ESS Big Data Event

Rome 2014

Technical Workshop Report

Editor:

José L. CERVERA (DevStat)

Authors:

José L. CERVERA and Paola VOTTA (DevStat)
Donatella FAZIO and Monica SCANNAPIECO (ISTAT)
Reg BRENNENRAEDTS and Tommy VAN DER VORST (Dialogic)

The logo was designed by artist Toni SIMARRO (http://tonisimarro.tumblr.com/)

Table of Contents

Acronyms

CROS	CROS Portal for Collaboration in Research and Methodology for Official Statistics (www.cros-portal.eu)
DGINS	Conference of the Directors General of the National Statistical Institutes
DIME	Directors of Methodology
EMOS	European Master in Official Statistics
ESAC	European Statistical Advisory Committee
ESGAB	European Statistical Governance Advisory Board
ESS	European Statistical System
ESSC	European Statistical System Committee
ESTP	European Statistical Training Programme
EU	European Union
HLG	High Level Group for the Modernisatoin of Statistical Production and Services
HR	Human Resources
IAOS	International Association of Official Statisticians
ICT	Information and Communication Technologies
ISI	International Statistical Institute
ITDG	Information Technologies Directors Group
MEETS	Modernisation of European Enterprise and Trade Statistics
MS	Member State of the EU
NSI	National Statistical Institute (generic denomination)
NSO	National Statistics Office
NTTS	Conference on New Techniques and Technologies for Statistics
OECD	Organisation for Economic Co-operation and Development
OS	Official Statistics
SHC	Scheveningen Memorandum Challenges
UNECE	United Nations Economic Commission for Europe
VIP	Vision Implementation Project

1. Introduction

This report[1] summarizes the presentations and discussions that took place during the 2014 ESS Big Data in Official Statistics event organized by Eurostat in Rome (31 March-1 April) and provides recommendations for the way ahead in the promotion of the use of Big Data for the production of European statistics.

The ESS Big Data Event was organised by Eurostat with the support of a consortium led by DevStat (ES) with VerbiVis (LU), Instituto Nazionale di Statistica-ISTAT (IT), and Dialogic (NL)[2] under a contract for the project: *"Facilitation of methodological cooperation in the ESS".*

- Section 2 presents a brief summary on the current situation of the using of Big Data in Official Statistics and the objectives of the event.
- Section 3 reports on the organizational aspects and recalls lessons learnt that could be of use for the following events.
- Sections 4 and 5 are organized following the sessions structure in plenary and parallel sessions, summarizing the discussions.
- Finally, Section 6 recalls the recommendations issued by the participants of the different parallel sessions, and relate them to the Scheveningen challenges described in Section 2.

The report has been prepared by the consultants of the consortium led by DevStat (ES) with VerbiVis (LU), Instituto Nazionale di Statistica-ISTAT (IT), and Dialogic (NL)[3], in particular by the facilitators of the parallel sessions. The contributions during the event of all participants are gratefully acknowledged.

The presentations and other event material (programme, logistic information, newsletters, etc) are available in the dedicated section[4] of the CROS-portal.

[1] A separate project administrative report is delivered separately.

[2] The Fachhochschule Nordwestschweiz (CH), a member of this consortium, did not provide expert input for this event, but for the ESSnet/ESS.VIP envet organized in Valencia under the same contract.

[3] The Fachhochschule Nordwestschweiz (CH), a member of this consortium, did not provide expert input for this event, but for the ESSnet/ESS.VIP envet organized in Valencia under the same contract.

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

2. Modernising European Statistics through Big Data

Official statistics play a fundamental role in today's society. The availability of impartial and objective statistical information is essential for all decision-makers. Statistical information underpins transparency and openness of policy decisions, and official statistics therefore represent a public good providing a basis for the smooth functioning of society. At the EU level, European statistics have become increasingly important for the development, implementation, monitoring and evaluation of EU policies.

European statistics are developed, produced, and disseminated on the basis of uniform standards and harmonised methods. The National Statistical Institutes (NSIs) of the Member States (MS) collect and produce harmonised data that are compiled by Eurostat to construct statistics at EU level.

This traditional way of producing statistics is, however, no longer fully adapted to the changing environment, as it was recognised in the "Vision for European Statistics"[5]. The European Statistical System (ESS) has indeed acknowledged the need for a thorough modernisation of the processes of production and dissemination of European Official Statistics. In particular, the Joint Strategy of the ESS, adopted to face the important challenges which surround the sector, asks for the integration of different data sources and domains, as well as for the development of new ICT tools are available for the production and dissemination of data, including the collection of 'Big Data' from Web 2.0.

In many sectors of society, Big Data is playing an increasingly important role. In the private sector, Big Data offers many companies the opportunity to increase their efficiency, for example by giving greater insight into customer demand. While there does not seem to be one single definition of what Big Data is (or is not) there seems to be a consensus that Big Data may be of real interest for Official Statistics.

2.1 Previous global initiatives on Big Data for Official Statistics

Many of the initiatives on the use of Big Data in Official Statistics have taken place under the umbrella of the High Level Group for the Modernization of Statistical Production and Services (HLG), in the context of the Conference of European Statisticians[6]. Several introductory documents have been produced. At a **High-Level Seminar on Streamlining Statistical Production and Services** (St Petersburg, October 2012), participants asked for a strategic document addressed to the heads of statistical offices. The paper "*What does Big Data mean for official statistics*"[7] states that "it is unlikely that National

[5] Communication of the Commission COM (2009) 404 final.
[6] See http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics,
[7] http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622.

Statistical offices will lose the official statistics trademark, but they could slowly lose their reputation and relevance unless they get on board" of the Big Data era.

At the 44th session (2013), the United Nations Statistical Commission organized a side event on “Big Data for Policy, Development and Official Statistics”[8]. The concept paper of this event launched a series of questions for NSIs;

- Should NSIs expand its business operations to take on the opportunities of using Big Data for official government purposes?
- Should NSIs take on a new mission as a trusted 3rd party whose role would be to certify the statistical quality of many of these newly emerging private sector sources?
- Should NSIs become a “clearing house” for statistics from non-traditional sources that meet their quality standards? Should NSOs use non-traditional sources to supplement (and perhaps replace) their official series?
- How might NSIs acquire people with the knowledge and skills to effectively take advantage of Big Data for official statistics purposes?

The issues of partnership with the private sector, certification of Big Data sets, integration of new sources and acquisition of skills were therefore put onto the agenda of NSIs.

In the process of modernization of European statistics, the ESS, both at the Eurostat and the EU NSIs level, has undertaken an approximation to Big Data, which is reflected in the Scheveningen Memorandum, and the establishment of a Eurostat Task Force on Big Data.

2.2 The ESS response: the Scheveningen Memorandum and the Eurostat Task Force on Big Data

At the Directors General of the EU NSIs (DGINS) meeting of September 2013, the heads of EU statistical offices recognised, as stated in the **Scheveningen Memorandum** (see Annex 1), the relevance of Big Data for the ESS and the need for adopting a related action plan considering the development of methodology, capabilities and a legislative framework, to be implemented in partnership with governments, academics and private sources.

The DGINS identified a number of challenges described in the Scheveningen Memorandum related to the use of Big Data in European official statistics, which are summarily described below (and referred to as “Scheveningen Challenges [SCH]):

[8] http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/concept_note.pdf

- [SCH1] Acknowledge that Big Data represent **new opportunities** and encourage the ESS to **examine its potential**.

- [SCH2] Recognise that an **"Official Statistics Big Data strategy"** has to be prepared, in coherence with government strategies at the national and EU level.

- [SCH3] The use of Big Data requires an adaptation of/to **legislation** regarding use aspects (i.e. with respect to the access and use of data), privacy aspects (i.e. managing public trust and acceptance of data re-use and its link to other sources, security of private data).

- [SCH4] **Sharing** the existing experiences within the ESS and establishing **collaborative** Big Data projects.

- [SCH5] Technical **skills** have to be built in the ESS, through training courses, dedicated communities and cooperation with academia.

- [SCH6] **Synergies** and partnerships with other branches of government, academia and private sector have to be developed. Presently a limited number of NSIs are actively engaged with technological aspects of Big data, but it is mainly the private sector which leads the work on Big data analytics tools and solutions. Synergies between NSIs and the private sector should be not limited to technological issues but also tackle sensitive issues such as privacy, trust and corporate competitiveness, as well as the legislation framework of the NSIs.

- [SCH7] New **methodological developments** are necessary for the use of Big Data in official statistics, including for **quality assessment** and **IT-related issues**. Aspects of Big Data that oblige to rethink the statistical methods to deal with it include the heterogeneity, lack of structure (requiring important work to prepare the data for statistical production), volume (which hampers the use of standard statistical tools), representativeness and coverage over the populations of interest for official statistics.

- [SCH8] Require a **follow-up of the implementation of the memorandum,** adopt an **action plan by mid-2014 and integrate it in the Statistical Annual Work Programmes** of Eurostat.

Eurostat set up an internal Task Force on Big Data with the lifespan 2014-2016 with the following objectives:

- to lead and co-ordinate developments within the ESS and the European Commission with regard to maximising the potential of Big Data for Official Statistics and evidence based policy making
- to develop — together with all members of the ESS — an ESS Big Data strategy along the lines of the Scheveningen Memorandum.

At the ESS level, an ESS Task Force on Big Data, composed of NSIs but in partnership with other international organisations (OECD, UNECE), Commission services (DG CNECT) and academics has been set up.

The Eurostat Task Force aims at network building (including public-private partnerships), cooperation at international level (e.g. with UNECE), stock-taking and transfer of know-how & skills, exploring concrete areas of application (including methodological issues such as coverage or completeness).

2.3 The objectives of the 2014 ESS Big Data event

The **2014 ESS "Big Data in Official Statistics"** event, organised in Rome on 31st March and 1st April 2014, aimed at contributing to the implementation of the Scheveningen Memorandum, by achieving the following objectives:

- Increasing awareness of the Big Data challenges in the ESS and propose ways to address them ([SCH1], [SCH2];
- Identifying and sharing emerging best practices concerning how to make use of Big Data in official statistics ([SCH4]);
- Identifying synergies within the ESS and with the private and academic sectors, where joint development efforts might serve the interest of multiple ESS NSIs or the entire ESS ([SCH4], [SCH5], [SCH6]);
- Identifying the legal, technical and methodological requirements to use Big Data for the production of official statistics ([SCH3], [SCH5], [SCH7]);
- Provide expert input to the preparation of an action plan by the ESS and its implementation in the Statistical Annual Work Programmes of Eurostat ([SCH8]).

The event therefore contributed to the aims of the ESS and Eurostat Task Forces.

Organisational details of the event are described in the next section. The issues discussed in plenary and parallel sessions, and the recommendations elaborated during the parallel sessions of the event, which are intended to contribute to address the Scheveningen challenges, are summarised in Sections 4, 5 and 6 respectively.

