ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

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

https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en

Work package K

Methodology and quality

Deliverable K7: Typification matrix for big data projects

Final version, 30.4.2020

Prepared by:
Sónia Quaresma (INE, PT)
Jacek Maślankowski (GUS, PL)
David Salgado (INE, ES)
Gabriele Ascari, Giovanna Brancato, Loredana Di Consiglio, Paolo Righi and Tiziana Tuoto (ISTAT, IT)
Piet Daas (CBS, NL)
Magdalena Six, Alexander Kowarik (STAT, AT)

Work package leader:

Alexander Kowarik (STAT, AT)

alexander.kowarik@statistik.gv.at

telephone : +43 1 71128 7513

Contents

Typification Matrix for Big Data Projects

Motivation

Big data projects often use more than one data source but even in the simplest case when one is confronted with the exploration of just a major data source it is hard to assert its overall level of maturity regarding its statistical exploitation. However, this is of great importance for NSIs, who must establish their strategy for statistical production and also the investment that should be devoted to a specific source that is required to attain a mature level.

After the previous ESSnet the big data projects were grouped pragmatically along three strands. Those are:

1. Big Data Exploratory Projects – in which data sources are probed, its potential, quality and methodological problems as well as possible uses are investigated;
2. Big Data Piloting Projects – in which pilot cases are developed to explore the uses devised and try to deal with the situations anticipated but in which many other problems may still be encountered;
3. Big Data Implementation Projects - in which at least part of the statistical production can make use of the big data source, thus the process is implemented and the practical difficulties dealt with;

Although the establishment of the three strands had an empirical base and reflects the different level of maturity of big data projects as observed during the first Essnet, this distinction was not clearly expressed or studied in a systematic way. For the purpose of the present grouping the implementation is the only strand expected to produce experimental or ongoing outputs to the wider public, while exploratory and pilot projects do not aim to disseminate to the general public.

It is the goal of the Typification Matrix to identify the questions that will make possible to assess the maturity level of a big data source to assist the future evaluation of big data sources/projects.

Having a tool to collect, in an organized way, information about the Big Data sources should enable not only their appraisal but also promote the re-use of solutions, and methodologies. It is conceivable that for different big data sources, even distinct in nature, some basic building blocks may be re-usable. In this sense the tool may be used as a guide, for example presenting a quality issue and the methodology adopted to tackle it.

Furthermore the documentation of strategies, and procedures, adopted to reach the outcome may in the future be used for similar Big Data projects, to adopt similar architectures to address analogous challenges, to shorten the times to progress from less to more developed stages and to achieve more mature solutions with less costs.

The Typification Matrix is therefore intended for all stages of a Big Data project: Exploratory, Piloting and Implementation. It should also be used during the whole project, documenting it and not only in an initial moment. The subsequent updates will reflect the maturity gains along the Big Data project. The next sections describe how such a tool was devised, built and tested.

Development

The description of Big Data in terms of the 3Vs model (Laney, 2001), which refers to: Volume (size of data), Velocity (speed of data transfers), and Variety (different types of data, ranging from video to data logs for instance, and with different structures) is not enough to evaluate the usability of Big Data for producing official statistics.

To be able to generate value for an organization out of big data exploration it is necessary to identify the available data sources, what type of data they provide, and how to treat these data (Desamparados et al, 2018). These simple questions address the source, the data itself and the metadata that prescribes, at least partially, the methodologies to use.

The same elements are identified as being crucial (Wirthmann et al, 2015) for the proposed accreditation procedure of a big data source. Elements describing the source, the metadata and the data should be collected as early as possible to decide upon the acquisition and eventual usage of the big data source. However even beyond this initial stage these are the elements that will determin the potential usage of the big data source, along with its quality issues. For this reason one of the axis of the typification matrix addresses the source, metadata and data elements.

Once these 3 layers,

- "Source" – Access, ownershop, legal and ethical issues – ,
- "Metadata" – Definitions on units, populations and identifiers – and
- "Data" – Type, size, format and structure of the data –

are established, it's still necessary to obtain an instrument that is on the one hand generic enough to be used for the appraisal of any kind of data source, be it web scraped, satellite images, sensor or administrative data, in nature; and on the other hand sufficiently specific to allow the identification of the problems that may be source dependent or intrinsic to that type of data. Not only that but the tool must be fit for documenting the project different phases.

As observed during ESSnet Big Data 1 (ESSnet BD1, IT Report 2018 & ESSnet BD1, Methodology Report 2018) one may become aware of problematic situations at different stages:

- Immediately upon its description;
- Once exploration started and challenges come up;
- When trying to devise and/or apply treatments;
- Due to the investment required (technical capabilities and know how, time, storage and procedure capabilities, contracts and legal requirements, money, etc.)

These four elements follow each other naturally. At least it's necessary to be able to describe a source, data or metadata. Each Big Data source/metadata/data poses its own challenges in order to be used and become a valuable asset. After the challenges are known a treatment[1] may be prescribed and applied. When a treatment is formulated an investment can be forecasted or once the treatment is performed it can be registered.

