Statistical Journal of the IAOS 31 (2015) 471–481
DOI 10.3233/SJI-150906
IOS Press

The opportunities, challenges and risks of big data for official statistics

Rob Kitchin
NIRSA, National University of Ireland Maynooth, County Kildare, Ireland
E-mail: Rob.Kitchin@nuim.ie

Abstract. The development of big data is set to be a significant disruptive innovation in the production of official statistics offering a range of opportunities, challenges and risks to the work of National Statistical Institutions (NSIs). This paper provides a synoptic overview of these issues in detail, mapping out the various pros and cons of big data for producing official statistics, examining the work to date by NSIs in formulating a strategic and operational response to big data, and plotting some suggestions with respect to on-going change management needed to address the use of big data for official statistics.

Keywords: Big data, official statistics, national statistical institutions (NSIs), opportunities, challenges, risks

1. Introduction

National Statistical Institutions (NSIs) are charged with producing and publishing official statistics across a range of domains and scales relating to a nation. Official statistics are used to report on the present state of play and unfolding trends with respect to society and economy to a domestic and international audience, with many statistics being collated into supra-national statistical systems. Over the last couple of hundred years, NSIs, both on their own initiative and in collaboration with each other, have developed rigorous and standardized procedures for sampling, generating through surveys, handling, processing, storing, analyzing, sharing and publishing official statistical data. During the past half century, NSIs have increasingly turned to exploiting administrative data sets produced by other state agencies to compile official statistics. In both cases, NSIs are the principle administrator of an official statistical system, in the first case controlling the whole data life cycle and in the second supported by legislative tools to ensure compliance with data provision.

The development of big data and its potential as a third source of data for official statistics poses a range of opportunities, challenges and risks to the work of NSIs. As with many new innovations which are driven by new technological developments, the term big data has become a buzz phrase that is variously understood, with many definitions making reference to a fundamental shift in the nature of some data with respect to the 3Vs of volume, velocity and variety [19]. Based on an extensive review of the literature and a conceptual comparison between small and big data (see Table 1), Kitchin [15,16] contends that big data have the following characteristics:

- Huge in *volume*, consisting of terabytes or petabytes of data;
- High in *velocity*, being created in or near real-time;
- Diverse in *variety*, being structured, semi-structured and unstructured in nature;
- *Exhaustive* in scope, striving to capture entire populations or systems ($n =$ all);
- Fine-grained in *resolution* and uniquely *indexical* in identification;
- *Relational* in nature, containing common fields that enable the conjoining of different data sets;
- *Flexible*, holding the traits of *extensionality* (can add new fields easily) and *scaleability* (can expand in size rapidly).

With some notable exceptions, such as financial, weather and remote sensing datasets, the occurrence

Table 1
Comparing small and big data

	Small data	Big data
Volume	Limited to large	Very large
Velocity	Slow, freeze-framed/bundled	Fast, continuous
Variety	Limited to wide	Wide
Exhaustivity	Samples	Entire populations
Resolution and identification	Course & weak to tight & strong	Tight & strong
Relationality	Weak to strong	Strong
Flexible and scalable	Low to middling	High

of big data is largely a post-millennium phenomena enabled by: advances in computational power; pervasive, ubiquitous and mobile computing; networked storage; new forms of database design; new modes of software-mediated communication, interactions and transactions; and data analytics that utilise machine learning and are able to cope with a data deluge. To date, official statistical data have been small data, holding some of the characteristics of big data but not all. For example, a census has volume, exhaustivity, resolution, and relationality, but has no velocity (generated once every five or ten years), no variety (usually c.30 structured questions), and no flexibility (once set a census cannot be altered mid data generation). Most other official statistical data lack exhaustivity using sampling frameworks to selectively represent populations. In comparison, mobile phone companies are logging millions of calls and associated metadata every hour, large supermarket chains are handling hundreds of thousands of customer transactions an hour, traffic sensors are tracking hundreds of thousands of vehicles a day as they navigate cities, and social media companies are processing billions of interactions a day. In each case the data relate to the entire population of that system, are often resolute relating to specific customers and transactions, and in the case of social media can be highly varied including text, photos, videos, sound files and weblinks.

Not unsurprisingly, given its scope, timeliness, and resolution, and the potential efficiencies it offers in the resourcing and compiling of data and statistics, big data have captured the interest of NSIs and related agencies such as Eurostat, the European Statistical System (ESS), the United Nations Economic Commission for Europe (UNECE), and the United Nations Statistical Division (UNSD). In 2013 the Heads of the National Statistical Institutes of the EU signed the Scheveningen Memorandum [25] to examine the use of big data in official statistics. However, a survey jointly conducted by UNSD and UNECE revealed that of the 32 NSIs that responded 'only a few countries have developed a long-term vision for the use of Big Data', or 'established internal labs, task teams or working groups to carry out pilot projects to determine if and how Big Data could be used as a source of Official Statistics' [31, p. 16]. Some are 'currently on the brink of formulating a Big Data strategy' ... 'but most countries have not yet defined business processes for integrating Big Data sources and results into their work and do not have a defined structure for managing Big Data projects' [31, p. 16]. As these organisations are discovering, whilst big data offer a number of opportunities for NSIs, they also offer a series of challenges and risks that are not easy to handle and surmount. Indeed, the use of big data needs careful consideration to ensure that they do not compromise the integrity of NSIs and their products. The rest of the paper discusses these opportunities, challenges and risks, which are summarized in Table 2.

