ESSnet Big Data

Specific Grant Agreement No 2 (SGA-2)

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata

Framework Partnership Agreement Number **11104.2015.006-2015.720**

Specific Grant Agreement Number **11104.2016.010-2016.756**

Work Package 8

Methodology

Results of Workshop on:

Important topics in the area of Methodology, Quality and IT when using Big Data for official statistics

Version 2017-06-19

Prepared by:

Piet Daas (CBS, Netherlands), Owen Abbott (ONS, United Kingdom), Ciprian Alexandru (INS, Romania), Eleni Bisioti (EL, Greece), Valentin Chavdarov (BNSI, Bulgaria), Marc Debusschere (SB, Belgium), Vesna Horvat (SURS, Slovenia), Jean-Marc Museux & Fernando Reis (Eurostat), Maiki Ilves (EE, Estonia), Øyvind Langsrud (SSB, Norway), Jacek Maślankowski (GUS, Poland), António Portugal (INE, Portugal), Marco Puts & Martijn Tennekes (CBS, Netherlands), Luis Sanguiao (INE, Spain), Magdalena Six (STAT, Austria) and Dan Wu (SCB, Sweden)

ESSnet co-ordinator:

Peter Struijs (CBS, Netherlands)
p.struijs@cbs.nl
telephone : +31 45 570 7441
mobile phone : +31 6 5248 7775

Table of contents

1 Executive Summary

In this paper the results are described of the workpackage 8 workshop held on the 25th and 26th of April at the Statistics Netherlands location in Heerlen. During these 2 days, a group of 18 people identified the main topics in the areas of Methodology, Quality and IT when using Big Data for official statistics in the context of WP 1-7 of the ESSnet on Big Data. The results of the workshop are three related lists. Two of them contain 11 identified issues for Methodology and IT and one has 7 issues for Quality. On these issues the work of WP8 will focus in the remainder of the ESSnet Big Data. The lists and their interrelations are shown below.

Overview of the issues identified for each of the three areas and their interrelations

IT	Quality	Methodology
Big Data processing Life Cycle ↔	Comparability over time ↔	Changes in Data Sources
Data source integration ↔	Linkability ↔	Data linkage
Metadata management	Coverage ↔	Unit identification problem
Format of Big Data processing	Process chain control ↔	Secure multi-party computation
Datahub		Data process architecture
Choosing the right infrastructure	Model errors and precision ↔	Inference
List of secure and tested API's	Measurement error	Assessing accuracy
Shared libraries and documented standards	Processing errors	What should our final product look like?
Data-lakes		Deal with spatial dimension
Training/skills/knowledge		Machine learning in official statistics
Speed of algorithms		Sampling

2 Introduction

Within the ESSnet on Big Data and also in many other Big Data initiatives there is a need to get a grip on the issues[1] relevant when using Big Data for statistical purposes. Within the area of official statistics this is felt essential as the data is being used for the official statistics of a country. For this reason and because of the need indicated by the leaders of workpackages (WP's) 1 until 7 in the ESSnet on Big Data, the second SGA of this European project includes an additional WP that focusses specifically on issue in the areas of methodology, quality and IT; this is WP 8. Experience has shown that identifying such issues is challenging for a number of reasons; the most important one in this context is the need to involve a considerable number of people actively studying the topic. To assure an as complete possible overview, a workshop was held on the 25th and 26th of April 2017 at the Heerlen location of Statistics Netherlands. In this workshop the most important issues, viewed upon from the work performed in WP 1 until 7 of the ESSnet on Big Data, were identified. The results of the workshop are described in this document.

2.1 Short description of WP1 until WP7

Because they are an important starting point for the workshop and to assure all readers are sufficiently informed, each WP in which Big Data is or will be studied within the ESSnet Big Data is very briefly described below. In the workshop the work performed in WP 1 until 7 was used as the starting point for the identification of the relevant issues in the area of methodology, quality and IT. This was done to assure that any of the practical issues the participants ran into during their work in the ESSnet Big Data project would be included in the workshop.