[1] A treatment may be a methodology, a combination of methods or a simple procedure. It's not assumed to be a single operation

Having gathered all this information, it should be possible not only to achieve an assessment on the Big Data project as well as trace the roadmap for the business case of the exploration of the data source. The 3 layers described above Source, Metadata and Data than have to be considered regarding its generic description, challenges, treatments, investments and roadmap. These two "dimensions" together define the conceptual typification matrix shown in figure 1.

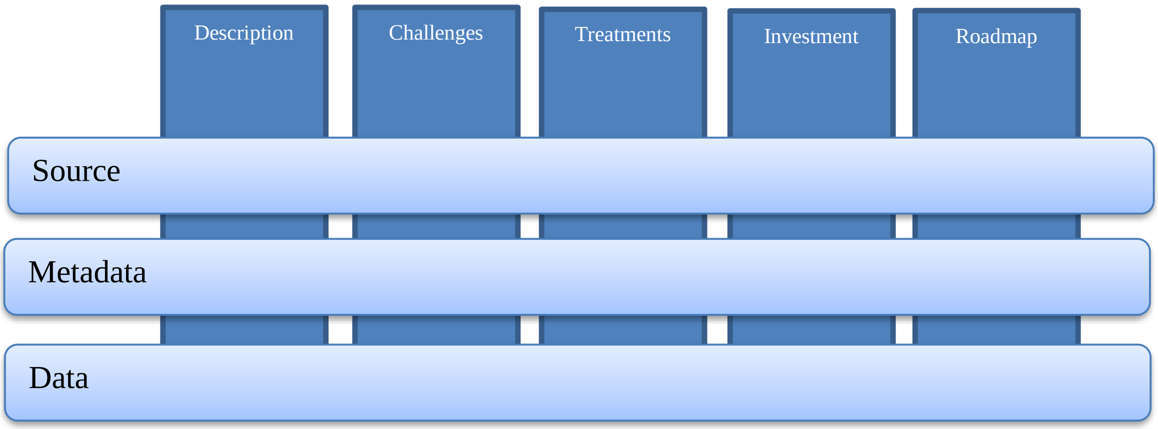


Figure 1 - Conceptual Typification Matrix

Based on the experience gathered during the first ESSnet (ESSnet BD1, IT Report 2018 & ESSnet BD1, Methodology Report 2018) the information on each of the areas Source, Metadata and Data can be collected through specific questions. While remaining open the need to address its description, challenges, treatments, investment and roadmap guides the users of the matrix and helps them being systematic and effective in their approach.

Although the areas where information is required are Source, Metadata and Data it is necessary to gather different elements for each of those. Several questions were devised to cover the most important aspect bearing in mind the business functions of the business architecture proposed by Big data REference Architecture Level (BREAL). In some cases the questions directly address a business function like "Is there a possibility to access the data to study its relevance?" which maps to the acquisition and recording business function. In other cases, because the business function groups several statistically important different activities, it provides a higher level of detail where it maps the application architecture of BREAL. For example "Are the variables that enable linking of the data already known?" maps to the data linking and enriching component of the application architecture, while "Do you have the necessary variables to reach the relevant granularity level for the statistical unit?" maps to the statistitical aggregation application component, both being part of the modelling and interpretation business function of BREAL.

Another important aspect is that the questions adopted focused not on asking details of the access, or the metadata or data structure but rather in asking if that aspect is known to the user. This prevents an extra effort on filling the matrix while still making sure that facts are known and if they present a challenge the user should then be able to identify it. Other good examples are asking if the range of the population is known and if the base unit in the data sets is known, instead of asking what it is. The users have the option of stating it for their own documentation purposes however it is not mandatory.

At the same time the matrix tries to be exhaustive, not only asking about the population range and base units present but to go beyond this and inquiring into the granularity of the data being appropriate. Answering these type of questions can in some case bind us to specific studies or purposes. This effect was considered beneficial keeping in mind that the targer is to assess a big data source regarding its statistical exploitation. The complete typification matrix is presented in the next section.

Typification Matrix

The matrix is available online at

https://webgate.ec.europa.eu/fpfis/wikis/display/EstatBigData/Big+Data+Typification+Matrix

Questions on the Big Data Source	**Description**	**Challenges**	**Treatments**	**Investment**	**Roadmap**
Source & Access					
Who owns the data? Public administration, one company, several companies?					
Is it possible to get access? Does it have to be paid?					
Are there legal or ethical problems to access the data?					
Are there limitations to the amount of data that can be accessed? a. What is the nature of this limitation? Legal, technical, financial, other? b. Which costs are involved? c. How can the costs be covered?					
Is there a possibility to access the data to study its relevance?					
Metadata	***Description***	***Challenges***	***Treatments***	***Investment***	***Roadmap***
Is the definition of the population known? If not do you already have a method to address this issue?					

