Will 'big data' transform official statistics?

Denisa Florescu, Martin Karlberg, Fernando Reis, Pilar Rey Del Castillo, Michail Skaliotis and Albrecht Wirthmann[1]

Abstract

Official Statistics, confronted by the dynamic explosion of big data, are in the early stages of a fundamental paradigm shift which will eventually usher in a new era in the statistical profession and in the role of national statistical institutes (NSIs). Will certification and accreditation of new data sources become a mark of quality for official statistics in the future? What will post-2020 population censuses and future surveys look like? Will NSIs build partnerships with owners of private data sources? This paper addresses some of the core issues at the intersection of Big Data and official statistics. Our aim is to provoke discussion on the potential power of big data to transform official statistics.

I. INTRODUCTION

1. Big data influences almost all aspects of everyday life in ways which would have been unimaginable a few years ago. Quantitative disciplines and sciences, and the policy-making which draws upon them, are in the front line. Big data has posed a challenge and won't be ignored. Often the only way to accommodate big data is by means of a dramatic restructuring and a radically changed approach towards the production of official statistics. Official statistics is one area where big data is making its presence felt. Recent initiatives, debates and literature on the subject of how official statistics should evolve in the age of big data if they are to remain relevant demonstrate that the official statistics community is about to undergo a significant change [5].

2. Whereas, up to a few years ago, big data was on the whole rarely discussed within NSIs, today, official statisticians are entering into the debate on big data with enthusiasm. Statistical agencies around the world are allocating resources to the area, setting up trial projects, and adopting programmes and strategies [12].

3. The leading statisticians of Europe expressed their commitment to collectively addressing the challenges presented by big data in the so-called 'Scheveningen Memorandum' last September, which addressed the relevance of big data for the European Statistical System (ESS).[2]

4. To kick start discussions on the various issues raised in the Scheveningen Memorandum, Eurostat has structured the programme of the 2014 Big Data Event[3] around the challenges presented in the Memorandum. In contrast to the wide-ranging discussions expected at the Big Data Event, the scope of this paper is more narrowly

[1] All authors are affiliated with **Eurostat, L-2920 Luxembourg**. Corresponding author: Denisa.Florescu@ec.europa.eu. The views expressed here are those of the authors and do not necessarily reflect the official views of the European Commission (Eurostat).

[2] http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM%20Final%20version_0.pdf.

[3] http://www.cros-portal.eu/content/big-data-event-2014.

focused, the main areas covered being the Scheveningen Challenge 6 (SCH6), concerning the need for synergies with *inter alia* the owners of private data sources, and the Scheveningen Challenge 7 (SCH7), concerning the need for developments in *inter alia* methodologies and quality assessment. Section II of this paper addresses the methodological aspect of SCH7, covering the potential ways of extracting valuable information from big data and of integrating big data into official statistics, while Section III relates to the quality assessment aspect of SCH7, considering in particular a possible accreditation procedure for big data sources, and the relevance of this to the synergies discussed in SCH6, in terms of the value of accreditation and certification procedures for big data providers.

5. Whilst drafting this paper, we discovered 'Big Data maturity model'[4] of the Netherlands Organisation for Applied Scientific Research (TNO), which can be used by other organisations to help them to determine their initial position and to develop a big data strategy by which to reach the desired maturity level. While there is a great deal of variation between various NSIs in terms of their big data maturity, we consider it realistic to develop an EU big data strategy for official statistics, given the underlying dynamics of a network of statistical agencies such as the ESS. This optimism is further strengthened by the fact that big data poses very similar challenges for any NSI, and requires a very similar approach in terms of organisational and technical capabilities.

II. INTEGRATING BIG DATA INTO OFFICIAL STATISTICS

6. Experience gained from using survey and administrative data, and the similarities between specific features of data from these sources and big data may help to show the way in terms of how to extract valuable information from the huge range of big data available and how to integrate big data into official statistics.

