****

**How big is Big Data?**

**Exploring the role of Big Data in Official Statistics**

(Version 0.1, March 2014)

DRAFT FOR REVIEW

**Please note** the development of this paper is a work in progress. This is not intended for official publication. It reflects the thoughts and ideas gathered at the first Virtual Sprint on this topic held in during March 2014 and will be used to further discuss how statistical organizations will adapt to this potential new future; realizing the opportunities and minimizing the risks. The purpose of the paper is to encourage others to join the debate and to identify the ‘big’ things statistical organisations need to tackle.

**Instructions for reviewers** and a **template for providing feedback** is available at http://www1.unece.org/stat/platform/display/bigdata/How+big+is+Big+Data

This work is licensed under the Creative Commons Attribution 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/3.0/>. If you re-use all or part of this work, please attribute it to the United Nations Economic Commission for Europe (UNECE), on behalf of the international statistical community.



# Table of Contents

[I. Introduction 3](#_Toc383013312)

[II. What is Big Data in the context of statistical organizations? 4](#_Toc383013313)

[A. Definition 4](#_Toc383013314)

[B. Big Data Sources 4](#_Toc383013315)

[III. How ‘big’ is Big Data for statistical organizations? 5](#_Toc383013316)

[A. The Demand for Big Data 5](#_Toc383013317)

[B. The Value Proposition 6](#_Toc383013318)

[C. Hype cycle 8](#_Toc383013321)

[IV. Are ‘big’ changes needed to build a Big Data Capability? 9](#_Toc383013322)

[A. Process 10](#_Toc383013323)

[B. Methods 11](#_Toc383013324)

[C. Technology 14](#_Toc383013325)

[D. People 15](#_Toc383013327)

[E. Standards, Frameworks and Policies 16](#_Toc383013330)

[V. Is Big Data too ‘big’ to tackle alone? 20](#_Toc383013334)

[VI. Conclusion 23](#_Toc383013335)

[VII. Annex A: A typology of Big Data sources 24](#_Toc383013336)

[VIII. Annex B: SWOT Analysis and identification of risks and benefits. 25](#_Toc383013337)

[IX. Annex C: Next steps for the HLG Big Data Project 29](#_Toc383013338)

# Introduction

*"Big Data is an increasing challenge. The official statistical community needs to better understand the issues, and develop new methods, tools and ideas to make effective use of Big Data sources."* [[1]](#footnote-1)

1. The High-Level Group for the Modernisation of Statistical Production and Services’ (HLG) strategy document[[2]](#footnote-2) states that “products and services must become easier to produce, less resource-intensive, and less burdensome on data suppliers” and that “new and existing products and services should make use of the vast amounts of data becoming available, to provide better measurements of new aspects of society”. The use of Big Data by statistical organizations is one way to achieve this aim.

2. Already a number of statistical organizations are investigating and reporting on their use of Big Data[[3]](#footnote-3). These have tended to be opportunistic uses of Big Data – driven by the increased availability of a data source or the need to provide statistics at short notice (for example in cases of a natural disaster). The Official Statistics community needs to take a strategic look at what Big Data means for the community.

3. In 2014, the HLG are sponsoring a project aligned with these aspirations with a focus on the role of Big Data in Official Statistics. The project aims to analyse the major strategic questions posed by the emergence of Big Data. The project also has a practical element. A web-accessible environment for the storage and analysis of large-scale datasets will be created and used as a ‘sandbox’ for collaboration across participating institutions.

4. In order to progress the work of the project, a Virtual Sprint was held during the week of 10 – 14 March, 2014. This paper summarises the discussions held during the Big Data Virtual Sprint. The purpose is not to find the answers, but to enumerate equally the opportunities, challenges, issues and questions that the use of Big Data by statistical organizations raises. The paper is comprised of four main parts:

* The definition and sources of Big Data in the context of statistical organizations
* The importance of using Big Data in terms of demand and value
* The changes to statistical organizations that may be needed for Big Data
* The partnerships that may be necessary to achieve the goal of using Big Data

5. The output of the Virtual Sprint is this paper. Following comments from the statistical community, the issues examined in this paper will be followed up at a workshop in Rome on 2-3 April. After that, a number of virtual task teams will be formed. These teams will work on the priority issues identified at the sprint and workshop. They will be required to produce outputs, typically in the form of guidelines and recommendations for statistical organizations, by November 2014. At that time, the results of the Big Data sandbox will also be released.

# What is Big Data in the context of statistical organizations?

## Definition

6. There are numerous existing definitions of Big Data available[[4]](#footnote-4) [[5]](#footnote-5) [[6]](#footnote-6). These definitions are usually split into two parts:

1. A breakdown of the different data sources that can be viewed as Big Data; and
2. The IT techniques and methodologies applied to the data that differ from the traditional treatment of data.

7. It is proposed that statistical organizations regard Big Data as:

*Data that is difficult to collect, store or process within the conventional systems of statistical organizations. Either, their volume, velocity, structure or variety requires the adoption of new statistical software processing techniques and/or IT infrastructure to enable cost-effective insights to be made.*

8. It is unlikely that survey data sources for the statistical industry could be considered Big Data - barring large Census data collections.

9. Many statistical organizations are currently exploring and/or using Administrative Data as a replacement for survey information.  Administrative data has the potential to enter into the Big Data problem domain. This will be the case when this data might expand over several Databases and the linking of such vast amounts of data leads itself to solutions that require Big Data type processing power and solutions.

## Big Data Sources

10. Big data arises from many sources, the following paragraphs set out three types of data sources that can be viewed as Big Data. In Annex A, a detailed typology can be found providing more information about these groups.

1. Human-sourced information (Social Networks )
2. Process-mediated data (Traditional Business Systems and Websites )
3. Machine-generated data (Automated Systems)

11. It is important to distinguish these different types of data sources as each of them brings a different set of considerations for a statistical organization. Table 1 sets out the different considerations for each type of data source.

***Table 1: Different characteristics of the data source types***

|  | Human-sourced | Process-mediated data | Machine-generated data |
| --- | --- | --- | --- |
| Origin | Human beings | Information Systems | Machines/sensors |
| Ownership | Providers/Companies and Institution  (Google, Facebook, Twitter, Flickr/Web site) | Companies  (internal/external) | New companies /  Government |
| Velocity  (frequency of new data) | Fast/ Depend on the phenomenon (goods,  services, prices) | Traditional | Faster |
| Volume | Big | Less big | Bigger |
| Variety | High | Low | Higher |
| Veracity: trust in ... | Provider/Companies and Institution | Companies | Hardware / Provider |
| Programs used to collect  (examples) | Scraping | Scraping / File transfer | Hardware / Dedicated |
| Statistical methods  (examples) | Automated Nonparametric Content Analysis[[7]](#footnote-7).  For quantitative variables: parametric/non parametric models[[8]](#footnote-8) | Traditional Business Intelligence | Data mining,  parametric/non parametric models |

# How ‘big’ is Big Data for statistical organizations?

12. In this section, the importance of Big Data to statistical organizations is examined. The demand for Big Data, the value proposition and the question of whether it is a short term trend or something that will become an integral to the work of a statistical organization are discussed.

## The Demand for Big Data

13. The demand that drives the need for Big Data can be broadly split into three categories: reputational drivers, efficiency drivers and information need drivers.

