rely upon (Landefeld, 2014; Kitchin, 2015). The technologies that produce 'tailpipe' data may change or become redundant, leading to changes in or disappearances of data. Changes in government social or taxation policy may lead to alterations or termination of important administrative datasets. Changes in data protection law, if it does not take the concerns of official statistics (and SDG compilers) into account, could

[25] See Eurostat (2014) for an excellent summary of how mobile phone data can be used to compile tourism statistics.

[26] Although the reference period for Agenda 2030 is strictly speaking 2015 - 2030, in many cases the time series for monitoring progress begins in 2010.

retard the development of statistics for decades[27]. These are risks that a NSO must carefully consider when deciding whether or not to invest in secondary data (administrative or big). Reliance on external data sources also introduces new financial and reputational risks. If a NSO is paying to access a big data set, there is always the risk, that the data provider realizing the value of the data will increase the price. There are also reputational risks. The first is that the public, learning the NSO is using or repurposing their social media, telephone, smart metering or credit card data without their consent, may react negatively. There may also be concerns or perceptions of state driven 'big brother' surveillance or what Raley (2013) terms 'dataveillance.' So an NSO must consider carefully how it communicates with the public to try and mitigate negative public sentiment. The other reputational risk is that of association. If an NSO is using particular social media data for example, and that provider becomes embroiled in a public scandal, the reputation of the NSO or IO may be adversely affected, through no fault of their own.

5.3 Data quality

As noted earlier, big data are essentially re-purposed data. Consequently a lot of contextual knowledge of the original generating system is required before the data can be recycled and used for statistical purposes. Developing that knowledge can be difficult as frequently data owners have no incentive to document changes or be transparent. Both the data and the algorithms are typically proprietary and often of enormous commercial value. But accessing accurate metadata is vitally important to using any secondary data. For example, understanding how missing data have arisen, perhaps from server downtime or network outages, is essential to assessing the quality of data and then using the data (Daas et al., 2015).

Big data can also be gamed or contain fake data (Kitchin, 2015; MacFeely, 2016) and so it is important to understand vulnerabilities in the data. There may also be challenges with regard to the representativity and accuracy of many big data: age; gender; language; disability; social class; regional; and cultural biases may exist. There are also concerns too that many social media are simply echo-chambers cultivating less than rigorous debate and leading to cyber-cascading, where a belief (either correct or incorrect) rapidly gains currency as a 'fact' as it is passed around the web (Weinberger, 2014). There are also concerns for veracity arising from the concentration of data owners. Reich (2015) notes that in 2010, the top ten websites in the United States accounted for 75 percent of all page views. According to Taplin (2017) Google has an 88 percent market share in online searches, Amazon has a 70 percent market share in ebook sales, Facebook has a 77 percent market share in mobile social media. Such concentration introduces obvious risks of abuse and manipulation, leaving serious questions for the continued veracity of any resultant data. The decision by the Federal Communications Commission (2017) in the United States in December 2017 to repeal Net Neutrality[28] raises a whole new set of concerns regarding the veracity of big data for statistical purposes. The United Nations Conference for Trade and Development (2015) noted that ambiguities exist for a range of issues connected with net neutrality, including traffic management practices and their effects on quality of service, competition, innovation, investments, and diversity, online freedom, and protection of human rights. Tim Berners-Lee (2014) has warned against the loss of net neutrality and the increasing concentration within the web: both trends that are undermining the web as a public good.

[27] For example, within the statistical community of the European Union there are concerns that the new General Data Protection Regulation (GDPR) has not fully taken the particular needs of official statistics into consideration. If this is the case then this new legislation may retard significantly the development of official statistics in that region. For this reason the UNECE Task Force on Statistical Legislation (2018) recommended that statistical authorities be exempt from restrictions on access to individual data laid down in other legislation. They note that the provision of personal data from other sources to statistical authorities is consistent with 'any other purpose' for which the data were primarily collected. They stress, correctly, that the intention of official statistics is not to use personal data for decisions regarding individuals, but rather to use statistical aggregates to inform broad policy decisions. They further note that the confidentiality of any data used by a NSO will be protected.

[28] Net Neutrality sets out the principles for equal treatment of Internet traffic, regardless of the type of service, the sender, or the receiver. In practice, however, the Internet service providers conduct a degree of appropriate traffic management aimed at avoiding congestion, and delivering a reliable quality of service. Concerns regarding the loss of net neutrality focus mainly on definitions of (in)appropriate and (un)reasonable management and discriminatory practices, especially those that are conducted for commercial (e.g. anti-competitive behaviour) or political reasons (e.g. censorship). Net neutrality has three important dimensions: (1) technical (impact on Internet infrastructure); (2) economic (influence on Internet business models); and (3) human rights (possible discrimination in the use of the Internet).

5.4 The digital economy

Apart from utilising big data, another challenge (and opportunity) for NSOs and IOs is to actually measure the digital economy itself (including big data) to help shed light on the importance of this rapidly emerging economy for trade, and other aspects of the 'traditional' economy and society. In the coming years, this will become an increasingly important subject and one where significantly more data and statistics will be required. From an SDG perspective, developments in information and communications technology are already relevant to several indicators[29]. More broadly, there is a growing demand for data to better explain the value added and benefits of the digital economy, the implications for taxation, how it facilitates trade, the contribution to GDP, and the likely influence on economic growth in the future. Countries are also wondering whether the digital economy can be harnessed to reduce economic, social and gender inequalities, and what are the regulatory challenges associated with trying to supervise, what is essentially a globalised phenomenon. Big data is of course only a sub-set of this bigger picture, but NSOs and IOs must address the data needs emerging from the complex world of the digital economy.

6. Privacy and confidentiality

For official statistics, safeguarding the confidentiality of individual data is sacrosanct and is enshrined in Principle 6 of the United Nations *Fundamental Principles of Official Statistics* (United Nations, 2014), which states 'Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.' The UN *Handbook of Statistical Organization* (United Nations, 2003) also underscores repeatedly the requirement that the information that statistical agencies collect should remain confidential and inviolate. The United Nations Office of the High Commissioner on Human Rights (2018) has also stressed the importance of protecting confidential data. The *Scheveningen Memorandum* (European Commission, 2013)[30] prepared by the Directors General of NSOs in the European Union identified the need to adapt statistical legislation in order to use big data - both to secure access but also protect privacy. The failure to treat individual information as a trust would prevent any statistical agency from functioning effectively. For a NSS to function, the confidentiality of the persons and entities for which it holds individual data must be protected i.e. a guarantee to protect the identities and information supplied by all persons, enterprises or other entities, and guarantee that their data are used for statistical purposes only. In short, everyone who supplies data for statistical purposes does so with the reasonable presumption that their confidentiality will be respected and protected[31]. In most countries, safeguarding confidentiality is enshrined in national statistical legislation. But with the increased volumes of big data being generated, and the potential to match those data, greater attention must be paid to data suppression techniques to ensure confidentiality can be safeguarded.