2. Opportunities

Clearly the key opportunity of big data is the availability of new sources of dynamic, resolute data that can potentially complement, replace, improve, and add to existing datasets and refine existing statistical composition, and produce more timely outputs. Indeed, Florescu et al. [11, pp. 3–4] detail that big data sources could be used in current statistical systems in five ways:

1. To entirely replace existing statistical sources such as surveys (*existing* statistical outputs);
2. To partially replace existing statistical sources such as surveys (*existing* statistical outputs);
3. To provide complementary statistical information in the same statistical domain but from other perspectives (*additional* statistical outputs);
4. To improve estimates from statistical sources (including surveys) (*improved* statistical outputs);
5. To provide completely new statistical information in a particular statistical domain (*new alternative* statistical outputs).

Table 2
Opportunities, challenges and risks of big data for official statistics

Opportunities	Challenges	Risks
– Complement, replace, improve, and add to existing datasets – Produce more timely outputs – Compensate for survey fatigue of citizens and companies – Complement and extend micro-level and small area analysis – Improve quality and ground truthing – Refine existing statistical composition – Easier cross-jurisdictional comparisons – Better linking to other datasets – New data analytics producing new and better insights – Reduced costs – Optimization of working practices and efficiency gains in production – Redeployment of staff to higher value tasks – Greater collaboration with computational social science, data science, and data industries – Greater visibility and use of official statistics	– Forming strategic alliances with big data producers – Gaining access to data, procurement and licensing – Gaining access to associated methodology and metadata – Establishing provenance and lineage of datasets – Legal and regulatory issues, including intellectual property – Establishing suitability for purpose – Establishing dataset quality with respect to veracity (accuracy, fidelity), uncertainty, error, bias, reliability, and calibration – Technological feasibility – Methodological feasibility – Experimenting and trialing big analytic – Institutional change management – Ensuring inter-jurisdictional collaboration and common standards	– Mission drift – Damage to reputation and losing public trust – Privacy breaches and data security – Inconsistent access and continuity – Resistance of big data providers and populace – Fragmentation of approaches across jurisdictions – Resource constraints and cut-backs – Privatisation and competition

To these, Tam and Clarke [28, pp. 8–9] add:

- Sample frame or register creation – identifying survey population units and/or providing auxiliary information such as stratification variables;
- Imputation of missing data items – substituting for same or similar units;
- Editing – assisting the detection and treatment of anomalies in survey data;
- Linking to other data – creating richer datasets and/or longitudinal perspectives;
- Data confrontation – ensuring the validity and consistency of survey data;
- Improving the operational efficiency and effectiveness of NSIs through use of paradata created and captured from its statistical operations.

Significantly, big data offer the opportunity to produce more timely official statistics, drastically reducing processing and calculating processes, and the ability to do so on a rolling basis [10]. For example, rather than it taking several weeks to produce quarterly statistics (such as GDP), it might take a few minutes or hours, with the results being released on the same timescale on a rolling basis. In this sense, big data offer the possibility for 'nowcasting', the prediction of the present [5, p. 1]. For Global Pulse [12, p. 39] the timeliness of big data enables:

1. "Early warning: early detection of anomalies in how populations use digital devices and services [which] can enable preventive interventions;
2. Real-time awareness: a fine-grained and current representation of reality which can inform the design and targeting of programs and policies;
3. Real-time feedback: real time monitoring makes it possible to understand where policies and programs are failing and make the necessary adjustments in a more timely manner."

In the developing world, where the resourcing of NSIs has often been limited and traditional surveys are sometimes viewed as cumbersome, expensive and of limited effectiveness, or they are affected by other external influences (political pressure, war, etc.), big data are seen as a means of filling basic gaps in official statistics and of by-passing political bottlenecks to statistical reform [2,12,17,21]. Such an aspiration is also relevant to the developed world in cases where official statistics are difficult to produce, or are methodologically weak, or lack adequate granularity and disaggregation (spatially, temporally). Indeed, big data offer a rich source of granular data, often at the level of unique individuals, households or companies, to complement and extend micro-level and small area analysis [23].

