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 K8: Evolution Roadmap Based on the Typification Matrix for Big Data Projects

Final version, 19.10.2020

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

Evolution Roadmap based on the Typification Matrix for Big Data Projects

During the first wave of big data projects for statistical production, the ESS members who participated detected several issues related either with the sources or with the challenges one faces when dealing with big data. During the second wave the big data projects were grouped pragmatically along three strands: exploratory, piloting and implementation projects. Concerning the current 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.

Although this classification implicitly evaluates a data source maturity level regarding its exploitation, several aspects are not clear. Namely:

- How the maturity assessment should be done
- Which characteristics catalogue one data source as being more suited to each of the 3 groups
- How to evolve from one less mature group to the next one, regarding quality, methodological or architectural issues.

This assessment 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 for it to attain a mature level. Such an appraisal is rather difficult to achieve, and subjective in nature, however since information can be gathered from the multiple and diverse work packages of ESSnet big data 1 and 2, it was thus collected and it is presented and discussed here to enable a deeper understanding of the current knowledge on big data sources issues, providing an heuristic assessment that may ultimately improve the future planning and development of big data sources usage.

In order to characterize and study in a more systematic way the big data sources being used, a Typification Matrix (TyM, 2020) was developed to collect, in an organized way, information about the big data sources. In the context of this tool, the data source maturity is always meant regarding its exploitation, for the production of official statistics.

Building on the Typification Matrix (TyM), the current document aims to provide an evolution roadmap based on the TyM for big data projects:

1. Covering the elements that characterize a big data source maturity
2. Mapping the required bus,iness architecture as proposed by Big data REference Architecture Level (BREAL)
3. Encompassing all the types/kinds of big data sources available in the projects of the ESSnet 2 on Big Data
4. Assessing a maturity level
5. Advising an evolution roadmap by:
    a. Comprehensively illustrating the connections with the Quality Guidelines for the Acquisition and Usage of Big Data
    b. Providing a link with the Methodological document for big data

The assessment into different levels of maturity was based on feedback from the work packages. This feedback in the form of filled out typification matrices is synthesized into the three levelled score (exploratory, piloting and implementation). Moreover, this score represents the current knowledge in the ESSnet Big Data II project team about the different data sources, so the assessment may change over time. Since such assessments rarely are really objective, it represents the authors' subjective view on this issue.

The next sections explain the way encountered to make such an assessment, using the big data source characterization collected via TyM, which is thus summarily described. To provide the maturity assessment we found necessary to know which processes and functions of the statistical production can be ensured by the new data source. For this purpose TyM is related to the BREAL architecture and some of the concepts necessary to understand the subsequent presentation of its 3 layers are introduced. In a similar fashion TyM’s relationship with the Quality Guidelines and Methodology Report will be presented as well as the connection to their structure, that establish how advice can be given and a possible evolution roadmap can be provided. All examples provided come from the work packages of Essnet on Big Data 2 and a recap of each will end the summary section.

The next 3 sections will be going in detail into the 3 layers of TyM, following the 5 points discussed above. The final section will present the project conclusions.

Typification Matrix Summary

To evaluate the usability of Big Data for producing official statistics several elements have been identified as necessary. These can be grouped according to their functionality in source, metadata and data. The characteristics regarding these elements 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 determine the potential usage of the big data source, along with its quality issues. For this reason one of the axis of the TyM is constituted by the source, metadata and data elements. TyM's 3 layers encompass:

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

The other axis focuses on the characteristic stages observed during ESSnet Big Data 1 (ESSnet BD1, IT Report 2018) (ESSnet BD1, Methodology Report 2018) that may empirically draw the boundaries of different maturity levels. Those:

- Upon an element 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 elements follow each other naturally. At least it's necessary to be able to describe a source, metadata or data. Each of which 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 usually be forecasted.

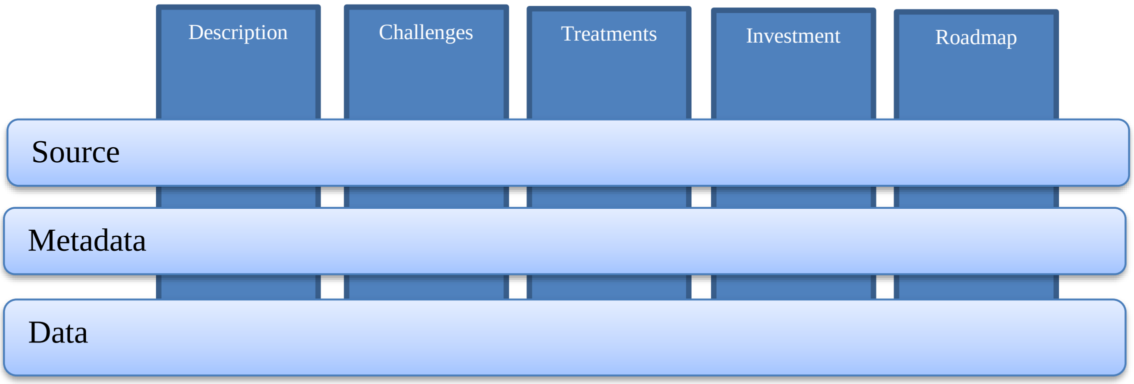


Figure 1 - Conceptual Typification Matrix

The 3 layers described above Source, Metadata and Data then have to be considered regarding its generic description, challenges, treatments, investments and roadmap. These two "dimensions", layers and columns, together define the conceptual typification matrix shown in figure 1.

