**Ethical implications of using Big Data for Official Statistics**

Albrecht Wirthmann, Eurostat, albrecht.wirthmann@ec.europa.eu

Fernando Reis, Eurostat, fernando.reis@ec.europa.eu

**Abstract**

*We are currently experiencing an all-embracing digitalisation of our societies and economies. The pervasive nature of information and communication technologies is leading to the 'datafication' of most of our activities and relationships. This development is producing so far unknown amounts of data. New developments in Information Technologies do not only allow capturing but also storing, linking and analysing these data to infer conclusions on data subjects in massive amounts. This new ability might impact persons' everyday life in different ways, positively or negatively. Ethical implications thereof have already been discussed in literature. The proposed paper will discuss the implications of using big data in official statistics. These implications will differ from risks and hazards related to big data usage in general. Firstly, Statistical offices are not targeting individual subjects but only aggregated results and identified patterns. Secondly, there are already principles and ethical guidelines concerning the statistical confidentiality of personal data. Therefore the paper will focus on issues of data quality, availability of data sources, dependence on third party sources, data manipulation influencing the results, transparency of data and methods of data analytics, or scientific approach to data analytics. The paper will identify issues and analyse possible consequences for official statistics and contrast them with the existing ethical frameworks (Fundamental principles of official statistics, European Statistics code of practice, Declaration on Professional Ethics). Finally the paper will define or emphasize relevant principles on the ethics of using big data that could be followed in the process of integrating these new data sources into Official Statistics.*

**Keywords:** Big Data, Ethics, Code of practice

**1. Introduction**

The main objective of this study is to review the ethical framework of European official statistics in the light of the envisaged use of big data for inclusion into producing official statistics. Before introducing big data that have a different nature from any other traditional source of official statistics, it is necessary to assess if the use of big data is compatible with professional ethics and whether it is necessary to update existing statistical principles that were based on statistical enquires and administrative data in the statistical production process.

**2. Conceptual Approach of the Study**

*2.1 Professional ethics of official statistics*

The ethical norms of statisticians working within a variety of economic, cultural, legal and political settings are expressed in the **Declaration on Professional Ethics** that defines the professional values of statisticians and principles derived from these values.

The ethical norms of statisticians working in government institutions that are entitled to produce and disseminate official statistics are set in national codes of conduct that are largely based on the **Fundamental Principles of Official Statistics** adopted by the UN General Assembly.

The development, production and dissemination of European statistics are governed by the statistical principles: professional independence, impartiality, objectivity, reliability, statistical confidentiality and cost effectiveness. The statistical principles are further elaborated in the **European Statistics Code of Practice**.

The European Statistics Code of Practice and the UN Fundamental Principles of Official Statistics based on shared professional values, constitute the ethical framework of European (official) statistics.

*2.2. Approach used to review professional ethics of European statistics*

Big data is different from traditional data sources of official statistics. Therefore, a first step of the Study should be to identify the characteristics of big data that may raise questions related to compliance with statistical principles at different stages of the statistical production process. These stages include data acquisition, data processing, dissemination of statistical output based on big data sources.

Secondly, the possible ethical issues are assessed against the existing principles of European Statistics Code of Practice and the Fundamental Principles of Official Statistics. In a case where the existing principles do not provide an immediate answer to ethical questions, they are evaluated from the point of view of shared professional values of statisticians expressed in the Declaration on Professional Ethics.

**3. Big Data and Official Statistics**

In order to be able to harness the potential of big data, statistical authorities need to be ready to face new challenges. Big data have not been designed for statistical purposes. Therefore, it does not necessarily meet statistical standards on the concepts and definitions. Big data may contain personal information that would require from statistical authorities not only to ensure the confidentiality of the data received but also to respect the privacy of the data subjects

Big data provides the ability to easily summarise and link various aspects of peoples’ behaviour, which can result in revealing the identity of people. The potential misuse of personal information can damage the reputation of individual(s) and may affect his or her personal life negatively.

For some companies, big data has an important commercial value. That could prevent sharing the data with statistical authorities or may have negative consequences of business performance.

*3.1. The use of big data for official statistics from professional point of view*

In order to produce statistical output and disseminate it as official statistics, NSIs follow a set of steps that constitutes a statistical business process model. Independently of the data sources (statistical questionnaires, administrative data etc.), the statistical business process covers data acquisition, data processing and dissemination of statistical indicators. These steps are reflected in the principles of European Statistics Code of Practice that are grouped into three main categories: institutional environment, statistical processes and statistical output.

The adherence to statistical principles ensures the trust of the public in the integrity of statistical systems and confidence in official statistics. The integration of big data into production of official statistics should ensure preserving and re-enforcing this trust in the statistical system. Therefore, the respect of the ethical norms is highly important. The principles of the European Code of Practice have been developed before the big data era and may not reflect the latest technological developments, particularly as far as the indicators of good practice are concerned.