WP1 Webscraping of job vacancies: This WP wants to demonstrate by concrete estimates which approaches (techniques, methodology etc.) are most suitable to produce statistical estimates in the domain of job vacancies and under which conditions these approaches can be used in the ESS. The intention is to explore a mix of sources including job portals, job adverts on enterprise websites, and job vacancy data from third party sources.

WP2 Webscraping enterprise characteristics: This WP investigates which webscraping, text mining and inference techniques can be used to collect process and improve general information about enterprises.

WP3 Smart meters: This WP wants to demonstrate by concrete estimates whether buildings equipped with smart meters (= electricity meters which can be read from a distance and measure electricity consumption at a high frequency) can be used to produce energy statistics but can also be relevant as a supplement for other statistics e.g. census housing statistics, household costs, impact on environment, statistics about energy production.

WP4 Automatic Identification System data: The aim of this WP is to investigate whether real-time measurement data of ship positions (measured by the so-called AIS-system) can be used 1) to improve the quality and internal comparability of existing statistics and 2) for new statistical products relevant for the ESS.

[1] The word 'issue' is used througout this document to identify any subject or problem people are thinking and talking about when using Big Data for official statistics. Within the context of the workshop the latter refers to work performed in WP's 1 through 7 of the ESSnet Big Data.

WP5 Mobile Phone data: The aim of this WP is to investigate how NSIs may obtain more or less 'stable' and continuous access to the data of mobile phone operators.

WP6 Early estimates: The aim of this WP is to investigate how a combination of (early available) multiple Big Data sources and existing official statistical data can be used in order to create existing or new early estimates for statistics.

WP7 Multiple domains: The aim of this WP is to investigate how a combination of Big Data sources and existing official statistical data can be used to improve current statistics and create new statistics in various statistical domains.

3. The workshop

Aim of the workshop was to generate as many issues relevant for methodology, quality and IT when using Big Data for official statistics. This was done by inviting a diverse group of participants from a large number of the partners involved in the ESSnet Big Data, from as many WP's as possible, and from Eurostat. Next, the people attending the workshop were creatively stimulated at the beginning of the workshop to obtain an as large as possible list of issues. These were subsequently checked, corrected, merged and prioritized. In the end this resulted in a list of essential issues for methodology, quality and IT, respectively.

3.1 Basic information of the workshop

The workshop was held in the afternoon of Tuesday April 25th and the morning of Wednesday April 26th. Three of Statistics Netherlands employees, trained to facilitate such workshops, lead the workshop to assure the final outcome would be as good as possible.

Schema of Day 1:

- Introduction & scope of the first day
- Energizer (Bingo)
- Visualize each workpackage & Present drawings
- Create groups & brainstorm on Methodology, Quality and IT issues
- For each area, add issues by the participants of the other groups
- Present, explain and improve issues identified per area
- Summarize and discuss results of the first day

Schema of Day 2:

- Scope of the second day & opportunity to add any important issues missing
- Analyse issues identified on first day per area per group
- Discuss, combine and prioritize issues per area (and optionally move them to another area)
- Check all issues with all participants
- Prioritize issues and identify for particular WP's by each participants ('stickering')
- Create final list of most important issues per area
- Discuss lists and relate issues across areas with all participants
- Wrap up of workshop.

The people attending the workshop and their affiliation are listed as co-authors of this paper (see the cover). The facilitators of the workshop were Frans Duijsings, Chantal Brakus and Daniël Herbers

(Statistics Netherlands). A PowerPoint presentation including the intermediate and final results and many other pictures is available on the Wiki-page of WP8 of the ESSnet on Big Data (https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/58/Report_ESSnet_Big_Data_workshop_2017_Heerlen.pdf).

3.2 Creating the gross lists

After an initial exercise to stimulate people to think 'outside the box', participants were asked to write down as much issues as they could identify regarding the methodology, quality and IT areas in relation to the work performed in WP1 until WP7 in the ESSnet Big Data. For each area, a group of 5-6 participants was formed to generate the starting list of issues. During this part of the meeting, it did not matter if a topic was already identified in another area. Ultimate goal was to obtain an as complete as possible overview. When in doubt, the issue was added to assure nothing was missed. People focusing on one area were subsequent asked to present their findings and look at the results for the other areas. They could add any issues they thought were missing. With this overview the first day ended. At the beginning of the next day, people had the opportunity to add any issues they felt missing to any of the areas.

