Commentary

BIG DATA & SOCIETY

Big Data & Society
April–June 2014: 1–6

DOI: 10.1177/2053951714538417
bds.sagepub.com
SAGE

Official statistics and Big Data

Peter Struijs, Barteld Braaksma and Piet JH Daas

Abstract

The rise of Big Data changes the context in which organisations producing official statistics operate. Big Data provides opportunities, but in order to make optimal use of Big Data, a number of challenges have to be addressed. This stimulates increased collaboration between National Statistical Institutes, Big Data holders, businesses and universities. In time, this may lead to a shift in the role of statistical institutes in the provision of high-quality and impartial statistical information to society. In this paper, the changes in context, the opportunities, the challenges and the way to collaborate are addressed. The collaboration between the various stakeholders will involve each partner building on and contributing different strengths. For national statistical offices, traditional strengths include on the one hand the position to collect data and combine data sources to statistical products, and on the other hand their focus on quality, transparency and sound methodology. In the Big Data era of competing and multiplying data sources, they continue to have a unique knowledge of official statistical production methods. And their impartiality and respect for privacy as enshrined in law uniquely position them as a trusted third party. Based on this, they may advise on the quality and validity of information of various sources. By thus positioning themselves, they will be able to play their role as key information providers in a changing society.

Keywords

Big Data, official statistics, European Statistical System

Introduction

The advent of Big Data is expected to have a big impact on organisations for which the production and analysis of data and information is core business. National Statistical Institutes (NSIs) are such organisations. They are responsible for official statistics, which are heavily used by policy makers and other important players in society. Arguably, the way NSIs take up Big Data will eventually have implications for all of society.

Official statistics play a key role in modern society. NSIs aim at providing information on all important aspects of society in an impartial way, and according to the highest scientific standards. Information that fulfils these demands is used in public discussion, forms the basis of policy decisions, is required for business use, feeds scientific research, is used in education and so on. Official statistics can only meet this demand if they can be trusted. In advanced societies, official statistics are often taken for granted, but where trust is lacking, society misses an important pillar for informed discussion and evidence-based policy making.

Professional standards play a vital role in securing trust in official statistics. Statisticians have their own ethics code (United Nations, 2013), which includes an absolute respect for the confidentiality of data provided by respondents. Data collected for statistical purposes may never be disclosed and may never be used for other purposes. At the level of the European Union (EU), quality norms have been codified in the so-called Statistics Code of Practice (Eurostat, 2014). The trust earned by respecting professional standards is also the basis for a privileged position of NSIs in respect of data acquisition. Many NSIs have access by law to government data sources and have the power to collect data from other parties, often without having to pay the provider. Moreover, for statistical purposes, many NSIs are allowed to link data from different sources.

Statistics Netherlands, Heerlen, The Netherlands

Corresponding author:
Peter Struijs, Statistics Netherlands, P.O. Box 4481, Heerlen, 6401 CZ, The Netherlands.
Email: p.struijs@cbs.nl

Given this role for NSIs, what does the emergence of Big Data mean for official statistics? This question is addressed in this contribution, but as we will see, there are many reasons why the role of NSIs in the Big Data era is not 'given'. In order to keep a sound and trusted basis of information for society to rely on, we argue that NSIs may have to adapt to the changing context in which they operate.

Official statistics in a changing context

In respect of information, society is changing rapidly. For example, there is an enormous growth of data that is gathered and recorded in myriad ways: from satellite and sensory data, to social network and transactional data and so on. The availability of data is also expanding and becoming the foundation of business models. Information is becoming more visual and interactive. Information and communication technology is becoming ever more advanced, processing power and data storage capacity is continuously rising, cloud solutions are emerging and applications are becoming more intelligent. These developments have been described in more depth and detail by many observers, such as Mayer-Schönberger and Cukier (2013).