14. **Reputational drivers** for Big Data centre around the notion that Big Data is an important development that has the potential to significantly impact the statistics industry. It is important that statistical organizations continue to demonstrate their relevance and remain competitive with other emerging sources of data if governments are to continue to see value in official statistics. These drivers seek to exploit new opportunities to keep pace with possibilities. This leads to a 'data-oriented' approach where statistical organizations ask how they can make use of new sources such as:

* Energy Consumption statistics and trends
* Credit card and Consumer Loyalty Information
* Web Search Information
* Satellite and ground sensor data
* Mobile device location data

15. **Efficiency drivers** arise from the challenges that are being faced by many statistical organizations to reduce costs while at the same time producing improved outputs. The emergence of data sources, technologies and methodologies in the world of Big Data are sought to solve existing issues or reduce aspects of the statistical business process that are associated with higher costs.  For example Big Data may be able to add value to:

* Sample frame and register creation – to identify and provide information about survey population units
* Full or partial data substitution – to replace survey collection, reduce sample size, or simplify survey instruments
* Data confrontation, imputation and editing – to ensure the validity, consistency and accuracy of survey data

16. **The need for new statistics or statistics with an improved timeliness or relevance** is also a key driver. This may include making use of new data sources that fill a particular information need to extend the existing measurement of economic, social, environmental phenomena to a high quality for use in policy making. Alternatively it may enable statistical organizations to produce statistics where high quality is less appropriate but can meet public demand on issues of the day. There may be a range of demands that can be assisted through the use of Big Data:

* Improve timeliness of outputs
* Enhance relevance or granularity of outputs
* Increase accuracy or consistency of outputs

## The Value Proposition

17. Big Data presents statistical organizations with a 'triple opportunity'; data sources, technologies and methodologies.  As described in the Definition section, Big Data is not a single 'thing' - it is a collection of data sources, technologies and methodologies that have emerged from, and to, exploit the exponential growth in data creation over the past decade.

18. This presents statistical organizations with the opportunity to utilise aspects that support new and existing business needs. This may take the form of new products based on large volumes of data from a non-traditional source, but also implementing a new methodology which improves efficiency or quality in an existing business domain.

19. Big Data has the potential to create a watershed moment (a critical turning point) for statistical organizations as it encompasses all three areas at once, unlocking the potential to provide analytical insights which were not previously feasible because of limitations or the associated cost.  All three of these areas are inextricably connected to the core business of statistical organizations.

20. There are two key perspectives from which to assess the value of Big Data; the value for the statistical organization in relation to its operational efficiency; and the value for the consumers of information products and services.  These perspectives are additionally informed by privacy, security and legislative considerations.

### Value for statistical production

21. Business value is assumed but not assured. To assist in capability investment decisions, statistical organizations need to be able to demonstrate that the use of new Big Data sources – either alone or in synergy with existing data sources – can actually improve end-to-end statistical outcomes. The value proposition can be expressed in terms of objective criteria such as the cost and sustainability of statistical outputs, and the accuracy, relevance, consistency, interpretability, and timeliness of those outputs stipulated in established quality frameworks. It is likely that each available source will involve a different set of compromises (e.g. more timely but less accurate statistics, more accurate statistics at a greater cost).

### Value for information consumers

22. Statistical organizations have highly qualified methodologists and subject-matter specialists with extensive experience in measuring and analysing ‘real world’ social, economic, environmental and demographic phenomena. They also have a strong base of robust techniques, mature frameworks and rigorous standards. In particular, statistical organizations are uniquely positioned to assess the bias and quality in Big Data sources, and they have sophisticated methods for protecting the privacy of individuals.

23. To obtain the full value from the opportunity that Big Data appears to offer statistical organizations, Big Data initiatives need to be able to clearly align with the strategic intent to provide Official Statistics (in the form of products and services) to government and citizens that are not available from the private sector.  Statistical organizations will additionally need to review and extend their organizational strategies to articulate their role in the Big Data world and ensure this role is clearly understood by government, citizens and commercial providers.  This alignment will ensure that Big Data initiatives are clearly associated with identified and measurable changes in outcomes delivered through improved or new products and services, not separate, isolated initiatives. Statistical organizations should seek to utilise the new data sources, technologies and methodologies presented by Big Data to meet a variety of their existing and emerging needs.

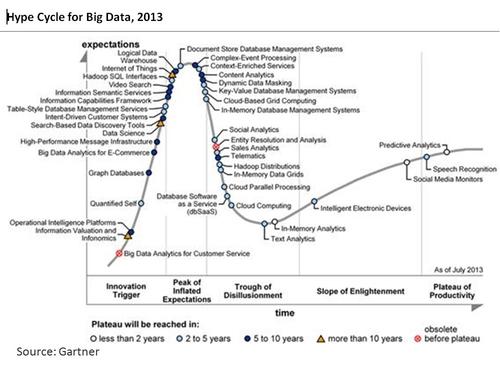
24. Traditionally a venture into any one of these areas would require the undertaking of a comprehensive assessment against known, robust frameworks to ensure there was a sufficient value proposition to warrant the investment required.  For Big Data however, there are no such models or evaluation frameworks to generate a quantifiable value proposition for the use of Big Data by statistical organizations.  Statistical organizations will need to ensure they have in place the capability, methodology and frameworks to undertake these evaluations.

25. The Virtual Sprint team undertook an initial SWOT analysis which can be found in Annex B. The SWOT analysis also includes the initial identification of what benefits are being aimed for and the challenges and risks to be managed, further discussions on these will help statistical organisations understand what it is they want to solve in relation to Big Data.

## Hype cycle

26. Having discussed in the previous section that there is a demand for and value to using Big Data, there is a question about the longevity of the Big Data trend. Statistical organizations need to know whether Big Data is a short term trend or if it is something that will become an important part of the work of statistical organizations.

27. The Gartner hype cycle TMis a research model used widely in the information technology industry to "provide a graphic representation of the maturity and adoption of technologies and applications, and how they are potentially relevant to solving real business problems and exploiting new opportunities".  Shown in Figure 1, the most recently published hype cycle for Big Data was published in July 2013, with Gartner predicting that a number of the technology related aspects of Big Data are currently at or past the 'peak of inflated expectations', with the suggestion that these will reach maturity between 2018 and 2020.



***Figure 1. Gartner Big Data Hypecycle 2013[[9]](#footnote-9)***

28. The Gartner hype cycle TMhas a specific technology focus, however, the model is useful as a mechanism for representing readiness or maturity more generally.  Big Data for Official Statistics encompasses much more than just the technology aspects. The strategy, data and methodology aspects are also key to Big Data having a significant future role to play in the production of Official Statistics.

29. Statistical organizations are currently working to understand exactly which of the data sources, analytical methods and technology trends will be of value or have a significant impact on their core business in the next 3-5 years.  It would be fair to say that for statistical organizations Big Data is also nearing the 'peak of inflated expectations'. There is wide spread scepticism about the value of Big Data in relation to the current activities of statistical organizations, and there is little evidence to suggest that statistical organizations will totally abandon their traditional data sources and methodologies in this timeframe.

30. Figure 2 is an initial attempt to represent the stages of the journey ahead for statistical organizations if Big Data is to become more than a passing trend.  It summarises how statistical organizations may progress on their Big Data journey over the 3-5 year period with respect to Big Data expectations and the capability of the Official Statistics Community.