The complete lists of the issues identified for the areas methodology, quality and IT are listed in table 1, 2 and 3, respectively. Compared to the original texts, the wording of some of the issues were rephrased to make more clear what was actually meant or merely to correct for spelling errors. For completion, the pictures of the boards on which all issues were listed are included in Annex A. These pictures were taken in the morning of the second day of the workshop.

Table 1. List of all issues identified for Methodology

Coverage assessment (representativity)	Locate objects
Identification of data generating process	Role of methodology (needed at all?)
Who processes? data architecture	How to infer movement from static data
Noise reduction	Why methodology, if data is perfect?
Uncertainty measures	Secure multi-party computation
Limited to what we already know/did	Sampling
What is the ground truth?	Benchmarking (gap between traditional and Big Data)
Changes in data sources	Data linkage
Unit of analysis?	Combining data (different representativeness)
How to aggregate spatial data	Advanced data analysis
Curation (cleaning)	Algorithms for different data sources
Linking units	Generalize methods
Feature recognition (attributes)	Web mining
What does the final product look like?	Bias measures
Assessment of user needs	Machine learning in official statistics
What is the reference?	Choosing the (best) level of aggregation
Reliable time series	NLP/Text analysis methods
New statistical units	

Table 2. List of all issues identified for Quality

Quantifying errors (getting something akin to a CI)	Linkability
Comparability over time	Managing expectations
Skills	Predictability (usefulness for early indicators)
Costs	Model errors & precision
Measurement error	Timeliness (availability over time)
Variable issues	Validation
Missing data	Burden
Control over data generating process	Incomplete metadata
Identifying bias	Unit issues
Processing errors	Communicating quality (people, organization etc)
Coverage	Relevance to the users
Different points of view (users vs producers)	Data coherence
Language	Quality framework
Classification errors	Reliability
Duplicates	Data noise
Lack of total population information	Data trust (data owners)
Conflicting quality dimensions	Sustainability of data source

Table 3. List of all issues identified for IT

How to deal with untrusted data (reliability updates)	Data transmission (transferring)
Shared libraries and standards for documents	Choosing the right infrastructure
Big Data archiving	Repository at GitHub
Web-scraping format/algorithms	Internet and IO (how to manage)
Metadata and data models	Datahub (access to big data)
Format of Big Data processing/Unified framework languages/libraries	Implementing methods in parallel
Data source integration	How to deal with unstructured data sources
List of secure, tested, allowed and used API´s	The need to experiment (multiple copies of data)
Training/skills/knowledge of the tools/environments in different phases	Don't move the data but move the process (data transportation)
Data management (No-SQL vs CSV-files)	Data at different locations
Money (IT and skills are expensive)	How to keep the data consistent
No legacy frameworks (python 2)	Data virtualization
Disclosure (anonymization of data)	Data vs. Goal driven way of working (&models)
Languages (R /Python/Scala)	Position of cloud (services)
Suggestions of IT-tools	Data storage
Different time slices?	Metadata management (ontology)
Data (pre)processing	How to cope with different conclusions/knowledge
Complexity of algorithms (over time & space)	Compare/benchmark the results of big data vs. traditional statistics
Experts to learn standards and tools for Big Data processing	Different IT needs for Big Data and register based statistics
Policies & data management versions to access data	Who is in charge? IT or Big Data teams
Speed of algorithms (for new and big data)	Change management
Joining/Matching data sets	Flexibility and new tools
Data-lakes (link with traditional sources)	

3.3 Combining and selecting.

After creating the gross list of issues, the focus was on identifying the most important ones in each of the three areas. The same groups that created the starting list, looked at the whole list of issues identified for that particular area. By discussing the issues, combining and/or renaming them, often

to make it more general, they prioritized issues with the aim to end up with a short list of most essential issues. These short lists were again discussed by all groups to make the outcome as generally acceptable as possible. In both the methodology and IT area a list of 11 most important issues were identified. For quality a list of 7 issues was found to suffice. These three list were all placed on a single board (see Annex B) and shown in subsections 3.3.1, 3.3.2 and 3.3.3, respectively.