These changes have many impacts on societies. For one, the increased gathering of data and the commercial and social possibilities of data usage influence public opinion on privacy. Some are concerned if their data is re-used without their consent, for commercial reasons or otherwise. Others do not mind so much, if this means that services are provided for free. Many people voluntarily share information on social networks without caring for privacy. People have less patience to fill in questionnaires, especially if the data requested have been registered somewhere else already. Government agencies are expected to be more forthcoming in providing data. Governments have reacted to the changes by formulating policies on, for instance, open data and availability of public sector information, also at the EU level (European Union, 2013).

How have NSIs responded? Until around the 1980s, data were essentially a scarce commodity with a high price. Before the era of Big Data, information was not readily available but had to be collected on purpose. Official statistical information based on survey data had a unique value: there simply was no alternative. For example, population census data, collected door to door, was immensely valuable to policy makers, researchers and other users. In the last few decades, data collected by public administrations have become increasingly accessible for statistical purposes, stimulated in part by IT developments. Statistical data collection by means of questionnaires was supplemented and increasingly replaced by administrative data sources. Nowadays, some countries do not conduct extensive population surveys anymore but compile census statistics by combining and analysing data from several administrative sources. NSIs became more integrated in the information architecture of the government. In this way, the burden on persons and businesses to respond to questionnaires was considerably reduced.

In the context of all of these developments, the information provided by NSIs still remained unique. In particular, the possibility of combining data from different sources made official statistics even more valuable, since in many countries no other organisation was positioned to do so. In parallel, efforts also increased to standardise and harmonise these various sources of official statistics, especially in the EU. Supported by legislation, official statistics in the EU are now considered a system, the so-called European Statistical System, or ESS.[1]

However, Big Data is changing the environment of the NSIs once more as data scarcity is becoming ever less an issue. For NSIs, there are potential benefits as new data sources and opportunities emerge. But it also makes the products of NSIs potentially less unique, since other players in the information market may start – and have actually started – producing statistics, for instance, on inflation, such as the Billion Prices Project of MIT.[2]

Let us first look at the opportunities for NSIs offered by Big Data. There is a huge potential for new statistics (Daas et al., 2013). Location data for mobile phones could be used for almost instantaneous daytime population and tourism statistics (De Jonge et al., 2012). Social media messages could be used for several types of indicators, such as an early indicator of consumer confidence. Inflation figures could be derived from price information on the web. And so on. In addition, Big Data sources may be used to substitute or supplement more traditional data sources, such as questionnaire and administrative data. For instance, data collection by questionnaire on road use may not be necessary anymore if detailed traffic loop data, i.e. data from sensors in roads, becomes available (Struijs and Daas, 2013).

However, in order to realise these opportunities, a number of challenges have to be overcome, which are generally applicable to all uses of Big Data as an information source and as such are not unique to NSIs.

Challenges and issues

Some of the biggest challenges that statisticians face in their use of Big Data concern methodology. Many Big Data sources, such as social media messages, are composed of observational data and are not deliberately

designed for data analysis, and thus do not have a well-defined target population, structure and quality. This makes it difficult to apply traditional statistical methods, based on sampling theory (Daas and Puts, 2014a). The unstructured nature of many Big Data sources makes it even more difficult to extract meaningful statistical information. For many Big Data sources, the interpretation of the data and its relationship with social phenomena of interest is far from obvious. For example, public Facebook messages in the Netherlands clearly reflect general sentiment in some sense, but it is far from clear how exactly (Daas and Puts, 2014b). Moreover, if such data are to be used as a source for a population sentiment indicator, one would like to know the relationship between the population of persons writing public messages on Facebook and the population at large. This is challenging without falling back to surveys. Furthermore, the population of persons using social media is likely to change over time, making a comparison to the population at large even more challenging.