3. Organization of the 2014 Big Data event

The ESS Big Data Event was organised by Eurostat with the support of a consortium led by DevStat (ES) with VerbiVis (LU), Instituto Nazionale di Statistica-ISTAT (IT), and Dialogic (NL)[9] under a contract for the project: *"Facilitation of methodological cooperation in the ESS".*

The objective of the contract was the organisation of two events on methodological issues as an opportunity for specialists and institutions of the ESS to meet and exchange information, good practices and future needs.

The first event, **the European Modernisation Workshop** (ESSnet/ESS.VIP workshop) was held in **Valencia** on 23 – 24 January 2014; the second event on **Big Data in Official Statistics**, wich took place in **Rome** last 31 March-1 April 2014, is the subject of this technical report.

The project team from the implementing consortium consisted in a Project Manager (José CERVERA, DevStat), an Event Manager (Lucia DEL VITA, VerbiVis), four Facilitators (Monica SCANNAPIECO and Donatella FAZIO of ISTAT, Reg BRENNENRAEDTS of Dialogic and José CERVERA as well) and one Assistant Project Manager (Paola VOTTA, DevStat). From Eurostat, a Project Manager was assigned (Willem KLOEK, Eurostat/B1).

The following stages took place between October 2013 and March 2014:

3.1 Preparation Phase

- A concept note was drafted by the consultants, describing the orientations of the event and each session in particular. This note was extensively discussed with Eurostat to establish the format of the event, select the topics for sessions, and propose experts, speakers and session chairs.
- An invitation letter was sent in January 2014 to all EU NSIs, providing logistic and technical information on the event.
- A logo for the event and all its visibility elements: banners, badges, folders was designed by the consultant and validated by Eurostat.
- All information was uploaded on the CROS-portal, especially for the occasion (http://www.cros-portal.eu/content/big-data-event-2014). On the website, participants were able to find the registration form for the event, all the information related to the venue, accommodation and

[9] The Fachhochschule Nordwestschweiz (CH), a member of this consortium, did not provide expert input for this event, but for the ESSnet/ESS.VIP envet organized in Valencia under the same contract.

reimbursements, as well as all the technical information, the programme, the concept paper and the Expert Lectures and presentations.

3.2 Implementation phase

The attendance included 107 participants from 27 different countries: 21 EU Member States plus one Candidate country (Turkey) and 5 non-EU countries (USA, México, Switzerland, Australia and Canada).

- The event was held over one day and a half, organised in plenary and parallel sessions. The agenda of the event is presented in Annex 2.
- In addition to technical sessions, networking among participants took place during coffee breaks between sessions as well as in a social dinner organised on the evening of the first day;
- Upon arival, delegates received the reimbursement for their travel expenses;
- During the evening of the second day of the event participants filled the evaluation survey previously distributed.

3.3 Follow-up Phase: reporting and evaluation

The follow-up phase included an evaluation of the event, based on the answers of participants[10] to a questionnaire that was filled in at the end of the event.

Organizational aspects where highly evaluated (information provided previous to the event, logistics of the event, etc.), rated above 6 (on a scale 1 to 7).

The expert lectures were rated above 5.7 on 7 showing the relevance of the topics and the quality of the presentations.

For each session, the relevance of the discussion, the function of the facilitator, the quality of the conclusions and an overall assessment were evaluated. Detailed results have been provided in a separate administrative report by the consultants, but the following lessons can be drawn:

- More time for discussion on **methodology** should be devoted, probably discussing concrete statistical issues by groups of specialists;

[10] The rate of response was low: only 36 out of 107 participants filled in the questionnaire, probably due to its distribution during parallel sessions.

- Discussions on **strategic issues** should be re-launched, at a higher level of decision. An event with participation of specialists rather than managers of NSIs is probably not the most suitable venue for dealing with issues that require being aware of ESS strategic environment.

4. Expert lectures and plenary sessions

During one afternoon and one full day, pasticipants assisted to an Opening Session, two Expert lectures, one panel discussion and four parallel sessions.

All sessions enjoyed the participation of speakers and experts from the EU NSIs, international institutions as Eurostat, the European Central Bank, the UN and UNECE Statistical Divisions, the OECD, the World Bank, as well as academic centres as University of Berkeley (USA), La Sapienza (Italy), Universitá di Napoli (Italy), Universitá de Pisa (Italy), Durham University (UK), London University (UK) and GENES (France).

4.1 Opening session: the current state of Big Data for official statistics

The opening session, inaugurated by Prof. Golini, president of ISTAT, introduced the issue of the use of Big Data in National Statistical Offices. ISTAT, as Mr Emanuele BALDACCI described, has put in place an inter-institutional technical commission on Big Data (with the National Research Council, Universities and private organizations) with three subgroups on statistical methodology, computer science and knowledge management (experiences learnt). Three projects have been selected as pilot ones: “Persons and Places” (on mobility using mobile phone data), “Labour market estimation” (on forecasting unemployment rates using Google Trends) and “Use of ICT by enterprises” (using webscraping techniques and text mining). It was recalled in the opening that Big Data are inserted in the ESS Vision for the modernization of European statistics.

Eurostat recalled the previous experiences in the framework of the High Level Group project (see section 2) and the expectations for the event:

- ✓ “crowd-sourcing” the ESS Action Plan for Big Data roadmap;
- ✓ finding inspiration and learning from best practices
- ✓ networking amon interested official statisticians and other statistical experts .

Mr Reg BRENNENRAEDTS, of Dialogic (NL), mentioned a previous project “Internet as a Data Source”. Since private companies are more pragmatic and flexible than NSIs, the use of Big Data may represent, he suggested, either an opportunity for Official Statistics, or a threat in the sense that they represent a disruption in the information industry. The Schumpeterian idea of creative destruction was recalled in the context of Official Statistics facing the challenge of Big Data, as it has happened with the introduction of digital technologies and in particular of web services and applications with respect to the music,

encyclopedias, travel bookings, the mapping industry, etc. The NSIs can offer quality stability and trust in this renewed infomraitn industry, but still can learn from private companies.

Mr Michail SKALIOTIS, leader of the Eurostat Task Force on Big Data, presented the timeline of ESS action plan, starting on 2013 with the adoption of the Scheveningen Memorandum, the set up of the Eurostat Task Force, the this ESS Big Data event and the expectation to launch during 2014 an action plan. In parallel, the ESS had participated in the HLG project (see Section 2), which had establishd a Big Data and a Sandbox Task Teams in 2013. He informed the participants of a parallel event which took place on 2nd and 3rd April 2014 in the form of a "Sprint meeting". As for the objectives of the ESS Big Data event, it was recalled that strategies and solutions for the integration of Big Data in Official Statistics should be preceeded by an exploratory in-house research. In the long term perspective, the ESS strategy for Big Data should reach out beyond the Official Statistics environment, considering the global initiatives (such as the Post2015 Development Agenda and the data revolution initiative[11]) and the possibility of a partnership with the private sector.

For Eurostat, the ESS Big Data event was an opportunity for inspiration and networking among interested official and academic statisticians. The event was itself an achievement towards the Scheveningen challeges (in particular [SCH8], calling for a follow-up action of the DGINS meeting).

4.2 Expert lecture 1: Opportunities and methodological challenges of Big Data for official statistics

Big data is an extremely interesting data source for statistics. Since more and more data is generated in our modern world and remains to be stored, this data has the potential to replace or provide additional information for official statistics. Especially in cases where the response of surveys declines, information gathered from Big Data could be an interesting addition. However, extracting statistical relevant information from Big Data sources is not an easy task.

Definitions for Big Data are diverse and depend on the target audience:

- ✓ "*Data that are 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*

[11] See for instance http://post2015.org/tag/data-revolution/.

processing techniques and/or IT infrastructure to enable cost-effective insights to be made." (HLG Virtual sprint paper)

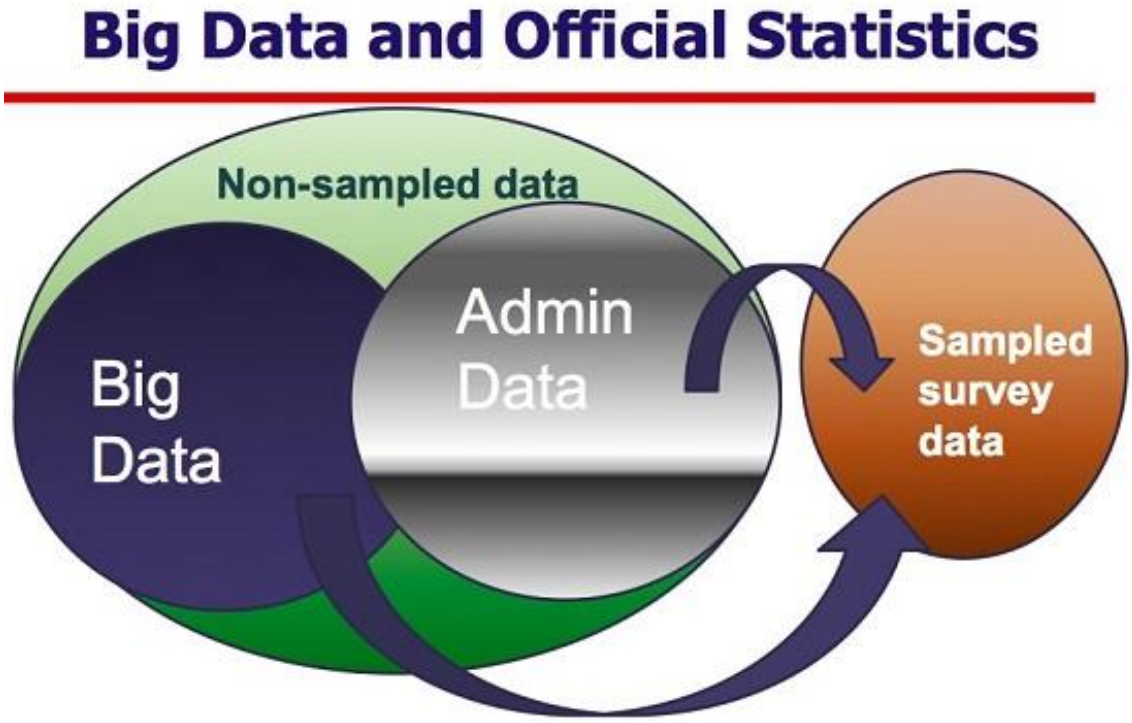
- ✓ "*Big Data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time*." (Wikipedia)

- ✓ "*Data sources that are awkward to work with.*" (given by a Big Data user).

The Expert Lecture was delivered by Dr Piet DAAS, of Statistics Netherlands, where an ad hoc group has been set-up to undertake exploratory methodological activities in the area of Big Data.

