CSW62
COMMISSION ON THE STATUS OF WOMEN
EMPOWERING RURAL WOMEN AND GIRLS
12–23
MARCH
2018

Commission on the Status of Women
Sixty-second Session

Participation in and access of women to the media, and information and communications technologies and their impact on and use as an instrument for the advancement and empowerment of women

INTERACTIVE EXPERT PANEL
Innovative data approaches for measuring progress on gender equality and women's empowerment

Monitoring the implementation of the SDGs: The role of big data

by

Steve MacFeely*

United Nations Conference on Trade and Development

Friday, 16 March 2018

3:00 – 6:00 pm

* The views expressed in this paper are those of the author and do not necessarily represent those of the United Nations.

Introduction

From a statistical perspective the population of the 232 SDG indicators is an enormous task. So much so it led Mogens Lykketoft, President of the seventieth session of the UN General Assembly, to describe it as an ‘unprecedented statistical challenge’.

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 but no data are available or Tier 3 (68 indicators), meaning that no internationally established methodology or standards are yet available. Five indicators have multiple tier classifications. In other words, as of December 2017, only 40% of the SDG indicators can be populated.

The question facing official statisticians is whether big data can help? Ever since the High-Level Panel of Eminent Persons have called for a data revolution in their report ‘A New Global Partnership’ and the Independent Expert Advisory Group on a Data Revolution for Sustainable Development explained what that might mean in their report ‘A World That Counts’ the potential of big data has excited much comment, debate and even evangelism. Described by Pat Gelsinger of EMC as the 'new science' with all the answers. A paradigm destroying phenomena of enormous potential. Is big data the panacea to our SDG indicator problems?

Defining Big Data

Stephens-Davidowitz argues that big data is 'an inherently vague concept'. Mayer-Schonberger and Cukier Note 'There is no rigorous definition of "big data"'. It is important to understand that Big Data is badly named – as size is not the defining feature of big data.

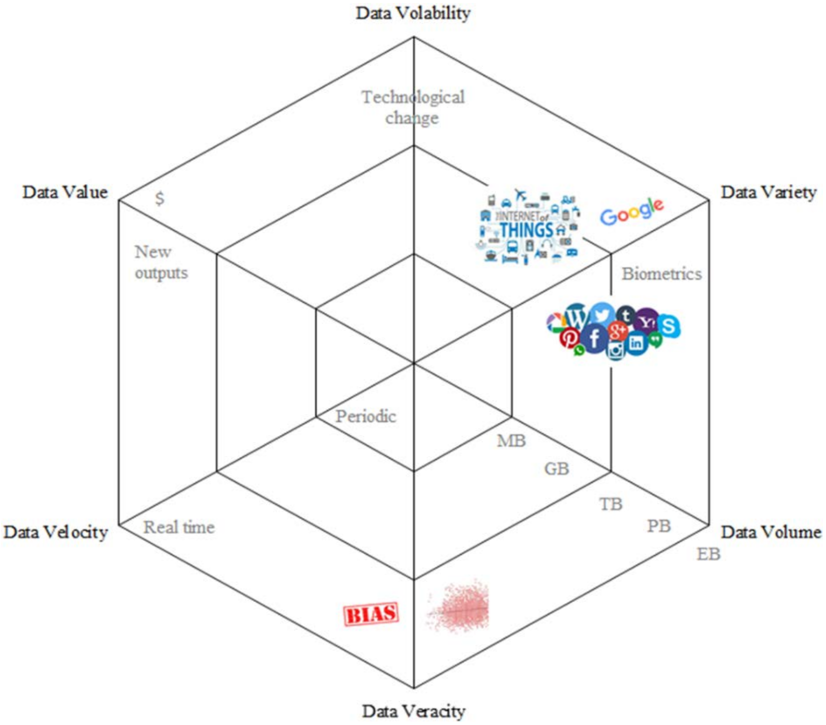
Gartner analyst Doug Laney provided the three 'Vs' definition in 2001, describing big data as high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. 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).

Tam and Clarke gave a more general definition, describing big data as 'statistical data sources comprising both the traditional sources and new sources that are becoming available from the "web of everything".'

In 2017 the IMF selected a 5V definition (the original 3Vs plus an additional two V's - volatility and veracity). Veracity refers to the 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'. 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 to note that using big data is 'a path with trap doors, land mines, misdirection, and false clues.'

