Title	The Big (data) Bang: Opportunities and challenges for compiling SDG indicators
Authors	MacFeely, Stephen
Publication date	2019-01-28
Original Citation	MacFeely, S. (2019) 'The Big (data) Bang: Opportunities and Challenges for Compiling SDG Indicators', Global Policy, 10(S1), pp. 121-133. doi: 10.1111/1758-5899.12595
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://onlinelibrary.wiley.com/doi/full/10.1111/1758-5899.12595 - 10.1111/1758-5899.12595
Rights	© 2019 University of Durham and John Wiley & Sons, Ltd. This is the peer reviewed version of the following article: MacFeely, S. (2019), The Big (data) Bang: Opportunities and Challenges for Compiling SDG Indicators. Glob Policy, 10: 121-133, which has been published in final form at https://doi.org/10.1111/1758-5899.12595. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions
Download date	2024-04-19 20:51:33
Item downloaded from	https://hdl.handle.net/10468/10636

The Big (data) Bang: opportunities and challenges for compiling SDG indicators

Short title: Big data and the SDGs

Steve MacFeely
United Nations Conference on Trade and Development, Switzerland
Centre for Policy Studies, University College Cork, Ireland

Abstract

Official statisticians around the world are faced with the herculean task of populating the Sustainable Development Goals global indicator framework. As traditional data sources appear to be insufficient, statisticians are naturally considering whether big data can contribute anything useful. While the statistical possibilities appear to be theoretically endless, in practice big data also present some enormous challenges and potential pitfalls: legal; ethical; technical; and reputational. This paper examines the opportunities and challenges presented by big data for compiling indicators to support Agenda 2030.

Keywords

Administrative data, national statistical offices, national statistical systems, global indicator framework

Policy implications

- The digital divide is creating a data divide
- Data misuse is likely to become a source of human rights abuse
- United Nations could introduce an accreditation system that would allow un-official compilers to be accredited as 'official' for the purposes of populating the SDG indicator framework.
- Removing Net Neutrality and platform concentration may compromise big data as a useable source for official statistics

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[i]. These universal goals cover all three key development pillars: economic, social and environment, as well as enablers such as institutional coherence, policy coherence and accountability. The 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) and statistical agencies of International Organisations (IOs) around the world and members of the Inter-agency and Expert Group on SDG Indicators (IAEG-SDGs), the group established by the UN Statistical Commission to develop and implement the global indicator framework (GIF) for the targets of the 2030 Agenda are faced with several questions, among them: whether that challenge can be met? And 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 (72 indicators) meaning the indicator is conceptually clear but the data are not regularly produced by countries or Tier 3 (62 indicators), meaning that no internationally established methodology or standards are yet available. Five indicators are determined as having several tiers (Inter-Agency and Expert Group on Sustainable Development Goals, 2018). In other words, as of May 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, 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 dramatic changes, the proportion of populated indicators for the SDG GIF 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. Statisticians must decide whether big data, which seem to offer rich and tantalizing opportunities to augment or supplant existing data sources or generate completely new statistics, will be useful for compiling SDGs. The jury is still out. On the one hand, some argue that big data needs to be seen as an entirely new ecosystem (Letouzé and Jütting, 2015) whereas others argue to the contrary that big data is just hype and that big data are just data (Thamm, 2017). Buytendijk (2014) argued that big data has already passed the top of the ‘Hype Cycle’ and moving towards the ‘Trough of Disillusionment.’ Beyond the hype of big data, and hype it may well be, 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.'

In simplistic terms, one can think of big data as the collective noun for all new digital data arising from our digital activities. Our day-to-day dependence on technology is leaving ‘digital footprints’ everywhere. 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 pits-falls: 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.

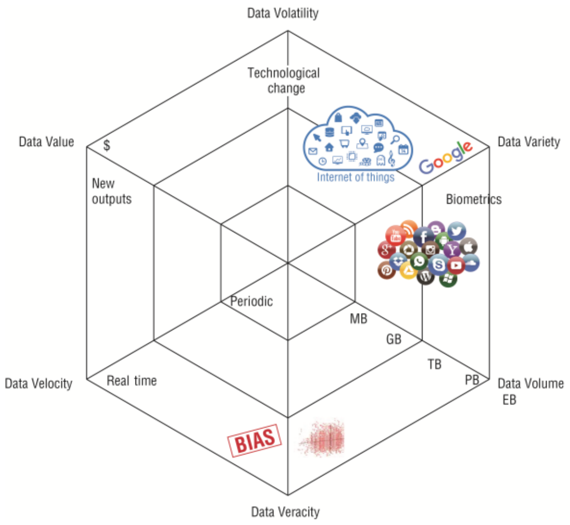
Defining Big Data

What are big data? While some, such as, Stephens-Davidowitz argue that big data is 'an inherently vague concept' (2017: p.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: p.6).