A. Main features of survey, administrative and big data

7. The following table compares the main features of all three data sources:

Table 1

Feature code	Survey data	Administrative data	Big data
F1	Statistical products specified ex-ante	Statistical products specified ex-post	Statistical products specified ex-post
F2	Designed for statistical purposes	Designed for other purposes	Organic (not designed) or designed for other purposes
F3	Lower potential for by-products	Higher potential for by-products	Higher potential for by-products

[4] http://repository.tudelft.nl/view/tno/uuid:6e0e4f90-13ce-4069-a365-5c787518270a/.

Feature code	Survey data	Administrative data	Big data
F4	Classical statistical methods available	Classical statistical methods available, usually depending on the specific data	Classical statistical methods not always useful
F5	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
F6	Weaker comparability between countries	Weaker comparability between countries	Potentially greater comparability between countries
F7	Representativeness and coverage known by design	Representativeness and coverage often known	Representativeness and coverage difficult to assess
F8	Not biased	Possibly biased	Unknown and possibly biased
F9	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 typical errors (e.g. missing data, reporting errors and outliers) although possibly less frequently occurring, and new types of errors
F10	Persistent	Possibly less persistent	Less persistent
F11	Manageable volume	Manageable volume	Huge volume
F12	Slower	Potentially faster	Potentially much faster
F13	Expensive	Inexpensive	Potentially inexpensive
F14	High burden	No incremental burden	No incremental burden

B. Strategies for integrating big data into official statistics

8. The ways in which big data sources can be used in current statistical systems can be classified as follows:

   U1: to entirely replace existing statistical sources such as surveys (*existing* statistical outputs);

   U2: to partially replace existing statistical sources such as surveys (*existing* statistical outputs);

U3: to provide complementary statistical information in the same statistical domain but from other perspectives (*additional* statistical outputs);

U4: to improve estimates from statistical sources (including surveys) (*improved* statistical outputs); and

U5: to provide completely new statistical information in a particular statistical domain (*new alternative* statistical outputs).

9. The first question which presents itself when considering the impact of big data on the use of statistical surveys in the production of official statistics is whether big data sources could entirely replace traditional statistical surveys (U1). With reference to the list of variables on which data is currently collected by the various surveys in the ESS, it is clear that big data sources do not yet provide an alternative for all of these variables. The same conclusion was reached by the 'Internet as a Data Source' project which closely assesses ICT surveys (Karlberg and Skaliotis, 2013).

10. With the exception of the 'efficiency' case, i.e. when big data sources are used as a replacement for existing sources (cases U1 and U2), the rigorous method of first formulating the need for information, and only then proceeding to look for suitable data sources, may be reversed. This is reflected in the accreditation procedure (Section III). This 'post-rationalisation' — tailoring needs to data availability — is not however entirely new. The principal users of statistics often formulate their information needs in terms of the existing statistics that they are already familiar with, meaning that the data available via existing surveys and other sources implicitly govern demand.

11. Big data has the potential to replace *some* statistical outputs entirely in the long term (U1) if: a) the (redefined) statistical outputs from big data meet the (evolving) needs for particular information; and b) other unbiased sources can be used for the purpose of benchmarking (adjusting for possible bias in big data).

12. Big data has the potential to partially replace statistical surveys (U2). Big data can replace some statistical outputs, either keeping their definitions unchanged or redefining them.

13. Big data can provide complementary statistical information in the same statistical domain, from other perspectives (U3). For example, instead of finding sources to replace the Household Budget Survey, a complementary approach would be to try to build indicators of trends.

14. U2 and U3 entail integrating big data into surveys. As is the case with administrative data, record linkage and statistical matching can be used.

    This then raises the question of how we can use big data to improve the statistics produced from surveys (U4). Combining big data with surveys will allow survey estimates to be improved by addressing the inherent weaknesses of surveys (Table 1).

    a) Flash estimates from big data can be used to improve timeliness of survey estimates. There are already some experiments (e.g. estimation of flu incidence based on Google query data)[5] which model a traditional statistical output using a much faster big data source in order to provide a timelier estimate. The question is

[5] http://www.google.org/flutrends/intl/en_gb/about/how.html

whether the relationships between variables which hold under normal conditions may be affected by extraordinary disruptive events, such as in a severe economic crisis (which is exactly when statistics are most important to guide policy intervention).

b) Big data can be used to calibrate survey results and for small area estimation. Although big data sources very often do not completely cover the traditional target populations of official statistics, they cover their own populations exhaustively. One way to transfer the power of big data onto traditional survey variables is to introduce some variables available from big data into surveys. The survey results can then be calibrated against the totals and breakdowns available from the big data. The gain in precision in the estimates relative to traditional survey variables will depend on the degree of correlation between the survey data and the big data for the variables in question.