The emergence of big data is forcing many challenging questions to be asked, not least with regard to privacy and confidentiality. Mark Zuckerberg, the founder of Facebook, famously claimed that the age of privacy is over (Kirkpatrick 2010). Scott McNealy, CEO of Sun Microsystems, too asserted that concerns over privacy are a 'red herring' as we 'have zero privacy' (Noyes, 2015). John McAfee (2015), founder of McAfee Associates has stated that the 'concept of privacy is fast approaching extinction.' Many disagree and have voiced concerns over the loss of privacy (see Pearson, 2013; Payton and Claypoole, 2015). Fry (2017) has

[29] Targets 4.b, 5.b, 9.c and 17.8

[30] Para 3 - Recognise that the implications of Big Data for legislation especially with regard to data protection and personal rights (e.g. access to Big Data sources held by third parties) should be properly addressed as a matter of priority in a coordinated manner.

[31] In effect this means that only aggregate data can be published for general release by official statistical compilers and those aggregates will have been tested for primary and secondary disclosure. Data that cannot be published due to the risk of statistical disclosure are referred to as confidential data. Primary confidentiality disclosure arises when dissemination of data provides direct identification of an individual person or entity. This usually arises when there are insufficient records in a cell to mask individuals or when one or two records are dominant and so their identity remains evident despite many records (this is a recurring challenge for business statistics where 'hiding' the identity of large multinational enterprises can be very difficult). Secondary disclosure may arise when data that have been protected for primary disclosure nevertheless reveal individual information when cross-tabulated with other data.

likened developments with regard to big data and the loss of privacy to the opening of Pandoras Box - what he terms, Pandora 5.0. The introduction in Europe of the new General Data Protection Regulation which comes into effect in May 2018, reinforcing citizen's data-protection rights, including among other things the right 'to be forgotten', suggests that privacy is still a real concern (European Parliament 2016) - at least in some regions of the world. Yet in the United States, despite reported growing anxiety over online privacy (United Nations Conference for Trade and Development, 2018), users who provide information under the *'third-party doctrine'* i.e. to utilities, banks, social networks etc. should have 'no reasonable expectation of privacy.'

This introduces two new challenges for official statisticians: one technical and one of perception. The technical challenge arises from the availability of large, linkable datasets which present a problem thought to have been solved in traditional statistics – anonymisation. It is now clear, owing to the availability of big data and enormous computing power, that simply removing personal identifiers and aggregating individual data is not a sufficient safeguard. A paper by Ohm (2010) outlining the consequences of failing to adequately anonymise data graphically illustrates why there is no room for complacency. Thus a problem that had been solved for traditional official statistics must now be re-solved, in the context of a richer and more varied data ecosystem. From an SDG perspective, compilers must push back against any attempts to access individual data. The argument that 'no one is left uncounted' is not a sufficient justification to access to confidential microdata.

The changing nature of perception is arguably a trickier problem. What if Zuckerberg and McNealy are correct and future generations are less concerned about privacy? There appears to be some evidence to suggest that they may be correct. It seems there are clear inter-generational differences in opinion vis-a-vis privacy and confidentiality, where those 'born digital' (roughly those born since 1990) seem to be less concerned about disclosing personal information than older generations (European Commission, 2011). Taplin (2017: 157) ponders this, musing 'It very well may be that privacy is a hopelessly outdated notion and that Mark Zuckerberg's belief that privacy is no longer a social norm has won the day.' If this is so, what are the implications for official statistics and anonymisation? If other statistical providers, not governed by the Fundamental Principles, take a looser approach to confidentiality and privacy, it may leave official statistics in a relatively anachronistic and disadvantaged position vis-a-vis other data providers. But moving away from or discarding principle 6 of the UN *Fundamental Principles for Official Statistics* would seem to be a very risky move, given the importance of public trust for NSOs.

A related and emerging challenge for official statistics is that of open data, or more specifically, the asymmetry in openness expected of private and public sector data. Many of the 'open data' initiatives are in fact drives to open government data[32]. This of course makes sense, in that tax payers should to some extent own the data they have paid for, and so those data should be public, within sensible limits. But arguably people also own much of the data being held by search engines, payments systems and telecommunication providers too. So why is there an exclusive focus on public or government data? Letouzé and Jütting (2015: 10) have highlighted this issue, remarking that 'Official statisticians express an acute and understandable sense of frustration over pressure to open up their data to private-sector actors, while these same actors are increasingly locking away what they and many consider to be "their" data.' SDG indicators, as a public good, should of course be open. But the philosophy of open data should be more evenly applied to avoid asymmetrical conditions. This is a complex challenge, as to some extent it feeds off poor understanding of privacy issues, statistical literacy and the data wars that are underway at the moment. As Rudder (2014: 241) notes 'because so much happens with so little public notice, the lay understanding of data is inevitably many steps behind the reality.'

[32] For example: the OECD Open Government Data (OGD) is a philosophy- and increasingly a set of policies - that promotes transparency, accountability and value creation by making government data available to all - see: http://www.oecd.org/gov/digital-government/open-government-data.htm. In the United States, Data.gov aims to make government more open and accountable. Opening government data increases citizen participation in government, creates opportunities for economic development, and informs decision making in both the private and public sectors - see: https://www.data.gov/open-gov/. In the European Union, there is a legal framework promoting the re-use of public sector information - Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information. See - http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013L0037&from=FR

Taplin (2017) argues that we trade our privacy with corporations in return for innovation or the benefits of improved servies but challenges the need to surrender information to government. MacFeely (2016) has warned that if the benefits of privacy are insufficiently clear to the public or policy makers, then it leaves official statistics vulnerable, and possibly facing a precarious and bleak future. Rudder (2014: 242) highlights this challenge too noting that 'the fundamental question in any discussion of privacy is the trade-off - what you get for losing it.' Like Taplin, Rudder also argues that the trade-off with the private sector is clear - better targeted ads! He argues that 'what we get in return for the government's intrusion is less straightforward.' McNealy too, who seems unconcerned about the lack of privacy in the private sector, takes a very different attitude when it comes to government, saying 'It scares me to death when the NSA or the IRS know things about my personal life…Every American ought to be very afraid of big government' (Noyes, 2015). Curiously, while there is a real fear of government Big Brother, there appears to be little concern regarding the emergence of a corporate Big Brother. To some extent there is ideology at play here, where a neo-liberal agenda is pushing to minimise the role of the public sector, but it also illustrates the challenge facing the UN and national governments and their agencies generally where their contribution of the SDGs to the wellbeing of economies and societies is poorly understood.