Gartner analyst Doug Laney provided what has become known as the '3Vs' definition in 2001. 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 3Vs definition.

It seemed that the 3Vs definition was generally accepted, within official statistics circles at least, with the UN Statistical Commission (2014: p.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.' 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. 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: p.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 for understanding the contribution that big data might make to compiling statistics and SDG indicators. The 5Vs definition is more balanced from an analytical perspective than the 3Vs as it flags some of the downside risks. But arguably a 6V definition that includes 'value', where value means that something useful is derived from the data, offers a superior definition, as it introduces the notion of cost-benefit i.e. the costs of investing in big data must be carefully weighed up against what they might deliver in practical terms - see Figure 1. Like volatility and veracity, value is not an intrinsic characteristic, but as above, including this dimension is nevertheless useful.

Figure 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 big data are conceptually quite different to traditional survey data. Big data are a collection of by-product data rather than data designed by statisticians for a specific purpose (Dass et al., 2015). In other words, the derivation of statistics is a secondary purpose. This difference is perhaps obvious but profoundly important.

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 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. Each transaction is leaving several footprints, from which new types of statistics can be compiled. In fact, as Stephens-Davidowitz (2017: p.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 or an info-glut. This deluge is also the result of an important behavioral change, where people now record and load content for free. Weigand (2009) described this phenomenon where people actively share or supply data directly to various social networks and product reviews as a 'social data revolution'.

Not only have 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: p.97). Now data includes text, sound and images, not just neat columns of numbers. Begging the question, in this digital age, how much data now exist. Definitional differences make this a difficult question to answer, and consequently there are various estimates. Hilbert and Lopez (2012) estimated that 300 exabytes (or slightly less than one third of a zettabyte[ii]) of data were stored in 2007. Waterford Technologies (2017) 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 now estimates we create an additional 2.5 quintillion bytes[iii] of data every day (IBM, 2017).

Despite the varying estimates, it is clear, that a massive volume of digital data now exists. But as Harkness (2017: p.17) wisely counsels, the 'proliferation of data is deceptive; we're just recording the same things in more detail'. Nor are all these data necessarily accessible or of good quality. As Borgman (2015: p.131) warns, big data must be treated with caution, noting that ‘as few as 35 percent of twitter followers may be real people, and as much as 10 percent of activity in social networks may be generated by robotic accounts.' Furthermore, Goodman (2015) states that 25% of reviews on Yelp are bogus. Facebook themselves have admitted that 3% of accounts are fake and an additional 6% are 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 not people - 'fake clicks.' The disruptive potential of these bots is so massive Goodman (2015) refers to them as WMDs - Weapons of Mass Disruption.

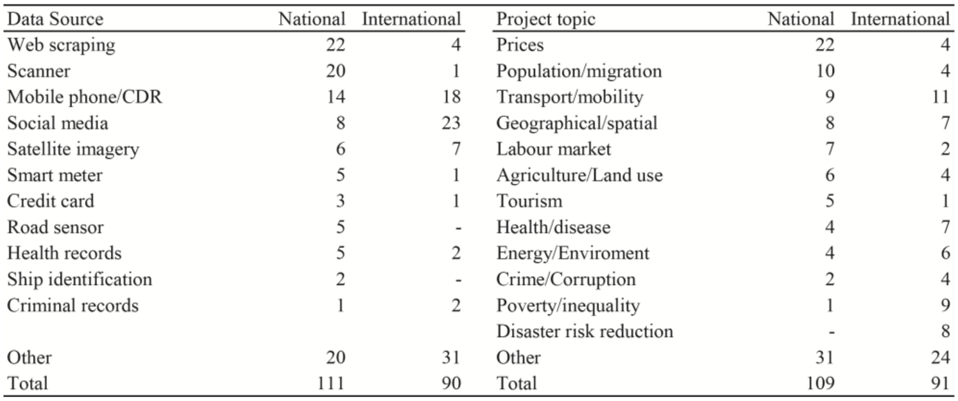
There are also issues of coverage. The International Telecommunication Union (2017) estimates that global Internet penetration is only 48 percent and global mobile broadband subscriptions 56 percent. Although global coverage is improving rapidly, it still means that in 2017 almost half of the world's population did not use the web. The digital divide (limited access and connectivity to the web and mobile phones) is creating a data divide. Anyone excluded will not have a digital footprint or at best, a rather limited one. To quote William Gibson (2003) 'The future is already here - it's just not very evenly distributed.' Even within countries, digital (and data) divides exist arising from a range of access barriers: social; gender; geographic; or economic strata, leading to important cohorts being excluded, with obvious implications for representativity (see Struijs and Daas 2014) and veracity. In the context of Agenda 2030 this is extremely important as the underlying rationale is that ‘no one gets left behind’ (United Nations, 2015) or put another way, no one gets left uncounted. The question being asked by NSOs is whether these data are representative and stable enough to be used to compile SDG indicators.