In this lecture, opportunities and challenges associated with using Big Data for official statistics were discussed. The findings presented were based on the Big Data studies performed at the Central Bureau of Statistics of The Netherlands (Statistics Netherlands), such as traffic loop detection records (e.g. for traffic statistics), mobile phone data (e.g. for mobility and tourism statistics) and Dutch social media messages (e.g. for consumer confidence). A graphical classification of data sources, presented in the lecture, is given in Figure 1.

Figure 1. Data sources.



A broad classification of Big Data sourecs was given already in the HLG Virtual Sprint paper[12]:

A) Human-sourced data ('Social Networks'), such as social media messages, blogs, web searches;

B) Process-mediated data ('Traditional Business Systems and Websites') such as credit card, bank or on-line transactions, CDR, product prices, page-views

C) Machine-generated data ('Automated Systems') such as rRoad or climate sensors, satellite images, GPS, etc.

The exploratory results achieved by Statistics Netherlands in the field of mobility or transport statistics reveal the potential use and the issues and challenges specific to the use of Big Data for official statistics.

In terms of skills, Statistical Netherlands has put in place a Big Data team. Its members have the necessary soft and hard skills, such as

- "open-mindedness" to work with data sets that do not comply with the usual statistical standards of sampling theory;
- IT skills and IT affinity, to learn and use sophisticated software;
- data-driven, pragmatic attitudes that facilitate the exploration of data sets.

Issues that had to be taken into account include the legal measures to protect privacy and security

4.3 Expert lecture 2: Big Data technologies and platforms for official statistics

This lecture, delivered by Dr Antonino VIRGILLITO from ISTAT, presented a reasoned overview of Big Data technologies and platforms, highlighting their mutual differences and describing the complementary tools and techniques that facilitate statistical users in data analysis.

Rather than providing a comprehensive coverage of all the software available on the market, the lecture focused on representative examples of Big Data

[12] "Final project proposal: The Role of Big Data in the Modernisation of Statistical Production" ("UNECE Sprint paper"), http://www1.unece.org/stat/platform/display/msis/Final+project+proposal%3A+The+Role+of+Big+Data+in+the+Modernisation+of+Statistical+Production.

tools, clarifying their target scenario and the technical and organizational requirements for their usage. Specific use cases and practical examples are discussed showing how Big Data technologies can integrate with statistical software and other typical components of a NSI enterprise architecture.

Specific IT tools and techniques have to be adopted to deal with Big Data. The huge size of data sets requires using distributed file systems to oversome physical limitations. Platforms for managing complex storage systems are therefore required (such as Hadoop HDFS[13]). Hadoop, which has become a standard for Big Data processing, is an open source supported by most major vendors and provides almost unlimited scalability.

To manage the data processing, the most widely used programming paradign is based on the MapReduce[14] algorithm which allows programmes to be executed in parallel on a cluster of computers. The combination of Hadoop and MapReduce allows working with a large number of storage facilities that process enormous amounts of data.

Other programming tools include high-level languages for data manipulation (such as Pig[15] to query data on Hadoop clusters, or Hive[16]) before using statistical software (R, SAS, SPSS or similar).

With respect to options for the deployment of IT infrastructure for Big Data, the expert mentioned that it can be an in-house solution, based on the cloud (which reduces hardware and software costs) or based on specific computing appliances (with higher costs). The skills necessary at NSIs to manage the complete process of working with Big Data should therefore include:

- Data anlysts, able to use statistical software and visual analytics tools;
- Data scientists, also able to manipulate complex data sets (using for instance MapReduce and Pig);
- Data engineers, designers of the IT architecture for collecting and processing data;
- Data integrators, running the ETL (Extract, Transform and Load) processes and
- Systems managers, setting up and managing the physical infrastructure.

[13] http://hadoop.apache.org/.

[14] See for instance http://research.google.com/archive/mapreduce.html.

[15] http://pig.apache.org/.

[16] https://hive.apache.org/.

5. Issues discussed in parallel Sessions

5.1 Parallel session on statistical methodology

A session on the methodological aspects of the integration of Big Data into the official statistical production was facilitated by Monica SCANNAPIECO (ISTAT). It consisted in a preliminary presentation by the facilitator, followed by open discussion by all the participants.

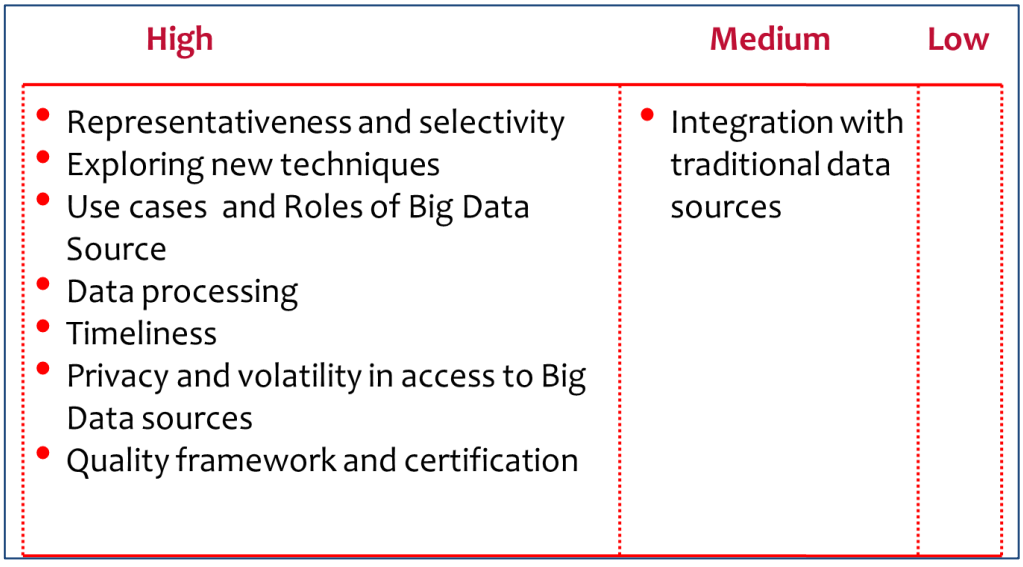
The Scheveningen Memorandum challenge that considers methodological issues is:

[SCH7] New developments in methodology and quality needed for the use of Big Data

Both statistical methodology and quality aspects are in the scope of the challenge.The principal issues that were identified and discussed during the workshop are described below (see Figure 2).

A list of recommendations is given in section 6.1:

Figure 2. Methodological issues discussed (by priority)



With **high priority**, the issues to be dealt with are:

Representativeness and selectivity

The populations covered by Big Data sources are not typically the target populations of Official Statistics and are often not explicitly defined.

Moreover, given that the Big Data generating mechanisms are not under the control of statistical institutions, deriving data from Big Data sources can be selective and not representative.

Dealing with the representativeness of Big Data sets is not easy, especially because it is not always feasible to assess the relationships between the covered population and the target population on one side, and to estimate the bias, on the other side.

Exploring new techniques for data analysis

This issue is related to the major paradigm shift in the analysis activities caused by the usage of Big Data. In particular, design-based and model-based approaches, which are traditionally used within Official Statistics, may not be directly applied to Big Data analysis. Approaches that proceed by exploratory analysis, like those based on data mining and machine learning, could be more appropriately applied. However, they are new for Official Statistics: though they are currently successfully applied in specific domains (e.g. customer profiling), their usage in the Official Statistics domain must be investigated.

Use cases and roles of Big Data Source

Big Data is a very broad concept that must be narrowed in two directions, namely:

- Identifying use cases that explain concrete usage of Big Data sources for OS purposes.
- Clearly identifying the role(s) of Big Data sources within the OS production process.

This issue has been identified among the *highest priority issues*.

Data processing

This issue relates to three different aspects that are very much important for dealing with Big Data in OS, namely: (i) data preparation, (ii) data filtering and (iii) data reconciliation.

With respect to (i) data preparation, Big Data sources are typically event-based rather than unit-based, as it traditionally happens for OS survey data (or for administrative data). Hence a first preparation (data manipulation) step is needed in order to deal with such new types of data.

With respect to (ii) data filtering, Big Data are often affected by "noise" with respect to the analytical purpose, which must be filtered. On one side, this noise is related to the fact the data generation process is not under a direct control of the statistician that cannot apply a design to the data collection phase. On the other side the noise can be related to particular nature of some sources, like unstructured data sources (e.g. Twitter data have high percentages of "pointless babbles").

With respect to (iii) data reconciliation, even when some metadata information is present in Big Data sources, such metadata need to be reconciled with the metadata driving the statistical production, hence a reconciliation step is needed. As a further observation, due to the great variety of scheme information that can derive from Big Data sources (e.g. Internet data), the reconciliation step can be very hard due to the sparsity/incompleteness of Big Data sources schemes.

Timeliness

Time-related dimensions are particularly relevant for Big Data sources characterization. Focusing on that type of Big Data that comes from the Web, two main aspects appear as particularly significant:

- The first one is related to the characterization of an intrinsic data quality dimension, namely data volatility i.e. a temporal variability of the information the data are meant to represent.
- The second aspect is more generally related to the time features of the data generating mechanism. For instance, some Web data spring up and get updated in an almost unpredictable fashion (let's think for instance to social data), so that their time dimension is not available in a direct way, but does need to be re-constructed, if wishing to use those data in any meaningful analysis. In a way, a dedicated activity for extracting time-related metadata, where these are not available directly, appears as an essential step.

Time-related dimensions are also important to characterize the output of OS. In particular, the timeliness of OS products can be improved by using Big Data sources, e.g. by using "Internet as a data source" instead of questionnaire-based survey to collect some information.

Some completely new products, like near-time indicators, could also be considered as possible future outputs of OS.

Privacy and volatility in access to Big Data sources

Privacy is recognized as one of the principal issues in dealing with Big Data. From a methodological perspective, it is very much important to understand if it is possible to use “privacy-by-design” approaches that allow giving privacy guarentees on the used data while preserving acceptable levels of effectiveness of statistical methods. Privacy-by design approaches include, for instance, privacy-preserving record linkage and privacy-preserving data mining techniques.

The volatility of Big Data sources is related to the fact that there are some Big Data applications for which input data vary with high frequency (e.g. traffic loop detection data). This may be a problem for statistical analyses that have to be aware and deal with such variability.

Quality framework and certification

A quality framework to assess and certificate the quality of Big Data both as an input and as an output of Official statistics is important.

The major issues identified with respect to this point are:

- Data quality is highly dependent on the specific Big Data source. As an example, sensor network's data streams can be quality characterized by the fact that data is often missing, and when not missing it is subject to potentially significant noise and calibration effects. As another example, the quality of Internet data is very much difficult to assess and is a current research topic.

- Data quality is also highly dependent on the specific application domain. As an example, the quality of BigDdata used within social statistics may have a different characterization with respect to the quality of data used withing economic statistics.