Further, big data are the outputs of direct measurement of a phenomenon and provide a reflection of actual transactions, interactions and behaviour of people, societies, economies and systems, rather than surveys which reflect what people say they do or think. Thus while big datasets can be noisy, and contain gamed and faked data, they potentially pose more ground truth with respect to social reality on some issues than cur-

Table 3
Potential use of big data in official statistics

Data source	Data type	Statistical domains
Mobile communication	Mobile phone data	Tourism statistics Population statistics
WWW	Web searches	Labour statistics Migration statistics
	E-commerce websites	Price statistics
	Businesses' websites	Information society statistics Business registers
	Job advertisements	Employment statistics
	Real-estate websites	Price statistics (real estate)
	Social media	Consumer confidence; GDP and beyond; information society statistics
Sensors	Traffic loops	Traffic/transport statistics
	Smart meters	Energy statistics
	Satellite images	Land use statistics; agricultural statistics; environment statistics
	Automatic vessel identification	Transport and emissions statistics
Transactions of process generated data	Flight movements	Transport and emissions statistics
	Supermarket scanner and sales data	Price statistics Household consumption statistics
Crowdsourcing	Volunteered geographic information (VGI) websites (OpenStreetMap, Wikimapia, Geowiki)	Land use
	Community pictures collections (flickr, Instagram, Panoramio)	–

Source: ESSC [9, p. 18].

rent instruments used for official statistics [13]. And since the big data being produced are an inherent part of the systems that generate them, they can compensate for significant survey fatigue amongst citizens and companies [27]. Moreover, since big data are generated from systems that often span or are deployed in many jurisdictions – unlike much data derived from surveys or administrative systems – they potentially ensure comparability of phenomena across countries.

An additional advantage is that big data offer the possibility to add significant value to official statistics at marginal cost, given that data are already being produced by third parties [1,7,18,27]. Indeed, it could lead to greater optimization of working practices, efficiency gains in production, and a redeployment of staff away from data generation and curation to higher value tasks such as analysis or quality assurance, communication or developing new products. It also has the potential to lead to greater collaboration with computational social science, data science, and data industries, leading to new insights and innovations, and a greater visibility and use of official statistics as they become more refined, timely and resolute. Further, new data analytics, utilising machine learning to perform data mining and pattern recognition, statistical analysis, prediction, simulation, and optimization, data visualization and visual analytics, mean that greater insights might be extracted from existing statistical data and new sources of big data, and new derived data and statistical products can be developed [24]. In a scoping exercise, the European Statistical System Committee (ESSC) [9, p. 8] has thus identified several official statistical domains that could be profitably augmented by the use of different kinds of big data (see Table 3).

3. Challenges

Whilst big data offer a number of opportunities their use is not without a number of significant challenges. A first issue is to gain access to the required big data in the first place for assessment, experimenting, trialing and adoption [10,12,28]. Although some big data are produced by public agencies, such as weather data, some website and administrative systems, and some transport data, much big data are presently generated by private companies such as mobile phone, social media, utility, financial, credit, insurance and retail companies [16]. These big data are valuable commodities to these companies, either providing a resource that generates competitive advantage or constituting a key product, and are generally not publicly available for of-

ficial or public analysis in raw or derived forms. For NSIs to gain access to such data requires forming binding strategic partnerships with these companies (so-called 'data compacts'; Krätke and Byiers [17]) or creating/altering legal instruments (such as Statistics Acts) to compel companies to provide such data. Such negotiations and legislative reform is time consuming and politically charged, especially when NSIs generally do not pay or compensate companies for providing data for official statistics.

Once data have been sourced, they need to be assessed for their suitability for complementing, replacing or adding to official statistics. This assessment concerns suitability for purpose, technological and methodological feasibility, and the change management required for implementation. From the perspective of both NSIs and the public, official statistics are generated: (a) with the purpose to serve the whole spectrum of the society; (b) based on quality criteria and best practices; (c) by statisticians with assured professional independence and objectivity [10]. However, unlike the surveys administered by NSIs, in most cases the big data listed in Table 3 are generated by commercial entities for their specific needs and were never intended to be used for the production of official statistics. The extent to which repurposed big data provide adequate, rigorous and reliable surrogates for more targeted, sampled data therefore needs to be established [27]. A key consideration in this respect is representativeness, both of phenomena and populations [6,12,28]. NSIs carefully set their sampling frameworks and parameters, whereas big data although exhaustive are generally not representative of an entire population as they only relate to whomever uses a service. For example, credit card data only relate to those that possess a credit card and social media data only relate to those using that service, which in both cases are stratified by social class and age (and in the latter case also includes many anonymous and bot accounts) and may represent a geographically uneven picture within a nation, potentially favouring urban areas (where there is a critical mass customers and infrastructure) over rural and remote areas. In cases such as the Consumer Price Index the same bundle of goods and services with statistically determined weights need to be tracked over time, rather than simply web-scraping a largely undefined unbundle [14]. There is a challenge then in using big data in the context of existing methodologies.