6.1 Summary of challenges

Thus, while big data may offer opportunities, they also present some real challenges for NSOs, NSSs and IOs. To some extent, these challenges are magnified versions of problems that already exist with other data sources, such as, uncertainty over the quality or veracity of data and dealing with a range of potential biases. Access to external secondary sources, such as, administrative data can already be challenging, and is not unique to big data. But big data do appear to present some rather unique challenges with regard to rapidly evolving and unstable data, ownership of data, data protection and safeguarding confidentiality. These are some of the issues that NSOs and IOs will need to carefully consider before committing resources to any big data projects.

7. Goverance issues

In considering whether big data provide a viable option, SDG compilers must carefully decide what governance systems will be required to ensure the SDG and official statistics brands are not compromised[33]. For the purposes of this chapter, governance systems can be defined as the policies, rules and monitoring mechanisms that allow the management of a NSO or IO to direct and control the activities of the office. That governance system should help decision makers to balance the often competing needs of new SDG demands with the rights of data owners and ensure public accountability.

At a global level, questions naturally arise as to whether some sort of global governance framework for the treatment of big data will be required or whether ad-hoc or bespoke national or regional agreements can work. In a world where big data are being used more extensively, the multinational enterprises generating many of these massive global datasets will effectively be setting many future data standards. What will this mean for the global statistical system and the SDG monitoring framework? What will it mean for the United Nations *Fundamental Principles of Official Statistics* and *Principles Governing International Statistical Activities*? These massive new globalized data also challenge the concept of sovereign data and the justification for national or local compilation, raising a host of legal, security and organizational questions.

At individual NSO, NSS or IO level, there are also governance issues to be considered. The issues identified here are not exhaustive, but give a flavor of the issues that a NSO or IO may need to be consider:

[33] The issue of modernising governance frameworks in the context of big data is specifically addressed in the *2017 Bogota Declaration on Big Data for Official Statistics* (United Nations Statistics Division, 2017b).

7.1 Ethics

Many big data are the exhaust from technology. Deriving SDG Indicators from big data will involve repurposing those data. The possibilities are exciting and may offer incredible opportunities. In the rush to compile new indicators, it may be easy to forget where those data came from. Thus it may be sensible to establish an ethics committee to consider whether the compilation of new indicators justify the potential 'intrusion' to citizens privacy. A board, not immediately involved in the compilation process, may be better able to weigh-up the pros and cons of a big data project and take a more balanced view on whether 'no harm' will be done. The guidance note *Data Privacy, Ethics and Protection. A Guidance Note on Big Data for Achievement of the 2030 Agenda* published by UN Global Pulse in 2018 is a useful reference (United Nations Development Group, 2018). A NSO may wish to consider also, that in using a particular big data set, it may be inadvertently taking an ideological or philosophical stance on a range of emerging debates, including for example, the ownership of data.

7.2 Legal

There will be many legal issues to be unpicked in the years to come with regard to big data. For example, can a NSO or IO access data, such as, credit card expenditure information or mobile phone location data without breaching data protection, statistical or other legislation? The ambiguous sovereignty of globalised datasets will no doubt raise very particular legal problems in the years to come, both for statistical offices, data protection commissioners and governments. It will probably be necessary (or, at the very least sensible) to establish a board of specialist legal experts who can adjudicate on these complex issues and provide comprehensive legal opinion to the management board of the NSO or IO. The correspondence between statistical and data protection legislation will be of paramount importance in the coming years.

7.3 Oversight and confidentiality

There will most likely be a growing need for a committee that deals specifically with the confidentiality and oversight of access to data held by a NSO or IO. Storing big data will present governance challenges. Who has access to those data and why? Who decides who should have access and using what criterion? How is confidentiality of published data being safeguarded? Answering these questions will require a mixture of statistical methodology and broader governance expertise. This board might also play a useful role, in coordination with ethics committee in deciding whether certain data sets should be linked, and if they are, what are the likely implications for protecting confidentiality?

7.4 IT and cyber-security

Storing large volumes of data, and providing sufficient processing power and memory, will present technical challenges too. Obviously sufficient space will be required. But new cyber-security protocols will also be required. 'Any data collected will invariably leak' - so warns Goodman (2015: 153)[34]. What does globalized data mean for storage location – does it make sense that NSOs, NSSs or IOs continue with the old paradigm where data are stored, locally, in-house? If data are stored locally, will the data be quarantined and stored offline (so that it cannot be hacked or corrupted). If not, will the NSO be required to introduce some types of randomized identifiers to suppress identities? But does storing global data and re-processing the same data many times over in different locations make sense? Would it be more efficient to store the data at source, or

[34] It appears he may be correct. In March, 2018 a UK company, Cambridge Analytica and Facebook were embroiled in scandal, arising from harvesting data from what was initially reported as 50 million Facebook account holders without permission (see Kleinman, 2018) but now appears to be 87 Million account holders (Kang and Frenkel, 2018). The same month the sports equipment manufacturer Under Armour, reported a data breach effecting 150 million customers (RTE News, 2018). In May, Amazon Echo has had to deal with breaches in privacy as it emerged that voice activated recorded a private conversation without the knowledge of those involved and send a recording of the conversation to a contact (Bond, 2018).

in some central location (in the cloud?). How then will the data be integrated with other data sources stored in different locations? The movement and transfer of data will require very secure pipeline systems and sophisticated encryption. This is a complex topic and NSOs and IOs must proceed cautiously. Purely technical or financial driven solutions may yield poor decisions - storage and movement solutions must take reputation and cultural norms in to consideration.

7.5 Quality assurance

Assessing the quality of big data is not the same as assessing the quality of traditional datasets. Firstly, quality must be defined from the perspective of big data and clear criterion for how these can be measured must be developed. The United Nations Economic Commission for Europe (2016) note that using big data may mean accepting 'different notions of quality.' Owing to new quality issues, for example, disorganized data management, more time and effort may be required to organize and properly manage data. Gao et al. (2016) identify a number of quality parameters unique to cleaning and organizing big data. They are: determining quality assurance; dealing with data management and data organization; and the particular challenges of data scalability; and transformation and conversion. Using big data may require an extended quality framework for official statistics and SDG compilation. Such a framework might put greater emphasis on risk management than is currently used.