After a discussion with the whole group, the participants were subsequently asked to indicate the issues they found most important. This was done by labelling the issues with stickers. The participants were give three stickers. Two green stickers to indicate the most important issues and one blue sticker, on which they could write one of the WP 1-7 numbers, to indicate the most important issue for a particular WP. This 'stickering' activity enabled three things: i) to get an indication of the overall importance of an issue, ii) to get an indication of the importance of the issue within methodology, quality or IT and iii) to get an indication of the importance of the topic for a particular WP. The number of green and blue stickers, including the WP they refer to, is shown between brackets for each issue in the three subsections below.

The reduced and prioritized complete lists of the topics identified for methodology, quality and IT are discussed below. For IT the numbering of the issues is an indication of their (relative) priority during the selection process of identifying the most important issues, for methodology and quality this is certainly NOT the case. The wordings of some of the items may be rephrased to make more clear what was actually meant or merely to correct for spelling errors. To each issue a short description is added to explain what the issue identified exactly entails. The picture of the board on which all issues were listed including all stickers is included in Annex B.

3.3.1 List of the most important methodological issues

1. Assessing accuracy
   How accurate are Big Data based the estimates. Both bias and variance need to be considered, but bias is expected to be more important.
2. What should our final product look like?
   Data driven statistics do not start with a predefined end product in mind. However, during this work it is important to start thinking about the product that can/will be delivered.
3. Deal with spatial dimension (WP4, WP5)
   Many Big data sources have a spatial component, such as a geolocation,. It is essential to make use of this kind of information. This means that attention has to be paid to the location of objects and aggregating spatial data.
4. Changes in data sources (3*)
   Many Big Data sources are by-products of new technological developments. The content of these sources may therefore change rapidly. This will also be the case for data sources produced by private companies. It is important to get a grip on these to enable the production of reliable time series.
5. Machine learning in official statistics (1*, WP7)
   Considering its rise in popularity, it is likely that in the future more and more (official) statistics will make use of machine learning based methods. It is vital to fully understand the implications of applying these kinds of methods in the production of official statistics. Estimation of variance, extracting features and knowing how to create good data sets for training purposes are examples of important considerations.

6. Data linkage (1*, WP3)
   To fully integrate Big Data in official statistics production it is essential that these data sources and/or their (intermediary) products can be combined with the data provided by other (more traditional) sources. Combining refers to the inclusion at either the unit or domain level here.
7. Secure multi-party computation (1*)
   A lot of interesting data are produced by other organisations, such as private companies. Combining data from different organizations is challenging as many of them may not want the others to have complete access to their data; i.e. access at the individual record level. Being able to combine sources with a method that keeps the individual inputs of each partner private for the other parties involved is key to unlock the full potential of Big Data. National Statistical Offices are organizations that are ideally suited to function as a trusted party for this.
8. Inference (4*)
   Methods that are able to reliably infer from Big Data and/or the combination of such data and other sources need to be developed. It is also essential to understand how this kind of inference is affected by the various sources of error that may occur.
9. Sampling (1*)
   Not all data are available or can be used. The ability to deal with subsets of Big Data and draw valid conclusions from it is important in unleashing its full potential.
10. Data process architecture (WP2, WP5)
    Big data are often produced by one of the partners in the chain and subsequently transferred and/or used by others. For statistics production it is essential to have an overview of the whole chain and the individual steps performed by each partner as each step affects the other.
11. Unit identification problem (2*, WP3)
    Big Data are produced by units of which (often) hardly any information is available in the source. This makes it challenging to identify the 'real world' units producing the data. For statistics it is essential to relate the units in Big Data with that of the (statistical) target population.

