Assessment of risks in the use of big data sources for producing official statistics – Results of a stakeholder survey

Wirthmann A,[1], Karlberg, M.[2], Kovachev B.[3], Reis F.[4], Di Consiglio L. [4]

[1] *European Commission – Eurostat, Luxemburg; Albrecht.Wirthmann@ec.europa.eu*
[2] *European Commission – Eurostat, Luxemburg; Martin.Karlberg@ec.europa.eu*
[3] *European Commission – Eurostat, Luxemburg; Bogomil.Kovachev@ec.europa.eu*
[4] *European Commission – Eurostat, Luxemburg; Fernando.Reis@ec.europa.eu*
[5] *European Commission – Eurostat, Luxemburg; Loredana.di-Consiglio@ec.europa.eu*

Abstract

An increasing number of statistical offices are exploring the use of big data sources for the production of official statistics. For the time being there are only a few examples where these sources have been fully integrated into the actual statistics production (Statistics Netherlands, 2015). Consequently, the full extent of implications caused by their integration is not yet known.

A first attempt to identify and structure risks related to using big data sources in the exploration and production phases of official statistics was made in the paper "Structuring risks and solutions in the use of big data sources for producing official statistics – Analysis based on a risk and quality framework" (Wirthmann et al., 2015). The main conclusion from the paper is that it is impossible to establish a single likelihood or impact for a given "big data risk" – typically, both measures depend heavily on the utilised big data source as well as on the type of statistical product. In order to gain more insight, a source-specific survey of the identified risks has therefore been conducted among stakeholders. The respondents were asked to quantify likelihood and impact of risks for a big data source of their choice (among a set list of eight bigdata sources), to provide a rationale for their assessments, and to suggest measures for prevention and mitigation of the identified risks. In order to be more complete, the respondents were also invited to identify additional risks in the exploration and use of Big Data sources for official statistics.

The paper analyses and presents the results of the stakeholder survey, contrasting the findings to the analysis of Wirthmann et al. (2015).