7.6 Professional development and training

Using big data will require a blend of different skills to that of the traditional statistician, with more emphasis on data mining and analytics. Given the demand for mathematically skilled graduates today, it will be necessary to retrain existing statisticians. This should be an on-going process in any event for professional statisticians. Nevertheless, big data may be the catalyst for some NSOs or IOs to consider establishing formal training or a Continuous Professional Development (CPD) programme. It may also provide an impetus to consider new partnerships and collaborations in order to bring in new skills.

7.7 Strategic partnerships

Using big data presents a range of technical challenges that may require new strategic partnerships. The decision for NSOs and IOs is whether it makes sense to try and develop all of the skills in-house or whether it will be better to partner with other entities that already have the required skills. These will be critical decisions, both in terms of costs and efficiencies, but also for legal and reputational reasons. In making these decisions, NSOs and IOs must ensure they do not compromise their adherence to the *Fundamental Principles of Official Statistics*, the *Principles Governing International Statistical Activities* or statistical legislation.

7.8 Communications and dissemination

Any NSO planning to use big data in the day-to-day compilation process should prepare carefully a communications strategy. How will repurposing be explained and communicated to the public? Will the NSO publish an inventory of administrative and big data being accessed, stored and used by the NSO? What is the plan, when and if, some scandal arises that embroils the NSO in a negative media story? NSOs and IOs must also carefully consider how to make new statistics available - in particular how to use technology to make the experience more interactive and user friendly for users.

7.9 Clear lines of responsibility

NSOs are typically headed by a national statistician who understands data issues and is responsible for the data governance of that office. But many of the institutions that make up a NSS are not statistical offices. These institutions may have a statistical unit, but may rely on the NSO for key statistical and methodological support. Nevertheless these institutions, such as tax authorities, cental banks or other government departments or agencies manage very large volumes of administrative and statistical data and thus may face some of the same challenges facing a NSO. These institutions are perhaps also investigating the use of big data. For these reasons, institutions may need to consider appointing a chief statistician or chief data officer to properly manage data governance and liaise with the NSO. This role is distinct from a chief information officer, who despite the title, is typically an IT specialist rather than a data specialist.

8. Conclusion

he purpose of the SDG global monitoring framework is to provide high quality, impartial and timely information that allow governments and their citizens to benchmark global progress towards implementation of the 2030 Agenda. It is not clear, as yet, whether big data will contribute anything special to the SDG global monitoring framework. It seems likely that Tamm was correct when he surmised that big data are just more data: the next phase in the evolution of data rather than a revolution. That said there are some unique aspects to big data. Perhaps the most unusual is the source – many new big data are created by and taken from people who are not necessarily aware that they have generated any data or that their data are being re-used. This raises some important ethical questions regarding the ownership of the data. It is likely that in the future, the argument that signing a 'terms of service' agreement means a citizen has signed over their data ownership rights will be tested in court. What that will mean for the compilation of SDG indicators is unclear at this juncture.

Big data, if they can be harnessed properly, would appear to offer some tantalizing opportunities - not least improved timeliness and the chance to better align SDG indicators with policy needs. Perhaps in some cases they can improve accuracy. The possibilities of matching different digital data sets may also allow us to dramatically improve our understanding of complex, cross-cutting issues, such as, gender inequality[35] or disability[36]. Advances, such as, the Internet of Things[37] and biometrics will all surely present opportunities to compile new and useful statistics. As yet, the implications of this 'big (data) bang' for statistics in general and the SDGs in particular is not immediately clear, but one can envisage a whole host of new ways to measure and understand the human condition and the progress of development. In relative terms, big data are still new. At the turn of the century, Scott Cook, the CEO of Intuit mused 'we're still in the first minutes of the first day of the Internet revolution' (Levington, 2000). Almost two decades later we are probably only in the first hours. Many norms and standards are yet to evolve. But it does not take a huge leap of imagination to foresee that in the not too distant future, the misuse of big data will be at the heart of a serious human rights abuse scandal. Official statistics must take the ethical dimension seriously. In trying to quantify human rights abuses, the United Nations must ensure they do not unwittingly create a new one. Just because something can be measured doesn't mean it should be. In assessing whether and how to use big data, IOs and NSOs must carefully consider the human rights of citizens in this digital age.

These developments will bring a myriad of new challenges too, not least the growth of unreliable information. It is already clear that big data will not be a panacea for statistical agencies confronting demands for more, better, and faster data with fewer resources (Landefeld, 2014). This may not be universally understood and so managing expectations will be an ongoing challenge for official statisticians. The challenges of how best to determine the quality and veracity of big data from a statistical perspective remain. The growing centralization or monopolization of the internet, the threat to net neutrality, and the growing volumes of 'bot' traffic are just some of the issues that may compromise the quality and impartiality of any resultant statistics.

[35] Goal 5

[36] Targets 1.3, 4.5, 4.a, 8.5, 10.2, 11.2, 11.7 and 16.7

[37] In 2006 there were some 2 billion 'smart devices' connected to each other. By 2020 it is projected that this 'internet of things' will compromise of somewhere between 30 and 50 billion devices (Nordrum, 2016). Goodman (2015) notes the result will be 2.5 sextillion potential networked object-to-object interactions.

Despite the abundance of information available today, the need for the SDG global monitoring framework has arguably never been greater. There are concerns that many social media channels are polarising social exchange and promoting 'echo chambers' and cyber-cascading. As David Eggers, in his wonderful book *The Circle* remarked, social media has 'elevated gossip, hearsay and conjecture to the level of valid, mainstream communication' (Eggers, 2013: 132). In a 'post-truth age', in a world awash with 'fake news' and 'alternative facts', where trust in information has been diminished, the SDG indicators must stand as a repository of robust and credible data. Official statisticians must ensure they can filter the wheat from the chaff. NSOs and IOs should act as gatekeepers, providing what Weinberger (2014) terms 'stopping points' i.e. information that can be taken as fact and used as the final resort in the case of disagreement.

There is a new gold rush underway - a data rush. Talk of big data and data revolution are everywhere. In that rush, NSOs and IOs are feeling the pressure to be seen to utilize big data. But as outlined above, it will be a bumpy road with many challenges along the way. It is of course often easier to see problems than opportunities, so NSO's and IOs must carefully weigh-up the likely costs and benefits of using big data, both now and in the future. In making that decision, they must not lose sight of their mission and their mandates.

References

Adams, D. (1979). 'The Hitchhikers Guide to the Galaxy.' Pan, London.

Aslam, S. (2015). 'Snapchat by the Numbers: Stats, Demographics and Fun Facts.' *Omnicore*, October 7, 2015. Retrieved from: http://www.omnicoreagency.com/snapchatstatistics/ [26 September, 2016].