3.3.2 List of most important quality issues

1. Coverage (2*, WP1)
   Information on the population included in a big data source is vital for reliable statistics. Important for this issue are the lack of information on the units included, their duplication and their selectivity.
2. Comparability over time (1*, WP6, WP7)
   To produce comparable statistics over time, it is essential that the source remains accessible, relevant and its content remain usable.
3. Processing errors
   During the processing of Big Data, errors may be introduced that negatively affect the quality of the data. Examples of this are the way outliers and missing values are treated.
4. Process chain control
   In a Big Data process it is very likely that multiple partners are involved. To assure a stable and timely delivery of data of high quality, the entire process needs to be controlled.
5. Linkability (1*)
   It is to be expected that Big Data needs to be linked or combined with to other data sources. During this process, errors may occur which affect the quality of the output.

6. Measurement error (1*)
   The values included in Big Data may not be all correctly measured; some may contain errors. This affects the outcomes produced, certainly when a systematic bias is introduced.
7. Model errors and Precision (3*, WP6)
   Big data based estimates are likely produced by models. The specifications of these models may be incorrect which negatively affects the reliability of the estimates.

3.3.3 List of most important IT issues

1. Metadata management (ontology) (1*, WP4)
   It is important to have (high quality) metadata available for big data. This is essential for nearly all uses of Big Data. Ideally, an ontology is available in which the entities, the relations between entities and any domain rules are laid down.
2. Big Data processing life cycle (4*)
   Continuous improvement of Big Data processing requires capturing the entire process in a workflow, monitoring and improving it. This introduces the need to design and adapt the process and determine its dependence on external conditions.
3. Format of Big Data processing
   Processing large amounts of data in a reliable and efficient way introduces the need for a unified framework of languages and libraries.
4. Datahub (1*)
   Sharing of multiple data sources is greatly facilitated when a single point of access, a so-called hub, is set up via which these sources are made available to others.
5. Data source integration (2*, WP2)
   There is a need for an environment on which data sources, including Big Data, can be easily, accurately and rapidly integrated.
6. Choosing the right infrastructure (1*)
   A number of Big Data oriented infrastructures are available. Choosing the right one for the job at hand is key to assuring optimal use is made of the resources and time available.
7. List of secure and tested API's (WP2)
   An application programming interface (API) is a set of subroutine definitions, protocols, and tools for building application software. It is important to know which API´s are available for Big Data and which of them are secure, tested and allowed to be used.
8. Shared libraries and documented standards (2*, WP1)
   Sharing code, libraries and documentation stimulates the exchange of knowledge and experience between partners. Setting up a GitHub repository (or something similar) would enable this.
9. Data-lakes (WP1, WP2)
   Combining Big Data with other, more traditional, data sources is beneficial for statistics production. Making all data available at a single location, a so-called data-lake, is a way to enable this.
10. Training/skills/knowledge
    Facilitating the exchange of skills and knowledge, for instance by offering trainings or assistance, will greatly stimulate the use of Big Data by many National Statistical Offices.
11. Speed of algorithms (WP4)
    Making use of fast and stable implementations of algorithms will greatly stimulate the application of Big Data as this speeds up the processing of large amounts of data tremendously.

3.3.4 Issues spanning across the areas

The final lists for each area have been listed above. Upon closer observation, it was noticed during the workshop that some of the issues occurred in more than one area. These are listed in table 4. The fact that the issues occur in more then one area makes clear that they point to a more general cross-topic