Is the base unit of the data set known?					
Do the units have an identifier?					
Do you have the necessary variables to reach the relevant granularity level for the statistical unit?					
Is there background information that you need to link the base units of the data set to the statistical unit, but that doesn’t have the base units of the data set?					
Is there other information to make the data set useful with auxiliary data (NSI or other sources)?					
Are there known quality issues with any of the variables?					
Does the data contain sensitive variable? (Meaning legal or ethical issues related to its use)					
Data Type	***Description***	***Challenges***	***Treatments***	***Investment***	***Roadmap***

Choose below the type/format of the data? a. Human-produced records (e.g. Administrative database) b. Machine-produced records (e.g. Sensor produced records in a database) c. Satellite Images (e.g. Sentinel 1 imagery) d. Web scraped text (e.g. Enterprises scraped data) e. Video files (e.g. Video surveillance cameras) f. Audio files (e.g. Sound detectors) g. Other. Which?					
Do you know the size of the dataset? Will it be a problem to treat it at once? Will you split it for processing?					
Do you know the structure of the dataset? Are many different files considered a collection? (e.g. Consumption file for a smart meter plus a file for the location of the device are part of the same collection) a. Do you have to relate several files to have the entire dataset? b. Are the variables that enable linking of the data already known? If not do you have already a proposal to test the linkability?					

The next section briefly describes the adjusments and tests made to the typification matrix and its usage.

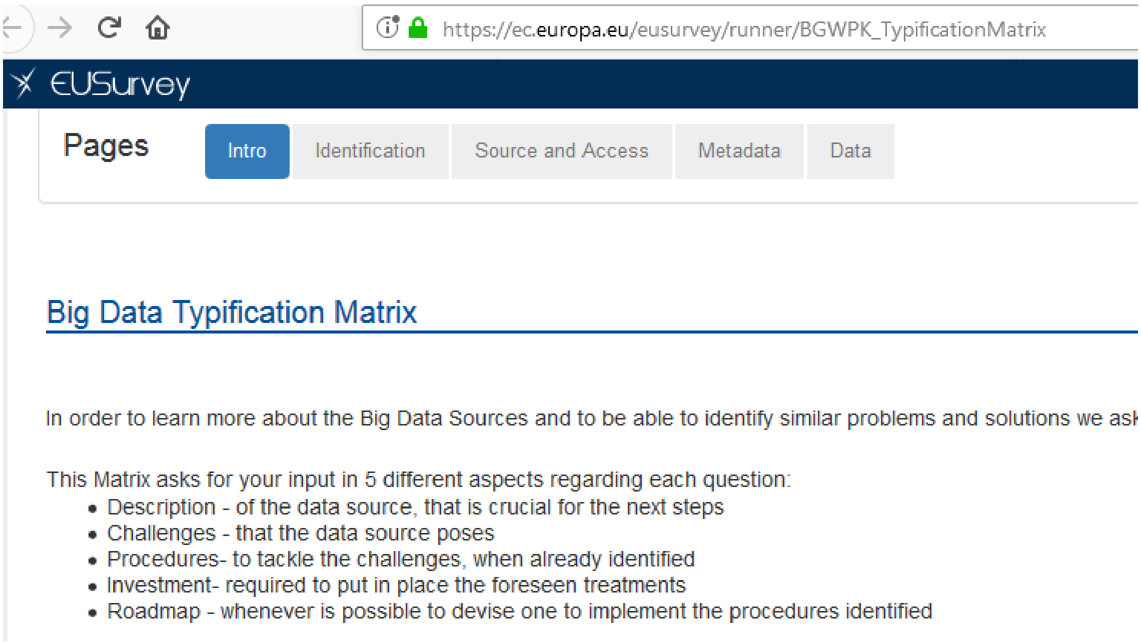
Usage and Testing

After design of the typification matrix it was pre-tested for easyness to follow, clarity and completion. It was filled with information from WP I on Mobile Network Operators and it was found that it was comprehensive enough to collect all the information needed to address the goal that was to assess a big data source/project level of maturity.

At this point it was intended that all work packages in the ESSnet could use the typification matrix not only to assess their own projects but also to provide WP K with information on quality issues detected and methodologies used.

For this usage it was determined that it could be confusing not asking the colleagues to describe the variables or structures, which is what they are used to do and expect. To overcome this difficulty a short (5 minutes) video was prepared and the filled example with WP I data (available in Annex A) was provided along with the Matrix.

The site on confluence was also not friendly enough when users wanted to provide longer answers or when the document was downloaded. To increase the legibility of the typification matrix it was made available in the EUSurvey platform[2].


https://ec.europa.eu/eusurvey/runner/BGWPK_TypificationMatrix
EUSurvey
Pages
Intro
Identification
Source and Access
Metadata
Data

Big Data Typification Matrix

In order to learn more about the Big Data Sources and to be able to identify similar problems and solutions we ask

This Matrix asks for your input in 5 different aspects regarding each question:

- Description - of the data source, that is crucial for the next steps
- Challenges - that the data source poses
- Procedures- to tackle the challenges, when already identified
- Investment- required to put in place the foreseen treatments
- Roadmap - whenever is possible to devise one to implement the procedures identified

Figure 2 - The Layers Source, Metadata and Data organized in different panels

The 3 layers were organized in different pages and some questions were taken out of the table format for easyness of use, allowing to pick who owns the data, public administration, one company or several, for example.