15. Big data can improve statistics produced using surveys, and surveys can also improve the statistics produced from big data, by correcting possible bias of big data.

16. The integration of big data into official statistical surveys will transform the way in which official surveys are conducted. Combining surveys with administrative sources and/or big data sources will eventually lead to greater use of statistical modelling within official statistics, which will mark a significant change in culture and practices within NSIs. The introduction of big data or of administrative data covariates in surveys would also constitute a fundamental change to the traditional survey-based approach used by NSIs. The extent to which big data sources will infiltrate NSIs depends on a number of factors, often different in nature. To justify the use of big data for official statistics, it must be demonstrated that this would bring clear benefits in terms of statistical quality dimensions and cost. NSIs would also need to acquire the necessary additional skills for using big data and would need to be able to work with owners of private data sources. The concerns about privacy associated with big data go far beyond formal legislation and include public trust in government bodies, the ethics of using data collected for other purposes, issues around what constitutes responsible analytics, and the competitive advantage that data can provide. We believe that the best way to address many of these issues is to be actively involved in research public-private partnerships (PPPs) where NSIs and other major stakeholders participate in cooperation with one another. In addition, NSIs must be actively involved in discussions about data protection, data ownership and access to data. The question has been raised of whether we are moving towards a 'new deal on data' [9], [13] defined by Professor Sandy Pentland as '*workable guarantees that the data needed for public goods are readily available while at the same time protecting the citizenry. The key to the New Deal is to treat personal data as an asset'*. Professor Pentland has further suggested clear, workable definitions around the issues of data ownership.

17. In a fully digital society, surveys could be designed in ways that incorporate reality mining technologies [8] using both administrative sources and traditional survey sampling techniques. It is unlikely that legislation capable of dealing with the privacy concerns arising from the continuously expanding forms of data collection will be adopted soon enough, and it is therefore crucial to obtain the consent of individuals on any occasion where such an approach is used.

18. NSIs have already begun preparations for post-2020 population censuses. Coordination at international level, within the EU, for example, where censuses are governed by legislation, is often focused on issues relating to content rather than data sources. In our opinion however, it is in the interest of NSIs to examine the potential

benefits of big data sources for post-2020 censuses. In a recent meeting of census experts held at Eurostat (February 2014) this issue was introduced in order to raise awareness amongst NSIs of the challenges that new big data sources can present. There are several reasons and examples which suggest that big data should be included in the post-2020 census agenda. An ever increasing number of NSIs are moving towards greater use of registers and administrative sources as an alternative, timelier, and more cost-efficient approach to census taking. In the era of big data, it is almost certain that, for a number of administrative sources, we are moving towards electronically observed administrative records in the same way as in other fields where a digital footprint exists. Another field of interest in recent research projects is the use of mobile phone network data to estimate populations of small areas [6], [7]. The Office for National Statistics (ONS) in the UK and the Central Statistics Office (CSO) in Ireland are experimenting, respectively, with the analysis of *internet search queries within migration statistics* [14], and the use of *electricity smart meter data to determine household composition* [1]. Research from Telefonica [4],[11] suggests that mobile phone records *can be used for forecasting socio-economic* ***trends*** *as well as predicting socio-economic* ***levels*** *of a population*, while a recent feasibility study commissioned by Eurostat provides valuable insights into the use of mobile positioning data for population statistics [3]. Emilio Zagheni [15] estimated global migration trends by analysing 43 million anonymous Yahoo! account holders' IP addresses.