Berners-Lee, T. (2014). 'Tim Berners-Lee on the Web at 25: the past, present and future'. *Wired*, 23 August, 2014. Retrieved from: http://www.wired.co.uk/article/tim-berners-lee [19 March, 2018].

Blackwell, J. (1985). 'Information for policy.' *National and Economic Social Council,* Report no. 78*.* Dublin: NESC. Retrieved from: http://files.nesc.ie/nesc_reports/en/NESC_78_1985.pdf [18 January, 2018].

Bond, S. (2018). 'Amazon Echo nightmare: private conversation sent to contact.' *Financial Times*, May 24, 2018. Retrieved from: https://www.ft.com/content/aebcc58c-5fa3-11e8-9334-2218e7146b04 [24 May, 2018].

Borgman, C.L. (2015). 'Big Data, Little Data, No Data - Scholarship in the Networked World.' Cambridge, MA: MIT Press.

Boyd, J and K. Crawford (2012). 'Critical Questions for Big Data - Provocations for a cultural, technological, and scholarly phenomenon.' *Information, Communication & Society*, Vol 15, No.5, pp. 662 - 679, DOI:10.1080/1369118X.2012.678878.

Brackstone, G. J. (1987). 'Statistical Issues of Administrative Data: Issues and Challenges'. Survey Methodology, Vol. 13, No. 1, pp. 29 – 43.

Buytendijk, F. (2014). 'Hype Cycle for Big Data, 2014.' *Gartner*. Retrieved from: https://www.gartner.com/doc/2814517/hype-cycle-big-data- [11 August, 2015].

Committee of the Chief Statisticians of the United Nations System (2018). 'UN Statistical Quality Assurance Framework'. Retrieved from: https://unstats.un.org/unsd/unsystem/documents/UNSQAF-2018.pdf [14 May, 2018].

Committee for the Coordination of Statistical Activities (2014). 'Principles Governing International Statistical Activities.' Retrieved from: https://unstats.un.org/unsd/accsub-public/principles_stat_activities.htm [21 February, 2018].

Cervera J.L., P. Votta, D. Fazio, M. Scannapieco, R. Brennenraedts and T, van der Vorst (2014). 'Big Data in Official Statistics.' *Eurostat ESS Big Data Event*, Rome2014 – Technical Event Report. Retrieved from: https://ec.europa.eu/eurostat/cros/system/files/Big%20Data%20Event%202014%20-%20Technical%20Final%20Report%20-finalV01_0.pdf [18 January, 2018].

Choi, H. and H. Varian (2011). 'Predicting the present with Google Trends.' Retrieved from: http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf [17 January, 2018].

Dass, P.J.H., M.J. Puts, B. Buelens and P.A.M. van den Hurk (2015). 'Big Data as a Source for Official Statistics.' *Journal of Official Statistics*, Vol. 31, No. 2, pp. 249 - 262.

Daas P.J.H, M.J. Puts, B. Buelens and P.A.M. van den Hurk (2015). 'Big Data as a Source for Official Statistics'. *Journal of Official Statistics*, Vol.31, No.2, pp.249 - 262.

Donkin, C. (2017). 'M-Pesa continues to dominate Kenyan market.' *Mobile World Live*, January 25, 2017. Retrieved from: https://www.mobileworldlive.com/money/analysis-money/m-pesa-continues-to-dominate-kenyan-market/ [20 February, 2018]

Eggers, D. (2013). 'The Circle.' Penguin Books, London

European Commission (2011). 'Attitudes on Data Protection and Electronic Identity in the European Union.' *Special Eurobarometer* No. 359, Wave 74.3 - TNS Opinion and Social. Published June 2011. Retrieved from: http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_359_en.pdf [16 January, 2018].

European Commission (2013). 'Scheveningen Memorandum on "Big Data and Official Statistics''.' Adopted by the European Statistical System Committee on 27 September 2013. Retrieved from: https://ec.europa.eu/eurostat/cros/content/scheveningen-memorandum_en [25 January, 2018].

European Commission (2014). 'Big Data.' *Digital Single Market Policies*. Retrieved from: https://ec.europa.eu/digital-single-market/en/policies/big-data [31 January, 2018].

European Parliament (2016). 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).' Retrieved from: http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf [16 Jan, 2018].

Eurostat (2014). 'Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics - Consolidated Report'. Eurostat Contract No 30501.2012.001- 2012.452, 30 June 2014. Retrieved from: http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf [3 April, 2018].

Federal Communications Commission (2017). 'Restoring Internet Freedom.' Retrieved from: https://www.fcc.gov/restoring-internet-freedom [24 January, 2018].

Fry, S. (2017). 'The Way Ahead'. Lecture delivered on the 28th May 2017, Hay Festival, Hay-on-Wye. Retrieved from: http://www.stephenfry.com/2017/05/the-way-ahead/ [19 March, 2018].

Gelsinger, P. in Whatsthebig data? (2012). 'Big Data quotes of the week'. Retrieved from: https://whatsthebigdata.com/2012/06/29/big-data-quotes-of-the-week-11/ [19 March, 2018].
Gibson, W. in The Economist (2001). 'Broadband Blues - Why has broadband Internet access taken off in some countries but not in others?' The Economist, June 21, 2001. Retrieved from: https://www.economist.com/node/666610 [19 March, 2018].

Global Partnership for Sustainable Development Data (2016). 'The Data Ecosystem and the Global Partnership.' Retrieved from: http://gpsdd.squarespace.com/who-we-are/ [19 January, 2018].

Goa, J., C. Xie and C. Tao (2016). 'Big Data Validation and Quality Assurance - Issuses, Challenges, and Needs.' 2016 *IEEE Symposium on Service-Oriented System Engineering* (SOSE), Issue Date: March 29 2016-April 2 2016. Retrieved from: http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=7473058 [20 February, 2018].

Goodbody, W (2018). 'Waterford researchers develop new method to store data in DNA.' *RTE News*, January 20, 2018. Retrieved from: https://www.rte.ie/news/ireland/2018/0219/941956-dna-data/ [20 January, 2018].

Goodman, M. (2015). 'Future Crimes - Inside the Digital Underground and the Battle for Our Connected World.' Anchor Books, New York.

Guerreiro, V., Walzer, M. and Lamboray, C. (2018). 'The use of Supermarket Scanner data in the Luxembourg Consumer Price Index', *Economie et Statistiques* - Working papers du STATEC, No. 97, Retrieved from: http://www.statistiques.public.lu/catalogue-publications/economie-statistiques/2018/97-2018.pdf [14 May, 2018].