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 offers the potential to contribute to the measurement of SDGs in several ways. According to the *Big Data Project*

Inventory 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 1). NSOs and other national agencies compiling statistics are attempting to use satellite imagery, aerial imagery, mobile phone data, data scraped from websites, smart meters, road sensors, ships identification data, public 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 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 catalogue of all big data activity, but it nevertheless provides a good overall picture of the types of activities that are underway. From Table 1 it is clear that NSSs are targeting web scraping, scanner data and mobile phone - 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 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 a clear focus and have been in development for many years and typically have fewer data access problems - See Guerreiro et al. (2018) or Nyborg Vov (2018) for some recent examples.

Table 1: Big Data sources and project topics registered by National and International Organisations on the UN Big Data Project Inventory

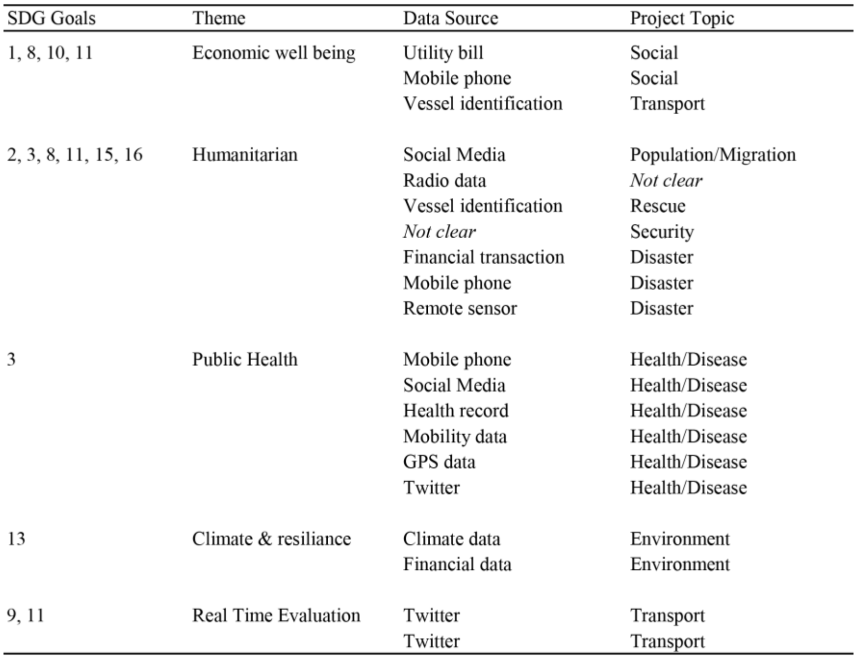


Data Source	National	International	Project topic	National	International
Web scraping	22	4	Prices	22	4
Scanner	20	1	Population/migration	10	4
Mobile phone/CDR	14	18	Transport/mobility	9	11
Social media	8	23	Geographical/spatial	8	7
Satellite imagery	6	7	Labour market	7	2
Smart meter	5	1	Agriculture/Land 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 https://unstats.un.org/bigdata/inventory/ [examined on 27 April, 2018][iv].

IOs, most particularly the World Bank, are also investigating big data - they have logged 87 projects on the *Big Data Project Inventory*. Here too, a wide variety of big data sources are being explored - mobile phone (CDR) call detail records , Wikipedia, Google Trends, scanner data, web scraping, road sensor data, satellite imagery, credit card transactions, bank machine (ATM) withdrawls data, online purchases, aerial imagery, financial transaction data, taxi global positioning system (GPS) data, freight data, medical insurance records, crime records, building certification data, OpenStreetMap, Twitter, public FaceBook data, social media, aid data, bus fleet automatic vehicle location (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 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 UN Global Pulse also listed 20 big data projects in their annual report - these projects are using similar big data sources to those listed in Table 1 and have similar stated objectives. In fact some of these projects may be the same as those registered with the *UN Big Data Project Inventory* – while care has been taken to avoid duplication, given the level of detail available it is impossible to be certain. Table 2 summarises the SDG goals towards which big data projects were focused in 2017 - goals 3, 8, 11 appear to be the most targeted, being included in at least 10 projects each. Goals 2, 15 and 16 enjoying somewhat less attention, included in 7 projects each.