Further, NSIs spend a great deal of effort in establishing the quality and parameters of their datasets with respect to veracity (accuracy, fidelity), uncertainty, error, bias, reliability, and calibration, and documenting the provenance and lineage of a dataset. The OECD [22] measures data quality across seven dimensions: relevance, accuracy, credibility, timeliness, accessibility, interpretability, and coherence. These qualities are largely unknown with respect to various forms of big data [23,30], though it is generally acknowledged that the datasets can be full of dirty, gamed and faked data as well as datasets being incomplete [6,16]. Further, their generators are reluctant to share methodological transparency in how they were produced and processed. In addition, the frames within which big data are generated can be mutable, changing over time. For example, Twitter and Facebook are always tweaking their designs and modes of interaction, and often present different users with alternate designs as they perform A/B testing on the relative merits of different interface designs and services. The data created by such systems are therefore inconsistent across users and/or time. These issues, created through the differences in characteristics of big data from the survey and administrative data usually used in official statistics (see Table 4), raise significant questions concerning the suitability of big data for official statistics and how they might be assessed and compensated for [28]. For some, the initial foray should only be to explore the potential of using big data to improve the quality of estimates within current methodological frameworks and to assess the levels and causes of sampling and non-sampling errors across data sources that threaten valid inference [14].

Once the suitability of the data is established, an assessment needs to be made as to the technological feasibility regarding transferring, storing, cleaning, checking, and linking big data, and conjoining the data with existing official statistical datasets (Scannapieco et al. [24,27,28]). As Cervera et al. [4] note, at present, there is a lack of user-friendly tools for big data that make it difficult to engage with and it is difficult to integrate big data into present workflows and big data infrastructure with existing infrastructure. In particular, there is a real challenge of developing techniques for dealing with streaming data, such as processing such data on the fly (spotting anomalies, sampling/filtering for storage) [24]. Moreover, there are questions concerning the methodological feasibility of augmenting and producing official statistics using big data and performing analytics on a constant basis as data are dynamically generated, in order to produce real-time statistics or visualisations.

Table 4
Characteristics of survey, administrative and big data

	Survey data	Administrative data	Big data
Specification	Statistical products specified ex-ante	Statistical products specified ex-post	Statistical products specified ex-post
Purpose	Designed for statistical purposes	Designed to deliver/monitor a service or program	Organic (not designed) or designed for other purposes
Byproducts	Lower potential for by-products	Higher potential for by-products	Higher potential for by-products
Methods	Classical statistical methods available	Classical statistical methods available, usually depending on the specific data	Classical statistical methods not always available
Structure	Structured	A certain level of data structure, depending on the objective of data collection	A certain level of data structure, depending on the source of information
Comparability	Weaker comparability between countries	Weaker comparability between countries	Potentially greater comparability between countries
Representativeness	Representativeness and coverage known by design	Representativeness and coverage often known	Representativeness and coverage difficult to assess
Bias	Not biased	Possibly biased	Unknown and possibly biased
Error	Typical types of errors (sampling and non-sampling errors)	Typical types of errors (non-sampling errors, e.g., missing data, reporting errors and outliers)	Both sampling and non-sampling errors (e.g., missing data, reporting errors and outliers) although possibly less frequently occurring, and new types of errors
Persistence	Persistent	Possibly less persistent	Less persistent
Volume	Manageable volume	Manageable volume	Huge volume
Timeliness	Slower	Potentially faster	Potentially must faster
Cost	Expensive	Inexpensive	Potentially inexpensive
Burden	High burden	No incremental burden	No incremental burden
Geography	National, defined	National or extent of program and service	National, international, potentially spatially uneven
Demographics	All or targeted	Service users or program recipients	Consumers who use a service, pass a sensor, contribute to a project, etc.
Intellectual Property	State	State	Private Sector

Adapted and extended from Florescu et al. [11, pp. 2–3].

A key challenge in managing these developments is the implementation of a change management process to fully prepare the organisation for taking on new roles and responsibilities. New data life cycle systems need to be established and implemented, accompanied by the building and maintenance of new IT infrastructure capable of handling, processing and storing big data [7]. These new systems need to ensure data security and compliance with data protection. They also need to be adequately resourced, creating demands for additional finance and skilled staff.

4. Risks

Given the various challenges set out above, along with general public and institutional perceptions and reactions to the use of big data, there are a number of risks associated with using big data in producing official statistics. The key risks relate to mission drift, reputation and trust, privacy and data security, access and continuity, fragmentation across jurisdictions, resource constraints and cut-backs, and privatisation and competition.

The key mission for NSIs is to produce useful and meaningful official statistics. Traditionally, the driver of what statistics have been produced has been a key concern or question; data have been generated in order to answer a specific set of queries. In the era of big data there is the potential for this to be reversed, with the abundance and cost benefit of big data setting the agenda for what is measured. In other words, official statistics may drift towards following the data, rather than the data being produced for the compilation of of-

ficial statistics. As well as having implications to the institutional work of NSIs, there is a clear threat to integrity and quality of official statistics in such a move. It is absolutely critical therefore that NSIs remain focused on the issues and questions data are used to address, assessing the suitability of big data to their core business, rather than letting big data drive their mission.