Keywords: big data, risks, quality, statistics

1. Introduction

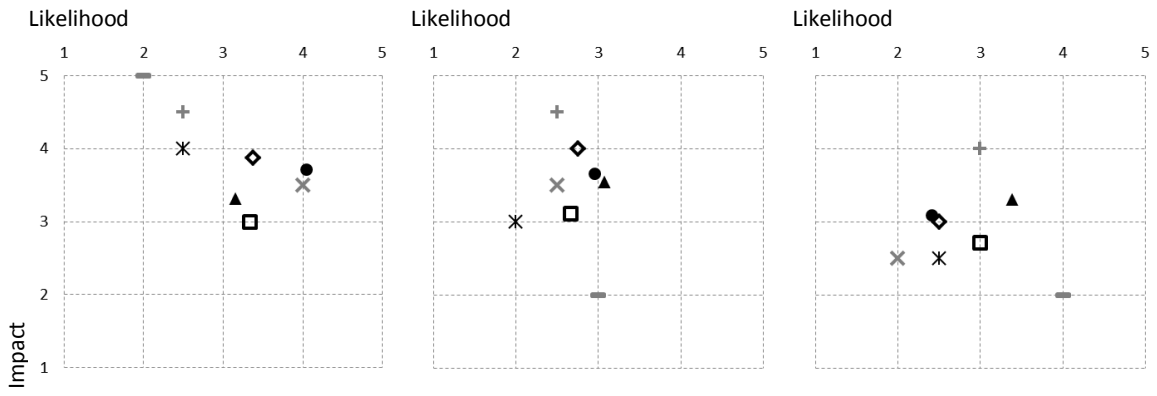
1.1. Methods and data

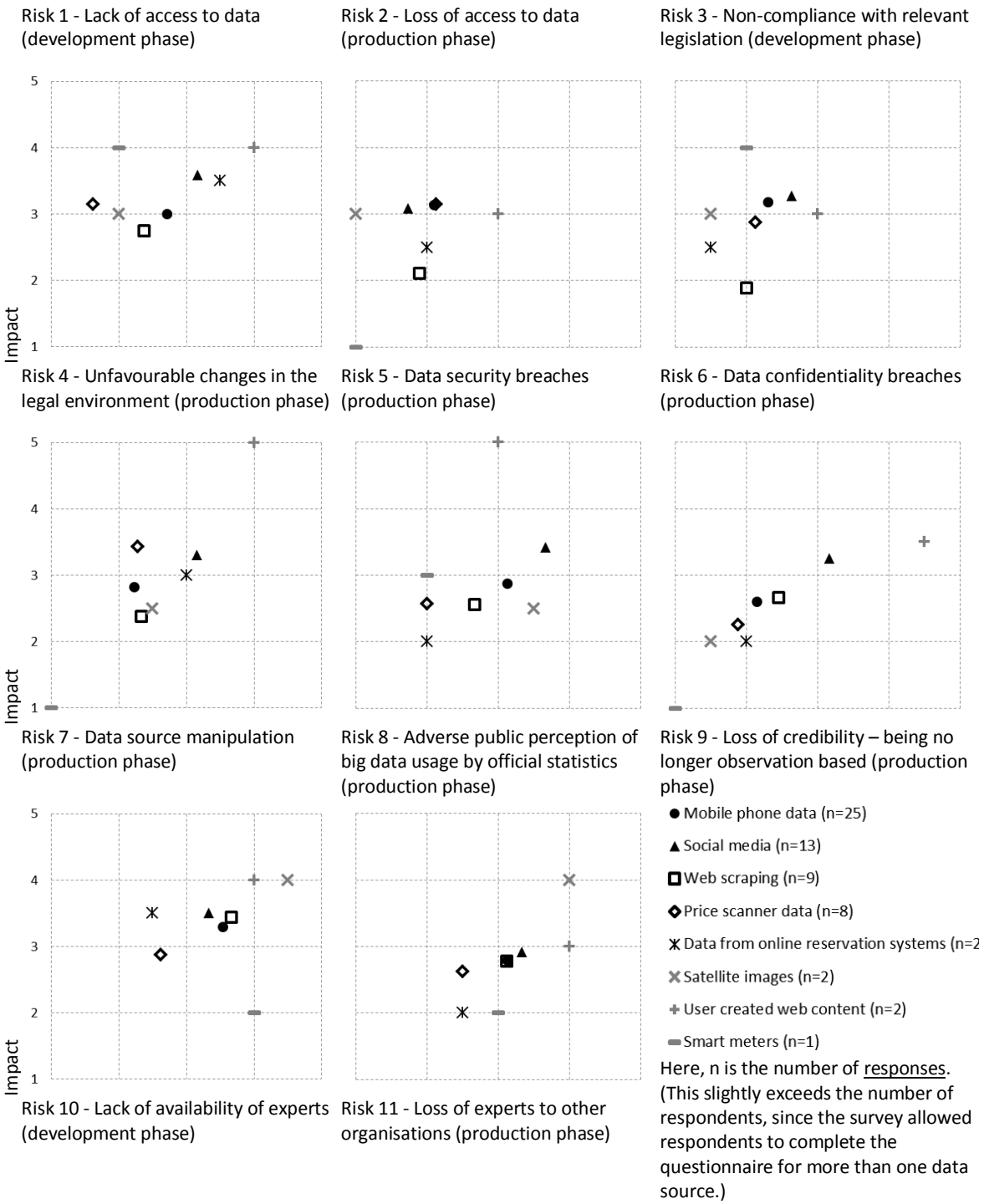
An online survey on Big Data Risks was launched on the CROS portal (http://ec.europa.eu/eurostat/cros/content/stakeholder-survey-big-data-risks). To keep the response burden reasonable, the survey was constructed so that respondents only had to indicate (and comment) likelihood, impact, prevention and mitigation actions for one Big Data Source; those who wished to do so could complete the survey multiple times (once for each Big Data Source)

Invitations to complete the survey went out to various stakeholders, including participants to recent ESS events related to Big Data. The survey should be viewed as exploratory /self-representing, as there is no specific target population to which the results could be extrapolated. During the period May-July 2015, a total of 62 valid responses were submitted. As the option to complete the survey multiple times was rarely used, and as the survey anyway isn't based on a probability sample, we conduct the analysis on response level rather than on respondent level, without any attempt at analysing intra-person phenomena.