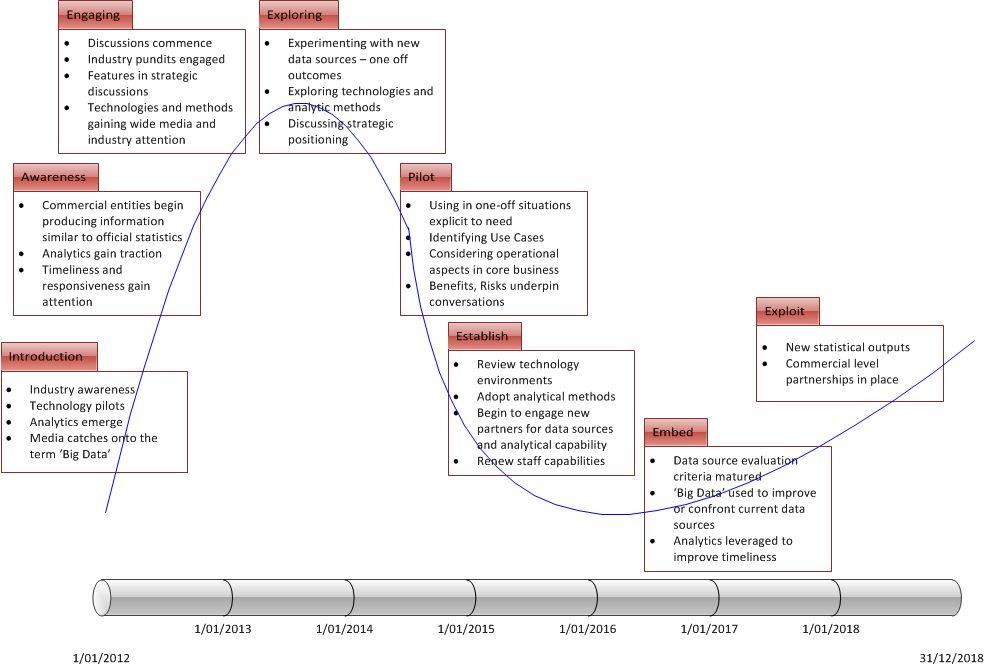


Figure 2. The Big Data journey for statistical organizations

# Are ‘big’ changes needed to build a Big Data Capability?

31. Big Data is difficult to collect, store or process within the conventional systems of statistical organizations. **If statistical organizations want to take advantage of the** data sources, technologies and methodologies that have emerged, there are a number of changes that will need to be adopted (for example, software processing techniques and/or IT).

**32. In the Common Statistical Production Architecture (CSPA), there is a design principle that states that statistical organizations should consider all capability elements to ensure the end result is well-integrated, measurable, and operationally effective.[[10]](#footnote-10)**

33. This section describes t**he changes needed and issues to be faced in building a Big Data capability in statistical organizations**. It examines the capability elements of process, methods, technology, people, and standards, frameworks and policies.

## Process

**34. There are a number of scenarios in which Big Data could be used in statistical organizations. Below**, the main potential Big Data use scenarios are identified and then assessed on how well each is represented in the Generic Statistical Business Process Model (GSBPM v5.0)[[11]](#footnote-11).

Scenario 1: use as auxiliary information to improve all or part of an existing survey

Scenario 2: supplementing/replacing all or part of an existing survey with Big Data

Scenario 3: producing a predefined statistical output either with or without supplementation of survey data

Scenario 4: producing a statistical output guided by findings from the data

35. For the majority of the scenarios where official statistical organizations will make use of Big Data, and in the cases which are most likely, the overarching process is largely supported by the GSBPM. Rather than the process being different, it varies in the focus and importance which is placed upon particular processes. Some key differences may include:

* each output may not have its own distinct supply of data, it is likely that multiple outputs will share and 'mix and match' data
* some steps are not relevant to Big Data sources and will not be undertaken/or replaced by other processes
* activities such as concept management become of greater importance as statistical organizations seek to understand comparability of data sources
* validation of outputs needing new approaches on how to measure accuracy
* structured field tests for surveys are likely to be replaced by feasibility studies of new sources
* more emphasis, as part of ongoing data and metadata management, to identify and describe new sources

36. In the case of producing non pre-defined outputs the process would be distinct from the GSBPM with much less focus on understanding detailed user need and designing precise outputs early in the process. There may be the need to introduce an initial exploration phase as insights are investigated and the potential of a Big Data source evaluated. Alternatively, the GSBPM might be considered to cover this exploration of data given that the processes are not necessarily sequential but this approach is fundamentally different from traditional statistical production.

37. Recently the Statistical Network group on Methodologies for an Integrated use of Administrative Data in the statistical process (MIAD) completed a mapping of Administrative Data processes to the GSBPM[[12]](#footnote-12). When Big Data processes are established, it would be interesting to map them to GSBPM at a lower level than described above.

## Methods

38. Big Data poses challenges in the production of quality statistical information. Existing methodologies must be adapted and new ones must be developed in order to take full advantage of the new data sources.

39. One of the key methodological concerns regarding Big Data is the representativity of the data. Statistical organizations do not have a lot of experience with non-probabilistic samples and they find it difficult to see the value in such unrepresentative data (issues such as self-selection and self-reporting are known to create bias, but the amplitude of the bias cannot be easily assessed in Big Data).

40. For example, it is possible, to a certain extent, to approximate the mood of tweeters by counting (sequences of) words. It would be interesting to view if such results apply to a larger population but it is not possible to do it since many of the characteristics of the tweeters, like gender or age, are not known.

41. While Human Generated data may suffer from several biases such as self-selection, self-reporting and representativity, other sources such as the passively generated data stemming from Automated Systems suffers from a placement bias (the sensor/satellite/log only picks up what it is set to measure and then does so without regards as to the actual occurrence). For instance if traffic accidents in a city were measured using recordings from cameras placed at intersections, it needs to acknowledge that the cameras were likely placed at these intersections because they were more prone to accidents.

42. Like administrative data, Big Data are typically “organic” data[[13]](#footnote-13) collected by others who have a purpose for the data that often differs from that of the statistical organization (that is, a statistical organizations might think that retail transaction data can provide prices for their CPI, while the data generator sees it as a way to track inventory and sales).

43. Beyond the descriptive statistics statistical organizations are used to, it will be necessary to determine under which conditions valid inferences can be made.

44. There are many more methodological challenges related to Big Data. Table 2 enumerates the needs and/or challenges in this area and some possible solutions to them. These methodological challenges will be investigated further in the virtual task teams to be start in April/May.