Table 2: SDG Goals being assisted by Big Data as reported by UN Global Pulse in 2017



SDG Goals	Theme	Data Source	Project Topic
1, 8, 10, 11	Economic well being	Utility bill	Social
		Mobile phone	Social
		Vessel identification	Transport
2, 3, 8, 11, 15, 16	Humanitarian	Social Media	Population/Migration
		Radio data	*Not clear*
		Vessel identification	Rescue
		Not clear	Security
		Financial transaction	Disaster
		Mobile phone	Disaster
		Remote sensor	Disaster
3	Public Health	Mobile phone	Health/Disease
		Social Media	Health/Disease
		Health record	Health/Disease
		Mobility data	Health/Disease
		GPS data	Health/Disease
		Twitter	Health/Disease
13	Climate & resiliance	Climate data	Environment
		Financial data	Environment
9, 11	Real Time Evaluation	Twitter	Transport
		Twitter	Transport

Source: Derived from UN Global Pulse 2017 Annual Report

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 realizing the lofty ambitions of leaving no one behind, but also from the perspective of realizing the general aim of target 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.'

Big data also offer the potential to compile datasets that are linkable, offering enormous potential to undertake cross-cutting and dynamic analyses that may help us to better understand causation, offering 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 sample data. 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) to other data. Consequently, it is not always possible subsequently to construct a comprehensive analyses or narrative for many complex phenomena. The ability to link data may also provide a solution to a common misconception – that all statistics can be disaggregated to reveal additional characteristics. For example, many datasets do not have a gender component, irrespective of the level of disaggregation. But perhaps a gender component could be added by linking datasets. In this respect and given the importance of inter-linkages between the SDG goals, the importance of being able to connect data cannot be overstated.

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 longstanding criticism of development indicators. In the words of the UN 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 voluminous and timely statistics - it is not always clear that that 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 (for example, target 17.13 - Enhance global macroeconomic stability, including through policy coordination and policy coherence) and reacting quickly to natural disasters (for example, targets 1.5, 2.4. 11.5, 11.b, 13.1). 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.

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 an 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 derived from satellite imagery (for example, targets 14.2, 14.3, 15.1, 15.2, 15.3, 15.4). Such an approach poses some difficult questions, not least legal. Globalized data present particular challenges as they escape 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. It is difficult to predict whether this will make it easier or more difficult for NSOs to access and use these data in the future.

For many developing countries, the provision of basic statistical information remains a real challenge. The Global Partnership for Sustainable Development Data (2016) note that much of the data that does exist is 'incomplete, inaccessible, or simply inaccurate.' As noted above, 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. It is clear therefore that despite significant progress, serious problems with data availability persist. Some (Long & Brindley, 2013; Korte, 2014; Ismail, 2016) have argued, that owing to the falling costs associated with technology, big data may offer developing countries the opportunity to skip ahead and compile next-generation statistics. 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 improved access to computers and the internet, significant development in numeric and statistical literacy, and in basic data infrastructure. 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.

Big data may in some cases be better data than survey data. Stephens-Davidowitz makes a compelling argument that the content of social media posts 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: p.4). He describes data generated from searches, views, clicks and swipes as 'digital truth.' Thus, big data may be able to provide more honest data with greater veracity than can be achieved from traditional survey data. Hand (2015) makes a similar argument, noting that as big data are transaction data they are closer to social reality than survey and census data that are based on opinions or statements that rely on recall.

Finally, big data may offer the UN 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 emerging data ecosystem. Access to data (discussed in the next section) is a challenge, not only for NSOs, but for all sorts of institutions hoping to use big data. The UN 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. At the global level, the UN would seem to the sensible body. But perhaps at national level, there is a role also for NSOs to act a trusted 3rd party, or middle man, 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 to 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 analytical purposes.

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, the UN 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). Cervera et al. (2014), Landefeld (2014), Kitchin (2015) and MacFeely (2016) have all presented similar arguments in the past. As noted in the introduction, the task of populating the SDG GIF will be nothing short of herculean. Experience suggests that adopting a 'business as usual' approach will bring only partial success. Instead, the UN 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 GIF. While UN Pulse 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 several different forms. But 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'[v], would review unofficial statistics to see whether they can be certified as 'official' for the purposes of populating the SDG GIF. 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 UN System, 2018) and the *Common Metadata Framework* (United Nations Economic Commission for Europe, 2013) 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; and 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. Nevertheless, it is a discussion worth having.