19. What becomes apparent from these research initiatives is that some core population statistics topics such as migration and usual residents, which are often difficult to measure, are now being explored by big data sources. In our opinion, this represents an unmissable opportunity for NSIs to join forces with the research community and private data owners (e.g. telecoms companies) to form PPPs with the specific aim of exploring the potential of big data in post-2020 censuses. There are numerous incentives for launching such partnerships.

C. Strategies to produce completely new statistical information

20. According to the proposed classification of the ways to use big data within the current statistical system, there is still an opportunity to provide completely new statistical information in a particular statistical domain (U5).

21. The other approaches to integrating big data into existing official statistics are rooted in traditional statistical models and tools but there is also a case for redesigning a new system with the aim of maximising efficiency in using big data. This approach would be specifically designed to improve the timeliness and it would not be subject to the limitations imposed by traditional statistical models. Integrating big data brings with it a certain number of new tasks, such as translating and linking the data to different classifications or standardising it so that it is line with such structures, which are not necessary when completely new statistical models are produced. On the other hand, official statisticians cannot design new procedures overnight, so using new models is also not always the answer.

22. A logical approach would therefore be to start studying the strategies of successful experiences of using big data with the perspective to apply them in statistical production. It would go beyond the scope of this paper to fully develop the strategy here, but a possible approach would be to start producing short-term indicators of the evolution of economic and social phenomena of interest without transposing big data structures onto statistical ones such as classifications and definitions. Machine

learning or data mining could then be used in addition to traditional statistical tools. The indicators thus produced may prove to complement classical and more detailed statistics, by providing the means to roughly update the figures until the next structural survey or census. When many different statistical figures are produced from different and independent big data sources, the coherence and agreement among them may serve as an argument to support the validity and representativeness of the whole set in a similar way to that in which the national accounts systems work.

23. Studying experiences of using big data would also help official statisticians to learn about possible new methods of data processing and to start redesigning new official statistics production systems which could accommodate the rich but heterogeneous and sometimes volatile data available to them.

24. At the same time, the role of some of the present statistical infrastructures that are resource heavy or time consuming needs to be reconsidered. In addition, before starting work on a complex process for producing statistics from a big data source, the potential gains should be analysed, taking into account the fact that the advantages may balance a possible decrease in accuracy or quality in general (see the cost-benefit analysis in Section III).

III. ACCREDITATION AND CERTIFICATION

25. The use of electronic systems is creating an ever increasing amount of data providing meta-information on a wide range of activities. This new data is collected mainly by private and public entities operating outside the statistical system. In order to use this data for the purpose of generating statistics, certain quality standards must be met. During its long history, the system of official statistics has established a generally accepted quality framework[6] that should be extended to cover statistics derived from big data sources. At the level of statistical outputs, the general acceptance that quality assessment is relative, i.e. that the level of quality required is dependent on the intended use of the data, has led to the development of quality dimensions[7], which were adopted by Eurostat and the ESS to describe the quality of statistical data. As most official statistics use surveys as the means of data collection, the related quality indicators refer largely to measures of accuracy of samples (F9). Big data are normally not based on sampling techniques and originate from non-official, mostly private sources. Concerns around big data relate to the measurement of the quality and to the fitness of big data for use in official statistics.

26. Drawing upon the approaches used for administrative data, a study commissioned by Eurostat (Petrakos, Sciadas and Stavropoulos, 2013) proposes an accreditation procedure which could guide statistical authorities in their selection of certain big data sources to produce statistical outputs conforming to the high standards of official statistics.

[6] At international level, official statistics are guided by the UN's Fundamental Principles. Within the European Statistical System, the quality framework is defined by the more detailed European Statistics Code of Practice, which sets out 15 principles covering the institutional environment, the statistical processes and the statistical outputs. The supporting Quality Assurance Framework of the ESS has been developed to guide and assist the implementation of the Code of Practice. It covers the principles of the code that relate to statistical processes and outputs as well as the commitment to quality within the institutional environment.

[7] The quality dimensions are relevance, accuracy, timeliness and punctuality, comparability and coherence, accessibility and clarity.