With **medium priority**:

Integration with traditional data sources

This issue is related to the usage of Big Data sources integrated with survey-based data or administrative data sources.Several problems were identified:

- Linking Big Data is hard because of privacy issues that prevent Big Data vendors to release identifiable data;
- The integration task requires having a precise and explicit structural metadata representation (scheme information) that is often not available for Big Data;
- Even when scheme information is available, this will need to be reconciled with traditional sources’ schemes (as discussed in issue 4).

The integration with traditional sources has been identified as a particularly relevant issue.

5.2 Parallel session on learning and development: capacity building and training for ESS human resources

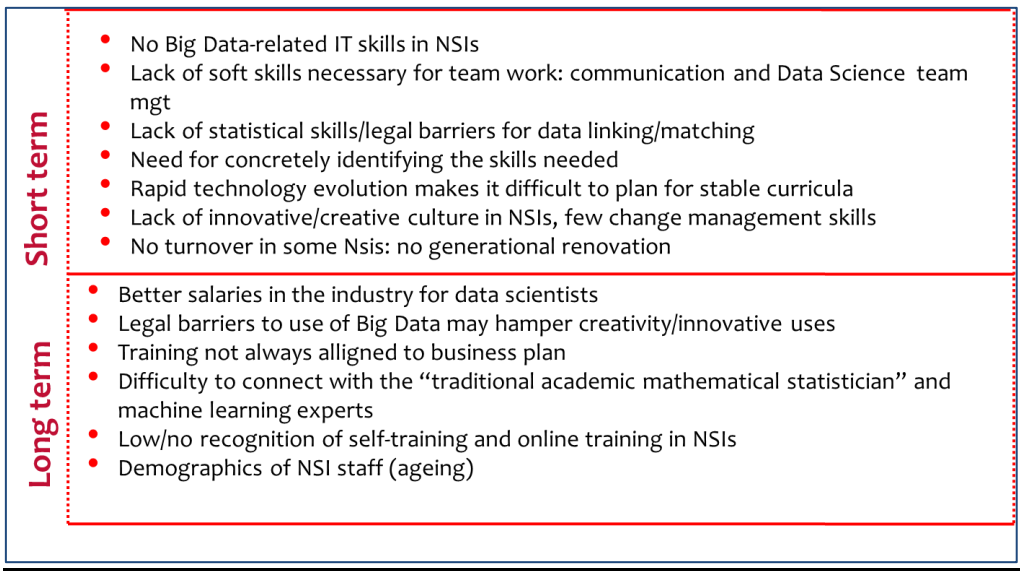
A parallel session on capacity building and training for ESS human resources (HR) to empower them in the use of Big Data was organized and facilitated by Mr José CERVERA (DevStat).

The session addressed the following Scheveningen challenges:

- [SCH5]- *Short-term Human Resources needs: recruitment, professional training, secondment/re-deployment*
- [SCH5]- *Long-term needs: academic curricula for Data Scientists*
- [SCH6]- *Collaboration with academia for training Data Scientists for official statistics*

Most participants were experts in statistical training from NSIs, but also from statistics departments of Universities, which allowed discussing [SCH6] from both viewpoints. The topics that were discussed included skills needed for using Big Data, opportunities for building those needed skills (see Figure 3). The participants aimed at making proposals related to capacity building within the roadmap to be elaborated. All issues were discussed by considering the different short- and long-term implications.

Figure 3. Issues discussed on learning and development



Skills required for Big Data

The existing skills in NSIs are indeed close to to “data science” skills required for Big Data: data cleaning, use of analytical software, data mining, etc. The staff of NSIs is well-trained in methodology and statistical domains. However, the Official Statistics community has less knowledge of Big Data than many important players like Google, and has limited – even outdated - skills and limited IT resources when it comes to the new, non-traditional, technologies used to gather process and analyse Big Data[17].

It is therefore needed to acquire new skills, either by hiring new staff, or by developing the capacities of those already working in NSIs.

As opportunities and threats to the acquisition of new skilled staff, it was mentioned that 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, but on the other side, the failrure to permit innovative methods may render NSIs less attractive workplaces for top talent. An additional constraint for NSIs is the difficulty for offering better salaries than the private sector, given the high demand for Big Data scientists (while NSIs can offer more job security as public institutions). The CBS of The Netherlands has, as an example, set up scholarships to trainees with new skills.

During the session, the skills necessary for Big Data were commented. The current “fashionable” term of Data Scientists was compared to the statistician: it requires bridging capacities between data-processing technologies and data-driven decision-making, and therefore, Data Scientists have to be strong in mathematics, statistics, IT, visualization techniques, but also subjet-matter knowledge.

In terms of software, Data Scientists need to be users of several of the following software tools: statistical packages, database management tools and ETL (Extraction, Transform and Load) techniques, as well as specific tools designed for the management of large data sets in distributed storage (such as MapReduce, see the Expert Lecture on IT tools in section 5.4.). The rapid evolution of technology may discourage developing traiing programmes in specific technoglogical solutions.

[17] “Final project proposal: The Role of Big Data in the Modernisation of Statistical Production” (“UNECE Sprint paper”), http://www1.unece.org/stat/platform/display/msis/Final+project+proposal%3A+The+Role+of+Big+Data+in+the+Modernisation+of+Statistical+Production.

In terms of statistical skills, Data Scientists should be proficient in computational statistics, analytical methods (correlation and causality, modelling, network analysis, information reduction, data matching/linking, text mining, machine learning) and in new methods for dissemination (data visualization, "data journalism").

The demanding skills for Data Scientists limit in fact the availability of such professionals: team should be built instead. Thus, soft skills are also of primary importance to grasp the benefits of Big Data: as multi-disciplinary teams have to be set up, interpersonal communication and management skills are required. It is indeed the "change management" skills which are strongly required, especially in senior NSI staff, when adopting a disruptive technology.

The case of CBS of The Netherlands was discussed with more detail. The Big Data team is not hierarchic, but has a facilitator to identify the necessary skills.

In addition to the technical skills for Data Scientists, those working in NSIs need also have knowledge of specific areas such as data protection and legislative aspects of Official Statistics, not necessarily the same as in the private sector environment. In fact, legal barriers may create frustrations to Data Scientists in NSIs. Legal advisers of NSIs should be included in the training programmes on Big Data.

Retaining the talent in NSIs is a challenge especially in a context of growing demand for data literate experts in the private industry. Thus, NSIs that have successfully achieved corporate plans that allow their staff to advance their careers while deepening their technical expertise (such as Statistics Canada and CBS Israel, with their "fast track" career programmes) may provide a good example for the ESS NSIs.

Opportunities in the ESS for developing HR for Big Data

There are currenly opportunities for the development of capacities in the ESS HR. The ESS Learning and Development Framework and its implementation though the European Statistical Training Programme (ESTP)[18] should be the frame for training in Big Data. An ESTP course on Big Data (initially planned on the management and use of large databases) has been designed by ISTAT. This course could probably be replicated at the national level and be accompanied by on-line teaching tools. However, it was recognised that complementary on-the-job training (learning by doing) should be considered in the realm of Big Data.

[18] See for instance http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/about_ess/estp.

Outside the official statistical system, there are plenty of training opportunities thanks to the diversity of academic programmes on Big Data, Business Analytics and Data Science (and other denominations of similar contents) and the offer from private companies. The preparation of the European Masters in Official Statistics (EMOS)[19] should take into consideration the certification of programmes offered by universities on this topic. The participants proposed that the issue of evaluating Big Data training programmes is brought to the next EMOS workshop.

The lack of attraction of Official Statistics for skilled students has been noted, compared to the appeal of Business Analytics and other related programmes which offer opportunities to well-paid jobs in the industry.

The Horizon 2020 programme also provides in the **medium-term** for opportunities for developing proposals fro training Data Scientists. NSIs may consider this framework for developing a tailored programme for official statisticians. In particular, the following initiatives within the Horizon 2020 were mentioned:

- Marie Sklodowska-Curie actions: support for innovative training networks, mobility of researchers, inter-sectoral cooperation;

- ICT 15 -2014: Big data and Open Data Innovation and take-up:
    - Objective: To contribute to capacity-building by designing and coordinating a network of European skills centres for big data analytics technologies and business development. The network is expected to identify knowledge/skills gaps in the *European industrial landscape* and produce effective learning curricula and documentation to train large numbers of European data analysts and business developers, capable of (co)operating across national borders on the basis of a common vision and methodology
    - Expected impact: Availability of deployable educational material for data scientists and data workers and thousands of European data professionals trained in state-of-the-art data analytics technologies and capable of (co)operating in cross-border, cross-lingual and cross-sector European data supply chains;

- The future call on “Training and educating Data Scientists”

[19] http://www.cros-portal.eu/content/emos-%E2%80%93-european-masters-official-statistics-markus-zwick-0.

Within the ESS, transfer of knowledge between NSIs on big Data experiences should be organized.

The collaboration with the Academia

The academic environment is not alien to NSIs, as many experts collaborate in methodological projects. It is also the source of training official statisticians at the origin. However, the academia offers other expertise which is not usually considered but could be relevant for Big Data, such as the existing skills in the treatment of large datasets (in astronomy, remote sensing, genetics or image processing).

The rapidly changing technological environment of Big Data is a barrier for elaborating academic programmes stable in time, adding a difficulty to the identification of sources of training. At the same time, some top universities provide, even free of charge, on-line courses in Big Data –related issues (such as those offered by *Coursera*[20] or *edX*[21].

It was agreed that **high priority** should be given to the identification of scientific areas that can contribute to methodological and quality issues on Big Data, as well as to setting up joint research and application projects.

Stable exchange programmes of academics statisticians to NSIs – and the other way back – are in place on some NSIs, and should be explored with **medium-priority** for the particular issue of Big Data. Some NSIs such as those of Finland or Romania already offer working opportunities to academic statisticians to collaborate part-time in projects.

During the discussion, the participants assigned a more important role to learning by doing, in the case of Big Data, than to academic training.

[20] https://www.coursera.org

[21] https://www.edx.org

5.3 Parallel session on programming and planning for Big Data in official statistics

Session 3 (organized in two sub-sessions of two hours each) on “Strategy, programming and planning for Big Data in official statistics” focused on some key issues related to the Scheveningen challenges:

- [SCH1]- *Acknowledge that Big Data represent new opportunities and challenges for Official Statistics, and therefore encourage the European Statistical System and its partners to effectively examine the potential of Big Data sources in that regard*;

- [SCH2] - *Recognize that Big Data is a phenomenon which is impacting on many policy areas. It is therefore essential to develop an 'Official Statistics Big Data strategy' and to examine the place and the interdependencies of this strategy within the wider context of an overall government strategy at national as well as at EU level*;

- [SCH6] - *Acknowledge that the multidisciplinary character of Big Data requires synergies and partnerships to be effectively built with experts and stakeholders from various domains including government, academics and owners of private data sources*.