All work packages, from implementation and pilot tracks, were able to fill in the typification matrix for describing their sources/projects.

Later WP G to J were asked to present their work package in the intermediate pilot meeting of the ESSnet Big Data Pilots II, regarding their findings when filling the typification matrix. It was deemed as useful and the comments are integrated in the conclusions presented in the next and final section.

[2] https://ec.europa.eu/eusurvey/runner/BGWPK_TypificationMatrix

Conclusions

The Typification Matrix can be used in multiple ways to support the development of Big Data projects. It was tested by all work packages of the second ESSnet on Big Data[3] and used to present the different projects in the intermediary meeting. We present these conclusions before our final remarks on the current typification matrix and on future work.

The typification matrix:

- Is not easy to fill at an early stage on the project, which can be interpreted as not having maturity enough to evolve to implementation. Thus it should be completed also during the development of the project.
- Highlights the challenges faced in smaller project areas, encouraging modularity that will in turn promote re-use.
- Calls the attention to the methodologies being used and therefore tested by several countries promoting robustness and reliablity
- Fosters some generalization from the specific use cases of the big data source to the more comprehensive big data classes, as depicted in the quality guidelines.
- Helps the users build a more holistic assessment of their source/project.
- Can be used as documentation to research which methodologies were adopted by others to address the same issues.

The multiplicity of usages reported by the users surpass the original goals of the typification matrix. Regarding those aims we find that it raises awareness to issues that should be appraised in the accreditation procedure of a big data source. It does so with questions that map the business functions of an architecture for big data in a transparent way, that doesn't require previous knowledge of BREAL. In the same way it connects with the Quality Guidelines[4] and the Methodology Report[5] through the columns Challenges and Treatments. While not all challenges relate to quality issues, as well as many treatments will not encompass statistical methodologies, many will.

The next developments of the typification matrix will consist in trying to relate areas/questions of the matrix to the Quality Guidelines and Methodology Report deliverables. Naturally there will be quality issues and methodologies for which no guidelines can be defined as recommended best-practice, as there is not enough information and the deliverables of WPK are being prepared simultaneously with the development of the other WPs in the ESSnet. Similarly the connections of some matrix areas with the BREAL business functions and application components will be made explicit, to guide the matrix users and foster the adoption of the BREAL architecture. In this way the typification matrix may be used as a guide and roadmap for big data projects.

The other expected development of the typification matrix will be to try to ascribe maturity levels to regions of the matrix, indicating if the available information about the data source is enough to include it in a Big Data implementation project, or instead if an exploratory or pilot. This characteristic will make the typification matrix a maturity assessment tool, as intended.

[3] The current document is a deliverable from work package K in ESSnet 2 on Big Data https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page

[4] Quality Guidelines for the Acquisition and Usage of Big Data – ESSnet 2 on Big Data WP K Deliverable https://webgate.ec.europa.eu/fpfis/wikis/pages/viewpage.action?pageId=324045012

[5] Methodology Report – ESSnet 2 on Big Data WP K Deliverable

References

Desamparados et al., 2018 - Desamparados B., Domenech J. Big Data sources and methods for social and economic analyses. Technological Forecasting and Social Change, Volume 130, May 2018, Pages 99-113 https://www.sciencedirect.com/science/article/pii/S0040162517310946 (accessed 11st February, 2020)

ESSnet BD1, IT Report 2018 - Report describing the IT-infrastructure used and the accompanying processes developed and skills needed to study or produce Big Data based official statistics. Work Package 8, Methodology Deliverable 8.3 (2018) https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/1/10/WP8_Deliverable_8.3_IT_Report_2018_03_05_final.pdf (accessed 11st February, 2020)

ESSnet BD1, Methodology Report 2018 - Report describing the methodology of using Big Data for official statistics and the most important questions for future studies - Work Package 8, Methodology Deliverable 8.4 (2018) https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/0/0d/WP8_Deliverable_8.4_Methodology_31_05_2018_final.pdf (accessed 11st February, 2020)

Laney D., 2001 - Laney D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Application Delivery Strategies, (2001) p. 949 http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed 11st February, 2020)

Wirthmann et al., 2015 - Wirthmann, A., Stavropoulos, P., Petrakos, M. & Petrakos, G. Proposal for an accredication procedure for big data source (2015) https://www.researchgate.net/publication/289344406_Proposal_for_an_accredication_procedure_for_big_data_source (accessed 11st February, 2020)

Annex A – Typification Matrix Test with WP I

The filled example of the typification matrix is available online at

https://webgate.ec.europa.eu/fpfis/wikis/display/EstatBigData/Big+Data+Typification+Example+-+MNOs