A critical risk for NSIs in implementing a new set of means and methods for producing official statistics is their reputation and public trust being undermined. A reputation as a fair, impartial, objective, neutral provider of high quality official statistics is seen by NSI's as a mission critical quality, and is usually their number one priority in their institutional risk register. Partnering with a commercial third party and using their data to compile official statistics exposes the reputation of a NSI to that of the partner. A scandal with respect to data security and privacy breaches, for example, may well reflect onto the NSI [7]. Further, failing to adequately address data quality issues will undermine confidence in the validity and reliability of official statistics, which will be difficult to re-establish. Similarly, given big data are being repurposed, often without the explicit consent of those the data represent, there is the potential for a public backlash and resistance to such re-use. It also has to be recognized, however, that it is a lack of trust in government both in the developed [3] and particularly the developing world [21] with respect to competence and motive which is driving some calls for the work of NSIs to be complemented or replaced by opening government data to enable replication and new analysis and the use of big data.

Related to reputation, but a significant risk in its own right is the infringement of privacy and breaching of data security. NSIs take privacy and security very seriously acting as trusted repositories that employ sophisticated systems for managing data, using strategies such as anonymisation and aggregation, access rules and techniques, and IT security measures, to ensure confidentiality and security. These systems are designed to work with carefully curated 'small' datasets. Big data increases the challenge of securing data by providing new forms of voluminous, relational data, new types of systems and databases, and new flows of data between institutions. There is therefore a need to establish fresh approaches that ensure the security integrity of the big data held by NSIs [1,4,18,27]. To this end, UNECE [30, p. 3] suggest that in addition to the dimensions used to assess administrative data, five new dimensions should be added: '*privacy* and *confidentiality* (a thorough assessment of whether the data meets privacy requirements of the NSO), *complexity* (the degree to which the data is hierarchical, nested, and comprises multiple standards), *completeness* (of metadata) and *linkability* (the ease with which the data can be linked with other data)'. The first two are important to prevent privacy being breached or data being stolen and used for nefarious ends. As the Wikileaks and Snowden scandals and other data breaches have demonstrated public trust in state agencies and their handling and use of personal data have already been undermined. Likewise a series of high profile breaches of private company data holdings, such as the stealing of credit card or personal information, has reduced public confidence in data security more widely. A similar scandal with respect to a NSI could be highly damaging, and potentially contagious to other NSIs.

At present, NSIs gain their data through dedicated surveys within their control and administrative databases which they access through legislative mandate. They have little control or mandate with respect to big data held by private entities, however. In partnering with third parties NSIs lose overall control of generation, sampling, and data processing and have limited ability to shape the data produced [18], especially in cases where the data are the exhaust of a system that are being significantly repurposed. This raises questions concerning procurement, intellectual property regimes, licensing, assurance, managing quality. A key risk is that access to the desired data on a voluntary or licensed basis is denied by companies who do not want to lose competitive advantage, share a valuable asset without financial compensation, or have the responsibility or burden of supplying such data, or that initially negotiated access is then discontinued [18]. The latter poses a significant risk to data continuity and time-series datasets if existing systems have been replaced by the new big data solution. It may be possible to mandate companies to provide access to appropriate data using legal instruments, but it is likely that such a mandate will be strongly resisted and legally challenged by some companies across jurisdictions. In cases where companies are compelled unwillingly to share data there has to be a process by which to validate and assure the quality of the data prepared for sharing.

A key issue for the compilation of supra-national statistics and benchmarking is finding comparable datasets. NSIs have traditionally been responsible for developing their statistical systems. While there has long been a swapping of knowledge and best prac-

tice, each NSI produced official statistics are defined by their statisticians, framed by public administrative needs and context. The result has been a patchwork quilt of different definitions, methods, protocols and standards for producing official statistics, so that while the data generated are similar, they are not the same. For example, how unemployment is defined and measured often varies across jurisdiction. There is a distinct risk of perpetuating this situation with respect to official statistics derived from big data creating a fragmented and non-comparable datasets.

Big data offer the potential to create efficiencies in the production of official statistics. There is a risk, however, of governments viewing the use of big data as a means of reducing staffing levels and cutting costs. This is particularly the case in a time of austerity and a strong neoliberal ethos dominating the political landscape of many jurisdictions. While there are some very real possibilities of rationalisation, especially with respect to casual and part-time staffing of censuses and surveys, the core statistical and technical staffing of NSIs need to be maintained, and may need to be expanded in the short-term given the potential to create new suites of statistics that need testing, validation, and continuous quality control checks. Indeed, there will be a need to develop new technical and methodological skills within NSIs, including creating expertise in new data analytics, as well as soft skills and knowledge related to big data and procurement, law, privacy, data protection, regulation, copyright, and intellectual property (IP), either through retraining or recruitment [1,4,20]. Without such investment, NSIs will struggle to fully exploit the potential benefits of utilising big data for official statistics. Any reductions in staffing and resources, especially before big data has been fully integrated into the workflow of NSIs, is likely to place serious strain on the organisation and threaten the integrity of the products produced.