Although the areas where information is required are Source, Metadata and Data, it is necessary to gather different elements for each of those. TyM includes several questions to cover the most important

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

aspects bearing in mind the business functions of the business architecture proposed by Big data REference Architecture Level (BREAL, 2020a).

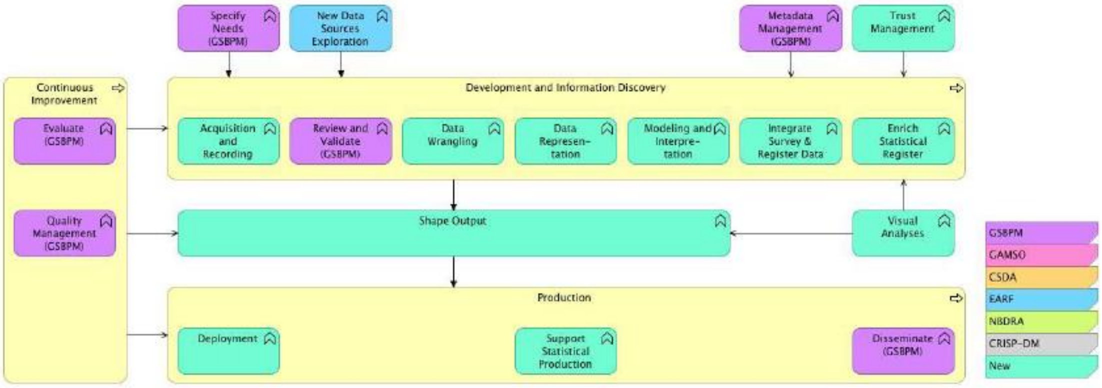

Specify Needs (GSBPM)
New Data Sources Exploration
Metadata Management (GSBPM)
Trust Management
Continuous Improvement
Evaluate (GSBPM)
Quality Management (GSBPM)
Development and Information Discovery
Acquisition and Recording
Review and Validate (GSBPM)
Data Wrangling
Data Represen-tation
Modeling and Interpre-tation
Integrate Survey & Register Data
Enrich Statistical Register
Shape Output
Visual Analyses
Production
Deployment
Support Statistical Production
Disseminate (GSBPM)
GSBPM
GAMSO
CSDA
EARF
NBDRA
CRISP-DM
New

Figure 2 - BREAL Big Data Life Cycle

In BREAL the business functions are grouped in business processes, for example the business functions Disseminate, Support Statistical Production and Deployment are all part of the Production business process. This business process is so far removed from the study of the data source that it's business functions were disregarded in TyM. The Shape Output and Visual Analysis being intermediary were also not taken into account.

TyM focuses rather on the development and information discovery business process as well as on all the business functions that may need to precede it, like new data sources exploration. Given the importance of a source allowing continuous improvement, even more with big data usage than with traditional data sources, those business functions were covered and will presently be considered to provide a maturity assessment.

In the next sections each of the 3 layers of TyM will be introduced, Source, Metadata and Data. For each, the questions that correspond to each business function will be presented, roughly corresponding to the first column of the conceptual typification matrix, description (see figure 1).

The examples collected during the testing of TyM from the second wave of pilots of the ESSnet2 on Big Data through the EUSurvey[2], will be discussed before presenting a corresponding maturity assessment; exploratory, piloting or implementation. This assessment will be based on the answers provided to TyM questions, independently of the actual strand where the big data project was placed in ESSnet 2 on Big Data.

The sections will finalize discussing the possible evolution based on the remaining columns of TyM. Challenges and treatments are closely related to the Quality Guidelines (Quality Guidelines, 2020), and Methodology (Methodological Report, 2020).

The Quality Guidelines are organized in a very clear and simple production process fashion, with only 3 phases: Input, Throughput and Output. Of course these could be further divided but for the present purpose these 3 are sufficient.