Table 4. List of common issues identified across topics

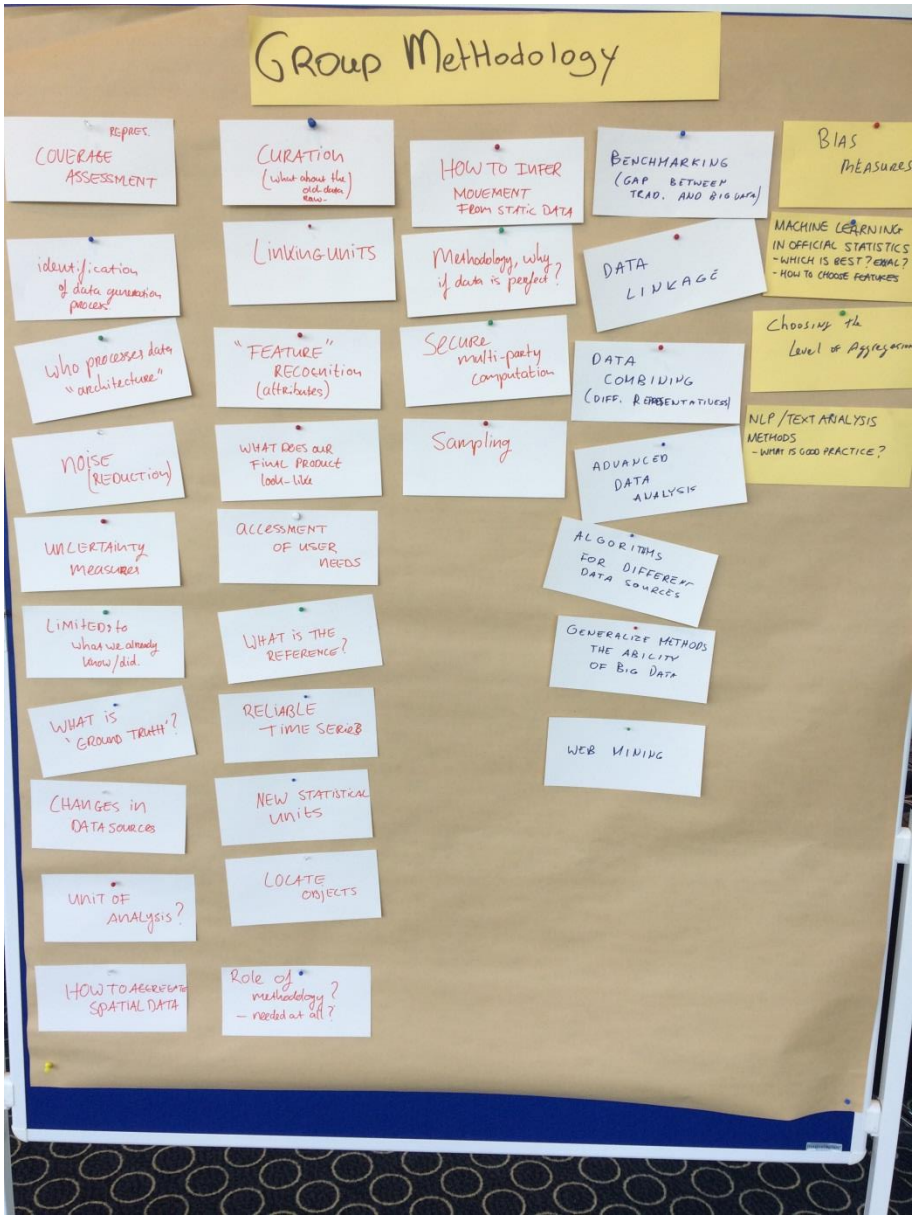
IT	Quality	Methodology
Big Data processing Life Cycle	Comparability over time	Changes in Data Sources
Data source integration	Linkability	Data linkage
	Coverage	Unit identification problem
	Process chain control	Secure multi-party computation
	Process chain control	Who processes architecture
	Model errors & precision	Inference