A final risk is competition and privatisation. If NSIs choose to ignore or dismiss big data for compiling useful statistical data then it is highly likely that private data companies will fill the gap, generating the data either for free distribution (e.g. Google Trends) or for sale. They will do so in a timeframe far quicker (near real-time) than NSIs are presently working, perhaps sacrificing some degree of veracity for timeliness, creating the potential for lower quality but more timely data to displace high quality, slower data [10]. The result may be a proliferation of alternative official statistics produced by a variety of vendors, each challenging the veracity and trustworthiness of those generated by NSIs [21]. Data brokers are already taking official statistical data and using them to create new derived data, combining them with private data, and providing valued-added services such as data analysis. They are also producing alternative datasets, registers and services, combining multiple commercial and public datasets to produce their own private databanks from which they can produce a multitude of statistics and new statistical products. For example, Acxiom is reputed to have constructed a databank concerning 500 million active consumers worldwide (about 190 million individuals and 126 million households in the United States), with about 1,500 data points per person, its servers processing over 50 trillion data transactions a year [26]. It claims to be able to provide a '360 degree view' of consumers by meshing offline, online and mobile data, using these data to create detailed profiles and predictive models [26]. Such organisations are also actively campaigning to open up the administrative datasets used by NSIs to produce official statistics, arguing that they and others could do much more with them, and in a much more efficient and effective way [3].

For NSIs that partially operate as using a cost recovery model , that is they generate additional income to support their activities from the sale of specialist derived data and services, opening data and the operations of data brokers will increasingly threaten revenue streams. As with other aspects of the public sector it may also be the case that governments will look to privatise certain competencies or datasets of NSIs. This has happened in some jurisdictions with respect to other public data agencies such as mapping institutions, notably the UK where Ordnance Survey is increasingly reliant on the sale and licensing of geospatial data and the postcode dataset has been recently privatised with the sale of Royal Mail. Such a neoliberal move has the potential to undermine trust in official statistics and threatens making and maintaining open datasets.

5. The way forward

The advent of relatively widely generated big data across domains has created a set of disruptive innovations from which NSIs are not exempt given their role as key data providers and authorities for official statistics. Indeed, Letouzé and Jütting [21] argue that "engaging with Big Data is not a technical consideration but a political obligation. It is an imperative to retain,

or regain, their primary role as the legitimate custodian of knowledge and creator of a deliberative public space." At the same time, as Cervera et al. [4, p. 37] argue "Big Data should reduce, not increase statistical burden ... Big Data should increase, not reduce statistical quality." As elaborated above, big data presents a number of opportunities, challenges and potential risks to NSIs and it is clear that they need to formulate a strategic and operational response to their production. Moreover, beyond the work of any individual NSI, responses need to be coordinated and aligned across jurisdictions so that the new official statistics produced by NSIs are comparable across space and time and can be conjoined to produce larger supra-national datasets. The challenge here is institutional and political in nature and requires significant levels of dialogue, collaboration and coordination across NSIs to establish a common strategic and operational position and new standardized approaches to leveraging big data for official statistics.

This work has already begun with the ESS, UNECE, Eurostat and UNSD taking leading roles. For example, the ESS have formulated a big data roadmap, UNECE have established a High Level Group for the Modernization of Statistical Production and Services focused on big data, with four 'task teams' (privacy, partnerships, sandbox and quality), and the United Nations Statistical Division (UNSD) have organized a Global Working Group on Big Data and Official Statistics (comprising of representatives from 28 developed and developing countries) [31]. The Scheveningen Memorandum [25] commits European NSIs to setting out a roadmap that will be integrated into the statistical annual work programmes of Eurostat. A very welcome development has been the creation of a big data 'sandbox' environment, hosted in Ireland by the Central Statistics office (CSO) and the Irish Centre for High-End Computing (ICHEC), that provides a technical platform to:

> "(a) test the feasibility of remote access and processing – Statistical organisations around the world will be able to access and analyse Big Data sets held on a central server. Could this approach be used in practice? What are the issues to be resolved?;
> (b) test whether existing statistical standards/models/methods etc. can be applied to Big Data;
> (c) determine which Big Data software tools are most useful for statistical organisations;
> (d) learn more about the potential uses, advantages and disadvantages of Big Data sets – "learning by doing";
> (e) build an international collaboration community to share ideas and experiences on the technical aspects of using Big Data." (UNECE 2014).

In 2014, approximately 40 statisticians/data scientists from 25 different organisations were working with the sandbox [8]. Over time, the sandbox could potentially develop into a Centre of Excellence and non-for-profit pan-NSI big data service provider, delivering comparable statistical information across jurisdictions [8].