***Table 2. Methodological challenges***

|  |  |
| --- | --- |
| **Needs and/or challenges** | **Possible solutions to explore...and more questions** |
| 1. Assess fitness for use given specific needs | Quality assessment should be tailored to the specific needs and intended use of the data. For example, lack of construct validity, coverage issues, volatility of the data may dramatically reduce the usability of the data in specific cases. |
| 2. Big Data = Big errors? | Errors can occur at various stages. The Total Survey Error concept applied for surveys[[14]](#footnote-14) could be reviewed to determine how it could be applied to Big Data.  The type of errors encountered could be source specific; others would potentially apply to all sources (i.e. construct validity, coverage error, measurement error, error due to imputation, non-response error).  Sampling error would apply to the specific cases where sampling techniques are used.  A few examples of the type errors are given below.   * **Human beings**: measurement error (mode effect, respondent related factors involving cognitive aspects and/or the will to declare the “true value” * **Information system:** lack of consistency in the construct used in different  files * **Machines/sensors:**measurement errors caused by malfunction, misuse, etc.   Quality requirements for each statistical organization are likely to depend on the intended use of the data and the risk/benefit analyses. |
| 3. A big cleaning job? | When processing data, the steps can include a reception function, where data are first verified and pre-treated, followed by a more in-depth analysis where erroneous data, outliers, and missing values are flagged to be processed.  Some databases may include records for which there is “total non response” meaning that limited information is available about the records (i.e. the addresses in the billing system) but all the values for the variables of interest are missing.  For other databases that involved self selection from the respondent (for example databases using information extracted from twitter) the non-respondents can essentially be associated to the non observed part of the population (which may lead to important coverage issues). All types of data sources can potentially suffer from partial non-response (values for specific variables are missing).  Points to consider   * Knowledge about the data and the associated metadata (including paradata), is a key factor in the development of efficient processing methods. * Given the bulk of the data, outliers may not be influential. * Processing could be limited to a portion of the file is sub-sampling is used. * Many files have few covariates, limiting potentially the efficiency of imputation methods. * Imputation methods are well known for numerical data, what are the options to impute character strings or other type of unstructured data? |
| 4. How to make Big Data small: or do statistical organizations really need all the data? | Reducing the bulk of the data by transforming the data into metrics of interest that meet the analytical needs (for example, calculating summary statistics over specific time periods during the day for data collected minute by minute). Only the summary statistics are used in subsequent analysis.  Another approach is “Downsampling” or subsampling idea to make BD small enough to manage (see the paper by John Dunne[[15]](#footnote-15)) |
| 5. Analysing unstructured data | Text Mining to gather data from non-structured resources (like PDF documents) |
| 6. Combining sources of data | A meta-analysis approach (that is comparing results from a number of studies) could be considered when analysing data from various sources.  Methods currently used to analyse large data sets such as linked open data could be reviewed. |
| 7. How to deal with coverage issues, selection bias, etc. | Consider using the estimation methodology being developed for web non-probabilistic surveys and also for traditional probabilistic surveys that are combined with non-probabilistic surveys (for example web opt-in panels).  Note that mixed results are observed in the literature when using these approaches. The AAPOR task force report on non-probability sampling gives an overview of these challenges and examines the conditions under which various survey designs that do not use probability samples might still be useful for making inferences to a larger population[[16]](#footnote-16). |
| 8. Data stability | How to deal with the volatility of Big Data sources in time series? Use modelling techniques? |
| 9. Dissemination: can the data stay Big? | What about confidentiality concerns, terms and conditions, privacy issues that would require applying disclosure techniques? |
| 10. IT capabilities | How can statistical organizations ensure that the tools available, especially open source tools, perform as they should? What procedures should be put in place to test and validate these tools?  How to adapt and evolve traditional methods and to create new methodologies with all the extra processing resources? |

## Technology

45. Big Data is a relatively new term that has emerged to describe large amounts of data that may have been there for a while, but, thanks to the internet and increased storage capabilities, they are being produced and saved at an accelerated rate. Conducting data analytics with these data could be very valuable and may produce new information, but, until recently, technologies to access and manipulate them were not available.

46. There are a lot of efforts to create platforms and services specifically built to handle this vast amount of data. Possible IT solutions of interest to statistical organizations are:

* Parallelization algorithms (like MapReduce, Sharding, consistent hashing) to make possible the computation on distributed environments
* Extraction, Transformation and Load methods to take unstructured data to a processable form, depending of the type of source real-time capabilities would be needed
* Non-iterative optimization algorithms, to avoid spending large amounts of resources, speed delivery of results and helping the parallelization procedures
* Machine learning, which is useful for classifications of large volumes of data in short times and translations from sources in different languages
* Algorithms for topic models, discourse processing, sentiment analysis, patterns recognition, etc.
* Move the processing to the source of the data, as the raw data is often too big to be moved in its unprocessed state

47. Big Data provides a collection of technologies and methodologies. This marriage of IT and methodology allows statistical organizations to undertake analyses that they have only dreamed of. For example, R code that takes weeks to run in the current IT environment of a statistical organization, could run significantly faster with the benefits of parallel processing.

### Exploring the possibilities

48. The Irish Centre for High End Computing (ICHEC) [[17]](#footnote-17) in conjunction with the Irish Central Statistics Office have volunteered to assist the HLG Big Data Project to implement a Big Data ‘sandbox’ for the testing and evaluation of Hadoop workflows and associated data analysis application software.

49. While individual statistical organizations can experiment with the production of official statistics from Big Data sources (and many are currently doing so or have already done so), and can share their findings and methods with other organizations, this sand box will be able to do the same in a more open and collaborative setting.

50. The sandbox will be a web-accessible environment for the storage and analysis of large-scale datasets. One or more free or low-cost, internationally-relevant datasets will be obtained and installed in this environment, with the goal of exploring the tools and methods needed for statistical production and the feasibility of producing Big Data-derived statistics and replicating outputs across countries. The sandbox will be demonstrated at the MSIS meeting to take place in Dublin on 14 – 16 April 2014.

## People

51. The use of Big Data in statistical organizations will require change to staff. The following sections outline some of the changes that needed in terms of culture and skills.

### Culture change required to exploit Big Data

52. Big Data sources (and regular administrative data sources) offer potentially significant benefits to statistical organizations as a replacement for data traditionally collected through census or sample surveys. These data sources will not replace existing surveys on a like for like basis. By definition these sources utilise the information available rather than being designed for the purpose to which statistical organizations wish to use them. This presents a difficulty for statistical organizations that are, in many cases, embedded in a culture that values high quality and accurate information and regards the best way to achieve this through use of methods where the design can be controlled. Big Data doesn't allow this luxury.

53. The shift to obtaining maximum value from Big Data requires thinking beyond what and how statistical organizations currently collect data and necessitates a return to considering what they are trying to measure. This requires statistical organizations to be much better at assessing the conceptual basis of data. The time required designing questions for use in surveys will reduce but assessing comparability of sources and ensuring accurate and sufficient metadata is captured about the sources will increase.

54. Effective use of Big Data requires a significant level of innovative thinking and an attitude of 'how can statistical organizations make best use of the data available'. This will require risk to be taken, unlike planning and designing a survey, investigations will need to be undertaken that may result in sources being assessed as unsuitable. These feasibility studies must become a core part of data collection activities in the future. Where they currently produce data based on surveys and may supplement this with Big or Administrative Data, in the future they should be looking to produce statistics based on Big Data supplemented by surveys to validate accuracy. This will require a significant change to the mindset needed and capabilities required within statistical organizations. This may result in a paradigm shift where they start to embrace what the data tells them rather than seeking their predefined statistics. An effective future statistical organization will value a wide variety of approaches to data collection and dissemination.

55. Part of this change will require a greater consideration of the quality required for intended use. Statistical organizations will need to ensure that they produce data that is fit for this purpose rather than always striving for the best possible quality. This will allow them to more easily embrace alternative means of statistical production.

56. In many ways the current culture is based upon false assumptions about the traditional methods of statistical organizations. Assumptions about the issues that arise from using Big Data are often based on the somewhat false assumption that the current methods statistical organizations used are the ideal way to produce statistics. In fact they are at a turning point, similar to that at the beginning of the 20th century when sample surveys[[18]](#footnote-18) were introduced, where they do yet have sufficient knowledge and experience dealing with this new approach to truly confirm the value and flaws it may bring to the production of official statistics.

### Skills required to exploit Big Data

57. Statistical production from Big Data will require a new skill set for staff in statistical organizations. These skills will be needed throughout the organization – from managers to methodologists to IT staff. It will also change the needs from new recruits.

58. Generally speaking, analysts in statistical organizations are not programmers so they cannot assess the validity of a particular program routines, the programmers are not methodologists and mathematicians and so they do not have the requisite background knowledge needed to code a new statistical routine or evaluate an existing one. In short, statistical organizations have no data scientists to show them that they are on the right path (using the right Big Data analytics solution).