4. Conclusions and future directions

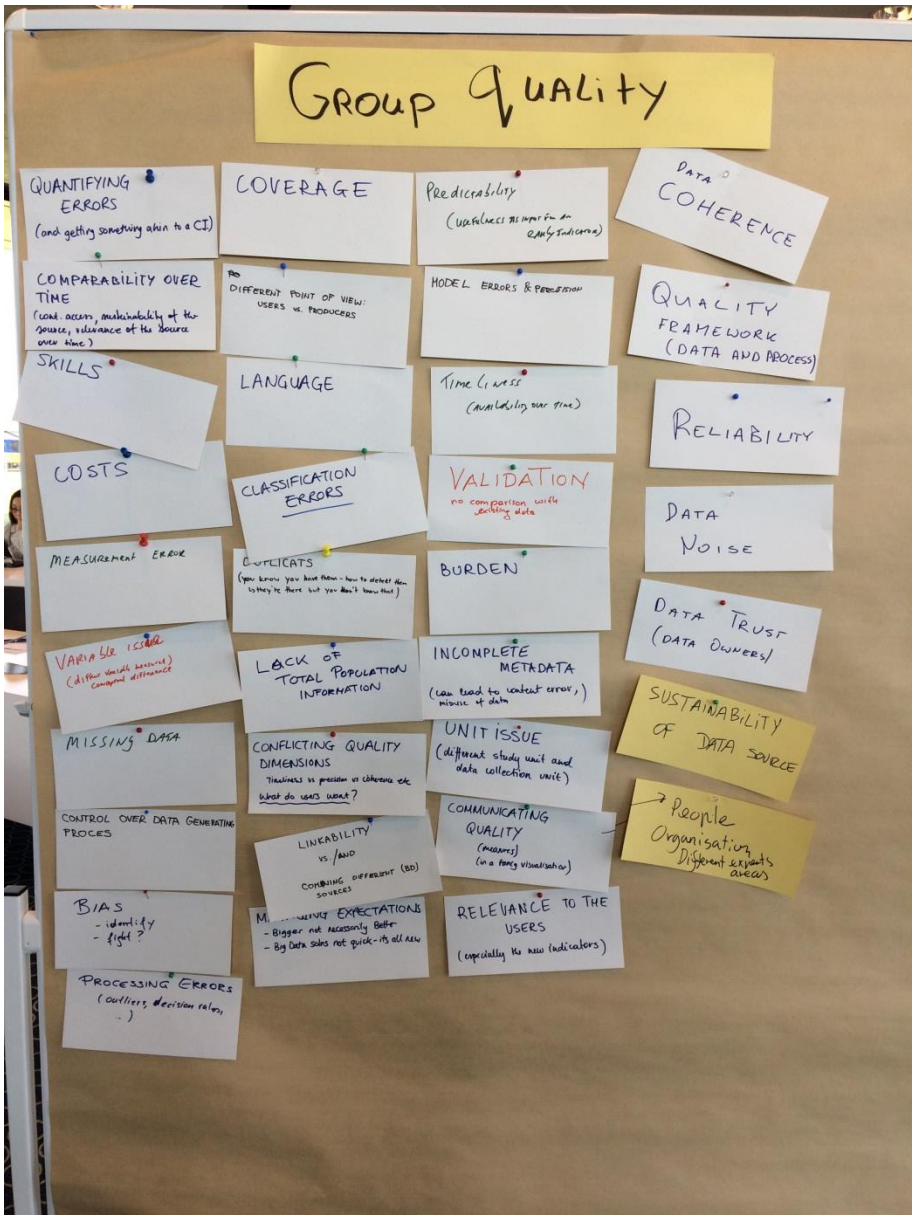
From the above it is clear that the workshop has enabled the identification of a list of 29 issues when using Big Data for official statistics. Each of the issues indicates a need for more knowledge on the topics identified. From the wordings used to describe each issue it is also clear that, certainly in a number of cases, these wordings may be improved. Some wordings even clearly indicate a solution, e.g. data-lakes, and as such do not directly refer to the underlying problem; in the case of data-lakes this is the need to have an environment in which all data can be accessed. The wordings used in this document have, intentionally, kept as close as possible to the ones generated during the workshop. The same holds for the fact that some issues could also be included in another area; e.g. secure multi-party computations identified under methodology could be included in IT as well. For the reader it is important to realize the findings of the workshop are merely a starting point for the work of WP8. From hereon the findings will be expanded upon, adjusted, renamed (if needed) and elaborated upon. This means that, for instance, the wordings used in the above mentined tables may well change. This is not a problem, it is just a natural evolution of the work in this exiting area. What is most important is the fact that during the workshop, a start has been made on the identification of the most important isseus that need to be tackled when producing Big data based official statistics. This day and age is an interesting time for statistics. Hopefully we have made an interesting addition to this during the workshop and we will certain do this during remainder of the ESSnet Big Data project.