Questions on the Big Data Source	**Description**	**Challenges**	**Treatments**	**Investment**	**Roadmap**
Source & Access					
Who owns the data? Public administration, one company, several companies?	Mobile phone data are generated in the telecommunication networks of MNOs. In principal, they own the data. Lately, according to their interpretation of the GDPR, in some cases they claim that the legal ownership of the data is that of the clients themselves, so that they are just some kind of depositary or trustee of their clients' data. For the statistical process it will be necessary to use some auxiliary information such as official data from NSIs themselves or from the National Telecommunication Regulator (e.g. market shares, penetration rates…).	The main challenge is to reach an agreement with MNOs to set up a partnership in the long range for a sustainable statistical production.	Two extreme options are in principle considered with very different scopes. On the one hand, we must keep on discussing and debating with MNOs to find the final working scenario. On the other hand, changes in the European and national legal regulations (Statistical Acts, …) can also be possibly considered, but this is a very long road to follow.	Financial investment will need to be considered as part of the working scenario, since marginal costs for data retrieval and preprocessing needs to be taken into account.	According to the treatments, we need either to provide a sustainable contact with MNOs working in the direction to reach agreements. Alternatively, tasks to prepare a legal change could be considered but this is not considered as a working option right now. The partnership choice is the preferred option

Is it possible to get access? Does it have to be paid?	Access to mobile phone data is only possible through an agreement with MNOs. The factors to be met to reach this agreement are currently unknown to us. For production conditions, there is no single agreement between MNOs and NSIs, only for research. One of the key factors is the compensation for the extraction and preprocessing costs. Currently, for research, in some cases no compensation is requested and in other cases a marginal quantity has been paid. In some other cases, the decision was not to pay and thus no access was granted. In the production scenario, this is an unsolved issue, but probably Official Statistics will need to consider using part of the data collection budget for this. Details need further work at different levels (business, technical…).	The challenge is to make a fair and realistic estimation of these marginal costs so that an agreement can be reached. In a broader sense, we are facing the challenge to incorporate other actors into the official statistical production process (in contraposition to traditional production).	So far, the only treatment considered is to work back-to-back with MNOs seeking for the optimal agreement	No concrete investment so far has been identified	
Are there legal or ethical problems to access the data?	Again, access to mobile phone data is only possible through an agreement with MNOs. Legal and ethical issues are also two factors blocking the way to the data. Although an internal study in the ESS by the ESS TF on Big Data concludes that National Statistical Acts seem to legally support the access to mobile phone data by NSIs, in practice this is not the case and a serious risk of a litigation process in Courts would be the way to the data if an agreement is not reached (or else no data at all).	The challenge is to find successful partnership models between NSIs and MNOs in the ESS that can be applied in standard production conditions to regularly produce official statistics.	The main course of action is to maintain the ongoing contacts with MNOs to find these partnership models.	It is not clear if an investment or what type of investment could be potentially needed for such partnership models.	Some of the possible concrete actions go beyond the scope of the ESSnet and will be coordinated with other European initiatives (communication strategy, …).

Are there limitations to the amount of data that can be accessed? a. What is the nature of this limitation? Legal, technical, financial, other? b. Which costs are involved? c. How can the costs be covered?	Limitations do possibly exist from different perspectives. Technical limitations arise mainly from the issue of confidentiality and privacy (on-premises access), but also from some partial collisions with Telecommunication Regulations. Moreover, this data is critical in MNOs’ production processes and thus there also exist business limitations. Costs indeed mean money, which we try to avoid as part of the partnership models. We are trying to find scenarios of win-win collaboration so that e.g. their data can be enriched with official data (in an anonymized way) in order to compensate them for their costs. In this way the value of their statistical products may increase. Regarding the volume of data, this strongly depends on the type of data to be used (CDRs or signalling data).	Again, the challenge regarding the cost is to reach a successful agreement with MNOs avoiding a direct payment in terms of money for the data. Marginal costs for extraction and preprocessing tasks need to be considered, but they are not clear at this point. Technological solutions to access data with full guarantee of privacy and confidentiality, especially regarding data integration, need to be found, tested, and deployed under a win-win agreement.	No specific treatment is under consideration, only to maintain the ongoing contacts with the MNOs to reach the needed agreements	At this moment, no concrete investment has been identified since this depends on the agreements.	Continue negotiating
Is there a possibility to access the data to study its relevance?	Only through an agreement with the MNOs. In some countries there is no contact with MNOs, hence no possibility to access the data whatsoever. Some other partners do have access currently to some kind of data, which will be used for the ESSnet.	See above.	No specific treatment is under consideration, only to maintain the ongoing contacts with the MNOs to reach the needed agreements.	At this moment, no concrete investment has been identified since this depends on the agreements.	Continue negotiating.

