UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT

June 2018

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

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

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

Abstract

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

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

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

UNITED NATIONS

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

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

© 2017 United Nations

Contents

Acknowledgements

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

Introduction

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

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

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

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

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

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

Organisation of the paper

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

1. Definitions

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

1.1 Big Data

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

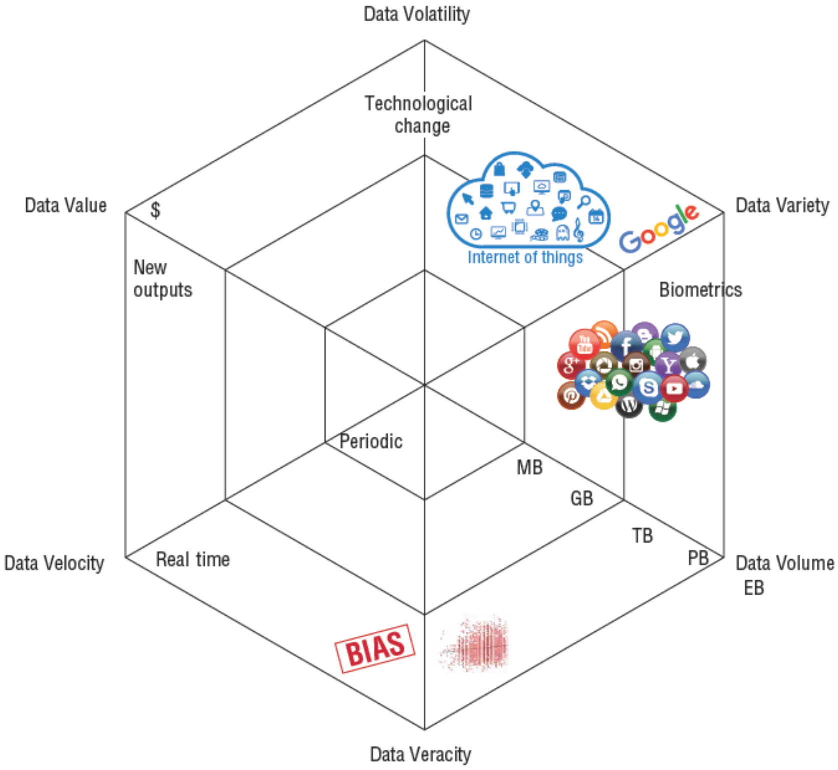
Gartner analyst Doug Laney (2001) provided what has become known as the three 'Vs' definition. He described big data as being high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision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

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

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

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

Figure 1.1
The 6Vs of Big Data for Official Statistics



Derived from Hammer et al. (2017)

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

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

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

1.2 Administrative Data

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

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

1.3 Official Statistics

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

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

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

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

2. Sources of Big Data

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

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

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

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

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

3. Opportunities for compiling SDGs

There will almost certainly be opportunities in the future to compile SDG indicators in new and exciting ways.

预览已结束，完整报告链接和二维码如下：
https://www.yunbaogao.cn/report/index/report?reportId=5_9102