ANNEX A

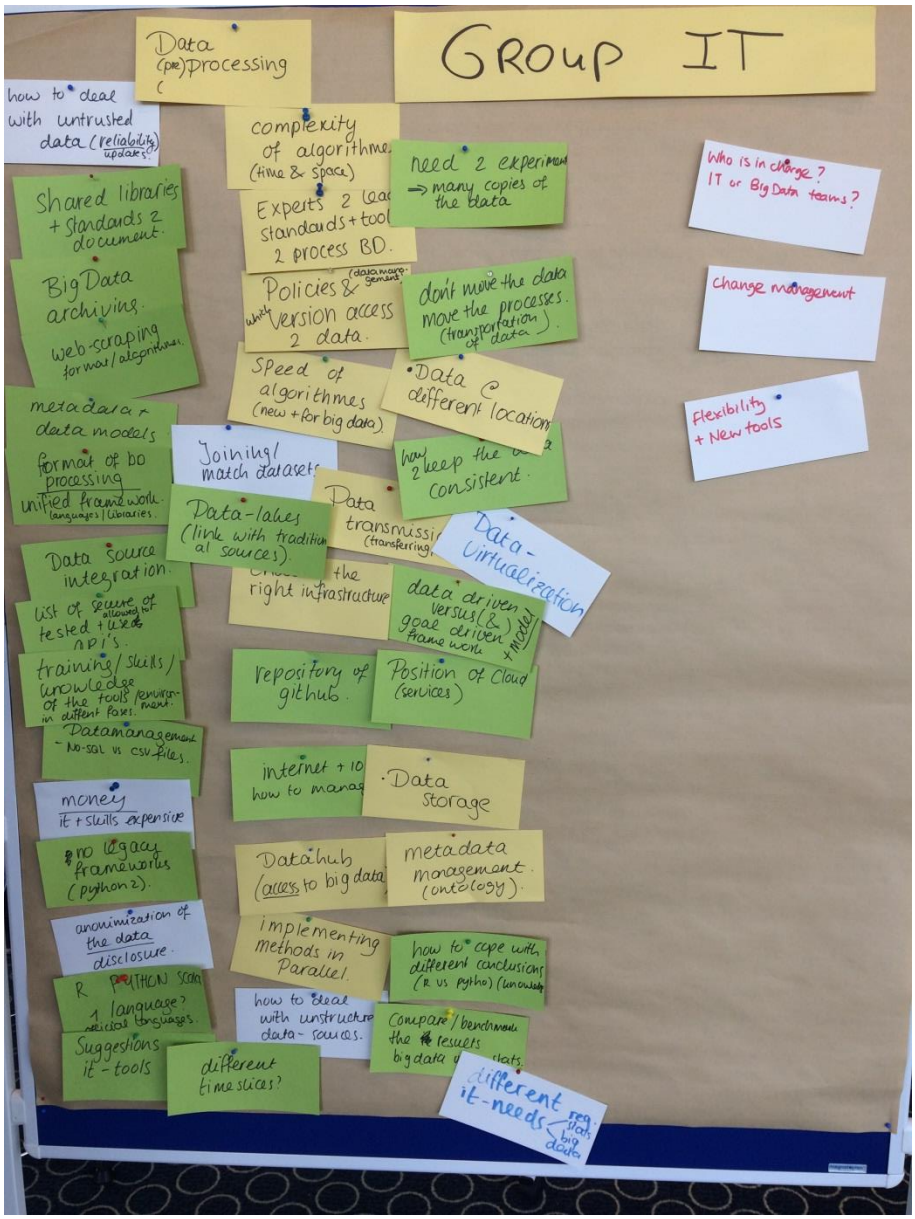
Overview of issues identified for Methodology



Overview of issues identified for Quality



Overview of issues identified for IT



ANNEX B

Overview of prioritized issues for Methodology, Quality and IT, including stickers