The quantitative information (likelihood and impact estimates) are presented in Figure 1 for all sources. In our analysis of this quantitative information, we limited ourselves to those sources for which at least 8 replies were given, i.e., mobile phone data, social media, web scraping and price scanner data.

Figure 1: Arithmetic mean of risk estimates provided in survey for various data sources





2. Risks related to data access

2.1. Lack of access to data

This risk consists of a project charged with developing a big data based official statistics product (BOSP) not getting access to a necessary Big Data source (BDS).

As could be seen from Fig. 1, the respondents assess the **likelihood** of this risk is to be probable (4) for mobile phone data, as noted by one respondents, there are "different legal systems in different countries. In some countries mobile phone data is easily accessible, while in others access is currently almost impossible." The likelihood is assessed to be somewhat lower (in the range 3.2-3.4, meaning occasional) for social media data, web scraping and scanner data; this could be attributed to certain social media and web-scrapable data being publicly available, and for scanner data, a respondent notes that the NSO (National Statistical Office) has "established a partnership with data providers for this project. Data are now in the NSO".

The **impact** is assessed to be critical (3.7-3.9) for mobile phone data and price scanner data, and somewhat lower somewhat lower (in the range 3-3.3, meaning major) for social media data and web scraped data (in the case of price statistics, one could develop a BOSP based solely on scanner data, making the need of web scraped data less urgent).

In terms of **prevention**, the most prevalent proposals from respondents concern legislation (e.g. "EU or national law that obliges providers to share data with NSIs. The access or procedures should be designed in such a way that individual data is used for statistical purposes only"), followed by win-win ("Make them benefit, e.g. share results of analysis on their data") partnerships ("Agreements detailing rights and obligations of each party, including continuous access and protection of individuals' privacy"). Concerning **mitigation** action one respondent notes that they "vary depending of the kind of change, they can go from adjustments in the software parameters to totally changing the big data source".

If there is no way to produce the BOSP without the BDS, and if it is not feasible to overcome the lack of access, the endeavour has to be terminated, and the new BOSP will not see the light of day.

2.2. Loss of access to data

This risk consists of a statistical office losing a BDS underlying a BOSP.

The respondents assess the **likelihood** of this risk is to be occasional (in the range 2.7-3.1 for all sources; see Fig. 1) than the risk of lack of access of data, one example (in the case of mobile phone data) of the rationale for this lower likelihood assessment being that "Once access to MNO (Mobile Network Operator) data is secured, there is a low probability that it will be lost." However, other respondents note the possibility of "Changing ownership or business models" of MNOs, and warn that "the legal access to the private company resources may be subject to change if not framed by the law."

Surprisingly, the **impact** estimate of the respondents is typically not higher than for the risk of lack of access of data; this runs counter to the assessment of Wirthmann et al. (2015) that "as the existing BOSP may be impossible to produce, a very high impact would often be the case".

The character of the **prevention** actions are legislative/contractual (e.g. "engage and make long term contractual relationships or regulate"), good management of partnerships (e.g. "take good and active care of relations with data provider") and diversification ("It is utmost important to keep several irons in the fire. In other words Statistical Offices should not rely on a single source but always have at least two alternatives.") The proposed **mitigation** actions include technological agility ("keeping track and moving in time to new but equivalent platform or application") as well as ex post attempts at diversification ("Use of alternative information sources. Identify alternative websites -> update the list frequently") and partnerships ("negotiate with data owners").

3. Risk related to the legal environment

3.1. Non-compliance with relevant legislation

The risk concerns the development phase of a statistical product based on big data sources. It is related to a project that fails to take relevant legislation into consideration, thereby rendering the BOSP non-compliant with relevant legislation. This could concern any piece of legislation that is relevant for using big data for official statistics, e.g. data protection legislation, regulations concerning processing of data from specific big data sources, etc.