Analysis of the relevance of the existing statistical principles to be applied at the main steps of statistical business production process, taking into account specific characteristics of big data as the data source of official statistics, is provided below.

*3.2. Access to big data sources*

In most cases, big data is collected by private companies. Access to these data sources by statistical offices is a priori not enforced by law. Private data holders may grant access to these data on a voluntary basis. Provision of big data to statistical offices will also depend on benefits realised or perceived by businesses.

Principle 2 of the European Statistics Code of Practice requires the existence of a legal mandate for data collection. The UN Fundamental Principles of Official Statistics (Principle 5) states that "Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents".

The statistical principles do not prevent the use of big data in official statistics, even if the law does not enforce access to the data. However, big data might be considered as an economic asset by private data holders, which can result in resistance to sharing the data with statistical offices without compensation. According to the Australian Bureau of Statistics (2014), "this raises the issue of how the NSO [National Statistical Offices] might acquire commercially valuable or sensitive data for statistical production, particularly if the statistics compete directly with information products created by the data owner or they compromise its market position. This issue is made more complex by the fact that there may be several parties with some form of commercial right in relation to a data set, either through ownership, possession or licensing arrangements”.

Varying or unclear conditions for data access may raise concerns of **professional independence** in a situation where private companies that have acquired big data put pressure on the statistical authorities for them to promote their business. Professional independence is essential for a statistical organization. It establishes credibility among its users and creates a relationship of respect and trust in statistical system.

*3.3. Selection of big data providers*

Some companies that collect big data provide the data to the brokerage companies in order to process the data and deliver the results to the users. Data brokers typically collect and share information about consumers without interacting directly with them. Consumers are largely unaware that data brokers are engaging in these practices and, to the extent that data brokers offer consumers explanations and choices on how they use their data, that information may be difficult to find and understand. Data may be collected from many sources and are linked by data subject and then aggregated to groups. The methods and algorithms for data processing usually are not disclosed. Information on quality of the results is limited and can only be assessed indirectly depending on the aggregation level. This may lead to concerns related to scientific approach, transparency and the quality of the data.

It may also raise a question of trust in big data providers when the data collected by them is going to be used for official statistics. The European Statistics Code of Practice as well as the UN Fundamental Principles of Official Statistics do not address the question of trust in the secondary data sources providers as they only have covered data collected from public administration so far. Trust is not only related to transparency that requires to disclose the data sources and methods used for producing official statistics but also to the respect of the data subjects.

Collaboration with the data providers that are lacking transparency may not only compromise the reputation of statistical authorities but also question the quality of official statistics if the origin of the data collected is questionable or unclear.

*3.4. Privacy of the data subjects*

The use of big data raises significant privacy concerns. According to the General Data Protection Regulation, personal data can only be gathered legally under strict conditions for a legitimate purpose. Persons or organisations that collect and manage personal information must protect it from misuse and must respect certain rights of the data owners which are guaranteed by law. The data subjects should receive clear and understandable information when their personal data is processed. Whenever the subjects’ consent is required, it will have to be given by means of a clear affirmative action before a company can process personal data. These legal requirements are fully in line with professional ethics of statisticians .

In order to work in line with professional ethics the data subjects should be informed that their personal data (or anonymised data) is shared with statistical authorities for the purposes of official statistics. It can be done by both big data holders and statistical agencies. Information to the data subjects could be provided personally, e.g. by including it into Terms of Services by data holders, or in general, e.g. publishing information on the use of data for statistical purposes on publicly accessible media. Furthermore, the public should be informed about the privacy and data security policy of the statistical agencies.

*3.5. Applying big data analytics*

Big data analytics refers to:

* Data analysis being undertaken that uses high volume of data from a variety of sources including structured, semi structured, unstructured or even incomplete data; and
* The phenomenon whereby the size (volume) of the data sets within the data analysis and velocity with which they need to be analysed has outpaced the current abilities of standard business intelligence tools and methods of analysis.
* The complexity of the relationships with complex structures embedded in the data has reached a level that cannot be handled by the current tools and models of statistics and analysis.[[1]](#footnote-1)

Siu-Ming Tam (2014) states, *“A significant unresolved issue is the threat of disclosure through data accumulation. Every individual is a unique mosaic of publicly visible characteristics and private information. In a data rich world, distinct pieces of data that may not pose a privacy risk when released independently are likely to reveal personal information when they are combined – a situation referred to in the intelligence community as the “mosaic effect”. The use of Big Data greatly amplifies the mosaic effect because large rich data sets typically contain many visible characteristics, and so individually or in composition enable spontaneous recognition of individuals and the consequential disclosure of their private information. This will be a significant issue when disseminating micro data sets from Big Data sources*”.