Metadata	***Description***	***Challenges***	***Treatments***	***Investment***	***Roadmap***
Is the range of the population known? If not do you already have a method to address this issue?	In the case of mobile network data, the goal is to reach all mobile subscribers in the national territory. Currently, the actual access to data depends on the ongoing agreements with MNOs. So far, these agreements grant access only partially and in restrictive and limited conditions. The range can be understood as that part of the population using a mobile device.	As a main challenge we face the goal to have access to mobile network data of all MNOs in each country in the ESS. This is overoptimistic and probably impossible, but it is the challenge regarding the range of the population. The goal to have access to all mobile network data must not mask the need for a rigorous statistical treatment to connect the datasets with the target populations under analysis.	As part of the ESS Reference Methodological Framework we are proposing to use Bayesian hierarchical models to deal with the representativity issue (dataset-target population connection), but also with the combination of existing official figures (through the choice of priors and hyperpriors) and with the production of quality indicators (such as credibility intervals and posterior variance and posterior coefficient of variations).	We need the tools to build the data simulator. We will use the BDTI to carry out this goal. Once in production, since Bayesian modelling is computationally demanding, we will assess these computational needs (still unknown). The BDTI or a similar structure within the ESS may play an important role in this regard.	We will construct a generator of semisimulated mobile network data and complete population data. Using these, we will use a collection of model variants to assess the performance of the models.

Is the base unit of the data set known?	In the case of mobile network data this depends on the type of data. If they're CDRs, the base unit is basically the network event producing a cost charged to the subscriber. If they're signaling data, the base unit is again a network event but now of a more complex nature (change of LA, mobile entrance in the network when it is switched on…). All in all, we can say they are network events of different nature.	The main challenge is to fully understand and assimilate the origin of each piece of network data so that we can generate the convergence data in the most meaningful and effective way for the upper-level statistical processing in the different statistical domains. The ESS Reference Methodological Framework aims at simplifying and standardizing this production stage. Everything points at a necessary partnership and understanding with telco engineers of the MNOs.	The main statistical process regards the transformation from event-based datasets to mobile device-based datasets. A priori, this should pose no real challenge (regarding the pseudonymised ID; other variables are different).	No investment detected	See above.

Do the units have an identifier?	The idea of unit here is crucial. If we are referring to base units (network events), in general they are identified by the mobile device and the time and spatial references (when and where the event took place). It is possible that technically, a given physical event (e.g. a long phone call during a journey) may be registered in the network as different events. Currently, NSIs do not have access to this kind of information and we assume that each network event corresponds to a physical event. If we are referring to transformed units (mobile devices – the transformation is a priori very simple), mobile devices have a persistent pseudoanonymised ID in the network which is used for the statistical analysis. However, in some countries like in Germany this ID must change every 24h (due to telco regulations), so that the longitudinal value of the data is reduced at the microlevel. In general, with mobile phone data technically this unique identification is always possible (otherwise, billing would be impossible), but from a legal perspective or of other nature (confidentiality and privacy, etc.) we may find barriers in univocally	To overcome legal restrictions regarding the ID variable of each mobile device. Statistical, data protection and telco legislations are not fully aligned in this regard (data protection are more restrictive). It is not clear that the challenge is feasible.	No methodological proposal has been constructed so far to overcome the loss of identifiers in different time periods.	No investment foreseen, only methodological analysis.	This will be part of the different elements of the ESS Reference Methodological Framework.

	identifying each device in a persistent way.				
Do you have the necessary variables to reach the relevant granularity level for the statistical unit?	In principal yes. For mobile phone data this question does not have currently a closed answer. There are evident aggregates as population counts, tourist trips, same-day visits, commuting flows, etc. But the list needs further exploration because intuitively more aggregates can be potentially incorporated.	The challenge is again methodological. We need to grow a list of use cases identifying statistical aggregates of interest in different domains	We are currently focused on present population and tourism aggregates as the two main use cases.	No investment foreseen, only methodological analysis.	This will be part of the different elements of the ESS Reference Methodological Framework.
Is there background information that you need to link the base units of the data set to the statistical unit, but that doesn't have the base units of the data set?	To univocally link base units (network events) with statistical units (individuals), we'd ultimately need an ID variable for both data sets (e.g. Passport No. or National Identity No. or similar). This is out of the question right now. Linkage must be done at some aggregate level (city district, grid cell, etc.).	The main challenge is to find auxiliary information with the same level of granularity as the one in the mobile phone datasets.	This will be part of the different elements of the ESS Reference Methodological Framework.	No investment foreseen, only methodological analysis.	This will be part of the different elements of the ESS Reference Methodological Framework.

Is there other information to make the data set useful with auxiliary data (NSI or other sources)?	Regarding complementary MNO data, there are other categorical variables that could be used for statistical purposes, such as the type of event (call, SMS, data connection…), roamer/non-roamer, device OS (iOS/Android/…). Apart, there could be auxiliary official data enriching mobile phone data potentially facilitating win-win collaborations with MNOs (e.g. sociodemographic variables, living conditions per cell, etc.).	The main challenge is to find auxiliary information with the same level of granularity as the one in the mobile phone datasets and possibly linkable to MNO data. Accessing auxiliary MNO data can also be challenging when reaching an agreement	This will be part of the different elements of the ESS Reference Methodological Framework.	No investment foreseen, only methodological analysis.	This will be part of the different elements of the ESS Reference Methodological Framework.