Arguably a '6V' definition that includes 'value', where value means that something useful is derived from the data offers the best overall definition, offering a balance between parsimony and utility – the idea of cost-benefit. This is extremely important, the costs of investing in big data must be weighed up against what it can deliver in practical terms. See Figure 1.

Figure 1 - The 6V's of Big Data for Official Statistics



Big data are conceptually different to traditional survey data. They are made available to us, rather than designed by us. They are a collection of by-product data rather than data designed by statisticians for a specific purpose. In other words deriving statistics is a secondary purpose.

As big data is a by-product of our interactions with technologies that are evolving quickly, we must accept that as a consequence big data are not a stable platform but a very dynamic one, and so the definition is likely to require further refinement in the not too distant future.

Sources of Big Data

Our day-to-day dependence on technology are leaving significant 'digital footprints' in their wake. Everything we think or do is now potentially a source of data. Spending and travel patterns, online search queries, reading habits, television and movies choices, social media posts.

Some examples, big data were generated by:

- 227.1 billion global credit/debit card purchase transactions
- 7.7 billion mobile telephone subscribers
- mobile phones generate 600 billion unique data events every day
- Every day we send 500 million tweets

- 8 billion snapchats
- upload 1.8 billion images
- conduct 3.5 billion google searches.
- Every minute of every day we upload 400 hours of video to YouTube

Today 'Everything is data.' It is described as a data deluge; data smog; info-glut or the original information overload. This deluge is also the result of an important behavioral change, where people now record and load content for free. Weigand described this phenomenon where people actively share or supply data directly to various social networks and product reviews and led to the evolution of the wiki model as a 'social data revolution'.

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. Now data includes text, sound and images, not just neat columns of numbers.

How much data now exist. Definitional differences make this a difficult question to answer, but:

- Hilbert and Lopez estimate that 300 exabytes (or slightly less than one third of a zettabyte[1]) of data were stored in 2007.
- 2017 Big Data factsheet, Waterford Technologies estimate that 2.7 zettabytes of digital data exist[2].
- 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 estimate we create an additional 2.5 quintillion bytes[3] of data every day.

Borgman warns, big data must be treated with caution:

- As few as 35 percent of twitter followers may be real people
- as much as 10 percent of activity is social networks may be generated by robotic accounts
- 11% of display ads, almost 25% of video ads are viewed by bots[4] not people - 'fake clicks.'
- 25% of reviews on Yelp are bogus.
- 3% of Facebook accounts are fake and an additional 6% are clones or duplicates (equivalent of 270 million accounts).

There are also issues of coverage, as sizeable digital divides exist.

ITU estimate that global Internet penetration is only 48% and global mobile broadband subscription 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 does not use the web.

[1] 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

[2] 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 larger than the google-able 'Surface Web'.

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

[4] Goodman refers to these bots as WMDs - Weapons of Mass Disruption.

Even within countries, digital divides exist arising from a range of access barriers across various social, geographic, or economic strata that may lead to important cohorts being excluded, with obvious bias implications for statistics - hence the importance of 'Veracity.'

Accessing Big Data

Much big data are proprietary and not publically available. For example, data generated from using credit cards, search engines, social media, mobile phones and store loyalty cards are all proprietary and may not be available for use.

There are also sensitivities around repurposing data 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. Furthermore, costs may also be prohibitive.

Changes to statistical legislation may be required to give NSOs or National Statistical Systems (NSSs) access to big data sources. MacFeely and Barnat argue that to future-proof statistical legislation, consideration should be given to mandatory access to all appropriate secondary data…where secondary data would be defined to include not only administrative or public sector data but also some important, commercially held data, such as for example, information on credit card transactions, information held by utilities or information regarding the movements of mobile phones.

Legal & ethical issues - NSOs must decide whether it is legally permissible, ethically wise or culturally acceptable to access and use big data. These are not always easy questions to answer. When it comes to accessing new sources of digital data, 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.

Opportunities for official statistics

Blended data - big data may be used in conjunction with or as a replacement for traditional data sources to improve, enhance and complement existing statistics. Big data may also entirely or partially replacing existing data sources to compile existing statistics in a more efficient, timely or more precise way or compiling entirely new statistics altogether.

Linkable data - Big data offer the potential for linkable data. Discrete sample based data cannot be easily connected or linked (other than at a fuzzy level) with other data. 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.