The group session ‘planning for Big Data in an NSI’ aimed to discuss the possibilities to plan for an ESS “Big Data in official statistics” strategy, including actions at the national and EU-wide (and even global) levels, following the Scheveningen Memorandum. The group session provided input for this strategy ([SCH1] and [SCH2])

NSIs have to face methodological as well as technical issues when blending Big Data in the construction of official statistics, as discussed above. Presently only a limited number of NSIs are actively engaged with the diverse issues related to Big Data. It is mainly the private sector who leads the work on Big Data analytics tools and solutions. Synergies between NSIs and the private sector ([SCH6]) could provide solutions to technological as well as socially sensitive issues such as privacy, trust and corporate competitiveness in the context of the legislative framework of the NSIs.

As governments move to Big Data strategies for the delivery of services (still in an initial level), NSIs have to get involved as one of the most intensive information provider in the public administration. Collaboration with other branches of the government (health, education, transport, etc.) has to be explored ([SCH6]).

Key issues were analysed through three main streams: 1) A Big Data strategy for official statistics; 2) Strategic partnership with other branches of government and with private data providers; 3) New opportunities in official statistics for addressing unmet user needs.

The issues launched for the discussion by the parallel session facilitator, Donatella FAZIO (ISTAT) dealt with a general level and with specific levels.

A Big Data strategy for official statistics

The discussion at the general level was animated by asking some basic but fundamental questions on the use of Big Data in official statistics:

- Why combining Big data with official statistics?
- Should some official statistics be replaced with Big data?
- How can new data gaps be filled, i.e. is it possible to develop new 'Big data - based' measurements to address emerging phenomena (not known in advance or for which traditional approaches are not feasible)?
- Should the issue be the object of International collaboration or treated at the national level?
- Which could be the frame for international collaboration? By the ESS, UNECE of global working groups as suggested by the United Nations Statistical Commission?

The discussion went on considering the multidimensional aspects of a Big Data strategy for NSIs. Indeed, the usage of Big Data implies, besides the methodological and technological aspects, more crucial topics:

- the legislative aspects for the access and use of Big Data;
- the privacy-related issues in relation to getting public trust and acceptance of data re-use and its link to other sources;
- the financial costs of using Big data for statistical production vs its benefits;
- the set-up of policies and directives about management and protection of the data;
- training issues, as new skills are needed for the usage of Big Data (this dealt in detail in another parallel session).

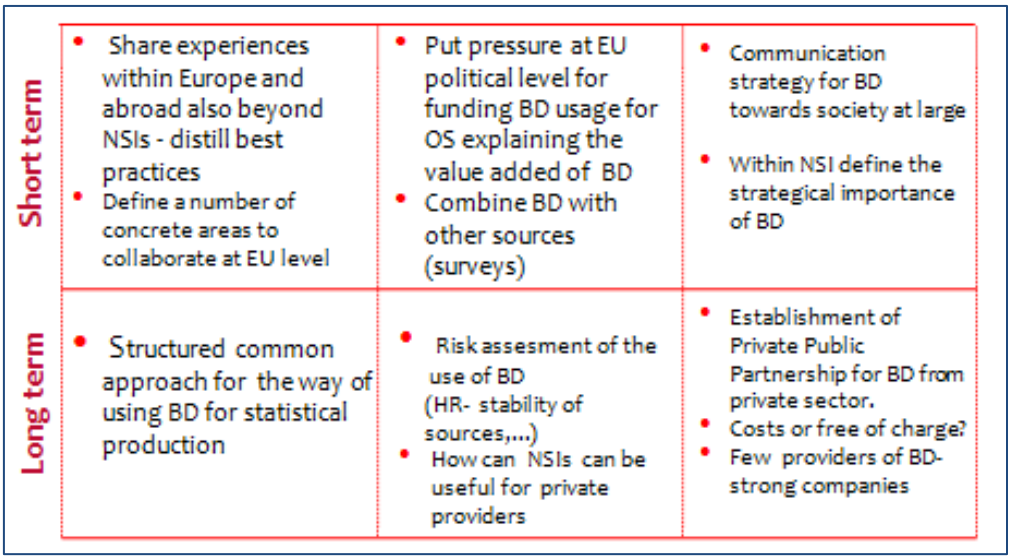
The debate on the specific levels focused on how to start to define a common ESS Big Data strategy. The issues discussed were:

- Pilot projects/isolated initiatives vs standard production: need of mapping the experiences carried out by NSIs and international organizations and distilling the best practices;
- Big Data sources with potential application to Official Statistics: need of establishing an inventory;
- Sources from private sector and from public sector as additional /supplementary sources for official statistics: for which domains can private sources complement official ones?
- Sources from private sector and from public sector as alternative / replacing sources for official statistics: for which domains can private sources replace official ones?

Finally, the discussion turned around the issue of the necessity to explore the Big Data sources separately by countries and on the role and place of Big Data initiatives in the different National and European Statistical Programmes.

Figure 4 synthetizes the participatory exercise of session 3 on “Strategy, programming and planning for Big Data in official statistics”, delineating the key issues pointed out during the discussion.

Figure 4- Key issues on programming and planning for Big Data in official statistics.



Strategic partnership with other branches of government and with private data providers

In order to identify the providers with which NSIs could establish partnerships, the second stream of the discussion started considering the classification of Big Data sources (structured and un structured) as proposed by the UNECE High Level Group[22] to have a common view of the different nature of data sets:

a. Administrative (arising from the administration of a programme, be it governmental or not), e.g. electronic medical records, hospital visits, insurance records, bank records, food banks, etc.

b. Commercial or transactional: (arising from the transaction between two entities), e.g. credit card transactions, on-line transactions (including from mobile devices), etc.

c. From sensors, e.g. satellite imaging, road sensors, climate sensors, etc.

d. From tracking devices, e.g. tracking data from mobile telephones, GPS, etc.

e. Behavioural, e.g. online searches (about a product, a service or any other type of information), online page view, etc.

f. Opinion, e.g. comments on social media, etc.

[22] http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622

Given the above classification, it was pointed out the need of the exploration by NSIs of governmental and private sources, before the identification of their potential outputs, and the necessity to carry out this analysis first within each NSI to seize the advantages and privileges of data access in each country context.

On the side of the **private and governmental providers** a multiplicity of sources were considered in their characteristics:

- Telecom companies: mobile phone location, mobile data content, text messages
- Utilities (electricity, water)
- Social media and aggregator companies (*Google, Facebook*, etc.)
- Blogs and comments
- Personal documents
- Pictures: *Instagram, Flickr, Picassa* etc.
- Videos: *Youtube* etc.
- Internet searches
- User-generated maps
- E-Mail
- Commercial transactions
- Banking/stock records
- E-commerce
- Credit cards
- Data from sensors
- Data from computer systems

Some crucial issues were debated on the usage of private Big Data sources:

There are few but large Big Data providers, creating a particular ecosystem for the collection/re-use of data. Establishing public-private partnerships between NSIs – governmental agencies – with these providers may require specific institutional treatment.

With respect to legal aspects, many questions arise on the frame to refer for privacy rules, namely, the consideration of national or international access to data. Most of these large providers collect and store (and utilize) data from many countries

On the financial side, the main issues discussed the feasibility for NSIs of accessing Big Data for free, providing fiscal incentives to the providers or buying them.

Finally the issue of the definition of ESS guidelines for standard agreements with private providers was discussed, raising the question of which modifications are needed prealably in statistical laws to allow for it.

New official statistics opportunities for addressing unmet user needs

Based on the awareness that Big Data can really represent a great opportunity for addressing unmet users' needs, the participants in the session stressed that it is essential to distinguish the context of the different needs for different users.

NSIs have to investigate which users' needs can be solved with Big Data. As examples, the following needs where mentioned:

- Covering information gaps, by producing data that do not exist yet;
- Reducing the response burden;
- Producing more detailed data;
- Processing real-time data;
- Increasing the spatial coverage and comparability;
- Disseminating more timely data.

5.4 Parallel session on IT and security issues

The discussions on IT issues were split into two sessions addressing on one side data protectin and security issues, and on the other side, IT infrastructure issues for Big Data in official statistics. They were facilitated by Mr Reg BRENNENRAEDTS (Dialogic) with the support of Mr Tommy VAN DER WORST (Dialogic). The discussions addressed the following Scheveningen challenges:

- [SCH3] - *The use of Big Data requires an adaptation of/to **legislation** regarding use aspects (i.e. with respect to the access and use of data), privacy aspects (i.e. managing public trust and acceptance of data re-use and its link to other sources, security of private data).*

- [SCH7] - *New **methodological developments** are necessary for the use of Big Data in official statistics, including for **quality assessment** and **IT-related issues**. Aspects of Big Data that oblige to rethink the statistical methods to deal with it include the heterogeneity, lack of structure (requiring important work to prepare the data for statistical production), volume (which hampers the use of standard statistical tools), representativeness and coverage over the populations of interest for official statistics.*

Data protection and security

In the parallel session on data protection and security, issues and solutions regarding data protection (privacy) as well as security were discussed.

Data protection and security mainly concerns personal rights and privacy, and confidentiality of data held by NSIs (e.g. confidential data gathered by an NSI should not be accessible to competitors, or even outside internal 'enclaves'). Although confidentiality is the major component, other aspects of security are relevant as well: integrity of data (auditing checks and balances) and availability of information comes to mind. Note that in the parallel session, the participants strictly ignored any aspect that is not specific to Big Data (for example, physical access control to an NSIs building is already in place, and does not differ for Big Data compared to small data, so it was deemed irrelevant for the discussion).

All aspects of data protection and security determine the requirements for the underlying IT infrastructure. In the second parallel session (on IT infrastructure for Big Data) this was further investigated.

During the parallel session on data protection and security, participants were asked to list issues that they perceived as most pressing in the context of Big Data for official statistics. The issues where then categorized and linked up to solutions (section 6.4).

Most issues were assessed as of high importance.

In the **short term**, the following should be considered:

- Security and privacy features, highly relevant in Official Statistics, are often not a part of the initial design of Big Data technologies. The current security mechanisms (technical as well as procedural) in NSIs are tailored to secure small-scale, static (as opposed to streaming) data analysis, and are inadequate for Big Data: NSIs will need custom components to ensure Big Data security;
- Big Data storage and processing in the context of official statistics requires mechanisms for granular access controls as well as functionality that supports auditing ("audit trail");
- When accessing data from private partners, special measures need to be taken in order to prevent that those private partners that deliver data to an NSI can cross-reference output data produced by that NSI to augment the data set they delivered to the NSI, and gain further insights about (individual) customers;
- While a common practice for current, small-scale data sets, encryption may not be feasible for Big Data, dynamically updated data sets.

IT infrastructure for Big Data

In the parallel session on IT infrastructure for Big Data in the context of official statistics, participants discussed various issues regarding underlying tools and architecture that enables Big Data processing and storage for an NSI. Aspects of data protection and security have a defining influence on this discussion.

During the parallel session on IT for Big Data, participants of the session were asked to list issues that they perceived as most pressing in the context of Big Data for official statistics. The issues where then categorized and linked up to solutions (see Section 6.4). They are also summarized in the Figures below.