D. Principles of the accreditation procedure

27. The design of an accreditation procedure is based on some principles setting out its desired, ideal and standard properties. Namely, in addition to the fact that the procedure must be fully compliant with the quality framework and principles of official statistics:

    a) **Flexibility**: It should be flexible enough to give consideration to sources which initial judgment would rule out, but which, after an in-depth examination, are shown to be fit for use.

    b) **Stepwise approach**: It should follow a stepwise approach designed to accommodate sequential decision-making. It should progressively approve or reject sources. This will allow gradual investment in worthwhile sources and will avoid a large initial investment being made in a source which then turns out to have significant shortcomings. Each step should provide for the subsequent decision on whether to continue the consideration of that source in later steps or not.

    c) **Assessment of input, process and output**. It should include an overall quality assessment which balances the input, process and output. Input refers to the source, metadata and data. Process is related to the methods of extracting and transforming information, aggregating data and producing statistics. The quality of the output should be assessed with reference to the quality dimensions. For new alternative statistical outputs (U5), *newness* can be regarded as an extra quality characteristic.

    d) **Risk assessment**. As the data source is not under the control of the statistical authority in question, a risk assessment should accompany the quality assessment.

    e) **Assessment by statistical authorities**. The statistical authorities should directly test the feasibility of using a big data source by obtaining original data (possibly in the form of sample data) from the same source. They should carry out this empirical assessment themselves and should not delegate it to any third party. If the quality assessment of an external data source is outsourced, additional risk is incurred due to the further loss of control on the part of the statistical authority.

    f) **Corporate criteria governing decisions on data usage.** The final decision on accreditation of a new data source must consider whether using it complies with the corporate standards and quality requirements of the statistical authority. The procedure of accrediting the source must be of value to the statistical authority in their work to compile related documentation (e.g. quality reports).

E. Accreditation procedure

28. The resulting accreditation procedure has five stages that generate information on which informed decisions can be based.

29. The first stage consists of the **preliminary examination of the source, data and metadata.** The main objective is to assess whether the data is potentially useful for statistical purposes, on the basis of information on:

    a) the ‘raison d’être’ of the source organisation, its activities and reputation;

    b) whether the data is structured or would need to be ‘treated’ prior to use in the production of statistics (F5);

    c) coverage of the target population, variables, units of measurement, frequency and timeliness,

d) whether the data is accompanied by metadata or requires a preliminary stage of metadata enrichment.

30. The second stage involves the **acquisition and assessment of (extracts of the) data.** Providing the source is willing to give access to (samples of) data and metadata (if available), the statistical authority should undertake an empirical assessment of its fitness for use and quality. When considering the fitness of the data, the statistical authority might identify a subset that is most meaningful for its purposes. Data quality cannot be fully controlled in the initial analysis of huge volumes of data, but a general assessment of the overall quality is possible as a minimum. The statistical authority should discuss with the source the means of, frequency and conditions for data access or transmission.

31. Assuming that the data fit the intended use and that the quality is acceptable, the authority should then, as a third stage, carry out **an in-depth investigation** into the data and its usability.

    a) The authority should identify errors and clean the data file. The quantity of errors may lead the statistical authority to abandon the data source.

    b) If the errors are not significant, after treating them, the authority would proceed to impute missing data and would use the data file to produce statistical outputs and to measure the quality. Quality requirements can vary depending on the role the big data source has in official statistics (cases from U1 to U5 in Section II).

32. Implicitly, the third stage reveals whether available statistical tools are suitable for storing, processing and analysing the volume (F11) and specific nature of this data or whether new tools would be needed.

33. The **decision of the statistical authority** forms the fourth stage. This should involve consideration of the usefulness and usability of data, the quality of the data and the impact of the data on all quality dimensions of the statistics produced, the resources and reputation of the statistical authority and the risks incurred by using the data. The use of a particular big data source may trigger a change in definitions, classifications and methodology which would affect time series and comparability over time. The comparison of big data results against benchmarks is critical to the decision.

34. In the fourth stage, the statistical authority should evaluate the feasibility of using the source data from a technical, methodological, social and legal point of view and also assess the cost-benefit ratio and the compliance of the data with the conditions of the ESS. Moreover, the statistical authority should run a cost-benefit analysis balancing the gains against the losses and risks. For example, for existing outputs (cases U1 and U2 in Section II), the gains are likely to relate to timeliness (F12) and burden reduction (F14), while the losses and risks may arise due to the decrease in accuracy (for example F8) and uncertainty over the continuity of the data source (F10).