The replies from the expert survey range between 2.4 (remote) to 3.4 (occasional) for the likelihood of occurrence of this risk and between 2.8 (major) to 3.6 (critical) for their

possible impact. As rationale for estimating the **likelihood**, respondents put forward that statistical offices include reviews on legality of new statistical products that should prevent the described risk. In addition, it is mentioned that statistical offices are very aware of this risk and are therefore very carefully verifying the relevant legislation. However, it is also stated that existing legislation might not be very clear and subject to interpretation. Another respondent relies on the fact that data providers have carefully verified the legal compliance of their products before supplying data to statistical offices. In addition, privacy commissions and other bodies would be powerful institutions that carefully review relevant initiatives and take appropriate actions in case of non-compliance. Related to data from social media a situation could occur that intended use of data would not comply with the specific terms and conditions of the data supplier.

Related to **impact**, an incident of non-compliance would have negative consequences on the reputation of the statistical office in general. The most frequently stated consequence would be to stop the project as soon as the breach would have been detected.

For risk **prevention**, respondents advised to carefully review existing legislation, to involve data protection agencies and lawyers from the start of the project, to restrict use of data to public information, if possible, and to prepare a good communication strategy. Some respondents additionally mentioned the need for harmonization at supranational level and a possible role of international organisations enabling access to data sources, e.g. from social media.

3.2. Unfavourable changes in the legal environment

This risk is related to changes of the legal environment when a production process is already in place. New legislation might be unfavourable and, in the worst case, might prevent access or use of specific data sources for further production of statistical products from big data sources.

Respondents considered the likelihood that this risk would materialize on average between 1.6 (remote) for price scanner data and 3.2 (occasional) for social media data. The impact of this event is assessed between 2.8 (major) for web scraping and 3.6 (critical) for social media data.

Motivations for assessing the **likelihood** of this risk as low are that during the development of a new legal act, existing jurisdiction would be screened and possible consequences assessed. Another respondent assumes that current production of statistics from big data sources would be enabled through a legal act that would assure legal compliance. On the other hand respondents put forward that the domain is very dynamic that requires adaptations of legal acts. Implementation of innovations might change attitude of citizens, e.g. as regards privacy, that might trigger legal changes. Some respondents claimed that new legislation is likely to be introduced especially for social media data to rebalance use of data with data protection. Respondents consider a change of the legal situation related to webscraping as unlikely.

Possible consequences (**impact**) in case of unfavourable legal changes would be changes in the production system, impacts on methodology to exclusion of respective data source from the production process. The impact is in general estimated as being lower that by the authors. As regards **prevention** of risks, respondents recommend a pro-active approach monitoring legal initiatives and trying to influence the legal initiatives stressing the public benefits of the current use of the specific big data source for official statistics.

4. Risks related to data confidentiality and security

4.1. Data security breaches

This is the risk refers to unauthorised access to data held by statistical offices. Third parties could obtain data that is held under embargo e.g. due to release schedule. This can be for example data that is highly anticipated by stock market investors.

For most of the data sources the respondents do not seem to think this risk is particularly likely. Some motivation for this seems to be trust in established security procedures and experience in dealing with confidential data.

Understandably the **impact** is rated higher for data that is not gathered from publicly available sources even if still quite some way below the evaluation of Wirthmann *et al*. (2015). As regards **impact**, reputational damage and loss of trust seem to be the main concerns.

In terms of **preventive** actions the respondents advocate appropriate IT security measures and procedures, staff awareness and training and risk assessment. There is also a suggestion to restrict analysis to aggregated and anonymised data.

As proposed **mitigating** measures we mostly see suggestions to handle communication correctly and improve the technical measures for protection which largely coincides with the measures advocated by the authors.

4.2. Data confidentiality breaches

This is the risk that the confidential information of one or more individuals from the statistical population is disclosed, either due to an attack on the IT infrastructure or due to pressure from other government agencies or due to inadequate statistical disclosure control measures.