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 what official statistics means and re-position the role of official statistics 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.

Challenges for compiling SDGs

Many big data are proprietary i.e. data that are commercially or privately-owned and not publicly available. Consequently, many big data are not currently accessible by NSOs, either because costs are prohibitive or proprietary ownership makes it impossible. For example, data generated from the use of credit cards, search engines, social media, mobile phones and store loyalty cards are all proprietary and are often not accessible. In many cases there are also legal impediments to accessing big data. MacFeely and Barnat (2017) have argued that to future-proof statistical legislation, changes to statistical legislation may be required to give NSOs or NSSs access to big data sources. Even if these data were publicly accessible, sensitivities around their repurposing to compile official statistics must be carefully considered (Daas et al., 2015). 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 from the culturally acceptable boundaries or norms of their country. So, a NSO must decide whether it is legally permissible, ethical and 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. 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 a 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. This tension does not appear to be well understood. MacFeely (2017) notes that regarding big data, a discernible mismatch exists between expectations and reality. The United Nations Economic Commission for Europe (2016) 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.'

Technology, the source of many big data, continues to rapidly evolve, raising 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: p.258) note 'The big data sources encountered so far seem subject to frequent modifications'. For example, social media may tweak their services to test alternative layouts, colors or design, which in turn may mutate or distort the underlying data. Kitchin (2015: p.9) warns 'the data created by such systems are therefore inconsistent across users and/or time'. Consequently, the United Nations Economic Commission for Europe (2016) caution that statisticians using big data will need to accept a general instability in the data. 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 GIF is to produce a time series for the 2010 - 2030 period[vi]. Volatility or instability of some big data sources introduces risks to continuity of data supply itself. NSOs must decide whether together, access and maturity are sufficiently stable to justify making an investment in big data. It is often said that data are the new oil. But data (just like crude oil) must be refined 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.

Ownership of source data is another issue of concern. As an NSO moves away from using survey-based data and becomes more reliant on administrative or other secondary data, such as big data, it surrenders control of its production system. As dependency on external source data increases, the NSO is progressively exposed 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 retard the development of statistics for decades[vii]. These are risks that a NSO must carefully consider when deciding whether to invest in secondary data. 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 the public, learning that 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.' NSO's 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 a NSO is using social media data for example, and the data provider becomes embroiled in a public scandal, the reputation of the NSO or IO may be adversely affected, through no fault of their own.

As noted earlier, big data are essentially re-purposed data and so, 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 be transparent. Both the data and the algorithms are typically proprietary and often of enormous commercial value. But accessing accurate metadata is essential 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 those data (Daas et al., 2015). Furthermore, as big data can be gamed or contain fake data (Kitchin, 2015; MacFeely, 2016) it is important to understand vulnerabilities in the data. There may also be representativity and accuracy deficits in many big data – for example, age, gender, language, disability, social class, regional and cultural biases. 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 (whether correct or incorrect) rapidly gains currency as a 'fact' as it travels through the web (Weinberger, 2014). There are also concerns for veracity arising from the concentration of data platforms. 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 e-book sales and 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[viii] raises a whole new set of concerns regarding the future veracity of big data for statistical purposes. 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. He could have added, and as a source of reliable data.

The emergence of big data is changing the information world. The digital revolution has created an abundance of data, challenging the monopolistic position enjoyed by official statistics for so long, to provide free, timely and high-quality statistics. Today's profusion of data has reduced the cost of entry into the statistics compilation business. Consequently today, there is a battle for the ownership of ‘facts’ – a battle that perhaps the global statistical community has not taken sufficiently seriously (MacFeely, 2017). Today a variety of compilers are producing statistics and although little is known about the quality of the input data or the compilation process, the allure of these statistics is seductive. The data deluge has contributed to the so called 'post-truth age' where virtually all authoritative information sources can be challenged by 'alternative facts' or 'fake news' with a consequent diminution of trust and credibility of all sources. It seems Huxley (1932) might have been correct when he predicted that truth would be drowned in a sea of irrelevance. As Fukuyama (2017) warns 'In a world without gatekeepers, there is no reason to think that good information will win out over bad.' In fact, there are mounting concerns at the weaponisation of data (O'Neill, 2016; Berners-Lee, 2018). Davies (2017) believes official statistics is losing this battle and argues ‘The declining authority of statistics is at the heart of the crisis that has become known as “post-truth” politics.’ Furthermore, these data are allowing new types of indicators and statistics to be compiled. So not only is the primacy of NSOs and NSSs being challenged, the legitimacy of many traditional statistics, such as GDP or unemployment statistics is also being diminished. National level statistics, based on international agreed classifications, are increasingly viewed by many as overly reductionist and inflexible. This criticism has been leveled at the global SDG indicators too. Letouzé and Jütting (2015: p.14) warn that the ‘proliferation of alternative "official" statistics' produced by a variety of outlets are challenging the veracity and trustworthiness of those generated by NSSs.’