35. Before taking the decision to use the data source, the statistical authority should make sure to have a well thought-out strategy for communicating and disseminating statistics based on big data and for addressing possible reactions from the public relating to concerns about privacy.

36. A favourable decision at the previous stage would lead to a **formal agreement with the source** at the fifth stage. The terms and conditions of the agreement should be set out clearly. For example, the source data should maintain at least the same level of quality, coverage and detail as is currently offered.

37. The agreement with the source should be for a period long enough to ensure the continuity of statistical outputs but short enough to allow the statistical authority the freedom to use any new and more competitive big data source which might become available in the near future. A non-binding agreement would also achieve this.

38. Competing big data sources (for example, those with the potential to replace the same outputs from a sample survey; U2) can be submitted at the same time to the procedure of accreditation with the purpose of running a comparative analysis and making an informed selection between competitors.

39. Statistical authorities should share information on previously assessed big data sources with multi-national coverage or with national coverage but similar characteristics to sources in other countries. This would make the process of integrating big data into official statistics more qualitative (thanks to the exchange of experiences and lessons learnt), quicker (as suitable methods and tools would be shared) and more consistent (as using sources with multi-national coverage would ensure ex-ante comparability; F6). The benefit from this would still be felt even if a national statistical authority chooses a national source over a multi-national one in the end because of it better meeting specific national needs.

F. Certification

40. The existence of an accreditation procedure supports the principle that quality standards of official statistics must be maintained, in order to assure users of the fitness of big data for use in official statistics. This approach could be extended beyond official statistics. It could be argued that it would be desirable to establish quality standards more widely to allow users of statistics to find their way in the 'jungle of information' with a certain degree of confidence, and to be able to distinguish reliable data from less reliable.

41. Most data produced today, whether intentionally or as a 'by-product', does not have to conform to established quality standards. Furthermore, the only standards in existence are those specific to NSIs. There is currently no widespread agreement, established approach or mechanism to address this matter, in a way which would be comparable to certification by the International Organisation for Standardisation (ISO). While the official statistical system has neither a monopoly on data nor can it take it upon itself to police the data world, it does have a moral authority by virtue of its longstanding record on quality and as such has an important role to play.

42. The impact of big data on the overall approach of statistical authorities can be illustrated through the example of an organisation with substantial data holdings and sufficiently advanced methods, and which is positively predisposed to accreditation. In such a case, the approach to take in the final stage of accreditation would be quite different. Rather than trying to establish whether the organisation was willing to cooperate and share their data, the starting point would be to determine whether the organisation wants to be certified as a data producer in that particular area. Taking a forward-looking approach in this type of situation would open up the option of certification as data producer in a particular statistical domain. This organisation would be responsible for processing the source data and producing statistical outputs. The questions asked would address the issues from a very different angle as compared to those asked for a standard accreditation. For example, they would be aimed at ascertaining whether the organisation would consider adopting the existing quality frameworks, issuing quality statements, adopting and abiding by provisions

regarding confidentiality including penalties for their breach, and generally adhering to the majority of principles that guide the work of the statistical system.

IV. CONCLUSIONS

43. The so-called 'big data V's' are often viewed as problems (the volume is too large, the velocity is too high, the variety is too complex, and the veracity is dubious), further exacerbated by the fact that the data require parallel processing, processing in situ and in real time. Instead, we would like to look at things differently and see this as an opportunity, by drawing attention to the other 'V' associated with big data – V for value.

44. For the official statistics community, big data can be regarded as having value as it represents an alternative source for official statistics which is large enough (and otherwise of a sufficient quality, vetted through an accreditation procedure as outlined in Section III) to make improvement of official statistics feasible (in the various ways described in cases U1 to U5 identified in Section II).