59. The data science associated with Big Data that is emerging in the private sector[[19]](#footnote-19) [[20]](#footnote-20) does not seem to have connected yet with the official statistics community. Statistical organizations may have to perform in-house and national scans (academic, public and private sector communities) to identify where data scientists are and connect them to the area of official statistics.

60. To meet these needs, statistical organizations may be interested in recruiting people from experimental physics, researchers in physical or social sciences, or other fields with strong data and computational focus. There are also opportunities for statistical organizations to work with academics and organizations that could provide the necessary expertise.

61. In the long term, perhaps the official statistics community could organise data science training and lectures in connection with Big Data players (for example, Google, Facebook, Amazon, ICT providers, Apache foundation) that would lead to a certification in data science.

## Standards, Frameworks and Policies

62. There are a number of standards, frameworks and policies that should be considered in relation to Big Data. The following section looks at the ones discussed during the Virtual Sprint. There are many other relevant standards, frameworks and policies that should be explored and these will be discussed in later stages of the HLG Big Data Project.

### Generic Statistical Information Model (GSIM)

63. During the development of GSIM, Big Data was raised as a use case that the model would likely need to support. Given the lack of experience using Big Data and therefore the corresponding lack of clear cases for what was needed to be supported no information pertaining directly to Big Data was modelled. For the key areas where there was likely to be a different approach required to support Big Data consideration was given to where this could be added in future e.g. types of dataset, methods of acquisition. In regard to unstructured datasets the GSIM documentation notes:

*“GSIM states that all Data Sets must have a structure associated with them. There are, however, cases where a Data Set has no structure – because it was not stored or lost, or it is not known. This type of data may become more prevalent for statistical organizations in the future. In order for a statistical organization to use this data, the data will need to go through a process of being structured. For example, in a case of investigation of new potential data sources for a new or changed Statistical Need, there will need to be a process where these new data are analysed to determine their content and structure. It is only after this process that these new Data Sets can be described using the Data Structure objects. This unstructured data is currently described by GSIM as a Process Input. Organizations could extend GSIM to capture this use case by creating a new subtype of the Information Set object.”[[21]](#footnote-21)*

64. Once clear use cases and processes for Big Data are developed within official statistical organizations GSIM can be expanded to support.

### Quality Framework

65. Principle 5 of the Fundamental Principles of Official Statistics[[22]](#footnote-22) 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****.*

66. This means that it is perfectly acceptable to use any type of Big Data source, as long as it meets the four requirements of quality, timeliness, cost and burden. Whether the data source is fit for use will be an important issue for the statistical industry. The paragraph below set out the considerations statistical organizations will need to take into account when looking at Big Data sources.

67. The **first step** is to identify potential sources of Big Data. This may be undertaken using internet searches, data markets or good practices from other organizations

68. The **second step** is a preliminary assessment of the suitability of these sources. For this, a suitable quality assessment framework is needed. Good starting points could be the [National Quality Assurance Framework](http://unstats.un.org/unsd/dnss/QualityNQAF/nqaf.aspx) developed by the UN Statistical Division, the [checklist for the quality evaluation of administrative data sources](http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CCoQFjAA&url=http%3A%2F%2Fwww.cbs.nl%2Fnr%2Frdonlyres%2F0dbc2574-cdae-4a6d-a68a-88458cf05fb2%2F0%2F200942x10pub.pdf&ei=CMQhU7SNG4fD0QXf-ICICw&usg=AFQjCNH8qVHNekCUnW9vJUzcEOGUptMwaQ&sig2=w5IIGE-DsMXPnb-QYi7fFg) published by Statistics Netherlands, and the checklists produced as [outputs 2a and 2b](http://essnet.admindata.eu/) of the ESSNet project on the Use of Administrative and Accounts Data for Business Statistics. This assessment is likely to include topics such as:

* Relevance - how close are the measured concepts and the concepts required for statistical purposes
* Coverage of the target population
* Representativeness - any biases?

69. The typology provides a good way of compartmentalising the sources of Big Data, and thus, evaluations of veracity can be addressed through this classification. Indeed, each subgroup in the classification can be said to generate data that ought to include some general caveats as to its uses and suitability. In this regards a framework to assess the fitness for use should be part and parcel of a Quality framework.

70. The **third step** is to get access to the Big Data source. In some cases, particularly where the source is freely available to anyone, this will be relatively easy, however other cases, e.g. commercially sensitive data, are likely to require much more effort. Four types of frameworks are needed to govern access:

* Legal framework - legislation to guarantee, permit or restrict access
* Policy framework - government or organizational policies, including codes of practice and data protection
* Organizational framework - data access arrangements with suppliers, e.g. contracts, informal agreements etc.
* Technical framework - data and metadata transfer formats and mechanisms

71. The **fourth step** is to produce pilot outputs for more in-depth assessment at the output stage. This may involve testing data integration methods and routines, in the case of multiple-source outputs.

72. The**fifth step** is to use this assessment to decide whether to use the source in practice.

73. The **sixth step** is repeat the third step, but with a view to longer-term sustainability of access (assuming the source will be used to produce a repeated rather than a one-off statistical output).

74. In the future, an accreditation procedure for Big Data sources should be developed. Eurostat have already begun investigations into this. Drawing upon the approaches used for administrative data, a study commissioned by Eurostat[[23]](#footnote-23) proposes an accreditation procedurewhich could be helpful to statistical organizations when deciding upon whether to use certain Big Data sources to produce statistical outputs conforming to the high standards of official statistics. The procedure comprises five stages (proceeding to the next stage only in case of a favourable outcome):

1. preliminary examination of the source, to assess whether it is potentially useful for statistical purposes (based on metadata and other information about the data and the provider);
2. acquisition and assessment of (extracts of the) data, thus continuing the assessment, but this time based on the actual data;
3. in-depth investigation into the data and its usability, including the production (and assessment) of statistical outputs;
4. decision of the statistical organization;
5. formal agreement with the data provider (sufficiently long to guarantee business continuity with respect to the production of the statistical output in question).

### Privacy Policies

75. Privacy concerns are not applicable to every source of Big Data. Satellite images do not typically include personal identifiable information, nor do most sensor information - they simply measure events. Other types such as retail transaction records (which at most may have a mode of payment section with some information about the payee) can be easily masked to conceal the sensitive identifiable information.

76. Indeed, where Big Data do collect such information, the confidentiality provisions of most statistical organizations are sufficiently advanced to flag such information for suppression. So, insofar as the information procured may contain sensitive personal information that can lead to the identification of an individual, Big Data poses threats similar to traditional data types and the privacy laws in each country provide ample legislation to protect those rights.

77. As with all statistical output, confidentiality has to be ensured when dealing with Big Data. In this respect, the following is important:

* Big Data concerns persons, households, business and institutions who may or may not be aware of their traces in Big Data sets and may or may not agree with the use of their data for statistical purposes. There is a legal issue under what circumstances agreement is required.
* When using Big Data, confidentiality may not be an issue at present, but it may become one at a later time, when other data become publicly available that allow the revelation of individual data.
* When dealing with Big Data sets, it may not be feasible to apply algorithms presently used to ensure confidentiality in respect of the data in more traditional datasets.

78. However, there are still privacy issues to be addressed, particularly regarding information gathered from the Internet. Most sites that collect information about respondents (name/address/email) do so in the course of their business and include a Privacy statement (typically embedded in a User’s Agreement). This details who will get their information and how it will be used. Should a statistical organization then ask for this information, it may be in contravention of this agreement. Unless statistical organizations manage to have that site change their privacy statement to include a "Notwithstanding clause" (for example, "Notwithstanding the possibility of sharing your information with statistical organizations to compile aggregate results and statistics of national importance, your data will not be shared with anyone..."), it may be difficult for statistical organizations to acquire this data.