Are there known quality issues with any of the variables?	There exist limitations in the use of mobile phone variables for statistical purposes because they were not designed and created for this purpose. But this is part of the challenge in using these data sources for the production of official statistics. As quality issues let us concentrate upon the following data quality dimensions and on original variables only: Consistency: consistency depends again on the underlying technology. Under the same technology, data can be considered consistent. They are presented in the same format and are compatible with past data using this same technology. For example, CDRs have in general a widely accepted format (although variations exist) making them compatible among different MNOs. Accessibility: mobile phone data are not easily available nor quickly retrievable. This depends on the agreement between NSIs and MNOs.	In the current state of affairs, we aim at providing an end to end statistical process following the ESS Reference Methodological Framework illustrated by some concrete examples. All the mentioned quality issues will be dealt with progressively depending on the agreements with the MNOs.	Accuracy will be dealt with using methodological analysis. The rest of quality issues depends strongly on the agreement and collaboration with MNOs.	No investment foreseen, only methodological analysis and ongoing contacts with MNOs.	This will be part of the different elements of the ESS Reference Methodological Framework.

	Accuracy: all variables are expected to be accurate regarding the network aspect they are related to from a statistical point of view. For example, the timestamp of a given network event is accurate enough for all statistical purposes. The spatial attributes (antenna cell, location area…) are also accurate enough for statistical purposes. This is not to be interpreted as accuracy in the final statistical analysis (e.g. the number of mobile devices in a given territorial cell may possibly not be given with enough accuracy but this is a consequence of the meaning itself of the original variables; this number of devices is not an original variable in the statistical process). Timeliness: technically the values of all variables could be possibly accessed on real time, but this is highly disruptive and costly. Thus, data sets have some time delay in their information. Indeed, for the research stage currently ongoing, data sets are in some cases rather old. In a future production environment, we hope that agreements will clearly improve the timeliness of official statistics products.				

	Reliability: technically the values of all variables can be considered reliable from a telecommunication point of view, since they enable MNOs to operate the network for its original purpose (telecommunications). From a statistical point of view, original telco data are highly technology-dependent and this may influence the upper-level statistical analysis (e.g. because spatial attributes are collected in a different way). In this sense, reliability should be monitored for the statistical analysis to prevent unforeseen changes. When considering not raw telco data but more elaborated variables as inputs for the statistical process, reliability will strongly depend on the preprocessing stage conducted by MNOs. Thus monitoring becomes more crucial. Completeness: the completeness of the data set strongly depends on the agreements between the NSI and the MNO. For example, some MNOs are not willing to share the location of their antennas. This data set can be considered incomplete since this has a strong influence on the subsequent statistical analysis.				

Are the variables sensitive?/ Does a data set contain any sensitive variable? Meaning legal or ethical issues related to its use	Mobile phone data are extremely sensitive both to legal and ethical reasons. Indeed these are two of the main issues regarding access to these data.	In the current state of affairs there exist two challenges regarding the sensitivity of the data. On the one hand, a communication strategy directed to MNOs, NSIs, and the general public opinion. This should communicate the privacy and confidentiality preserving policies and measures in using these data for Official Statistics. On the other hand, the measures assuring legal restrictions for privacy and confidentiality of citizens posed by DPAs must be adopted.	The communication strategy goes beyond the scope of the ESSnet and is under consideration in the ESS Task Force on Big Data. The legal measures are considered in a national basis depending on the concrete agreements with the MNOs.	No investment is identified	The communication strategy will be dealt with outside this ESSnet. Legal issues regarding DPAs are dealt with in a national basis between the corresponding NSI and DPA, thus no overall roadmap has been proposed.
Data Type	***Description***	***Challenges***	***Treatments***	***Investment***	***Roadmap***

Choose below the type/format of the data? a. Human-produced records (e.g. Administrative database) b. Machine-produced records (e.g. Sensor produced records in a database) c. Satellite Images (e.g. Sentinel 1 imagery) d. Web scraped text (e.g. Enterprises scraped data) e. Video files (e.g. Video surveillance cameras) f. Audio files (e.g. Sound detectors) g. Other. Which?	Mobile network data in any of its form (CDR, radio network signaling data, core network signaling data, …) are always generated by the network. There possibly exists some complementary information for each subscriber coming from the subscription contract which is more admin-like. However, currently MNOs are not willing to share these data. Data are machine-generated. As auxiliary data sources, official figures coming from either admin or survey data can (should) be used. These can also be potentially complemented with data from the National Telco Regulator.	The main challenge regards the access to mobile network data, which impinges on many different aspects (not only technical). An agreement for routine production has not been achieved in any European country. The first internal issue regarding this challenge is to determine exactly what data to be used for statistical production (signaling data seem to offer wider possibilities for further statistical analyses, but CDRs are common use since they require less investment).	We are constructing the ESS Reference Methodological Framework in which we are identifying a technology-independent form of data to be used as the basis for any upper-level statistical analyses in different statistical domains (tourism, population, etc.). As part of this framework, we need to make a data processing design together with the MNOs to produce these convergence data and probably the first steps (pre-aggregation) for the different statistical analyses.	Reusing mobile network data for statistical analysis requires a minimum amount of investment by the MNOs. The deeper the data inside the network (i.e. closer to the antennas), the higher the investment. Since MNOs seek profitability of their investments and operations (they are a private business), the investment return is key in this decision and an alignment between private and public goals is needed but no solution has been found so far.	In the past, we considered to request access and then to develop the methodology, the IT aspects, and the rest of elements of the statistical process. Experience has shown that we should build the ESS Reference Methodological Framework (at least a skeleton of it) carrying us from the network events (calls, SMS, data connection, ...) to the final statistical outputs to negotiate with MNOs details about both the access and their preprocessing. We plan to have this skeleton by the end of 2019. Contacts with MNOs are still ongoing in different countries but not reaching concrete agreements.