Improved timeliness - this has been a long standing criticism of official statistics. In the words of the Data Revolution Group 'Data delayed is data denied…The data cycle must match the decision cycle.' Big data offers the possibility of publishing very current indicators, using what Choi and Varian describe as 'contemporaneous forecasting' or 'nowcasting.

New production model? - many digital data are global or at least supra-national in scope. This offers the potential to switch from a national to a collaborative international production model. In the case of global digital data, the most logical and efficient approach might be to centralise statistical production in a single centre rather than replicating production many times over in individual countries.

Data divide – for many developing countries, the provision of basic statistical information remains a real challenge. The Global Partnership for Sustainable Development Data note that much of the data that do exist are incomplete, inaccessible, or simply inaccurate. UNCTAD noted that at the end of the fifteen year MDG life cycle, developing countries could populate, on average, only two thirds of the Millennium Development Goal (MDG) indicators. If this is a barometer for data availability in developing countries, then we it is clear, that despite significant progress, that serious problems with data availability persist. Can big data help developing countries to skip ahead, and compile next-generation statistics? Countries will still need access to computers and internet, significant development in numeric and statistical literacy, and in basic data infrastructure. There may also require strengthened statistical legislation and data protection legislation.

Better data? - Seth Stephens-Davidowitz argues that content of social media posts, social media likes and dating profiles is no more (or less) accurate than social surveys. But also notes that people's search for information is, in itself, information. He describes data generated from searches, views, clicks and swipes as 'digital truth.' So Big data may be able to provide more honest data or greater *veracity* than we can ever achieve from survey data. Hand 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, statements or recall.

Global leadership - big data may offer NSSs and IOs an opportunity to exercise some leadership and regain some control over an increasingly congested and rapidly fragmenting information space. 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.

Reduce burden - 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 a mechanism to relieve survey fatigue and burden to households and businesses.

Improved registers - 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 instruments.

New/more disaggregated statistics - big data also offer the chance for new flow or dynamic statistics to be derived, offering the potential for more policy-relevant, outcome-based statistics. The sheer *volume* of data may also allow greater disaggregation of some statistics, or greater segmentation or granular analyses. It also offers the chance to measure a much wider variety of new statistics.

Challenges for official statistics

Instability - technology, the source of many big data is evolving rapidly, raising questions for their practicality as a data source for the compilation of official statistics. UNECE caution that official statisticians using big data will need to accept a general instability in the data. 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.

Ownership of data - 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.

*Reputational risk*s - reliance on external data sources may introduce 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 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 may be adversely affected, through no fault of their own.

Representativity - There are concerns over how representative many big data sets are. There may be age, gender, 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 (either correct or incorrect) rapidly gains currency as a 'fact' as it is passed around the web. There are also concerns for veracity arising from the concentration of data owners. Reich notes that in 2010, the top ten websites in the United States accounted for 75 percent of all page views. Google has an 88% market share in online searches, Amazon has a 70% market share in e-book sales, Facebook has a 77% market share in mobile social media. Such concentration introduces obvious risks of abuse and manipulation, leaving serious questions for the veracity of the data. The decision by the Federal Communications Commission (2017) in the United States in December 2017 to repeal Net Neutrality raises a whole new set of concerns regarding the veracity of big data for statistical purposes.

Privacy (now) - Mark Zuckerberg, the founder of Facebook claimed that the age of privacy is over. Scott McNealy, CEO of Sun Microsystems, too asserted that privacy is a 'red herring' and that we 'have zero privacy'. Many disagree. In Europe the new General Data Protection Regulation reinforces citizen's data-protection rights and suggests that privacy is still a real concern - at least in some regions of the world. 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.'

Confidentiality - safeguarding the confidentiality of individual data is sacrosanct for official statistics and enshrined in Principle 6 of the United Nations Fundamental Principles of Official Statistics. The failure to treat individual information as a trust would prevent the statistical agency from functioning effectively. For a NSS to function, confidentiality of the persons and entities for which it holds individual data must be protected. 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.

Privacy (future) - what if Zuckerberg and McNealy are correct and future generations are less concerned about privacy? 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) are less concerned about disclosing personal information than older generations. Taplin muses that '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 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 vis-a-vis other data providers.