Big data analytics can create new information that is not known in advance and can reveal individual data. That means, there is a risk of disclosure of personal information that can potentially be used against the data subjects intentionally or unintentionally. Working in ethical terms, statistical authorities have to protect the identity and the reputation of the data subjects. Any possibility to compromise individuals or organizations should be excluded.

*3.6. Dealing with quality of big data*

Big data are not designed for statistical purposes and do not meet statistical standards on concepts and definitions. The studies conducted so far have pointed out a number of quality issues that challenge the use of big data for official statistics. These issues include the **selectivity** of the data (representativeness), the lack of guarantee in the **continuity and stability** of the data and the data structure, the risk of **data manipulation**, the **adequacy of statistical concepts**, the **selection of models**, **imputation techniques**, **statistical inference**, etc. Some examples are given below.

Thus, Couper (2013) has described two types of bias regarding big data. “The first is **selection bias**. … and while almost all of us are users of the new media, we must remind ourselves not to generalize from our own experiences, and remember that while the number of active Facebook users (for example) is enormous, not everyone is on Facebook. …That is, we should make a distinction between the producers of social media and the consumers of such media. The former may not be representative of the latter, and neither may be representative of the general population. … We also need to understand the limits of transaction data – not everyone uses loyalty cards (for example) or credit or debit cards. Mobile phone (and especially smart phone) penetration is not at 100%. Not everyone communicates by e- mail, and those who do may use different accounts for different purposes. Selection bias can occur at the individual level (e.g., those still using cash) and at the transaction level (e.g., some types of purchases may be more likely to be paid for in cash). There are still many ways in which transactions can be conducted without leaving a trace, and many tips and techniques for avoiding being traced (e.g., Singer, 2013) ….The second type of bias is **measurement bias**. Despite the stories one reads about the things people post on Facebook or other social media sites, social media is primarily about impression management (see Boyd & Ellison, 2008). To what extent do people’s posts represent their “true” values, beliefs, behaviors, etc.? Similarly, if we counted the number of Facebook friends one has as an indicator of true social network size, we may be seriously wrong”.

Siu-Ming Tam (2014) underlined the volatility of the sources. “Much Web content is also unstructured and ungoverned – the metadata describing its usage and provenance (origin, derivation, history, custody, and context) are either incomplete or incongruous. Indeed, the long-term reliability of Big Data sources may be an issue for ongoing statistical production. Reputable statistics for policy making and service evaluation are generally required for extended periods of time, often many years. However, large data sets from dynamic networks are volatile – the data sources may change in character or disappear over time. This transience of data streams and sources undermines the reliability of statistical production and publication of meaningful time series.”

The analysis of the use of big data for official statistics by Wirthmann et al (2015) points out that social network data bears the risk of being manipulated. This can be done either by the data provider itself or by third parties. “For example many spurious social media messages could be generated in order to push a statistical index derived from these data in one or another way in case it is known that the index is calculated from such data”

Couper (2013) points out: “It is hard to find evidence, but it is believed that the more people realize that analysis of big data can influence decision-making, the more likely we are to see attempts to manipulate the system – e.g., to generate interest in a topic or produce the desired results by directly manipulating social media, to create multiple accounts, to generate buzz by re-tweeting, and so on. With increased visibility and importance of big data may come increased attempts to manipulate the data for financial or political gain, or merely to make mischief.”

In summary, exploring new data sources and learning new methods to overcome the data quality problems is a professional duty of statisticians. Big data may require more efforts from statistical agencies to understand the nature of the data sources and to learn what kind of effect it might have on the statistical end-product. Collaboration with data holders is required to obtain knowledge on the provenance of big data. Collaboration with academia could help to develop new methodologies on the use of these sources for producing official statistics.

*3.7. Dissemination of statistical output*

Big data requires complex techniques to produce the statistical output. Sometimes, we can only observe the inputs and see the outputs, but the transformation (process) may not be understood. Engineers use the term *black box* to describe a system in which one can watch inputs going in and see outputs coming out, but in between exists some opaque, intangible process that transforms the inputs into outputs under a veil of secrecy. Google sometimes refers to its black box algorithm as the secret source that gives them a competitive advantage. Businesses are increasingly relying on this secrecy strategy when it comes to data, using the rationalization that people would game the system if they knew how it worked (Pasquale (2015)).

In order to work in line with the professional ethics, NSIs using big data analytics to produce statistical output should not only describe the data sources but also explain clearly methods and models applied. That is absolutely necessary for building confidence and trust of users in official statistics.