Hammer, C.L., D.C. Kostroch, G. Quiros and STA Internal Group (2017). 'Big Data: Potential, Challenges, and Statistical Implications.' *IMF Staff Discussion Note*, SDN/17/06, September 2017. Retrieved from: http://www.imf.org/en/Publications/SPROLLs/Staff-Discussion-Notes [11 January, 2018].

Hand, D. J. (2015). 'Official Statistics in the New Data Ecosystem.' presented at the *New Techniques and Technologies in Statistics conference*, Brussels, March 10-12, 2015. Retrieved from: https://ec.europa.eu/eurostat/cros/system/files/Presentation%20S20AP2%20%20Hand%20-%20Slides%20NTTS%202015.pdf [23 January, 2018].

Harkness, T. (2017). 'Big Data: Does size matter?' Bloomsbury Sigma, London, UK.

Hayak, F. A. (1944). 'The Road to Serfdom.' The University of Chicago Press, Chicago, MA.
Hilbert, M. and P. Lopez (2012). 'How to Measure the World's Technological Capacity to Store, Communicate and Compute Information.' *International Journal of Communication*, Vol 6, pp. 956–979.

IBM (2017). '10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations.' *IBM Marketing Cloud*. Retrieved from: https://public.dhe.ibm.com/common/ssi/ecm/wr/en/wrl12345usen/watson-customer-engagement-watson-marketing-wr-other-papers-and-reports-wrl12345usen-20170719.pdf [25 January, 2018].

Ismail, N. (2016). 'Big Data in the developing world.' Information Age, 8 September, 2016. Retrieved from: http://www.information-age.com/big-data-developing-world-123461996/ [19 January, 2018]

International Telecommunications Union (2017). 'ITU Key 2005 - 2017 ICT data.' Retrieved from: https://idp.nz/Global-Rankings/ITU-Key-ICT-Indicators/6mef-ytg6 [12 Jan, 2018].

Jerven, M. (2015). 'Africa - Why economists get it wrong.' Zed Books, London.

Kang, C. and Frenkel, S. (2018). 'Facebook Says Cambridge Analytica Harvested Data of Up to 87 Million Users'. *The New York Times*, April 4, 2018. Retreived from: https://www.nytimes.com/2018/04/04/technology/mark-zuckerberg-testify-congress.html [4 April, 2018].

Keating, J. (2009). ' Why did One Laptop Per Child fail?' Passport - Foreign Policy. September 9, 2009. Retrieved from: http://foreignpolicy.com/2009/09/09/why-did-one-laptop-per-child-fail/ [20 March, 2018].

Kirkpatrick, M. (2010). 'Facebook's Zuckerberg Says the Age of Privacy is Over.' *Readwrite.Com*, 9 January 2010. Retrieved from: https://readwrite.com/2010/01/09/facebooks_zuckerberg_says_the_age_of_privacy_is_ov/ [21 February, 2018].

Kitchin, R. (2015). 'The opportunities, challenges and risks of big data for official statistics.' *Statistical Journal of the International Association of Official Statistics*, Vol. 31, No. 3, pp. 471-481.

Kleinman, Z. (2018). 'Cambridge Analytica: The story so far.' BBC News, Technology. 21 March, 2018. Retrieved from: http://www.bbc.com/news/technology-43465968 [21 March, 2018].

Korte, T. (2014). 'How Data and Analytics Can Help the Developing World.' *Huffington Post - The Blog*. 21 September, 2014. Retrieved from: https://www.huffingtonpost.com/travis-korte/how-data-and-analytics-ca_b_5609411.html [19 January, 2018].

Krikorian, R. (2013). 'New Tweets per Second Record, and How!' *Engineering Blog*. Retrieved from: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how [26 September, 2016].

Kulp, P. (2017). 'Facebook quietly admits to as many as 270 million fake or clone accounts.' *Mashable*, November 3, 2017. Retrieved from: https://mashable.com/2017/11/02/facebook-phony-accounts-admission/#UyvC2aOAmPqo [20 February, 2018].

Landefeld, S. (2014). 'Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues.' Discussion paper presented at the United Nations Global Working Group on Big Data for Official Statistics, Beijing, China 31 October, 2014. Retrieved from: https://unstats.un.org/unsd/trade/events/2014/beijing/Steve%20Landefeld%20-%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf [January 18, 2018].

Laney, D. (2001). '3D Data Management: Controlling data volume, velocity and variety.' Meta Group, File 949, 6 February, 2001. Retrieved from: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf [11 January, 2018].

Lebada, A. M. (2016). 'Member states, statisticians address SDG monitoring requirements.' Retrieved from: http://sdg.iisd.org/news/member-states-statisticians-address-sdg-monitoring-requirements/ [19 March 2018].

Letouzé, E. and J. Jütting (2015). 'Official Statistics, Big Data and Human Development.' Data Pop Alliance, *White Paper Series*, Retrieved from: https://www.paris21.org/sites/default/files/WPS_OfficialStatistics_June2015.pdf [16 Jan, 2018].

Levington, S. (2000). 'Internet Entrepreneurs Are Upbeat Despite Market's Rough Ride.' *The New York Times*, May 24, 2000. Retrieved from: http://www.nytimes.com/2000/05/24/business/worldbusiness/internet-entrepreneurs-are-upbeat-despite-markets.html [20 February, 2018].

Long, J. and W. Brindley (2013). 'The role of big data and analytics in the developing world: Insights into the role of technology in addressing development challenges.' Accenture Development Partnerships. Retrieved from:https://www.accenture.com/us-en/~/media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Strategy_5/Accenture-ADP-Role-Big-Data-And-Analytics-Developing-World.pdf [19 January, 2018]

MacFeely, S. and J. Dunne (2014). 'Joining up public service information: The rationale for a national data infrastructure.' *Administration,* Vol.61, No.4, pp. 93–107,

MacFeely, S. (2016). 'The Continuing Evolution of Official Statistics: Some Challenges and Opportunities.' *Journal of Official Statistics*, Vol. 32, No. 4, 2016, pp. 789–810.

MacFeely, S. (2017). 'Measuring the Sustainable Development Goals: What does it mean for Ireland?' *Administration*, Vol.65, No.4, pp. 41 - 71.

MacFeely, S. and N. Barnat (2017). 'Statistical capacity building for sustainable development: Developing the fundamental pillars necessary for modern national statistical systems.' *Statistical Journal of the International Association of Official Statistics*, Vol.33, No. 4, pp. 895 - 909.

Mahajan, S. (2018). 'New Data Sources and Recent Innovations in Producing National and International Accounts Data Science, Big Data and Economic Statistics'. Economic Commission for Europe - Conference for European Conference: Group of Experts on National Accounts, Geneva, Switzerland, 25 May, 2018. Retrieved from: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.20/2018/mtg1/Item_8_UK_Data_Science_Campus_EN.pdf [14 May, 2018].