Data wars - asymmetry of open data between public and private sector data. There is also a battle for ‘facts’ in the post truth era – fake news/alternative facts. Risks of privatizing truth. Diminution of trust and credibility of all sources. ‘The declining authority of statistics is at the heart of the crisis that has become known as “post-truth” politics.’ Primacy of NSOs being challenged and the legitimacy of many traditional statistics.

Governance Issues

Big data presents a range of challenges for NSOs, NSSs and IOs. In considering whether big data provides a viable option, statistical offices and systems must carefully decide what governance systems will be required to ensure the official statistics brand is not compromised. Governance systems can be defined as the policies and rules, and the 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 statistical 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 those massive global datasets will effectively be setting many of the future data standards. What will this mean for the global statistical system? What will it mean for the United Nations Fundamental Principles of Official Statistics? These massive new globalized data also challenge the idea of national or local compilation, raising a host of legal, security and organizational issues.

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 might need to be consider:

Ethics - many big data are the exhaust from a variety of technologies. Deriving statistics involves repurposing those data. The possibilities will be exciting and may offer incredible opportunities to derive new and exciting statistics. In the rush to compile new statistics, it may be easy to forget where those data came from. Thus it may be sensible to establish an ethics committee that considers whether the compilation of new statistics 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 ensure 'no harm' is done. 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 debates, including for example, the ownership of data.

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 sources, such as, credit card expenditure information or mobile phone location data without breaching data protection, statistical or other legislation? It will probably be necessary (or, at the very least wise) to establish a small board of specialist legal experts who can adjudicate on these complex issues and provide comprehensive legal opinion to the management board of the NSO. The correspondence between statistical and data protection legislation will be of paramount importance in the coming years.

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 new 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? This is a mixture of statistical methodology and broader governance. 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?

IT and 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. What does globalized data mean for storage location – does it make sense we continue with the old paradigm where data are stored, locally, in-house? If it is stored locally, will the data be quarantined and stored offline (so that it cannot be hacked or corrupted). If not will the NSO require some types of randomized identifiers to mask identifiers and 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 in some central location (in the cloud?). But how then will the data be integrated with other data sources stored in different locations. The movement and transfer of data will require secure pipelines system, requiring encryption.

Quality Assurance - Assessing the quality of big data is not the same as assessing traditional datasets. Firstly, quality must be defined from the perspective of big data and clear criterion for how these can be measured. United Nations Economic Commission for Europe 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. 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, transformation and conversion. Using big data may require an extended quality framework for official statistics. A framework that perhaps puts greater emphasis on risk management than that currently used.

Continuous Professional Development & 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 some 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.

Strategic Partnerships - as noted above, using big data presents a range of technical challenges. A 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 have the required skills. These will be critical decisions, both in terms of costs and efficiencies, but also for legal and reputational reasons.

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.

Conclusion

Ubiquitous technology has created a deluge of digital data that are now being used to compile a variety of new informational indicators and statistics. But it is not clear, as yet, whether big data offers anything special for the compilation of SDG indicators. What makes big data so intriguing is the fact that they simultaneously present both threats and opportunities for official statistics.

There is a new gold rush underway - a data rush. In that rush, NSOs and IOs are feeling the pressure to be seen to utilize big data. 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 mandates. Above all else, irrespective what data sources used, official statisticians must supply independent and impartial information that allow citizens to challenge stereotypes, governments, public bodies and private enterprises and hold them to account.

Big data cannot really help with Tier 3 indicators - for the moment. Until the concept is clear, potential data sources are not the main concern. However it is possible that big data might offer some help for Tier 2 indicators. Perhaps too, some Tier 1 indicators could be compiled in new ways using big data.

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). Even today we are probably only in the first hour. 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. Just because something can be measured doesn't mean it should be. In assessing whether to, and how to use big data, NSOs must begin to carefully consider what are the human rights of citizens in this digital age?

But 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 the availability of statistics 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 or the challenges of being disabled.

Developments, such as, the Internet of Things[5], biometrics and behaviometrics will all surely present opportunities to develop new and useful statistics. As yet, the implications of this 'big (data) bang' is for statistics is not immediately clear, but one can envisage a whole host of new ways to measure human interactions and experiences. But 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'. This may not be universally understood and so managing expectations will be an ongoing challenge for official statisticians. Challenges regarding 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 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'. Official statisticians must ensure they can filter the wheat from the chaff.

[5] 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 notes the result will be 2.5 sextillion potential networked object-to-object interactions.