79. Privacy is therefore inextricably linked with acquisition and these concerns must be addressed during the negotiations with data suppliers...it isn't just their reputation at stake, but ours as well.

80. Even if confidentiality can be ensured, the perception of privacy has to be taken into account. The use of Big Data by businesses and non-statistical government organizations, such as for intelligence purposes, may influence the public attitude towards the possible use of Big Data by statistical organizations. Even if the use of Big Data sources for statistical purposes is legitimate, this may reduce public trust in statistical institutes.

81. If Big Data is used for statistical purposes, the following is important:

* The statistical institute has to observe complete transparency about the source of the data used, the legitimacy of the use, the purpose of the use, the way the data is used and the way confidentiality is guaranteed.
* A pro-active public information policy is advisable, not only to generate public support for the intended use of Big Data, but also to create, in the public mind, a distinction between the use of Big Data by statistical institutes and by businesses.

# Is Big Data too ‘big’ to tackle alone?

82. The US government's "Big Data Research and Development Initiative" launched in 2012 outlined that innovation and collaboration and may act as real catalysts for change, "by improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some the Nation’s most pressing challenges"[[24]](#footnote-24).  On November 12th 2013 the Whitehouse sponsored the 'Data to Knowledge to Action' event highlighting innovative collaborations to benefit Americans.  The event showcased "dozens of public and private organizations describing their contributions to an inspiring array of collaborations that embrace a common theme: sharing resources and drawing upon sophisticated tools to plumb the depths of huge data sets in order to derive greater value for American consumers and grow the Nation’s economy"[[25]](#footnote-25).

83. "America is rich with institutions that are expert at generating data, but as a Nation we have not fulfilled our potential to make the most of these data by merging pre-competitive resources, partnering on analytics, and sharing lessons learned," said John P. Holdren, Assistant to the President for Science and Technology and Director of the White House Office of Science and Technology Policy. "Today’s announcements show that we are maturing in this respect, finding synergies and collaborative opportunities that will accelerate progress in a wide range of scientific, social, and economic domains."

84. The advent of Big Data, and its potential impact on the core business of statistical organizations, signals that partnership models such as those outlined above are the most logical, and maybe the only, way forward. This has been identified by the official statistical community at senior levels (for example in the “Scheveningen Memorandum[[26]](#footnote-26)). The reality is that no statistical organization can take advantage of the opportunities, or respond to the challenges alone; even together the industry would struggle to develop the access to data sources, analytical capability and technology infrastructure needed to deliver to Big Data strategies.

85. Partnership opportunities are easily identified, Table 3 outlines an initial set identified during the Big Data virtual sprint, the list is not exhaustive and presents the opportunity with an associated set of discussion points which should be explored further at either the Big Data workshop in Rome or by task teams set up under the HLG Big Data project.

***Table*** ***3. Partnership*** ***opportunities***

| Partnership Opportunity | Partner Options | Discussion points raised |
| --- | --- | --- |
| New providers and sources  Alternative providers and sources  New sources from existing providers:  Examples:   * Satellite Data (ground usage) * Telecommunication Companies * Social Media * Utility Companies | * Public/Private * National   Multi or Inter national | * Should statistical organizations jointly develop best practices and processes related to; gaining permission to access datasets, gaining permission for different data uses, confirming 'ownership' of the data, determining how, and how often, the data will be accessed? * Are there privacy and confidentiality issues statistical organizations are not yet aware of and how will they test for these during processing and dissemination, particularly if they integrate or match datasets? * Does using the data from these sources in official statistics indicate a 'default seal of quality' for those sources?  Does this 'default seal of approval' provide a competitive advantage to this provider that was not intended?  (For example, there are some private companies very eager to make data available on the proviso that they can make this public knowledge) * Do statistical organizations pay?  How does this compare with the costs of collection and would this be perceived as a 'cost of collection' or commercialisation of data? * Would paying for data impact negatively on reputation? * Can statistical organizations (How can they) leverage their collective bargaining power to enter into these relationships with the multi-national companies, eg Google, Facebook, Twitter, Vodafone)? * Can statistical organizations influence their data streams and output to negotiate 'standardised' data feeds and how does the micro-data?  How much 'overhead' can statistical organizations ask these providers to take on? |
| Data providers undertaking upstream processing so that statistical organizations do not need to invest in the processing infrastructure, capability or storage.   * Instead of ingesting these large datasets, offload part of the pre-processing and ingest aggregates | * Public/Private * National * Multi or Inter national | * Payment for the work the data provider would be undertaking? * What requirements would statistical organisations place on this pre-processing? * What legislative barriers are there, do these vary significantly between countries and jurisdictions? |
| Big Data standards, processes, methodologies | * Academic and 'knowledge' institutions * Scientific communities * Others in the wider statistical community * Researchers | * It is very important in a Big Data world that Statistical Organizations are seen as a significant contributor, if not leader, in specific methodological areas.  How do statistical organizations change the perception that they are only producers, not 'thinkers'? * Are statistical organizations able to distinguish clearly between their actual reputation (as perceived by others) and their desired reputation and how will they action the change? |
| Commercial arrangements, partnership models and ongoing contractual relationships | * Legal | * Statistical organizations do not have experience in establishing these long term, contract based relationships.  How do they ensure they have appropriate legal representation and support from the beginning? |
| Analytical capability | * Private/Public * Others in the wider statistical community | * How well positioned do statistical organizations think they really are in relation to having the capability required to utilise advances in analytical capability, can they accelerate the learning required? * How do statistical organizations engage with private companies to develop and evolve analytical research tools, statistical methods and algorithms that will benefit both the private and public organization(s)? |
| Technology | * Private | * Can statistical organizations develop shared technical platforms? * Can they work with public technology providers to deliver the accessibility, scalability and responsiveness required from the technical environment required to support Big Data initiatives? Securely? |
| Statistical Organizations | * Public | * Can statistical organizations collaborate to develop a virtual centre of excellence around the various aspects of Big Data? * Can statistical organizations develop a Big Data strategy that will result in action within their organizations? |

86. Statistical organizations do not have a strong history of collaborating widely outside of their 'industry sector'.  Commercial, academic and public/private collaborations do occur, however these tend to be for specific pieces of research or one off events, statistical organizations do not have the principles and guidelines for long term collaborative partnerships in place.  They need to develop a culture that is more willing to bring others on board, to trial things and engage with the wider community in constructive debates around the value of official statistics. Statistical organizations need to clearly outline what specific value they will bring to these partnerships, this is a significant gap, as often the very benefit a partner would capitalise on is at odds with the core elements of their reputation; independence, trust and integrity.

87. The individual and collective reputation of statistical organizations is founded on the Fundamental Principles of Official Statistics[[27]](#footnote-27).  These principles provide a robust framework for the discussion points outlined above and will allow for the identification and assessment of the likely impacts to their reputation.  Negative impacts to the reputation of an official statistics organization can affect the continuity of data supply, respondent and customer trust, independence and overall confidence in the quality of official statistics products and services.  In a Big Data world statistical organizations cannot afford to use 'risk to their reputation' as a 'show stopper' as relevance, timeliness and responsiveness are also key.

88. Statistical organizations will need to develop risk mitigation strategies as these impacts are identified, however, as with security they cannot expect that incidents or negative outcomes will never occur.  They should look to adopt reputation (or brand) management practices so that they are in a position to respond quickly and minimise any ongoing negative outcomes.  Other industries have examples of reputation (or brand) management practices that the statistical industry can learn from[[28]](#footnote-28).