For NSIs, a key question concerns how the quality of official statistics can be guaranteed if they are based on Big Data. To address this, new methodologies and forms of interpretation need to be developed. Take for example mobile phones. If data from mobile phone providers are used for statistics on, say, population mobility, the statistician has to interpret anonymised detailed call records from individual phones and derive information about the behaviour of the people using them. That means dealing with the fact that measureable phone activity may vary during the day, some persons may have multiple mobile phones or none, children carry mobile phones which are registered to their parents, phones may be switched off, etc. For social media, even more questions arise such as who is the author of a message. While some methodological remedies have already been developed to some extent, such as deriving the gender and age of a social media user by the known correlation between sex, age and choice of words, these still pose a challenge, as explained above.

Privacy and legal issues form another challenge. The prevention of the disclosure of the identity of individuals is an imperative, but this is difficult to guarantee when dealing with Big Data. Since legislation typically lags behind the emergence of new social phenomena, the legal situation for cases involving Big Data is not always clear. In such cases, one may have to fall back on ethical standards to decide on whether and how to use Big Data. Other legal issues relate to copyright and the ownership of data. Even if data may legally be used, this does not imply that it is wise or appropriate to do so. Of critical importance is the implication of any use of Big Data for the public perception of an NSI as this has a direct impact on trust in official statistics. These concerns have been heightened by the revelations that intelligence agencies are among the most active Big Data users. For NSIs, it is critical that these concerns be addressed through practices such as being transparent about what and how Big Data sources are used. Other mechanisms could also be developed. For example, in some cases it might be feasible to adopt informed consent approaches. Some mobile phone subscription contracts, for instance, offer an opt out to the subscriber for using their data for other purposes than providing the phone service. If the opt out rate is not too high, this does not seriously affect the usability of mobile phone data for statistical purposes.

Another obvious challenge is the processing, storage and transfer of large data sets. Technological advances like increases in computing power, larger storage facilities and high bandwidth data channels may partly solve these issues. Having data processed at the source, thus preventing the transfer of large data sets and the duplication of storage, may also be considered. These technological challenges include mechanisms for ensuring the security of data, which is of the utmost importance because of privacy and confidentiality concerns and makes for example cheap cloud-based solutions less attractive.

Another issue is the possible volatility of Big Data sources, given the fact that official statistics often take the form of time series analyses. For many users, the continuity of these series is of the utmost importance. Still another issue is the skills required for dealing with Big Data. Modern data scientists may be better equipped than traditionally trained statisticians. Probably more important is the need for a different mind-set as the use of Big Data may imply a paradigm shift, including an increased and modified use of modelling techniques (Daas and Puts, 2014a; Struijs and Daas, 2013).

Collaboration

Faced with these challenges, NSIs have recognised the necessity of not working in isolation but collaborating with each other and others outside the community of official statistics. This collaboration is often exploratory and may be aimed at sharing knowledge and experiences, but there are already examples of collaboration that go further.

From the perspective of NSIs, several types of partners are of interest. First of all, the potential providers of Big Data are essential partners: if they do not grant access to their data, the story is over before it starts. Data owners have their own concerns and, like NSIs, they are subject to privacy rules. This may complicate collaboration even if they have a positive outlook and approach. But since Big Data sources are not designed

for statistical use, such collaboration is also essential in order to obtain good knowledge of the provenance of such sources. Additionally, for statistical production, it may be more efficient to have data processed at the site of collection and storage. In such cases, the assumption that data can be provided for free may no longer hold. On the other hand, statisticians also have much to offer such as providing analytic insights that may help data owners understand their data better. Doing complex statistical analyses is core business for NSIs, but not for, say, a mobile phone company. In these and other ways, the relationship with data providers could potentially become true partnerships. For example, one specific role that NSIs could play is that of a trusted third party. In a competitive market, competitors will be reluctant to share sensitive data among each other. But they might be willing to share it with an NSI who compiles statistical information that is beneficial to all.

Collaboration between NSIs and academia may grow as well. Universities have historically been natural partners for NSIs. It stands to reason that such collaboration will extend to the field of Big Data, for instance, in solving methodological problems, developing technical solutions and training future data scientists. Such collaboration is also being supported by public funders who are facilitating research and innovation partnerships through targeted grants.[3] By working in partnership, researchers in universities and NSIs could better leverage such opportunities.