45. Admittedly, the statement that big data is 'large enough' is somewhat vague, and would, if not otherwise qualified, also hold true for 'traditional' administrative data. However, this hopefully also helps to convey the general idea: just because a new source, fairly structured, rather small (a mere terabyte?), with a low (only daily?) updating frequency fails to qualify as "Big Data" according to some of the definitions on the market, this wouldn't be grounds to reject it out of hand.

46. This approach means that the official statistics community must remain open minded in order not to miss out on opportunities or to be rendered obsolete. Providing methods (see Section II) exist to extract valuable information from a new big data source, and providing there is a procedure for assessing the quality of this new source (see Section III), the source should be explored. We would neither refuse to consider big data because they are at the high extreme of the four 'V's (volume, velocity, variety and veracity), nor would we refuse to consider data that don't seem to embody these characteristics. Any new big data source with the potential to increase efficiency or offer users of statistics something new or better should be considered as part of exploratory research into big data.

47. While big data potentially has great promise in terms of its value for official statistics, the changes required to the practices and ways of working of statistical authorities and also to the statistical techniques used would be significant. In particular, statistical authorities would assume greater and different responsibilities and functions. As described above, some of these changes will demand great flexibility and adaptability of approach.

48. The transformation of official statistics has already started, with many NSIs, as well as other official statistics producers and international statistical bodies, entering the big data market with conviction. The key to success is to engage with big data as one would any other data, and to give it due recognition as part of our core business – making data meaningful.

REFERENCES

[1] Dunne, J., MacFeely, S., *Big data coming soon …..to an NSI near you*, Big Data Collaboratory, Centre for Creative Collaboration, 3-4 February 2014.

[2] European Statistical System Committee (2013), *Scheveningen Memorandum on 'Big Data and Official Statistics'*, http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM%20Final%20version.pdf (accessed 17 January 2014).

[3] EUROSTAT (2014), *Feasibility study on the use of mobile positioning data for tourism statistics*, [section on Population, Migration and Commuting Statistics].

[4] Frias-Martinez, V. et al, *Forecasting Socioeconomic Trends With Cell Phone Records*, *http://dev3.acmdev.org/papers/dev-final4.pdf* (accessed 16 February 2014).

[5] Karlberg, M. and Skaliotis, M. (2013), *Big Data* for *Official Statistics — Strategies and Some Initial European Applications*, WP30 presented at the Seminar on Statistical Data Collection, http://www.unece.org/stats/documents/2013.09.coll.html (accessed 16 February 2014).

[6] Loibl, W., Peters-Anders, J. (2012), *Mobile Phone Data as Source to Discover Spatial Activity and Motion Patterns*, *http://gispoint.de/fileadmin/user_upload/paper_gis_open/537521028.pdf*, (accessed 16 February 2014).

[7] Makita, N et al, *Can mobile phone network data be used to estimate small area population? A comparison from Japan*, Statistical Journal of the IAOS 29 (2013) 223-132.

[8] MIT Technology Review, *Reality Mining*, http://web.media.mit.edu/~sandy/tr10pdfdownload.pdf, (accessed 16 February 2014).

[9] Pentland, A., *A new Deal on Data*, Scientific American, October 2013, p. 69.

[10] Petrakos, M., Sciadas G. and Stavropoulos, P. (2013), *Accreditation procedure for statistical data from non-official sources*, (study commissioned by Eurostat).

[11] Soto, V. et al, *Prediction of Socioeconomic Levels using Cell Phone Records*, http://www.vanessafriasmartinez.org/uploads/umap2011.pdf, (accessed 16 February 2014).

[12] United Nations Statistical Commission, Forty-fifth session, 4-7 March 2014, *Big data and modernisation of statistical systems*, http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf, (accessed 16 February 2014).

[13] World Economic Forum, *The Global Information Technology Report 2008-2009* http://hd.media.mit.edu/wef_globalit.pdf, (accessed 16 February 2014).

[14] Williams, S., Office for National Statistics, *Internet Search queries within migration statistics*, Big Data Collaboratory, Centre for Creative Collaboration, 3-4 February 2014.

[15] Zagheni, E., *You are where you e-mail: Global migration trends discovered in email data*, http://www.mpg.de/5868212/internet_demographics, (accessed 16 February 2014).