# Conclusion

89. This paper has set out a number of challenges and issues regarding Big Data that face statistical organizations. Further work will be undertaken during 2014 to provide guidelines that address these issues. In Annex C, there is a list of potential ideas for further work to be undertaken by the HLG Big Data Project.

# Annex A: A typology of Big Data sources

**1. Human-sourced information (Social Networks)**: this information is the record of human experiences, previously recorded in books and works of art, and later in photographs, audio and video. Human-sourced information is now almost entirely digitized and stored everywhere from personal computers to social networks. Data are loosely structured and often ungoverned.

  1100. Social Networks: Facebook, Twitter, Tumblr etc.

  1200. Blogs and comments

  1300. Personal documents

  1400. Pictures: Instagram, Flickr, Picasa etc.

  1500. Videos: Youtube etc.

  1600. Internet searches

  1700. Mobile data content: text messages

  1800. User-generated maps

  1900. E-Mail

**2. Process-mediated data (Traditional Business systems and Websites)**: these processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected is highly structured and includes transactions, reference tables and relationships, as well as the metadata that sets its context. Traditional business data is the vast majority of what IT managed and processed, in both operational and BI systems. Usually structured and stored in relational database systems. (Some sources belonging to this class may fall into the category of "Administrative data").

  21. Data produced by Public Agencies

      2110. Medical records

  22. Data produced by businesses

      2210. Commercial transactions

      2220. Banking/stock records

      2230. E-commerce

      2240. Credit cards

**3. Machine-generated data (Automated Systems)**: derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world. The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed is beyond traditional approaches.

31. Data from sensors

      311. Fixed sensors

         3111. Home automation

         3112. Weather/pollution sensors

         3113. Traffic sensors/webcam

         3114. Scientific sensors

         3115. Security/surveillance videos/images

    312. Mobile sensors (tracking)

3121. Person (Mobile phone location)

3122. Road (Car, Trucks)

3123. Rail (Trains)

3124. Air (planes)

3125. Nautical (Ships)

   313. Satellite data

         3131. Topographic

         3132. Thermal

         3133. Surveillance

3134. Meteorological

3135. Other

32. Data from computer systems

      3210. Logs

       3220. Web logs

# Annex B: SWOT Analysis and identification of risks and benefits.

The scope of this SWOT analysis is proposed to be “SWOT for the Official Statistics Community (OSC) in the light of Big Data issues”, operationalised as:

* The strengths of the OSC when faced with Big Data challenges
* The weaknesses of the OSC when faced with Big Data challenges
* The opportunities offered to the OSC by the emergence of Big Data
* The threats to the OSC resulting from the emergence of Big Data

|  |  |  |
| --- | --- | --- |
| **Strengths** |  | **Weaknesses** |
| OSC organizations are trusted. This is based on professional ethics, institutional independence, and legal safeguards such as for privacy.  Driven by desire to improve and provide new statistics rather than making money.  The OSC is perceived as an objective arbiter.  Institutional stability (compared to private sector). | Image and culture | The culture of the OSC may be less tuned to what is required in the Big Data era than the culture of the private sector.  Statistical domain silos may be restricted in using innovative statistical methods |
| OSC organizations are in a better position than businesses to combine data sources.  OSC organizations are authorised by law to collect data and has exclusive access to some administrative data sources and other benchmarking sources of primary data. | Data access and data integration | Data access is much more difficult for Big Data than for traditional data sources. For private business this is the other way round. |
| The products of the OSC comply with high standards. | Quality | OSC organizations have long and slow programming and budget cycles compared to the flexibility and responsiveness to developments – and new data needs – that characterises private business. |
|  | Resources | OSC organizations don’t have much control over total budget available for Big Data investment, no venture capitalists  Even within the public sector, OSC organizations are marginal players in the field of Big Data. |
| The OSC toolbox is uniquely positioned to achieve representativity (in some situations). | Methods | Methods for providing reliable official statistics based on Big Data sources lacking – major development needs exist. |
| Mature IT infrastructure and expertise  Being in control of large probability surveys, OSC organizations could adapt them to generate calibration opportunities (e.g. “do you tweet” items in questionnaires) | Systems | Traditional IT infrastructure lacks scalability required for Big Data – new infrastructure is required. |
| The OSC maintains a large body of standards. | Standards | Major adaptation of existing standards |
| Trained statisticians and IT staff in statistics are already close to the “data science” skills required for Big Data (data cleaning, cubes, analytical software, data mining, etc.). Staff well-trained in methodology and statistical domains. | People and skills | The OSC has less knowledge of Big Data than many important players like Google.  The OSC has limited skills and limited IT resources when it comes to the new, non-traditional, technologies used to gather, process and analyse Big Data. |
| **Opportunities** |  | **Threats** |
| By applying knowledge, expertise and funding to Big Data practises, the OSC could help improving the overall open/Big Data community with statistical methodology | Community | The OSC risks being out-competed by other Big Data actors, as such actors could produce faster, more comprehensive, more attractive alternatives to official statistics. If the appearance of said "alternative statistics" is superior, but the true quality is inferior, this will render worse evidence for use by citizens and policymakers. |
| OSC organizations are trusted third parties. A business may be prepared to share information with statistical institutes but not with other businesses.  Statistical information produced by businesses, based on Big Data, may be assessed, reviewed or even certified by OSC organizations. | Image and status | OSC organizations may be perceived to lose relevance. There is a risk of losing budget as a consequence.  External companies could push so far ahead to make OSC existing dissemination irrelevant  There is a risk of external companies leveraging Big Data to cover the same stats domains as OSC and these non-official datasets being more popular and accessible than official datasets because Big Data enables real-time stats which can be very attractive. However, those more popular datasets may well have less quality and metadata than OSC datasets, and may affect policy decisions, or at least poor information |
| Young staff coming in from universities may be very innovative and already have a personal relationship with Big Data (Facebook, Google, Twitter trends) and less constrained by traditional IT and analysis | People and skills | Failure to permit innovative methods might render OSC organizations less attractive workplaces for top talent |
| In the field of statistical cooperation/aid: a young or evolving organization, without legacy systems, could proceed directly to cutting-edge technology instead of building up traditional survey operations | Systems | Continued investment in “legacy systems” (or systems otherwise lacking Big Data capabilities) will crowd out investments in Big Data capabilities |

While benefits and risks are closely related to the opportunities and threats, there is a slight shift in meaning. The benefits and risk identified here are related to what happens when a statistical organization embarks on a Big Data initiative. Thus:

* The benefits potentially resulting from OSC Big Data initiatives
* The risks associated with OSC Big Data initiatives.

A full-blown risk analysis would also include aspects such as likelihood and impact, and perhaps also be expanded to outline strategies to mitigate and manage risks, but for the time being, this analysis is limited to identifying risks. This is also the case for benefits; it is far too premature to try and quantify them.