However, it is evident that NSIs and associated agencies are only at the start of the process of engaging with, testing and assessing, and thinking through the implications of big data to the production of statistics and the organisation and work of NSIs. Consequently, while there has been some notable progress since 2013, as set out above, there are a still a number of open issues that require much thinking, debate, negotiation, and resolution. And, as made clear in the sessions and discussion at the New Techniques and Technologies for Statistics conference in Brussels in March 2015, there is a wide divergence of opinions across official statisticians as to relative merits of big data and its potential opportunities and risks and how best for NSIs to proceed.

It is clear that the initial approach adopted needs to continue apace, with the international community of NSIs working through the challenges and risks presented in this paper to find common positions on:

- Conceptual and operational (management, technology, methodology) approach and dealing with risks;
- Other roles NSIs might adopt in the big data landscape such as becoming the arbiters or certifiers of big data quality within any emerging regulatory environment, especially for those used in official statistics, or become clearing houses for statistics from non-traditional sources that meet their quality standards [4,18,23,27];
- Resolving issues of access, procurement, licensing, and standards;
- Identifying and tackling privacy, ethics, security, legal, and governanceissues;
- Undertaking experimentation and trialing;
- Establishing best practices for change management from the short to long-term which will ensure stable institutional transitions, the maintenance of the high standards of quality, and continuity of statistics over time and across jurisdiction; and,

– Political lobbying with respect to resourcing.

To this end, alliances could be profitably forged with other international bodies that are wrestling with the same kinds of issues such as the Research Data Alliance (RDA) and the World Data System (WDS) to share knowledge and approaches.

At present, NSIs are in reactive mode and are trying to catch up with the opportunities, challenges and risks of big data. It is important that they not only catch up but get ahead of the curve, proactively setting the agenda and shaping the new landscape for producing official statistics. There is, however, much work to be done before such a situation is achieved.

Acknowledgements

The author is grateful to Padraig Dalton, Richie McMahon and John Dunne from the Central Statistics Office, Ireland, for useful discussion concerning big data and NSIs, Tracey Lauriault and Pat O'Hara for reading and commenting on an initial draft, and also audience feedback from the New Techniques and Technologies for Statistics conference held in Brussels, March 10–13, 2015 where this paper was presented as a keynote talk. The research for this paper was funded by a European Research Council Advanced Investigator award (ERC-2012-AdG-323636-SOFTCITY).

References

[1] AAPOR (2015) *AAPOR Report on Big Data*. American Association for Public Opinion Research, Deerfield, IL, http://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15.pdf (last accessed 1 April 2015).

[2] J.R.G. Albert, (2013) *Big Data: Big Threat or Big Opportunity for Official Statistics?* 18th October. Philippines Statistical Authority, National Statistical Coordination Board. http://www.nscb.gov.ph/statfocus/2013/SF_102013_OSG_bigData.asp (last accessed 16th March 2015).

[3] B. Casselman, (2015) Big government is getting in the way of big data. *Five Thirty Eight Economics*. March 9th. http://fivethirtyeight.com/features/big-government-is-getting-in-the-way-of-big-data/ (last accessed 16th March 2015).

[4] J.L. Cervera, P. Votta, D. Fazio, M. Scannapieco, R. Brennenraedts and T. Van Der Vorst, *Big data in official statistics*. ESS Big Data Event in Rome 2014, Technical Workshop Report. http://www.cros-portal.eu/sites/default/files/Big%20Data%20Event%202014%20-%20Technical%20Final%20Report%20-finalV01_0.pdf (last accessed 20 March 2015).

[5] H. Choi and H. Varian, Predicting the present with Google Trends. *Google Research*. http://people.ischool.berkeley.edu/∼hal/Papers/2011/ptp.pdf (last accessed 1 April 2015), (2011).

[6] P.J.H. Daas, M.J. Puts, B. Buelens and P.A.M. van den Hurk, (2013) *Big Data and Official Statistics*. http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_76.pdf (last accessed 1 April 2015).

[7] J. Dunne, (2013) Big data coming soon ... to an NSI near you. Paper presented at the World Statistics Congress, 25th-30th, Hong Kong. http://www.statistics.gov.hk/wsc/STS018-P3-A.pdf (last accessed 1st April 2015).

[8] J. Dunne, (2014) Big data . . . now playing at "the sandbox". Paper presented at The International Association for Official Statistics 2014 conference, 8-10 October, Da Nang, Vietnam.

[9] ESSC (2014) *ESS Big Data Action Plan and Roadmap 1.0*. European Statistical System Committee, 26th September 2014. http://www.cros-portal.eu/sites/default/files//ESSC%20doc%2022_8_2014_EN_Final%20with%20ESSC%20opinion.pdf (last accessed 8 May 2015).

[10] Eurostat (2014) *Big data – an opportunity or a threat to official statistics?* Paper presented at the Conference of European Statisticians, 62nd plenary session, Paris, 9–11 April 2014. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2014/32-Eurostat-Big_Data.pdf (last accessed 20 March 2015).