The challenges of Privacy and Confidentiality

For official statistics, safeguarding the confidentiality of individual data is sacrosanct and enshrined in Principle 6 of the UN 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: p.2), too, ‘underscores repeatedly the requirement that the information that statistical agencies collect should remain confidential and inviolate’. The Scheveningen Memorandum (European Commission, 2013)[ix] prepared by the Directors General of NSOs in the European Union identified the need to adapt statistical legislation to use big data - both to secure access but also protect privacy. For a NSS to function, 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. In short, everyone who supplies data for statistical purposes does so with the reasonable presumption that their confidentiality will be respected and protected[x]. 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 surfacing many challenging questions, not least regarding 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 famously asserted that concerns over privacy are a 'red herring' as we 'have zero privacy' (Noyes, 2015). Many disagree and have voiced concerns over loss of privacy (see Pearson, 2013; Payton and Claypoole, 2015). Fry (2017) has likened developments in the big data space and the loss of privacy to the opening of Pandora's Box – what he terms ‘Pandora 5.0’. The introduction in Europe of the new General Data Protection Regulation (European Parliament 2016) 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 - at least in some regions of the world. By contrast, in the United States, 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 make anonymization of individual data very difficult. Big data, combined with the enormous computing power available today, mean that simply removing personal identifiers and aggregating individual data is no longer a sufficient safeguard. A paper by Ohm (2010) outlining the consequences of failing to adequately anonymize data graphically illustrates why there is no room for complacency. Thus, a problem that had been solved in the context of traditional official statistics must now be re-solved, in the context of a richer and more varied data ecosystem. In terms of SDGs, compilers must push back against any attempts to access individual data under the guise that 'no one is left uncounted' automatically grants access to 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 is some evidence to suggest this might be the case. There appear to be inter-generational differences in opinion vis-a-vis privacy and confidentiality, where those ‘born digital’ (roughly those born since 1990) are less concerned about disclosing personal information than older generations (European Commission, 2011). Taplin (2017: p.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 anonymization? If other statistical providers, not governed by the UN fundamental principles, take a looser approach to confidentiality and privacy, it may leave official statistics in a relatively anachronistic and disadvantaged position. 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[xi]. This of course makes sense, in that tax payers should to some extent own the data they have paid for with their taxes, 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 opening 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 a poor understanding of privacy issues and weak statistical literacy. Rudder (2014: p.241) notes that 'because so much happens with so little public notice, the lay understanding of data is inevitably many steps behind the reality.'

Taplin (2017: p.157) argues that we trade our privacy with corporations in return for benefits, 'but it is one thing to forfeit our privacy as individuals to a company that we believe is delivering a needed service and another to open our personal lives to the federal government.' MacFeely (2016) has warned that if the benefit of official statistics, and the importance of privacy, is insufficiently clear to the public or to policy makers, then it leaves official statistics vulnerable, and possibly facing a precarious future. Rudder (2014: p.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 argues that the trade-off benefit 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 and how I vote…Every American ought to be very afraid of big government' (Noyes, 2015). Who could argue? But complacency about the growth of a substitute private sector Big Brother seems naive. To some extent there is ideology at play here, where a neo-liberal agenda is pushing to minimize the role of the public sector, but it also illustrates the challenge facing IOs and NSOs generally where their contribution to the wellbeing of economies and societies is poorly understood. The challenge for NSOs and IOs is how to highlight the benefits of official statistics as a public good.

Thus, while big data may offer opportunities, they also present some real challenges for NSOs and NSSs. 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 regarding 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 deciding that big data is an appropriate source for compiling SDG indicators.

Conclusion

The purpose of the SDG GIF 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, yet, whether big data will contribute anything special to the SDG GIF. It seems likely that Tamm is correct and big data are just more data: another 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 or taken from people who are not necessarily aware 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.

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[xii]. Official statistics must take the ethical dimension seriously. In trying to quantify human rights abuses, the UN 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.