Overall with reasonable preventive measures the **likelihood** could be kept to reasonable levels, and the evaluation of likelihood of this risk was on average between 2 and 2.6 (remote to occasional), with differences related to the type of data. For example, a low risk is on average for web scraped data and a low impact, as data can be obtained in alternative ways quite easily. The highest risk is envisaged for social media, in relation to the nature of this source of data.

Statistical Offices have already in place measures to prevent their sources from external attacks to keep confidential data in secure environment and in most cases to preserve their independence from other governmental agencies.

Impact of confidentiality breach was seen higher for mobile data and social media, in terms of Statistical Offices' credibility and in terms of the impact on the agreement with the private operators,

On the other hand improving IT systems, enhancing methods for guarantee reducing the risk of disclosure, testing the risk of disclosure against different data sources, and finally ensuring independence of statistical offices are among the list of possible mitigation measures.

4.3. Data source manipulations

This is the risk for data provided from third parties, for example social network data or voluntarily contributed data being manipulated. This could 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.

For most data sources the respondents' average **likelihood** evaluation varies between the remote and occasional. For social media it is slightly higher. In general it is considered unlikely that individuals would be able to manipulate any of the data source. As an exception to this it has been pointed out that social movements could try to manipulate, though the expectation is that such cases would become known which would allow the statistical agency to deal with the situation. Reputational risk for anyone involved in such a manipulation is seen as the main reason for the moderate likelihood score together with the expectation that only market sensitive statistics would be at any risk at all.

The average **impact** estimate is visibly higher than the likelihood for mobile phones data and particularly scanner data. The main reason for this is the damage to public trust. The fact that such a manipulation would be difficult to detect and could potentially continue for longer periods is also pointed out. In Wirthmann *et al*. (2015) the impact estimate of this risk is slightly lower – the reputational risk is acknowledged however more trust is put in the effects of adequate communication.

Comparing, where possible, data from different providers has been pointed out as a way to protect the statistical office against this risk.

4.4. Adverse Public Perception of big data usage by official statistics

This risk refers to a situation where there is a negative public perception of big data usage by official statistics which might lead to additional restrictions or even impede use of certain big data sources. The likelihood of such a risk is assessed on average between 2 (remote) for price scanner data and 3.7 (probable) for social media data. For mobile phone data the likelihood of the risk is considered as being occasional (3.1). The impact of an

event ranges on average from 2.6 (minor - major) for web scraping and price scanner data to 3.4 (major - critical) for social media data.

Motivations for assigning a higher **likelihood** of this risk are a general distrust of the public in governmental organisations and that the public does not distinguish between actors (businesses or government bodies) in case of negatively perceived incidences. The risk would be lower if the public would be informed extensively on the purpose, the final statistical product and safeguards for preventing misuse of the data.

The **impact** would be a general loss of reputation of the statistical office that might negatively influence the general attitude of persons to collaborate with statistical offices. A negative public opinion might inhibit the use of specific big data sources for official statistics. A reason for low impact is the fact that agreements on the use of big data sources are concluded between data providers and statistical offices without involvement of the general public.

For **preventing** this risk Statistical Offices should prepare a suitable communication strategy before going into production. The communication should stress the benefits of big data usage for the citizens, e.g. lower burden on respondents and improved statistical data while assuring data security and privacy. Communication campaigns should involve relevant stakeholders with the purpose of raising awareness and informing the public on the purpose of the big data usage for statistics. In this context, respondents consider transparency as key element of the communication strategy.

4.5. Loss of credibility – being no longer observation based

Users of official statistics have high confidence in accuracy and validity of statistical data. This is based on the fact that statistical data production is embedded in a sound and publicly available methodological framework as well as the documentation of quality of a statistical product. In addition, most statistical data are observation based, i.e. are derived from surveys or censuses, which establish an easily understandable relationship between observation and statistical data.

On average the **likelihood** of this risk was evaluated as remote (around 2) for sources such as mobile phone data and scanner data, and as occasional (almost 3) for sources such as web scraping and social media.

The **impact** of occurrence of the risk is correlated with the likelihood of the respective risk.