**4. CONCLUSION**

Professional ethics of statisticians do not preclude the use of big data sources in producing official statistics. The exploration of new data sources is a professional duty of statisticians, especially if they have the potential to better satisfy user needs, to reduce the reporting burden, to extend the number of statistical products, etc. Nevertheless, in order to retain the trust and confidence in official statistics and to adapt to the digital world, statistical agencies need to pay attention to possible ethical questions that may arise when big data is used. These questions may concern the relations with the big data holders, the protection of the privacy of data subjects, the transparency of the methods used to obtain statistical outputs. In order to use the potential of big data sources in official statistics, it seems indispensable to set up collaboration between statistical agencies, academia, big data holders, and private companies possessing required skills.

**5. References**

European Statistical System, (2014), The ESS vision 2020, Building the future of European statistics, [http://ec.europa.eu/eurostat/documents/7330775/7339647/ESS+vision+2020+brochure/4baffcaa-9469-4372-b1ea-40784ca1db62](http://ec.europa.eu/eurostat/documents/7330775/7339647/ESS%2Bvision%2B2020%2Bbrochure/4baffcaa-9469-4372-b1ea-40784ca1db62)

Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics

European Statistics Code of Practice, for the National and Community Statistical Authorities, Adopted by ESSC 28 September, 2011, <http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15>

International Statistical Institute, (2010), Declaration on Professional Ethics, <https://www.isi-web.org/images/about/Declaration-EN2010.pdf>

United Nations, (2014), Fundamental Principles of Official Statistics, <https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf>

United Nations, (2015), Fundamental Principles of Official Statistics, Implementation guidelines, <https://unstats.un.org/unsd/dnss/gp/Implementation_Guidelines_FINAL_without_edit.pdf>

Eurostat, (2015), Business Case, VIP BIGDATA, <https://ec.europa.eu/eurostat/cros/system/files/Big%20Data%20Business%20Case%20final.pdf>

United Nations Global Working Group on Big Data for Official Statistics, (2015), Towards Principles for Access to Data for Official Statistics; Discussion note for the Joint OECD and PARIS21 Workshop on Access to New Data Sources for Statistics, 17-18 December 2015;

Wirthmann A, Karlberg, M., Kovachev B., Reis F., (2015), Structuring risks and solutions in the use of big data sources for producing official statistics – Analysis based on a risk and quality framework, Eurostat working paper, <https://ec.europa.eu/eurostat/cros/system/files/Big%20Data%20Risk%20Paper%20Version1.pdf>

Wirthmann A, Karlberg, M., Kovachev B., Reis F., Di Consiglio L., (2016), Assessment of risks in the use of big data sources for producing official statistics – Results of a stakeholder survey, <http://www.ine.es/q2016/docs/q2016Final00283.pdf>

Steven Vale, UNECE, (2015), International collaboration to understand the relevance of Big Data for official statistics, Statistical Journal of the IAOS 31 (2015) 159–163

UNECE, (2014), How big is Big Data? Exploring the role of Big Data in Official Statistics, <https://statswiki.unece.org/pages/viewpage.action?pageId=99484307>

Robin Nicholas, Klein Thilo, Jütting Johannes, (2016), Public-Private Partnerships for Statistics: Lessons Learned, Future Steps, OECD Development Co-operation Working Papers No. 27

Daas Piet, Van der Loo Mark, (2013), Big Data (and official statistics), Statistics Netherlands, <http://www.unescap.org/sites/default/files/1-Big%20Data%20%28and%20official%20statistics%29-Netherlands.pdf>

Struijs Peter, Braaksma Barteld, Daas Piet, (2014), Official statistics and Big Data, Sage Journals, Big Data & Society, <http://journals.sagepub.com/doi/pdf/10.1177/2053951714538417>

UNECE Big Data Quality Task Team, (2014) A suggested framework for the quality of Big Data, <https://statswiki.unece.org/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2>

Siu-Ming Tam, Frederic Clarke, (2014), Small steps towards Big Data - Some initiatives by the Australian Bureau of Statistics

Commonwealth of Australia, (2014), Better practice guide for big data, V2.0, ISBN 978-1-922096-31-9

Committee on Professional Ethics of the American Statistical Association, (2015) Ethical guidelines for statistical practice, <http://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf>

Rein, Ahas, Tiru, Margus, Saluveer, Erki, (2014) Mobile telephones and mobile positioning data as source for statistics: Estonian experience, <https://ec.europa.eu/eurostat/cros/system/files/S19P4_0.pdf>

Vayena, Effy, Urs Gasser, Alexandra Wood, David O'Brien, and Micha Altman. (2016), Elements of a New Ethical Framework for Big Data Research, Washington and Lee Law Review Online 72 (3): Article 5

SOGETI, (2015) Privacy, technology and the law big data for everyone through good design; VINT research report 3,

Davis, Kord, Patterson, Doug, (2012), Ethics of Big Data

1. Australian Commonwealth (2015) [↑](#footnote-ref-1)