Furthermore, there are many commercial partners with which NSIs could collaborate. Google and Facebook are two examples for which Big Data form the core of their business model. Their knowledge and the data to which they have access may be very relevant to NSIs. IT companies also possess relevant knowledge on Big Data processing and storage, security, cloud processing, etc. Apart from the provision of paid services, collaboration may be of interest to them with a view to obtaining statistical expertise and for benchmarking or validating their information products.

Collaboration between NSIs in the field of Big Data has already started. Big Data has become a prominent subject at many statistical meetings and conferences in Europe, such as the 2013 New Techniques and Technologies for Statistics (NTTS) conference,[4] a scientific conference organised by Eurostat, and the 'ESS Big Data event 2014' in Rome.[5] The directors-general of all European NSIs met in Scheveningen in September 2013 to learn about Big Data and adopted the Scheveningen Memorandum (DGINS, 2013). This memorandum calls for an international strategic approach to Big Data and plans for the adoption of an action plan and roadmap by mid-2014.

For some time already, Big Data has been an important topic for the UNECE, the United Nations Economic Commission for Europe. Collaboration at that level resulted in an overview paper about the implications of Big Data for official statistics (UNECE, 2013a). Seminars have been held, facilitating the exchange of knowledge, for instance, on statistical data collection.[6] In 2014, the UNECE has gone one step further in facilitating cross-national work through a project with the following stated objectives:

a. to identify, examine and provide guidance for statistical organisations to act upon the main strategic and methodological issues that Big Data poses for the official statistics industry;
b. to demonstrate the feasibility of efficient production of both novel products and 'mainstream' official statistics using Big Data sources, and the possibility to replicate these approaches across different national contexts;
c. to facilitate the sharing across organisations of knowledge, expertise, tools and methods for the production of statistics using Big Data sources (UNECE, 2013b).

The future of official statistics

What does the advent of Big Data mean for official statistics? As we have argued, it provides many opportunities. But in order to make optimal use of Big Data, a number of issues have to be addressed. This calls for increased collaboration with private and academic partners who have access to specific Big Data sources and knowledge, but also between NSIs. The relationship between the various stakeholders will involve each partner building on and contributing different strengths and likely result in flexible networks. Such networks are flexible in the sense that membership of the network and the contribution of partners depend on actual needs instead of being fixed in advance for a long time.

Seen from the viewpoint of NSIs, there are also potential risks. Official statistics are facing more competition. In a time of growing data abundance, generating statistical information that is potentially relevant to society is no longer an activity intrinsically restricted to NSIs. And even the traditional advantage of NSIs, being legally allowed to collect data and combine data sources, is eroding. It may not be possible to combine survey data and administrative data with Big Data sources at the micro-level, which reduces the relative disadvantage traditionally faced by the competition.

For some statistics, Big Data sources cannot be easily envisaged as alternatives to more traditional

data sources. This certainly holds for official figures on government finance and economic growth, which are heavily used for decision making at both the national and international level. But, given the increasing competition that data generated by other sources is presenting to the role of NSIs as bearers of official statistics, a strategic reassessment is needed. This could include fundamental questions such as whether statistics based on Big Data sources should be a core activity of NSIs, or if some data and information should be provided by other market actors or if NSIs can or should provide new services in this context.

But by posing these questions, we return to the basic premise that society's access to impartial statistical information must be maintained at all times, either by NSIs or other parties. In choosing a position, NSIs could build on and promote their strengths and unique position. Especially at a time of competing and multiplying data sources, their impartiality and respect for privacy as enshrined in law uniquely position them as a trusted third party. They also have unique knowledge of official statistical production methods. Finally, they continue to have privileged access to government data sources that provide unique information and knowledge and have the authority to collect data for statistical purposes that because of privacy considerations will never be available to businesses.