Figure 5. Issues on security and data protection

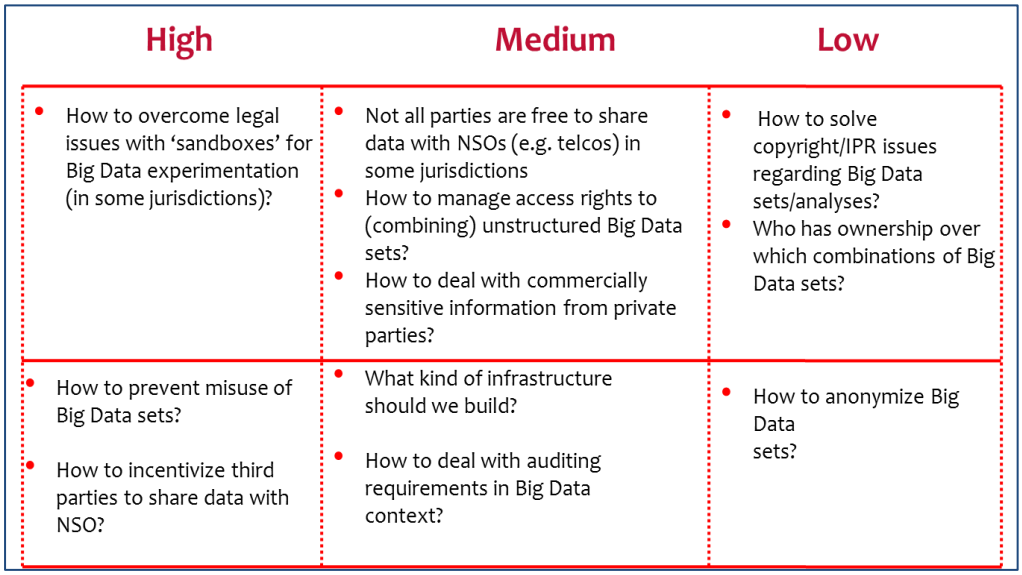
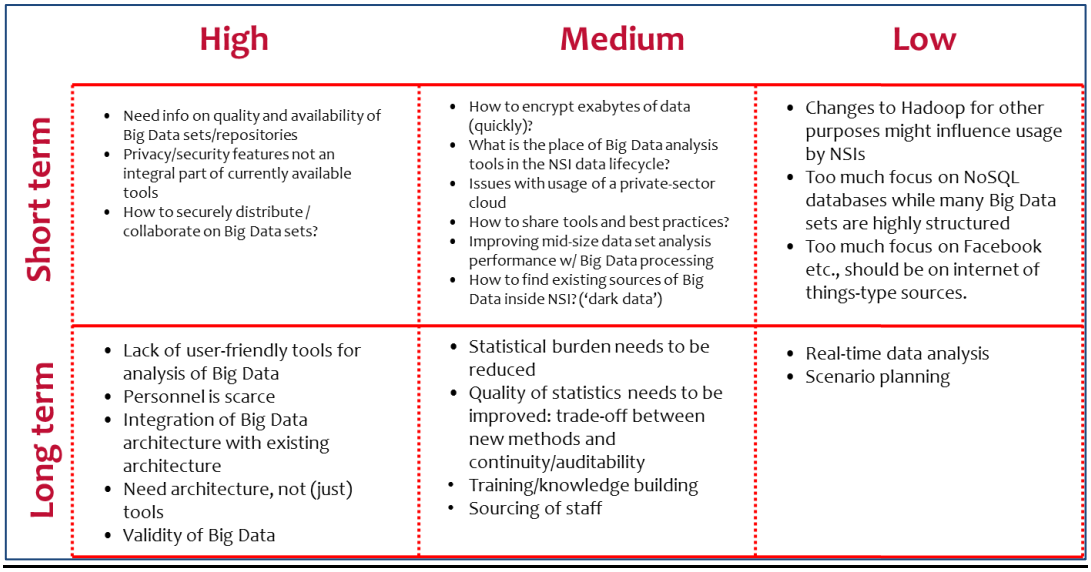


Figure 6. Issues in IT infrastructure



In the **short term**, the following IT infrastructure issues should be considered:

- NSIs have difficulties finding Big Data sets and keeping metadata for these sets. A repository for Big Data sets, which includes basic metadata about the contents of these sets, is necessary, but does not currently exist. Such a repository would foster discovery and documentation, as well as sharing of Big Data sets between NSIs;

- Big Data should reduce, not increase statistical burden. NSIs are continuously looking for ways to reduce statistical burden without reducing statistical quality. When working with Big Data, more data is used, but it is often of a lower (raw) quality. Care should be taken not to increase statistical burden when employing Big Data methods;

- Big Data should increase, not reduce statistical quality. When working with Big Data, more data is used, but it is often of a lower (raw) quality. Care should be taken not to reduce statistical quality when employing Big Data methods;

- NSIs have no infrastructure to share Big Data tools and best practices. NSIs should be able to re-use algorithms, *bots*[23], etc. developed by other NSIs, and there should be a central place for finding tem.

[23] See http://en.wikipedia.org/wiki/Internet_bot.

In the **medium term**,

- NSIs are lacking a clear vision on architecture vs. tool development. In the phase of experimentation, Big Data tools are developed without an underlying vision for a future Big Data architecture;

- Mid-size data processing should be improved with Big Data techniques. NSIs are looking for ways to improve current mid-size data processing jobs using Big Data techniques.

In the **long term**, the following should be taken into account:

- There is a lack of user-friendly tools for Big Data. In order for Big Data analysis to gain traction and adption in an NSI, it needs to be accessible to more than just the highly technical users. Tools that provide a user-friendly interface for performing Big Data analysis are however scarce and/or expensive;

- NSIs have difficulties finding and documenting Big Data analysis workflows. A place that collects and displays Big Data analysis tools and workflows in a standardized, process-oriented NSI data lifecycle structure would be of high value to NSIs, but currently does not exist;

- Find ways to employ commercial infrastructure without the downsides. NSIs (and governments in general) are hesitant to place sensitive data on private-sector cloud infrastructure. Yet commercial cloud structure may be more cost-effective than private (NSI-maintained) infrastructure. NSIs therefore need to find a way to make use of private-sector cloud (e.g. through certain 'government-certified clouds');

- NSIs rely on Hadoop, but have different goals than other users. NSIs may require that changes be made to the Hadoop Big Data processing framework (e.g. related to security or privacy). However, the other users of Hadoop may not have interest in making these changes. NSIs need to have the required development capacity to make these changes nonetheless, and should actively attempt to align their goals with those of other Hadoop users, or otherwise its usage of Hadoop will be inefficient;

- Big Data infrastructure needs to be integrated with existing infrastructure. There are many data processing infrastructures in place today, and Big Data infrastructures need to 'fit in' the current environment once it leaves the realm of experimentation.

6. Recommendations for an ESS roadmap on Big Data in Official Statistics

This final section recalls the recommendations issued of the "crowd-sourcing" exercise carried out during the Big Data even. The participation of a large number of specialists from NSIs, international organizations and universities provided a wide spectrum of opinions and experiences.

The ESS, in the preparation of a roadmap for the integration of Big Data in the production of official European statistics, may consider the recommendations of this event, which addressed the Scheveningen Memorandum challenges as stated by the DGINS.

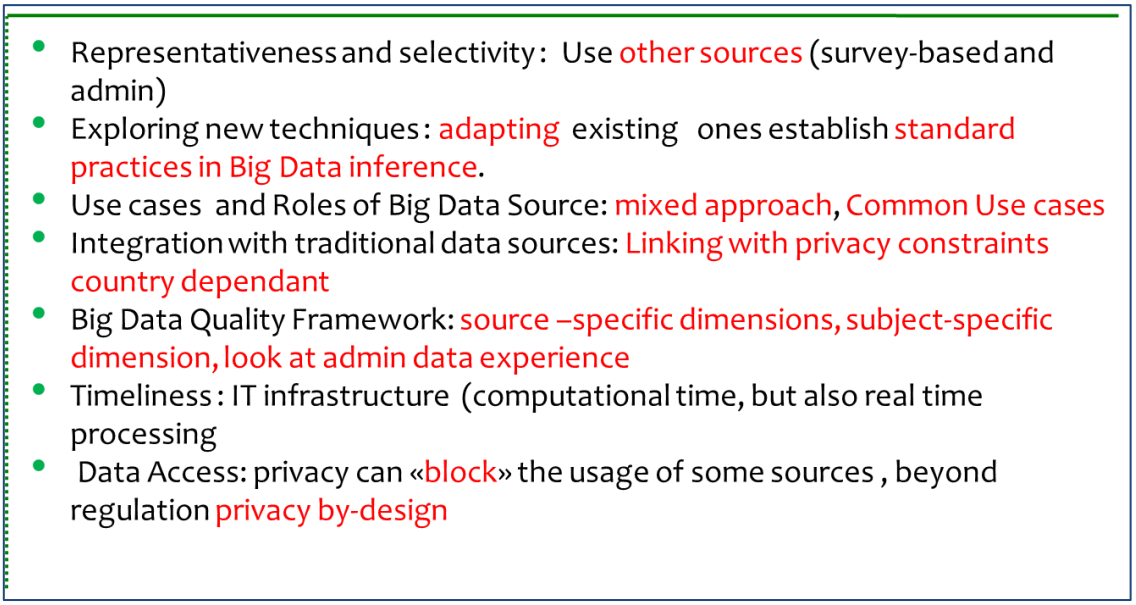
The intervention of the decision-making structures of the ESS, in particular, the ESS Committee (ESSC), but also the consultative bodies on governance (ESGAB) and on users' perspective (ESAC), will provide guidance to such roadmap in the context of the current legal and institutional framework of European statistics.

A final summary links the Scheveningen challenges to the recommendations.

6.1 Recommendations on methodological issues

Figure 7 illustrates the key actions identified by the participants of session 1, proposing initiatives to develop the methodology for using Big Data in official statistics.

Figure 7- Recommendations on methodological development for the use of Big Data in official statistics.



- ✓ *Recommendation M1:* A possible action to address the representativeness of Big Data sets is to integrate them with more traditional sources used within Official Statistics, namely survey-based sources and administrative data sources, through data matching techniques.

- ✓ *Recommendation M2:* In developing Big Data analysis, it is recommended to start from standard practices and understand how they can be adapted or complemented by exploratory techniques such as data mining and machine learning.

- ✓ *Recommendation M3:* Concrete actions that should be engaged with respect to the development of use cases include:
  - A mixed approach that combines use cases' identification in a top-down way with bottom-up feedbacks deriving from a practical implementation of them.
  - Common use cases should be identified that indeed address issues common to different NSIs.

- ✓ *Recommendation M4:* A major action to address data processing issues is to start putting in place IT and statistical methods and tools to deal with this "completely new" step of the statistical production process.