Concerning the three emerging themes:

* The well-known potential advantages of increased efficiency/reduced resource have to be weighed against the risks of failed investments (wasted resources).
* Similarly, the potential quality increases (timeliness, relevance) have to be weighed against the risk of major quality issues (including discontinuation of statistics – or unwittingly producing statistics of inferior quality, which sometimes may be worse than producing no statistics at all).
* Concerning privacy, the balance is slanted against Big Data, with few apparent advantages.

|  |  |  |
| --- | --- | --- |
| **Benefits** |  | **Risks** |
| Increased efficiency (if a statistical product is possible to produce at lower cost if Big Data sources are used)  Big Data techniques may make some data processing less expensive than traditional techniques. Examples are trade and patents data that are housed in huge relational databases but which easily scale to Big Data repositories | Efficiency and resources | Big Data could be a bubble where statistical organizations invest a lot of R&D time without much to gain  The focus on real-time, “cool” and new datasets may take resources from the established, non-Big Data projects  Training staff in Big Data techniques and purchasing new IT infrastructure may drain resources from core business  Sourcing and processing Big Data source data may require the maintenance of a lot more “exceptional” referential metadata. This will require extra resources |
| Wider product range (if “completely new statistics” based on Big Data are used)  Increased quality (if a statistical product could be improved [timeliness, completeness, relevance, accuracy,...] if Big Data sources are used)  Using Big Data techniques could increase quality and timeliness of existing statistics, through scanning for duplication, and enabling real-time updates  Faster adaptability (if the phenomenon that official statistics tries to capture “moves”, the Big Data source may possibly “move with it”, including new, relevant variables as part of an expanding business) | Quality and business continuity | Pulling Big Data from new, untested data source may be a risk in data quality. Not only quality but resources are required to analyse that risk.  Lack of risk analysis of new commercial Big Data sources may lead to deteriorating quality, increased costs and business continuity issues  Business continuity of statistical production (if the owner of a Big Data Source goes out of business, changes its data structure - or simply decides to start charging for the data) |
| Provision of Big Data based official statistics, produced in compliance with sound SDC principles, may reduce the general public’s use of “non-compliant”, “alternative” statistics produced by other actors. | Privacy, disclosure (and image) | New privacy issues triggered because of increased granularity (maybe leaving old SDC tools useless)  Business secrecy of data providers jeopardised (in particular in the case of oligopolies, such as the mobile phone operator market).  Since privacy is an issue for Big Data (one of the biggest users of Big Data is NSA), there are image risks to statistical institutes that work with Big Data.  Perception of the general public (“guilt by association” of OSC organizations in case a data owner conducts intrusions of privacy). |

# Annex C: Next steps for the HLG Big Data Project

Value Proposition

* Develop framework for the business value of Big Data for statistical organizations[[29]](#footnote-29)

SWOT Analysis

* Review and finalise

Hype cycle

* Review validity of presentation (there are other models which could be used).
* Review and finalise

Methods

* Prioritise issues in Table 2 to determine which are the most important to work on during 2014

Technology

* Compile table of Big Data technologies currently used by statistical organizations to indicate:
  + what a statistical organization should consider or not,
  + help speed up the technology platform analysis, and
  + potentially enable partnerships with organizations to reduce costs
* Finalise specific technologies and methodologies to be tested in the sandbox

People

* Investigate possibilities for developing data scientist skills

Standards, Policies and Frameworks

* Review current quality frameworks and adapt to Big Data
* Review current privacy policies and determine next steps
* Consider what other frameworks are needed
* Consider metadata management frameworks and related standards

Partnerships

* Investigate questions raised in Table 3

1. http://www.unece.org/stats/documents/2012.10.hls.html [↑](#footnote-ref-1)
2. *Strategy to Implement the Vision of the HLG* available at http://www1.unece.org/stat/platform/display/hlgbas/HLG+Strategy [↑](#footnote-ref-2)
3. What does “big data” mean for official statistics? available at http://www1.unece.org/stat/platform/download/attachments/58492100/Big+Data+HLG+Final.docx?version=1&modificationDate=1362939424184 [↑](#footnote-ref-3)
4. http://strata.oreilly.com/2012/01/what-is-big-data.html, [↑](#footnote-ref-4)
5. <http://www.gartner.com/it-glossary/big-data/> [↑](#footnote-ref-5)
6. http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html [↑](#footnote-ref-6)
7. D. J. Hopkins and G. King (2010)  A Method of Automated Nonparametric Content Analysis for Social Science [↑](#footnote-ref-7)
8. H. Variant (2013) Big Data: New Tricks for Econometrics; H. Choi, H. Varian (2011) Predicting the Present with Google Trends; T. Hastie, R. Tibshirani and J. Friedman (2009)  The Elements of Statistical Learning  [↑](#footnote-ref-8)
9. <http://www.informationweek.com/big-data/software-platforms/big-data-meets-trough-of-disillusionment-gartner/d/d-id/898939> [↑](#footnote-ref-9)
10. http://www1.unece.org/stat/platform/display/CSPA/CSPA+v1.0 [↑](#footnote-ref-10)
11. http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0 [↑](#footnote-ref-11)
12. http://www1.unece.org/stat/platform/download/attachments/99484305/MIAD%20Output%20A1%20%281%29.docx?api=v2 [↑](#footnote-ref-12)
13. R. M. Groves (*Three Eras of Survey Research*, Public Opinion Quarterly 75: 861-871, doi:10.1093/poq/nfr057; http://poq.oxfordjournals.org/content/75/5/861.full.pdf) [↑](#footnote-ref-13)
14. http://poq.oxfordjournals.org/content/74/5.toc [↑](#footnote-ref-14)
15. [www.statistics.gov.hk/wsc/STS018-P3-A.pdf](http://www.statistics.gov.hk/wsc/STS018-P3-A.pdf)  [↑](#footnote-ref-15)
16. <http://www.aapor.org/AM/Template.cfm?Section=Reports1&Template=/CM/ContentDisplay.cfm&ContentID=6055> [↑](#footnote-ref-16)
17. http://www.ichec.ie/ [↑](#footnote-ref-17)
18. R. M. Groves (*Three Eras of Survey Research*, Public Opinion Quarterly 75: 861-871, doi:10.1093/poq/nfr057; http://poq.oxfordjournals.org/content/75/5/861.full.pdf) [↑](#footnote-ref-18)
19. T. Davenport, D. Patil, Data Scientist: The Sexiest Job of the 21st Century,

    http://128.255.244.58/strategic/articles/data\_scientist-the\_sexiest\_job\_of\_the\_21st\_century.pdf [↑](#footnote-ref-19)
20. M. Schaklett, In-demand Big Data skills: a mix of old and new, http://www.zdnet.com/in-demand-big-data-skills-a-mix-of-old-and-new-7000020949/ [↑](#footnote-ref-20)
21. http://www1.unece.org/stat/platform/display/gsim/III.+Foundational+Information#III.FoundationalInformation-\_Toc375075368 [↑](#footnote-ref-21)
22. http://www.unece.org/stats/archive/docs.fp.e.html [↑](#footnote-ref-22)
23. Petrakos, M., Sciadas G. and Stavropoulos, P. (2013), *Accreditation procedure for statistical data from non-official sources*, (study commissioned by Eurostat; currently not available in the public domain). [↑](#footnote-ref-23)
24. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big\_data\_press\_release\_final\_2.pdf [↑](#footnote-ref-24)
25. http://www.whitehouse.gov/sites/default/files/microsites/ostp/Data2Action%20Press%20Release.pdf [↑](#footnote-ref-25)
26. Scheveningen Memorandum on “Big Data and Official Statistics” (2013), http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp\_ess/0\_DOCS/estat/SCHEVENINGEN\_MEMORANDUM%20Final%20version.pdf [↑](#footnote-ref-26)
27. http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf [↑](#footnote-ref-27)
28. [http://knowhownonprofit.org/how-to/how-to-manage-your-organizations-reputation](http://knowhownonprofit.org/how-to/how-to-manage-your-organisations-reputation) [↑](#footnote-ref-28)
29. An example can be found here: The Business Value of Big Data for Financial Services

    <http://alexandria.tue.nl/extra2/afstversl/tm/Vries_de_2013.pdf> [↑](#footnote-ref-29)