As a consequence, in the context of the challenges of Big Data sources, NSIs will remain important providers of official statistics. And where other organisations are able to provide statistical information to the public, rather than competing, NSIs could build on their position as an impartial, trusted third party and their expertise to advise on the quality and validity of information of these various sources. Possibly then, providers of Big Data may even seek validation of their data from NSIs, thereby opening up yet another possibility for new partnerships.

The future of official statistics in the age of Big Data is still a matter of some deliberation and experimentation. But what is clear already is that the international statistical community needs to adapt to a new reality and respond to the opportunities and challenges it provides. To do so calls for greater collaboration with players inside and outside the statistical community, through the formation of flexible networks that can forge new ways of generating statistical data. For all engaged with statistics, we think the Big Data era is a most exciting time.

Acknowledgements

The views expressed in this contribution are those of the authors and do not necessarily reflect the position of Statistics Netherlands. The authors wish to thank the editors for their valuable suggestions for improvements.

Declaration of conflicting interest

The authors declare that there is no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/ess/ess_news
2. http://bpp.mit.edu/
3. The current EU framework programme for research and innovation, Horizon 2020, is an example (European Commission, 2013), which mentions Big Data specifically.
4. http://www.cros-portal.eu/content/ntts-2013
5. http://www.cros-portal.eu/content/big-data-event-2014
6. http://www.unece.org/stats/documents/2013.09.coll.html

References

Daas PJH and Puts MJH (2014a) Big Data as a source of statistical information. *The Survey Statistician* 69: 22–31 (Available at: http://isi.cbs.nl/iass/N69.pdf (accessed 22 May 2014).

Daas PJH and Puts MJH (2014b) Social media sentiment and consumer confidence. In: *Paper for the workshop on using Big Data for forecasting and statistics*, Frankfurt, Germany, 7–8 April. Available at: http://www.ecb.europa.eu/events/pdf/conferences/140407/Daas_Puts_Sociale_media_cons_conf_Stat_Neth.pdf?409d61b733fc259971ee5beec7cedc61 (accessed 22 May 2014)..

Daas PJH, Puts MJ, Buelens B, et al. (2013) Big Data and official statistics. In: *Paper for the 2013 NTTS conference*, Brussels, Belgium, 5–7 March. Available at: http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_76.pdf (accessed 22 May 2014).

De Jonge E, Van Pelt M and Roos M (2012) *Time patterns, geospatial clustering and mobility statistics based on mobile phone network data*. Discussion paper 201214, Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/010F11EC-AF2F-4138-8201-2583D461D2B6/0/201214x10pub.pdf (accessed 22 May 2014).

DGINS (2013) Scheveningen memorandum on Big Data and official statistics. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM%20Final%20version.pdf (accessed 22 May 2014).

European Commission (2013) Horizon 2020, the EU framework programme for research and innovation. Available at: http://ec.europa.eu/programmes/horizon2020/ (accessed 22 May 2014).

European Union (2013) 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. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:EN:PDF (accessed 22 May 2014).

Eurostat (2014) European statistics code of practice. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice (accessed 22 May 2014).

Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution that will Transform How We Live, Work, and Think*. London: John Murray Publishers.

Struijs P and Daas PJH (2013) Big Data, big impact? In: *Paper presented at the seminar on statistical data collection*, Geneva, Switzerland, 25–27 September 2013. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2013/mgt1/WP31.pdf (accessed 22 May 2014).

UNECE (2013a) What does "Big Data" mean for official statistics? In: *Paper prepared on behalf of the high-level group for the modernisation of statistical production and services*, 10 March. Available at: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614 (accessed 22 May 2014).

UNECE (2013b) The role of Big Data in the modernisation of statistical production. Project plan. Available at: http://www1.unece.org/stat/platform/display/msis/Final+project+proposal%3A+The+Role+of+Big+Data+in+the+Modernisation+of+Statistical+Production (accessed 22 May 2014).

United Nations (2013) Fundamental principles of official statistics. Available at: http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf (accessed 22 May 2014).