- ✓ *Recommendation M5:* A suggested action to address the timeliness/velocity of Big Data is putting in place IT infrastructure(s) able to deal on the input side with the extraction of time-related metadata, and on the output side with the production of quasi real-time information.

- ✓ *Recommendation M6:* With respect to privacy, an important action is to investigate the possibility of using privacy-by design techniques for Official Statistics.

- ✓ *Recommendation M7:* With respect to volatility, it is recommended to investigate statistical methods that are aware of the volatility of measured data.

- ✓ *Recommendation M8:* It is recommended to investigate quality frameworks for Big Data that are source-specific and subject-specific. A suggestion is to start from the work already done for administrative data.

- ✓ *Recommendation M9:* It is recommended to consider the integration of Big Data sources with survey-based and administrative data, together

with regulatory issues (probably at the national level), as well as with privacy issues that involve necessarily dedicated partnerships with Big Data providers.

6.2 Recommendations on training

Based on the discussion at the parallel session, the participants elaborated a list of recommendations for the ESS to be explored, both in the short and long term. The actions recommended (listed as *Txx,* below, and summarized in Figure 7) were further classified as difficult or of medium difficulty, based on the existing resources and constraints.

In the **short term:**

- ✓ *Recommendation T1:* Adapt the ESTP course provided by ISTAT to allow blended (face-to-face plus on-line) learning to replicate it. Consider the possibility of organizing a summer school based on it.
- ✓ *Recommendation T2:* Study the possibility of including the topic of Big Data in the "Statistical week" of Eurostat[24], scheduled for 13-17 October 2014, to increase the awareness on the issue among young staff of NSIs.
- ✓ *Recommendation T3:* Identify staff in NSIs that can start with selected on-line, free courses on Big Data (such as those provided by Coursera or edX).
- ✓ *Recommendation T4:* Develop ideas for a proposal by a consortium of NSIs, private companies and universities to be submitted to Horizon 2020 calls.

In the **long term,**

- ✓ *Recommendation T5:* Design instruments for the exchange of Big Data experts within the ESS, to facilitate on-the-job training of colleagues.
- ✓ *Recommendation T6:* Identify the current training opportunities in the ESS and European universities that can be certified by the EMOS.

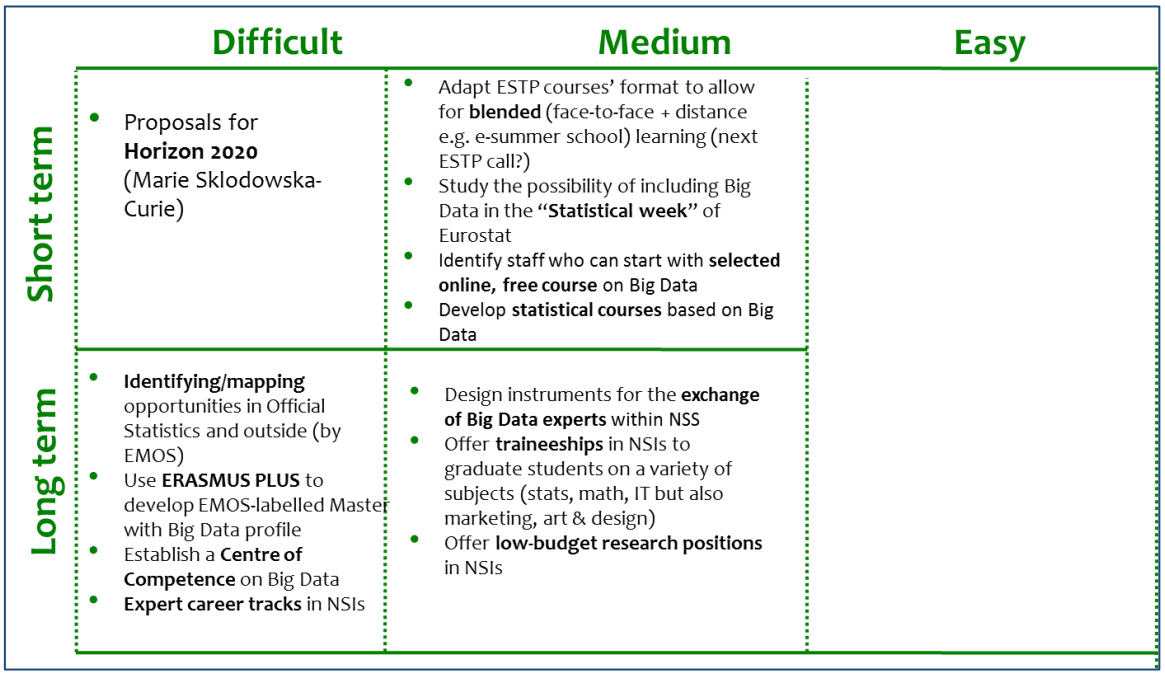
[24] http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/news/ess_news_detail?id=168475090&pg_id=2737&cc=ESTAT_EUROSTAT

Explore the use of ERASMUS PLUS[25] for the exchange of young professionals.

- ✓ *Recommendation T7:* Establish, once the best practices have been identified, a Competence Centre in the ESS.

- ✓ *Recommendation T8:* Consider, within the career plans in NSIs, the possibility of implementing a "fast track" for data scientists which allows advancing their careers while deepening their technical skills (in a complementary way of career plans which require to progressively drop technical roles to undertake managerial ones).

Figure 8: Recommendations on training and development of capacities for Big Data in the ESS.

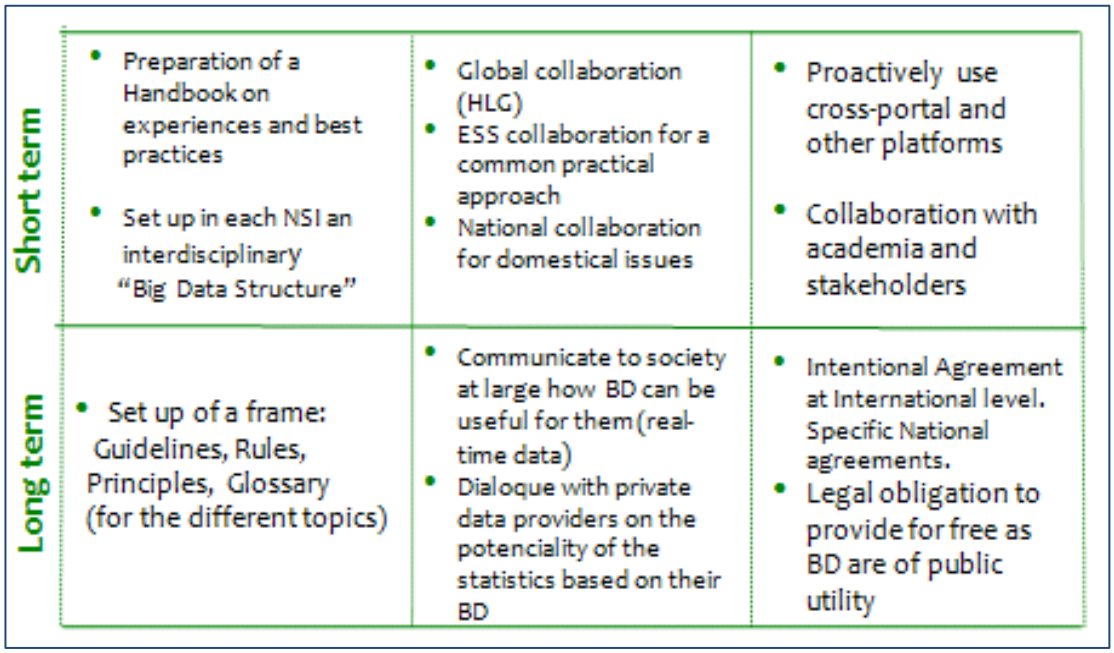


6.3 Recommendations on planning

Figure 9 illustrates the key actions identified by the participants of session 3, proposing concrete initiatives to plan and programme the NSIs activities in the field of Big Data.

[25] http://ec.europa.eu/programmes/erasmus-plus/discover/guide/index_en.htm

Figure 9- Recommendations on programming and planning for Big Data in official statistics.



Combining the issues discussed and the ideas proposed, it is possible to give some reccomendations (identified as *Px* for "planning and programming"), in the short and long term, for the definition of an ESS action plan and roadmap for the usage of Big Data in official statistics:

In the **short term:**

- ✓ *Recommendation P1:* Eurostat should prepare an ESS handbook on experiences and best practices in the use of Big Data in EU NSIs to share information on the relevance of gernerated data (assessment of the value for users), the methodological and technical issues, the organisational aspects within NSIs and in relation to the partnership with data providers, and the lessons learned.

- ✓ *Recommendation P2:* Set up in each NSI an interdisciplinary "Big Data structure" and coordinate the work of them across the ESS, to create synergies.

- ✓ *Recommendation P3:* NSIs, Eurostat and other international statistical producers should collaborate at different institutional levels of the international and national statistical systems to empower the debate and the research on usage of Big Data for official statistics. In particular, maintain the collaboration at the global level under the HLG initiative, set up a ESS-wide collaboration for a common practical approach, and

develop the national collaboration for “domestic” issues (such as access to sources, privacy laws, etc.).

- ✓ *Recommendation P4:* NSIs should set up collaborative projects with the academic institutions using the funding opportunities given by the European Commission and Eurostat (such as Horizon 2020).

- ✓ *Recommendation P5:* NSIs should open the dialoque with private data providers on the potenciality of the statistics based on their Big Data, in order to build public-private partnerships.

- ✓ *Recommendation P6:* NSIs should evaluate the usage of Big Data in terms of costs and benefits, and assess the related risks (such as the instability of sources).

In the **long term**:

- ✓ *Recommendation P7:* At the ESS level, a common approach should be adopted for the way of using Big Data for official statistical production, by developing principles, guidelines (including a glossary) and rules, as well as a “starter’skit” to facilitate the launching of Big Data experiences in all EU NSIs.

- ✓ *Recommendation P8:* NSIs should develop a communication strategy, communicating to citizens how the re-use of Big Data can be useful for society at large (e.g. real-time data).

- ✓ *Recommendation P9:* Lobbying for Big Data in official statistics: the ESS should put pressure at EU political level for funding initiatives on Big Data usage in official statistics, explaining the value added.

- ✓ *Recommentation P10:* Define the legal aspects for the use of Big Data from private providers, (1) setting up an international agreement as an umbrella for national specific agreements, and (2) foreseeing legal obligations to provide access by NSIs to Big Data sources which can be considered of public utility.

6.4 Recommendations on IT aspects

With respect to IT infrastructure, data protection and security, the parallel sessions as well as general discussions during the event have lead to the following (high-level) recommendations (listed as *Tx*):

- ✓ *Recommendation IT1:* Pay attention to the specific needs of NSIs regarding privacy and security when it comes to using third-party (commercial or open source) tools. Prefer tools that were designed with security in mind from the beginning. Align the goals of the NSI with the goals of other end-users through collaboration with vendors (or open source communities) of Big Data tools.