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 (Goal 5) or disability (see targets 1.3, 4.5, 4.a, 8.5, 10.2, 11.2, 11.7 and 16.7). Advances, such as, the Internet of Things[xiii] and biometrics will all surely present opportunities to compile new and useful statistics. 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. The projects listed in the *UN Big Data Inventory* are impressive. But few of these have yet borne fruit; some have not yet moved beyond the planning stage, others are bogged down in legal wrangling with Data Protection Commissioners. Work on developing price statistics does however appear to be advancing well. Only time will tell how the other projects progress.

These developments will bring a myriad of new challenges, not least the growth of unreliable information. It is already clear that big data are not a panacea for statistical agencies confronted with the challenge of compiling SDG indicators. 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. There are concerns too, that many social media channels are polarizing social exchange and promoting echo chambers and cyber-cascading. Official statisticians must ensure they can filter the wheat from the chaff.

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. They are also under pressure to populate the SDG GIF. 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 spot 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 missions and their mandates.

References:

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

Berners-Lee, T. (2018). 'The web is under threat. Join us and fight for it.' World Wide Web Foundation. March 12, 2018. Available from: https://webfoundation.org/2018/03/web-birthday-29/ [Accessed 19 March, 2018].

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

Boyd, J and Crawford, K. (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.

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

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

Committee of the Chief Statisticians of the United Nations System (2018). United Nations Statistics Quality Assurance Framework. Available from: https://unstats.un.org/unsd/unsystem/documents/UNSQAF-2018.pdf [Accessed 10 May, 2018].

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

Choi, H. and Varian, H. (2011). 'Predicting the present with Google Trends.' Available from: http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf [Accessed 7 January, 2018].

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

Davies, W. (2017). 'How statistics lost their power – and why we should fear what comes next.' R Available from: https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy?CMP=share_btn_link [Accessed 19 January 2018].

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

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. Available from: http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_359_en.pdf [Accessed 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. Available from: https://ec.europa.eu/eurostat/cros/content/scheveningen-memorandum_en [Accessed 25 January, 2018].

European Commission (2014). 'Big Data.' *Digital Single Market Policies*. Available from: https://ec.europa.eu/digital-single-market/en/policies/big-data [Accessed 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).' Available from: http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf [Accessed 16 Jan, 2018].

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

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

Fukuyama, F. (2017). 'The Emergence of a Post Fact World.' *Project Syndicate*, August 21, 2017. Available from: https://www.project-syndicate.org/onpoint/the-emergence-of-a-post-fact-world-by-francis-fukuyama-2017-01 [Accessed 4 January, 2018].

Gelsinger, P. in Whatsthebig data? (2012). 'Big Data quotes of the week'. Available from: https://whatsthebigdata.com/2012/06/29/big-data-quotes-of-the-week-11/ [Accessed 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. Available from: https://www.economist.com/node/666610 [Accessed 19 March, 2018].

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

Goodbody, W (2018). 'Waterford researchers develop new method to store data in DNA.' *RTE News*, January 20, 2018. Available from: https://www.rte.ie/news/ireland/2018/0219/941956-dna-data/ [Accessed 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., Kostroch, D.C., Quiros, G. and STA Internal Group (2017). 'Big Data: Potential, Challenges, and Statistical Implications.' *IMF Staff Discussion Note*, SDN/17/06, September 2017. Available from: http://www.imf.org/en/Publications/SPROLLs/Staff-Discussion-Notes [Accessed 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. Available from: https://ec.europa.eu/eurostat/cros/system/files/Presentation%20S20AP2%20%20Hand%20-%20Slides%20NTTS%202015.pdf [Accessed 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 Lopez, P. (2012). 'How to Measure the World's Technological Capacity to Store, Communicate and Compute Information.' *International Journal of Communication*, Vol 6, pp. 956–979.

Huxley, A. (1932). 'A brave new world'. Chatto and Windus, London.

IBM (2017). '10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations.' *IBM Marketing Cloud*. Available 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 [Accessed 25 January, 2018].

Inter-Agency and Expert Group on Sustainable Development Goals (2018). ‘Tier Classification for Global SDG Indicators.’ Available at: https://unstats.un.org/sdgs/iaeg-sdgs/ [Accessed 25 June, 2018]

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

Ismail, N. (2016). 'Big Data in the developing world.' Information Age, 8 September, 2016. Available from: http://www.information-age.com/big-data-developing-world-123461996/ [Accessed 19 January, 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/ [Accessed 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. Available from: http://www.bbc.com/news/technology-43465968 [Accessed 21 March, 2018].

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

Kulp, P. (2017). 'Facebook quietly admits to as many as 270 million fake or clone accounts.' *Mashable*, November 3, 2017. Available from: https://mashable.com/2017/11/02/facebook-phony-accounts-admission/#UyvC2aOAmPqo [Accessed 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. Available from: https://unstats.un.org/unsd/trade/events/2014/beijing/Steve%20Landefeld%20-%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf [Accessed January 18, 2018].