		In any form of data, the situation is complex and different per country and MNO. Since private investments are need to reuse these data for statistical purposes, private business interest and public interest need to be aligned and a solution for this has not been found yet.	These first steps in the statistical process will be executed by MNOs within their own premises. Only some form of pre-aggregated data will reach the NSIs' information systems. Methodological details (hence also computational and technological) are being worked out.	We defend not to pay for the data, but a compensation of marginal costs for the needed operations for public use must be considered. These marginal costs are not clear right now. Among other things, they depend on exactly what data are to be used to produce the convergence data aforementioned. Complementarily, an identification of the type of outputs is also needed, among other things, to assess the investment return both from the private and the public points of view.	

Do you know the size of the dataset? Will it be a problem to treat it at once? Will you split it for processing?	For mobile network data this question is not easy since it depends on the geographical and time range of the data (for one week we have obviously less data than for one month). Regarding the geographical scope, the consideration is similar (although it seems natural to have data for the whole national territory). Thus, without a time reference this question is difficult to answer. In the case of mobile network data, our access is completely partial and limited and we do not have a comparable answer in each case. For example, Italy has access to 5 weeks of CDRs for one province (although now negotiating an extension), The Netherlands are working with signaling data for the whole national territory, France is accessing an old set of CDRs from 2007 (5 months) for the whole national territory, Germany has received aggregate data for few weeks for only one Federal State (Hesse). Also, the data ecosystem within a mobile telecommunication network is extremely complex, thus the size of the dataset will depend on which concrete data we are referring to.	The first challenge is the identification of the concrete data we need to generate in each stage of the statistical process starting from the raw telco data. The convergence data will still need to be of high volume (although ready for statistical use), since they need to be used for many different statistical domains in later stages. Technical and computational details are not clear at this moment since the ESS Reference Methodological Framework is under construction. Mobile network data are generated daily and some form of splitting will certainly be needed.	The exact treatment depends on details of the ESS Reference Methodological Framework. In any case, the main goal of the data preprocessing stage is to generate the convergence data, which hopefully will have a standard structure	As explained above, some minimal form of investment by MNOs is needed to reuse telco data for statistical purposes. This investment will depend on the concrete datasets to reuse. The deeper in the network (i.e. closer to the antennas), the higher the volume and hence the investment. The role of NSIs is this investment is not clear at all. Some marginal costs will need to be considered but details are completely unknown right now.	See above.

	In the scheme describe above, raw telco data has the highest volume and they will never leave MNOs' premises. They have to be processed in-situ (they can be in the order of Pb). Once convergence data are produced, we estimate to have a much lesser volume of data, although still very large (no numeric estimation available yet). (Pre-)aggregated data reaching NSIs' premises will probably be not so much.	Apart from the design, a great deal of the execution will presumably be carried out by the MNOs themselves at their own information systems. This is a novelty in the traditional statistical process which will need careful addressing.			

Do you know the structure of the dataset? Are many different files considered a collection? (e.g. Consumption file for a smart meter plus a file for the location of the device are part of the same collection) a. Do you have to relate several files to have the entire dataset? b. Are the variables that enable linking of the data already known? If not do you have already a proposal to test the linkability?	Raw telco data are highly dependent on the concrete technology used by the MNOs, hence we face a diversity of structures. Besides, depending on the concrete raw telco data (CDRs, signaling, etc.), data structures can be different even at a given MNO. Different datsets need to be considered (e.g. CDRs, radio cell plans,...). Data are pseudonymised, time-referenced and geolocated. Basically the linkage is carried out using different form of these variables (pseudonymised ID, time interval, geocode).	The main challenge is double. On the input side, the diversity of data structures in different MNOs will require to adapt computational details to these differences. On the output side (regarding the convergence data), a standard structure needs to be found across MNOs and countries.	On the input side, we do not have a strategy yet. On the output side, we aspire to provide a (proto)standard for the convergence data so that all upper-level statistical analyses in different domains and in different countries will start from the same input.	See above.	See above.