Mayer-Schonberger, V. and K. Cukier (2013). 'Big Data: A Revolution That Will Transform How We Live, Work and Think.' London: John Murray.

McAfee, J. (2015). Untitled posting on Facebook. 8 May, 2015. Facebook. Retrieved from: https://www.facebook.com/officialmcafee/posts/464114187078100:0 [16 May, 2018].

Meeker, M. (2017). 'Internet Trends 2017.' Presented at the *Code Conference*, Rancho Palos Verdes, California, May 31, 2017. Retrieved from: http://www.kpcb.com/internet-trends [12 Jan, 2018].

Mutuku, L. (2016) in Serra, C. (2016). 'The big data challenge for developing countries.' *The world academy of sciences*, 2 September, 2016. Retrieved from: https://twas.org/article/big-data-challenge-developing-countries [19 January, 2018].

Nilson (2018). 'Global Cards - 2015: Special Report.' *The Nilson report*, Jan 12, 2018. Retrieved from: https://www.nilsonreport.com/publication_special_feature_article.php [12 January, 2018].

Nordrum, A. (2016). 'Popular Internet of Things Forecast of 50 Billion Devices by 2020 Is Outdated.' *IEEE Spectrum*, August 18, 2016. Retrieved from: https://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-things-forecast-of-50-billion-devices-by-2020-is-outdated [21 February, 2018].

Noyes, K. (2015). 'Scott McNealy on privacy: You still don't have any.' *PC World*, IDG News Service, June 25, 2015. Retrieved from: https://www.pcworld.com/article/2941052/scott-mcnealy-on-privacy-you-still-dont-have-any.html [29 January, 2018].

Nyborg Vov, K. (2018). 'Using scanner data for sports equipment', Paper written for the joint UNECE/ILOs Meeting of the Group of Experts on Consumer Price Indices, 7-9 May 2018, Geneva, Switzerland. Retrieved from: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Norway_-_session_1.pdf [14 May, 2018].

Ohm, P. (2010). 'Broken promises of privacy: Responding to the surprising failure of anonymization.' UCLA Law Review, 2010, Vol. 57, pp.1701-1777. Retrieved from: http://www.uclalawreview.org/pdf/57-6-3.pdf [19 March, 2018].

Payton, T. and T. Claypoole (2015). 'Privacy in the Age of Big Data - Recognising the Threats Defending Your Rights and Protecting Your Family.' Lanham, MD: Rowman & Littlefield.

Pearson, E. (2013). 'Growing Up Digital.' Presentation to the *OSS Statistics System Seminar Big Data and Statistics New Zealand*: A seminar for Statistics NZ staff, Wellington, 24 May 2013. Retrieved from: https://www.youtube.com/watch?v=lRgEMSqcKXA [19 December, 2017].

Raley, R. (2013). 'Dataveillance and countervailance' in Gitelman, I. (Ed) '"Raw Data" is an Oxymoron.' MIT Press, Cambridge.

Reich, R. (2015). 'Saving Capitalism: For the Many, Not the Few.' London: Icon Books Ltd.

RTE News (2018). 'Under Armour says 150m affected in data breach'. Updated 30 March, 2018. Retrived from: https://www.rte.ie/news/2018/0330/951116-under-armour/ [4 April, 2018].
Rudder, C. (2014). 'Dataclysm: What our online lives tell us about our offline selves.' 4th Estate, London.

Runde, D. (2017). 'The Data Revolution in Developing Countries Has a Long Way to Go.' *Forbes*, February 25, 2017. Retrieved from: https://www.forbes.com/sites/danielrunde/2017/02/25/the-data-revolution-in-developing-countries-has-a-long-way-to-go/2/#3a48f53e482f [19 January, 2018].

Statistica (2018a). 'Average daily number of trades on London Stock Exchange (UK order book) in the United Kingdom from January 2015 to February 2018'. *Statistica - The Statistics Portal.* Retrieved from: https://www.statista.com/statistics/325326/uk-lse-average-daily-trades/ [15 May, 2018].

Statistica (2018b). ' Number of user reviews and opinions on TripAdvisor worldwide from 2014 to 2017 (in millions). *Statistica - The Statistics Portal.* Retrieved from: https://www.statista.com/statistics/684862/tripadvisor-number-of-reviews/ [16 May, 2018].

Stephens-Davidowitz, S. (2017). 'Everybody lies - What the internet can tell us about who we really are.' Bloomsbury, London, UK.

Struijs, P., B. Braaksma and P.J.H. Daas (2014). 'Official statistics and Big Data.' *Big Data & Society*, April–June 2014: 1–6, DOI: 10.1177/2053951714538417. Retrieved from: http://journals.sagepub.com/doi/pdf/10.1177/2053951714538417 [21 January, 2018].

Tam, S. and F. Clarke (2015). 'Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics.' *International Statistical Review*, doi: 10.1111/insr.12105. Retrieved from: https://www.researchgate.net/publication/280972848_Big_Data_Official_Statistics_and_Some_Initiatives_by_the_Australian_Bureau_of_Statistics [11 January, 2018].

Taplin, J. (2017), 'Move Fast and Break things - How Facebook, Google and Amazon cornered culture and undermined democracy.' Little, Brown and Company, New York.

Thamm, A. (2017). 'Big Data is dead.' LinkedIn, November 23, 2017. Retrieved from: https://www.linkedin.com/pulse/big-data-dead-just-regardless-quantity-structure-speed-thamm/ [16 January, 2018].

United Nations (2003). 'Handbook of Statistical Organization – 3rd Edition: The Operation and Organization of a Statistical Agency.' Department of Economic and Social Affairs Statistics Division Studies in Methods Series F No. 88. United Nations, New York, 2003. Retrieved from: https://www.paris21.org/sites/default/files/654.pdf [25 January, 2018].

United Nations (2014). 'Resolution adopted by the General Assembly on 29 January 2014 - Fundamental Principles of Official Statistics.' *General Assembly*, A/RES/68/261. Retrieved from: http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf [26 September, 2016].

United Nations (2015). 'Transforming our world: the 2030 Agenda for Sustainable Development.' Resolution adopted by the General Assembly on 25 September 2015: 70/1. Retrieved from: http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E [7 February 2017].

United Nations (2017). 'Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development'. Resolution 71/313 of the UN General Assembly *(A/71/L.75)*. 10 July, 2017. Retrieved from: http://ggim.un.org/meetings/2017-4th_Mtg_IAEG-SDG-NY/documents/A_RES_71_313.pdf [20 March, 2018].