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

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

Letouzé, E. and Jütting, J. (2015). 'Official Statistics, Big Data and Human Development.' Data Pop Alliance, *White Paper Series*, Available from: https://www.paris21.org/sites/default/files/WPS_OfficialStatistics_June2015.pdf [Accessed 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 [Accessed 20 February, 2018].

Long, J. and Brindley, W. (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. Available 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 [Accessed 19 January, 2018]

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 Barnat, N. (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.

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

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

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

Noyes, K. (2015). 'Scott McNealy on privacy: You still don't have any.' *PC World*, IDG News Service, June 25, 2015. Available from: https://www.pcworld.com/article/2941052/scott-mcnealy-on-privacy-you-still-dont-have-any.html [Accessed 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. Available from: https://www.uclalawreview.org/pdf/57-6-3.pdf [Accessed 19 March, 2018].

O'Neill, C. (2016). 'Weapons of Math Destruction - How big data increases inequality and threatens democracy.' Allen Lane, London.

Payton, T. and Claypoole, T. (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. Available from: https://www.youtube.com/watch?v=lRgEMSqcKXA [Accessed 19 December, 2017].

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

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

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. Available from: https://www.forbes.com/sites/danielrunde/2017/02/25/the-data-revolution-in-developing-countries-has-a-long-way-to-go/2/#3a48f53e482f [Accessed 19 January, 2018].

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

Struijs, P., Braaksma, B. and Daas, P.J.H. (2014). 'Official statistics and Big Data.' *Big Data & Society*, April–June 2014: 1–6, DOI: 10.1177/2053951714538417. Available from: http://journals.sagepub.com/doi/pdf/10.1177/2053951714538417 [Accessed 21 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. Available from: https://www.linkedin.com/pulse/big-data-dead-just-regardless-quantity-structure-speed-thamm/ [Accessed 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. Available from: https://www.paris21.org/sites/default/files/654.pdf [Accessed 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. Available from: http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf [Accessed 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. Available from: http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E [Accessed 7 February 2017].

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

United Nations Economic Commission for Europe (2013). 'Common Metadata Framework.' UNECE Virtual Standards Helpdesk. Available from: https://statswiki.unece.org/display/VSH/The+Common+Metadata+Framework [10 May, 2018].

United Nations Economic Commission for Europe (2016). 'Outcomes of the UNECE Project on Using Big Data for Official Statistics.' Available from: https://statswiki.unece.org/display/bigdata/Big+Data+in+Official+Statistics [Accessed 15 February, 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. Available from: https://unstats.un.org/unsd/statcom/47th-session/documents/2016-2-IAEG-SDGs-E.pdf [Accessed 19 March, 2018].

United Nations Global Pulse (2017). ‘Annual Report 2017 - Harnessing Big Data for Development.’ Available from: https://www.unglobalpulse.org/sites/default/files/UNGP_Annual2017_final_web.pdf [Accessed 7 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. Available from:

http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf [Accessed 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. Available from: https://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf [Accessed 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. Available from: https://unstats.un.org/unsd/statcom/49th-session/documents/BG-Item-3a-IAEG-SDGs-DataFlowsGuidelines-E.pdf [Accessed 20 March, 2018].

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

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

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

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

Endnotes

[i] These indicators were adopted by the UN Statistical Commission in March 2017 (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 06 July 2017 (A/RES/71/313).

[ii] 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

[iii] A quintillion bytes is the equivalent of 10^{18} bytes or 1 exabyte.

[iv] 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.

[v] 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.

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

[vii] 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 new legislation may retard significantly the development of official statistics in that region.

[viii] 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).

[ix] Para 3 - Recognize 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.

[x] 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.

[xi] 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

[xii] At the time of writing, a UK company, Cambridge Analytica and Facebook are embroiled in scandal, arising from harvesting data from 50 million Facebook account holders without permission (see Kleinman, 2018). No doubt there will be worse to come.

[xiii] 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.

~~Biography~~

~~Dr. Steve MacFeely is the Head of Statistics and Information at UNCTAD. Steve is also Adjunct Professor at the Centre for Policy Studies at University College Cork in Ireland and deputy-director of the IASE International Statistical Literacy Program. Before joining UNCTAD, he was the Assistant Director-General at the Central Statistics Office (CSO) in Ireland.~~

Formatted: Footnote Text,single space,FOOTNOTES,fn,5_G, Justified