- ✓ *Recommendation IT2:* Create a repository for Big Data knowledge and tool sharing. In particular, a repository for Big Data algorithms (e.g. crawlers, bots, Hadoop snippets, et cetera) can be of great use and should not be too difficult to realize. A repository should specifically be tailored to the processes and lifecycles of NSIs.

- ✓ *Recommendation IT3:* Create an ESS body for standardization of various aspects of Big Data for official statistics. Such a standards body should decide on:

    - technical aspects (e.g. preferred file formats, storage methods)
    - procedural aspects (best practices, security levels, access control methods, et cetera).
    - methodological aspects (statistical methods)
    - legal framework (for using private data, using private cloud infrastructure, et cetera)

    Standardization opens the door for (future) sharing of data, methods and infrastructure. Through standardization, the efforts of the NSIs with respect to Big Data are more focused, and NSIs can be of greater influence in the development of (open source) tools and infrastructure.

- ✓ *Recommendation IT4:* The NSIs together need to invest in research in several aspects of data protection and security that are of high relevance to the NSI context. Specific examples are:

    - (Real-time) encryption of large Big Data sets
    - Guaranteeing secrecy (preventing cross-referencing of Big Data) with espect to data provided by private parties ('business secrecy').
    - Methodology for assessing degree of sensitivity of Big Data.

- ✓ *Recommendation I51:* NSIs need to partner with each other and with national government bodies when it comes to (cloud) infrastructure. Partnerships with private parties ('government certified clouds') as well as research institutes are possible as well.

- ✓ *Recommendation IT6:* NSIs need to develop a vision that addresses how Big Data analysis will be integrated in existing practice. In addition, Big Data opportunities for improving current mid-size data analysis should be planned for.

At a lower level, participants of the parallel session drafted the following solutions (or 'best practices') regarding concrete, operational issues:

- ✓ *Recommendation IT6:* When experimenting with Big Data (tool or process testing), start using small data sets (while paying attention to future scaling)

- ✓ *Recommendation IT7:* Use 'improper' tools for Big Data analysis; there is highly specialized software out there that can perform tricks that conventional Big Data tools cannot. These tools may be highly valuable for particular needs of NSIs.

- ✓ *Recommendation IT7:* Organize international training events, shared between NSOs.

- ✓ *Recommendation IT8:* NSIs should create small (3-4 person), multidisciplinary research teams focusing on Big Data. NSIs should give the teams room to 'discover and experiment', preferably using a 'safe proof of concept'.

- ✓ *Recommendation IT9:* Don't reinvent security methods and practices, but first attempt to apply the existing ones to the Big Data context

- ✓ *Recommendation IT10:* In addition to Big Data analysis, have some attention for Big Data visualization.

- ✓ *Recommendation IT11:* For each analysis, evaluate whether the NSI should perform all of the Big Data analysis, or that a data provider can perform basic aggregation (or other processing) steps before the data is transferred to the NSI. This way, costs and hassle for Big Data analysis are greatly reduced. Do pay attention to the quality of preprocessing steps, however.

Annex 1. The Scheveningen Memorandum

Big Data and Official Statistics

The DGINS

CONSIDERING

1. Recent innovations in the information and communication technologies have been leading to an increasing degree of digitization of economies and societies at all levels that offer new opportunities for the compilation of statistics.

2. The use of Big Data for statistical purposes challenges the European Statistical System to effectively address a variety of issues.

3. The demand for timely and cost efficient production of high quality statistical data increases, as well the need for new solutions to declining response levels.

4. Official statistics should incorporate as much as possible all potential data sources, including Big Data, into their conceptual design.

5. The distinguishing aspect of many Big Data sources is that they are not confined to national borders and, as such, represent unique opportunities for collaboration at European level as well as on global level.

6. Many European initiatives have a link to Big Data, including the European Commission's ambition for developing a strategy for the European data value chain, the on-going EU Data Protection reform and the Horizon2020 program.

7. The implementation of new methods of production of European statistics represents an objective of the European Statistical Programme 2013-2017[26] and aims at efficiency gains and quality improvements, including increased timeliness.

The DGINS

1. Acknowledge that Big Data represent new opportunities and challenges for Official Statistics, and therefore encourage the European Statistical System and its partners to effectively examine the potential of Big Data sources in that regard.

2. Recognise that Big Data is a phenomenon which is impacting on many policy areas. It is therefore essential to develop an 'Official Statistics Big Data strategy' and to examine the place and the interdependencies of this strategy within the

[26] Regulation (EU) No 99/2013 of the European Parliament and of the Council of 15 January 2013 on the European statistical programme 2013-17, OJ L 39, 9.2.2013, p. 12–29

wider context of an overall government strategy at national as well as at EU level.

3. Recognise that the implications of Big Data for legislation especially with regard to data protection and personal rights (e.g. access to Big Data sources held by third parties) should be properly addressed as a matter of priority in a coordinated manner.

4. Note that several NSIs are currently initiating or considering different uses of Big Data in a national context. There is a momentum to share experiences obtained from concrete Big Data projects and to collaborate within the ESS and beyond, on a global level.

5. Recognise that developing the necessary capabilities and skills to effectively explore Big Data is essential for their integration into the European Statistical System. This requires systematic efforts like appropriate training courses and establishing dedicated communities including academics for sharing experiences and best practice.

6. Acknowledge that the multidisciplinary character of Big Data requires synergies and partnerships to be effectively built with experts and stakeholders from various domains including government, academics and owners of private data sources.

7. Acknowledge that the use of Big Data in the context of official statistics requires new developments in methodology, quality assessment and IT related issues. The European Statistical System should make a special effort to supports these developments.

8. Agree on the importance of following up the implementation of this memorandum by adopting an ESS action plan and roadmap by mid-2014 that should be further integrated into the Statistical Annual Work Programmes of Eurostat.

Annex 2. Agenda of the 2014 ESS Big Data event

<table>
<tr><th colspan="5">Big Data in Official Statistics</th></tr>
<tr><td colspan="5">Programme</td></tr>
<tr><th colspan="5">Monday 31 March 2014</th></tr>
<tr><td>11:30-13:00</td><td colspan="4">Registration, Networking</td></tr>
<tr><td>13:00-14:00</td><td colspan="4">Buffet Lunch</td></tr>
<tr><td rowspan="3">14:00-15:00</td><td colspan="4">Opening Session
Chair: Mr Roberto BARCELLAN, Head of Unit for "Methodology and corporate architecture", Eurostat</td></tr>
<tr><td colspan="4">Welcome address
Mr Antonio GOLINI, President of ISTAT</td></tr>
<tr><td colspan="4">Opening adress: Outcomes of the Scheveningen Memorandum
Mr Emanuele BALDACCI, ISTAT
How to deliver on the Scheveningen Memorandum
Mr Michail SKALIOTIS, Eurostat
The Impact of Big Data on Official Statistics
Mr Reg BRENNENRAEDTS, Dialogic
Introduction to the parallel sessions
Mr José CERVERA, Devstat</td></tr>
<tr><td>15:00-16:00</td><td colspan="4">Expert lecture: Opportunities and methodological challenges of Big Data for official statistics
Speaker: Dr. Piet DAAS, CBS</td></tr>
<tr><td>16:00-16:30</td><td colspan="4">Coffee Break</td></tr>
<tr><td>16:30-18:30</td><td colspan="4">Panel discussion
Panelists: Mr Piet DAAS (CBS); Mr Antonino VIRGILITO (ISTAT); Mr Reg BRENNENRAEDTS (Dialogic); Mr Emmanuel LETOUZÉ (Berkeley University); Mr Michail SKALIOTIS (Eurostat)
Moderator: Mr Emanuele BALDACCI, ISTAT</td></tr>
<tr><td>20:00</td><td colspan="4">Social Dinner</td></tr>
<tr><th colspan="5">Tuesday 1st April 2014</th></tr>
<tr><td>9:00-9:30</td><td colspan="4">Registration, Networking</td></tr>
<tr><th>SESSIONS</th><th>PARALLEL SESSION 1A +1B</th><th>PARALLEL SESSION 2A+2B</th><th>PARALLEL SESSION 3A+3B</th><th>PARALLEL SESSION 4A+4B</th></tr>
<tr><td rowspan="3">9:30-11:30</td><td>Methodology, quality issues and accreditation of Big Data sets
Facilitator: Ms Monica SCANNAPIECO (ISTAT)</td><td>Learning & Development
Facilitator: Mr José CERVERA (DevStat)</td><td>Strategy, programming and planning for Big Data in Statistical Programmes
Facilitator: Ms Donatella FAZIO (ISTAT)</td><td>IT & Security
Facilitator: Mr Reg BRENNENRAEDTS (Dialogic)</td></tr>
<tr><td>Session 1A</td><td>Session 2A</td><td>Session 3A</td><td>Session 4A</td></tr>
<tr><td>New developments in methodology needed for the use of Big Data in the context of Official Statistics

Strategic partnerships with academia</td><td>Short-term HR needs: recruitment, professional training, secondment/redeployment</td><td>A Big Data strategy for Official Statistics (and beyond)

Strategic partnerships with other branches of goverment</td><td>Data protection and security issues</td></tr>
<tr><td>11:30-12:00</td><td colspan="4">Coffee Break</td></tr>
<tr><td>12:00-13:00</td><td colspan="4">Expert lecture: Big Data technologies and platforms for official statistics
Speaker: Mr. Antonino VIRGILITO, ISTAT</td></tr>
<tr><td>13:00-14:15</td><td colspan="4">Buffet Lunch</td></tr>
<tr><td rowspan="3">14:15-16:30</td><td>Session 1B</td><td>Session 2B</td><td>Session 3B</td><td>Session 4B</td></tr>
<tr><td rowspan="2">Quality assessment, accreditation</td><td>Long-term HR needs: academic curricula for Data Scientists</td><td>New official statistics opportunities addressing unmet needs</td><td rowspan="2">IT for Big Data in Official Statistics</td></tr>
<tr><td>Collaboration with the academia for training Data Scientist for official statistics</td><td>Strategic partnerships with private data providers</td></tr>
<tr><td>16:30-17:00</td><td colspan="4">Coffee Break</td></tr>
<tr><td rowspan="4">17:00-18:00</td><td colspan="4">Closing Session
Looking forward: the shifting paradigm of Official Statistics</td></tr>
<tr><td colspan="4">Conclusions from group sessions
Ms Monica SCANNAPIECO, Mr José CERVERA, Ms Donatella FAZIO, Mr Reg BRENNERAEDTS. Moderator: Mr José CERVERA</td></tr>
<tr><td colspan="4">Conclusions from the Event
Mr José CERVERA, DevStat</td></tr>
<tr><td colspan="4">Closing address
Mr Roberto BARCELLAN, Eurostat</td></tr>
</table>