Suggested **preventive** actions were to complement big data sources with surveys and to compare results with results from traditional sources. But some sources are also perceived as a more accurate measurement instrument than survey (e.g. smart meters). Before engaging into statistical production, BOSP could be published as experimental and stakeholders could be encouraged to contest the BOSP in order to confirm or enhance the BOSP.

In addition, Statistical Offices should invest in communication, develop strategy and publish scientifically sound methodology which is recognised by the scientific community.

Enrichment of data with metadata on quality, ensure consistency of the BOSP with non BOSP can preserve public trust.

5. Risks related to skills

5.1. Lack of availability of experts

The risk of lack of availability of experts consists of upon receiving data from one of these new big data sources, the statistical office not having the possibility of processing and analysing it properly, due to its staff not having the required skills. The use of big data requires skills on model based inference and machine learning, skills in natural language processing, audio signal processing and image processing and a good understanding of distributed computing methodologies.

The risk **likelihood** attributed by the respondents to the survey, occasional (2.6) to probable (3.7) is lower than the one attributed previously by the authors, probable (4) to frequent (5). The lowest likelihood is assigned to price scanner data. However, it is evident from the qualitative answers that some respondents already factored in the effect of prevention measures such as training and cooperation. New factors pointed out were the

constraints posed by resources shortages and the Statistical Office inability to mobilise eventual existing internal human resources. The survey respondents considered the **impact** of this risk, major (2.9) to critical (3.5), a bit lower than the authors' initial assessment, which was critical (4), and considered the impact of the risk lower for price scanner data than for other sources.

Besides training and recruitment of new staff identified initially by the authors as **prevention** measures, the survey respondents added some other. Cooperation with the academia and other Statistical Office, proposed by the authors as a mitigation measure, was pointed out as a prevention measure also, where knowledge could be transferred to existing staff before the lack of skills becomes a problem. Financial measures, for example directed to more attractive salaries, was also proposed and in relation to this, raising awareness of decision makers to the importance of using these new data sources. In terms of **mitigation** measures, the survey respondents confirmed the ones proposed by the authors, sub-contracting and cooperation, and added the smart pooling of existing resources, by integrating the few staff with the required skills in teams working on the implementation of big data sources in the several statistical domains.

5.2. Loss of experts to other organisations

This risk consists of statistical offices losing their staff to other organisations after they have acquired big data related skills.

The survey respondents agreed with the authors' initial assessment of the **likelihood** of this risk as being occasional (3.1-3.3), although considered it to be slightly lower for those skills related to price scanner data (2.5). One additional factor identified by the respondents which increases this likelihood was the type of data products being developed by organisations other than the Statistical Office, which are more engaging. However, the respondents identified the increasing supply of data scientists and the attractiveness of big data for existing staff as mitigating factors. Although the authors considered the **impact** of this risk to be the same as for the lack of skills, the respondents considered it to be lower. Even if new impact factors were pointed out, namely having to constantly to train new staff and the disruption that staff turnover causes, two reasons for the impact of losing skills being lower than not having them to start with, were that by the time big data moves into

production the Statistical Office have had developed more capability and that established production systems require less expertise to maintain than to develop.

Besides the **prevention** measures identified previously by the authors, namely offering learning opportunities, being open to new projects and ideas and identification of staff able and willing to work on big data, the respondents added provision of better salaries and campaigning for emphasising the value (social good) of working in official statistics. In addition to sub-contracting and cooperation, pointed out previously as **mitigation** measures, the respondents identified improved and faster recruitment procedures and continuous training on big data.

6. Additional risks proposed by the respondents

A total of 13 additional risks were proposed by the respondents. Setting aside risks that are more to be considered as causes to the risks presented above (e.g. "Law not updated to specifics of Big Data" or "cost increases from source"), the proposals These could largely be grouped into the four categories. First, there are risks related to the **volatility** of the data source. Whereas the framework of Wirthmann et al. (2015) already includes a "Data source manipulations" risk, changes in data sources typically take place for operational reasons, without statistics in mind. To quote one respondent, "In September 2014, we observed a 25 per cent (change) in the number of geolocated tweets. This was eventually traced to the release of the iOS8 operating system which included increased flexibility for managing privacy settings in relation to location." This is a compelling argument for either adding an "unintentional volatility" to the risks or extending the "Data source manipulations" by removing the restriction to malicious intent.