United Nations Conference for Trade and Development (2015). 'Mapping of international Internet public policy issues.' E/CN.16/2015/CRP.2, *Commission on Science and Technology for Development*, Eighteenth session, Geneva, 4-8 May 2015. Retrieved from: http://unctad.org/meetings/en/SessionalDocuments/ecn162015crp2_en.pdf [24 January, 2018].

United Nations Conference for Trade and Development (2016). 'Development and Globalization: Facts and Figures 2016.' Retrieved from: http://stats.unctad.org/Dgff2016/ [19 January, 2018].

United Nations Conference for Trade and Development (2018). 'Data privacy: new global survey reveals growing internet anxiety.' Retrieved from: http://unctad.org/en/pages/newsdetails.aspx?OriginalVersionID=1719 [19 April, 2018].

United Nations Development Group (2018). 'Data Privacy, Ethics and Protection. A Guidance Note on Big Data for Achievement of the 2030 Agenda.' Retrieved from: https://www.slideshare.net/unglobalpulse/data-privacy-ethics-and-protection-a-guidance-note-on-big-data-for-achievement-of-the-2030-agenda [15 May, 2018].

United Nations Economic Commission for Europe (2000). 'Terminology on Statistical Metadata.' Conference of European Statisticians Statistical Standards and Studies, No.53. Retrieved from: http://ec.europa.eu/eurostat/ramon/coded_files/UNECE_TERMINOLOGY_STAT_METADATA_2000_EN.pdf [26 February, 2018].

United Nations Economic Commission for Europe (2011). 'Using Administrative and Secondary Sources for Official Statistics - A Handbook of Principles and Practices'. Retrieved from: https://unstats.un.org/unsd/EconStatKB/KnowledgebaseArticle10349.aspx [26 February, 2018].

United Nations Economic Commission for Europe (2013). 'Classification of Types of Big Data'. Created by the Task Team on Big Data, in June 2013. Retreived from: https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data [5 April, 2018].

United Nations Economic Commission for Europe (2013b). 'The Common Metadata Framework'. UNECE Virtual Standards Helpdesk. Retrieved from: https://statswiki.unece.org/display/VSH/The+Common+Metadata+Framework [14 May, 2018].

United Nations Economic Commission for Europe (2016). 'Outcomes of the UNECE Project on Using Big Data for Official Statistics.' Retrieved from: https://statswiki.unece.org/display/bigdata/Big+Data+in+Official+Statistics [15 February, 2018].

United Nations Economic Commission for Europe (2018). 'Guidance on common elements of statistical legislation.' *Conference of European Statisticians*, Sixty-sixth plenary session, Geneva, 18–20 June 2018. Retrieved from: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2018/CES_6_Common_elements_of_statistical_legislation__Guidance__for_consultation_for_upload.pdf [17 May, 2018].

United Nations Economic and Social Council (2016). 'Report of the Inter-Agency and Expert Group on Sustainable Development Goal Indicators', 47th Session of the Statistical Commission (8-11 March). E/CN.3/2016/2/Rev.1 2016. 2016. Retrieved from: https://unstats.un.org/unsd/statcom/47th-session/documents/2016-2-IAEG-SDGs-E.pdf [19 March, 2018].

United Nations Global Pulse (2017). 'Annual Report 2017 - Harnessing Big Data for Development and Humanitarian Action.' Retrieved from: https://www.unglobalpulse.org/sites/default/files/UNGP_Annual2017_final_web.pdf [15 May, 2018].

United Nations Office of the High Commissioner on Human Rights (2018). 'A Human Rights-Based Approach to Data: Leaving no one behind in the 2030 Agenda for Sustainable Development'. Retreived from: http://www.ohchr.org/Documents/Issues/HRIndicators/GuidanceNoteonApproachtoData.pdf [24 May, 2018].

United Nations Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development (2014). 'A World that Counts: Mobilizing the Data Revolution for Sustainable Development.' Report prepared at the request of the United Nations Secretary-General, by the Independent Expert Advisory Group on a Data Revolution for Sustainable Development. November 2014. Retrieved from: http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf [17 January, 2018].

United Nations Statistical Commission (2014). 'Big data and modernization of statistical systems; Report of the Secretary-General.' E/CN.3.2014/11 of the forty-fifth session of UNSC 4-7 March 2014. Retrieved from: https://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf [11 January, 2018].

United Nations Statistical Commission (2018). ' Guidelines on Data Flows and Global Data Reporting for Sustainable Development Goals'. Background document prepared by the Inter-Agency and Expert Group on Sustainable Development Goal Indicators. 49th Session of the Statistical Commission, 6 - 9 March, 2018. Retrieved from: https://unstats.un.org/unsd/statcom/49th-session/documents/BG-Item-3a-IAEG-SDGs-DataFlowsGuidelines-E.pdf [20 March, 2018].

United Nations Statistics Division (2008). Millennium Development Goals Indicators – The official United Nations site for the MDG indicators. Effective 15 January, 2008. Retrieved from: http://mdgs.un.org/unsd/mdg/Host.aspx?Content=Indicators/OfficialList.htm [19 March, 2018].

United Nations Statistics Division (2017). 'Tier Classification for Global SDG Indicators - version 15 December, 2017.' Retrieved from: https://unstats.un.org/sdgs/files/Tier%20Classification%20of%20SDG%20Indicators_15%20Dec%202017_web%20final.pdf [19 March, 2018].

United Nations Statistics Division (2017b). 'Bogota Declaration on Big Data for Official Statistics'. Agreed at the 4th Global Conference on Big Data for Official Statistics, Bogota Columbia, 8-10 November 2017. Retrieved from: https://unstats.un.org/unsd/bigdata/conferences/2017/Bogota%20declaration%20-%20Final%20version.pdf [14 May, 2018].

Van Loon, K. and Roels, D. (2018). 'Integrating big data in the Belgian CPI', Presented to the Meeting of the UNECE Group of Experts on Consumer Price Indices, Geneva, Switzerland: 7-9 May 2018. Retrieved from: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Belgium.pdf [14 May, 2018].

Waterford Technologies (2017). 'Big Data Statistics & Facts for 2017.' Posted on February 22, 2017: Retrieved from: https://www.waterfordtechnologies.com/big-data-interesting-facts/ [3 Jan, 2018].

Weigand, A. (2009). 'The Social Data Revolution(s).' *Harvard Business Review*, May 20, 2009. Retrieved from: https://hbr.org/2009/05/the-social-data-revolution.html [24 April, 2017].

Weinberger D. (2014). 'Too Big to Know.' Basic Books, New York.