[11] D. Florescu, M. Karlberg, F. Reis, P.R. Del Castillo, M. Skaliotis and A. Wirthmann, (2014) *Will 'big data' transform official statistics?* http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf (last accessed 1 April 2015).

[12] Global Pulse (2012) *Big data for development: Challenges and opportunities*. May 2012. Global Pulse, New York. http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf (last accessed 1st April 2015).

[13] D.J. Hand, (2015) Official Statistics in the New Data Ecosystem. Paper presented at the New Techniques and Technologies in Statistics conference, Brussels, March 10–12. http://www.cros-portal.eu/sites/default/files//Presentation%20S20AP2%20%20Hand%20-%20Slides%20NTTS%202015.pdf (last accessed 7 April 2015).

[14] M.W. Horrigan, (2013) Big Data: A Perspective from the BLS. 1st January. *Amstat News*. http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/ (last accessed 16th March 2015).

[15] R. Kitchin, Big data and human geography: Opportunities, challenges and risks, *Dialogues in Human Geography* **3**(3) (2013), 262–267.

[16] R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage, London, (2014).

[17] F. Krätke and B. Byiers, The Political Economy of Official Statistics Implications for the Data Revolution in Sub-Saharan Africa, (2014). *PARIS21, Partnership in Statistics for Development in the 21st Century Discussion Paper No. 5*, December. http://ecdpm.org/publications/political-economy-official-statistics-implications-data-revolution-sub-saharan-africa/ (last accessed 1 April 2015).

[18] S. Landefeld, Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues. Discussion Paper at International Conference on Big Data for Official Statistics, Beijing, China, 28–30 Oct 2014 (last accessed 1 April 2015).

[19] D. Laney, (2001) 3D Data Management: Controlling Data Volume, Velocity and Variety. *Meta Group*. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Manageme

nt-Controlling-Data-Volume-Velocity-and-Variety.pdf (last accessed 16th January 2013).

[20] T. Lauriault, Critical Analysis of the Irish Big Data Skills Report. *Programmable City blog*, May 2014, http://www.maynoothuniversity.ie/progcity/2014/05/big-data-series-critical-analysis-of-the-irish-big-data-skills-report/ (last accessed 8 May 2015).

[21] E. Letouzé and J. Jütting, Official Statistics, Big Data and Human Development: Towards a New Conceptual and Operational Approach. *Data-Pop Alliance White Paper Series*. (2014), Paris 21. 14th December. http://www.datapopalliance.org/s/WhitePaperBigDataOffStatsNov17Draft.pdf (last accessed 16th March 2015).

[22] OECD (2011) *Quality Framework and Guidelines for OECD Statistical Activities*, updated 17th January 2012. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=std/qfs%282011%291&doclanguage=en.

[23] C. Reimsbach-Kounatze, The Proliferation of "Big Data" and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis, *OECD Digital Economy Papers, No. 245*, (2015), OECD Publishing. http://dx.doi.org/10.1787/5js7t9wqzvg8-en (last accessed 16th March 2015).

[24] M. Scannapieco, A. Virgillito and D. Zardetto, Placing Big Data in Official Statistics: A Big Challenge? Paper presented at New Techniques and Technologies in Statistics, (2013). http://www.cros-portal.eu/sites/default/files//NTTS2013fullPaper_214.pdf (last accessed 1 April 2015).

[25] Scheveningen Memorandum (2013) *Big Data and Official Statistics*. http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13 (last accessed 1 April 2015).

[26] N. Singer, (2012) You for Sale: Mapping, and Sharing, the Consumer Genome. *New York Times*, 17$^{\text{th}}$ June, http://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html (last accessed 11th October 2013).

[27] P. Struijs, B. Braaksma and P.J.H. Daas, Official statistics and Big Data, *Big Data & Society* **1**(1) (2014), 1–6. (last accessed 20 March 2015).

[28] S.-M. Tam and F. Clarke, Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics. Paper presented at International Conference on Big Data for Official Statistics, Beijing, China, 28–30 Oct 2014 http://unstats.un.org/unsd/trade/events/2014/Beijing/documents/other/Australia%20Bureau%20of%20Statistics%20-%20Some%20initiatives%20on%20Big%20Data%20-%2023%20July%202014.pdf (last accessed 1 April 2015).

[29] UNECE (2014a) *Sandbox*. 24th April. United Nations Economic Commission for Europe. http://www1.unece.org/stat/platform/display/bigdata/Sandbox (last accessed 16th March 2015).

[30] UNECE (2014b) *A Suggested Framework for National Statistical Offices for assessing the Quality of Big Data*. United Nations Economic Commission for Europe. http://www1.unece.org/stat/platform/download/attachments/108102944/Big Data Quality Framework Abstract for NTTS.docx (last accessed 16th March 2015).

[31] UNESC (2015) *Report of the Global Working Group on Big data for official statistics*. United National Economic and Social Council. http://unstats.un.org/unsd/statcom/doc15/2015-4-BigData.pdf (last accessed 16th March 2015).