Second, there are also a couple of suggestions regarding **IT infrastructure** ("Lack of appropriate IT equipment for adequate data processing"). Just as for the other enable (skills), the framework should perhaps be extended to include a "lack of adequate IT resources" risk. Considering that risks are unforeseen events, there is need for further discussion if the lack of adequate IT infrastructure constitutes a risk or an *issue*.

One respondent proposes a risk related to **competition** ("The competition, besides the Statistical Office other data collectors (banks, other) are dealing with big data on the same

item (i.e. prices)". The emergence of alternative providers of statistics similar to official statistics is indeed a *threat* to official statistics in general, in case these statistics are presented in an attractive way, but are of substandard quality – but not necessarily a *risk* in the context of production of official statistics based on big data.

Finally, there are proposals related to the immaturity of **methodology** ("how to measure precision of data", "consistency and reliability", "errors of linkage EAN/PLU with COICOP", "quality changes of products"). This could also be regarded as existing, already materialised, challenges to be tackled; it is not immediately evident what risks (if any) this gives rise to.

7. Conclusions

While the responses to the survey show that the selection of risks in Wirthmann et al. (2015) was relevant, respondents also proposed additional risks, which should be considered for big data based official statistics products in the future. The highest figures for likelihood and impact are assigned to the risk "access to data" and "lack of skills". In our opinion the statistical community should put emphasis on prevention and mitigation measures for these risks.

The data sources that most respondents chose to express themselves on are mobile phone data, social media, web scraping data and price scanner data. According to the comments received, these seem to be the most frequent data sources being investigated in current big data projects; this is consistent with the findings of Consiglio et al. (2016).

In general, the likelihood and impact of risks are rated lower by the respondents to the survey than by Wirthmann et al (2015).

The assessment of likelihood and impact of risks are dependent on the data sources. Both estimates seem to be lower for data sources that are already used by statistical offices such as scanner data compared to those sources where there is less experience.

The comments by respondents suggest that National Statistical Institutes Offices have already started to define and implement mitigation and preventive actions in order to manage risks related to these data sources.

The results of the survey should be used to create and update the list of risks for big data projects in official statistics so that they can be better managed.

8. References

Daas, P., M. Puts, B. Buelens and P. van den Hurk. 2015. "Big Data as a Source for Official Statistics". *Journal of Official Statistics* Volume 31, Issue 2, Pages 249-262, ISSN (Online) 2001-7367, DOI: 10.1515/jos-2015-0016

Di Consiglio, L, M. Karlberg, M. Skaliotis and I. Xirouchakis (2016; forthcoming), paper for the invited overview lecture "Overview of big data research in European statistical agencies" to be delivered at ICES V

Eurostat (2014), "Accreditation procedure for statistical data from non-official sources" in Analysis of Methodologies for using the Internet for the collection of information society and other statistics, http://www.cros-portal.eu/content/analysis-methodologies-using-internet-collection-information-society-and-other-statistics-1

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

Reis, F., Ferreira, P., Perduca, V. (2014) "The use of web activity evidence to increase the timeliness of official statistics indicators", paper presented at IAOS 2014 conference, https://iaos2014.gso.gov.vn/document/reis1.p1.v1.docx

Statistics Netherlands (2015), "A first for Statistics Netherlands: launching statistics based on Big Data", https://www.cbs.nl/NR/rdonlyres/4E3C7500-03EB-4C54-8A0A-753C017165F2/0/afirstforlaunchingstatisticsbasedonbigdata.pdf

UNECE (2014), "How big is Big Data? Exploring the role of Big Data in Official Statistics", http://www1.unece.org/stat/platform/download/attachments/99484307/Virtual%20Sprint%20Big%20Data%20paper.docx?version=1&modificationDate=1395217470975&api=v2

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", http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2015/mtg1/WP18-Wirthmann_AD.pdf .