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

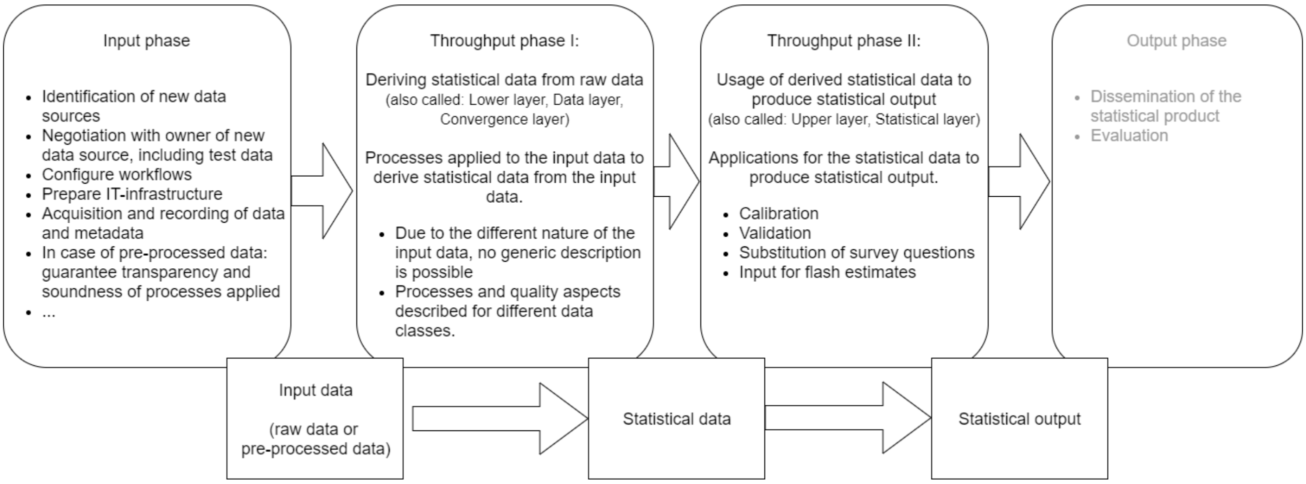

Input phase
Identification of new data sources
Negotiation with owner of new data source, including test data
Configure workflows
Prepare IT-infrastructure
Acquisition and recording of data and metadata
In case of pre-processed data: guarantee transparency and soundness of processes applied
...
Throughput phase I:
Deriving statistical data from raw data
(also called: Lower layer, Data layer, Convergence layer)
Processes applied to the input data to derive statistical data from the input data.
Due to the different nature of the input data, no generic description is possible
Processes and quality aspects described for different data classes.
Throughput phase II:
Usage of derived statistical data to produce statistical output
(also called: Upper layer, Statistical layer)
Applications for the statistical data to produce statistical output.
Calibration
Validation
Substitution of survey questions
Input for flash estimates
Output phase
Dissemination of the statistical product
Evaluation
Input data
(raw data or pre-processed data)
Statistical data
Statistical output

Figure 3 - Input, Througput and Output phases of the Statistical Production Process

The input phase includes every process that may provide data, i.e. input to the production, may it be received, collected, accessed or otherwise acquired. The througput phase, with one or more subphases encompasses all the processing and treatment required to transform the input data into statistical data, that may or may not be a statistical output. Finally the output phase of the production process is related to making the statistical products available and it's not important for the present discussion, since the usage of new data sources does not alter the typical processes of the output phase like dissemination and evaluation.

The Methodology report follows the same phases and the methods presented are organized accordingly. Despite the novelty or appeal of a given method, the criteria for its inclusion in the methodolgy report is its usage on published or soon to be published big data projects for the production of official statistics, laying the ground for a big data methodology. These are the only methods that may be refered to in the evolution roadmap.

The relationship between the Quality Guidelines provided and the Methodologies reported has traditionally been made in the statistical world. Its less evident relationship with the business functions of BREAL will be made clear along each of TyMs 3 layers and the diverse examples collected during the tool testing.

During November 2019 TyM was tested and filled by all the projects participating in the ESSnet Big Data 2. A brief summary of the activities of each work package will be provided below. For further detail please refer to the wikipage[3]. The ESSnet consists of 12 workpackages (WP), A to L. Only 4 workpackages are not exploring a specific class of data source. Apart from:

- WP A (Coordination supporting and Coordinating the project),
- WP F (Process and Architecture – Implementation Strand),
- WP K (Methodology and Quality – Pilot Strand)
- WP L (standalone WP on smart statistics)

all the other WPs filled TyM with the information for their respective source after 12 months of work with it, testing the tool usability regardless of the data class at hand. For further detail see TyM report (TyM,2020).

[3] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page

The data classes are defined by methodological, technical and governance aspects. They are here introduced because of the connection made with the Quality Guidelines and Methodological Report. Big data encompasses data that is so fundamentally diverse (from webscraped text to satellite images) that giving generic guidelines, that could target all of them simultaneously, would ultimately not be useful. The following table presents the aim of each WP and the respective data class.

Workpackage	Focus	Data class
WP B	Produce statistical estimates on Online Job Vacancies from Web advertisements	Webscraping
WP C	Enterprise information collected from the Web to improve or update information held by the national business registers	Webscraping
WP D	Usage of continuosly monitored elctricity smart meters for statistical production	SmartMeters Data
WP E	Real-time measurements of ship positions derived from Automated Identification System data (AIS) for statistical production	AIS Data
WP G	Explore the potential of data on financial transactions for contributing to official statistics	Financial Transactions Data
WP H	Earth observation data (satellite data and aerial photography) for statistical use	Earth Observation Data
WP I	Mobile networks data, generated through the use and moving about of mobile phones and similar devices for statistical production	MNO Data
WP J	Integration of various online big data sources with administrative registers on the Tourism Domain	Webscraping

Table 1 - Essnet on Big Data 2 workpackages description and data class

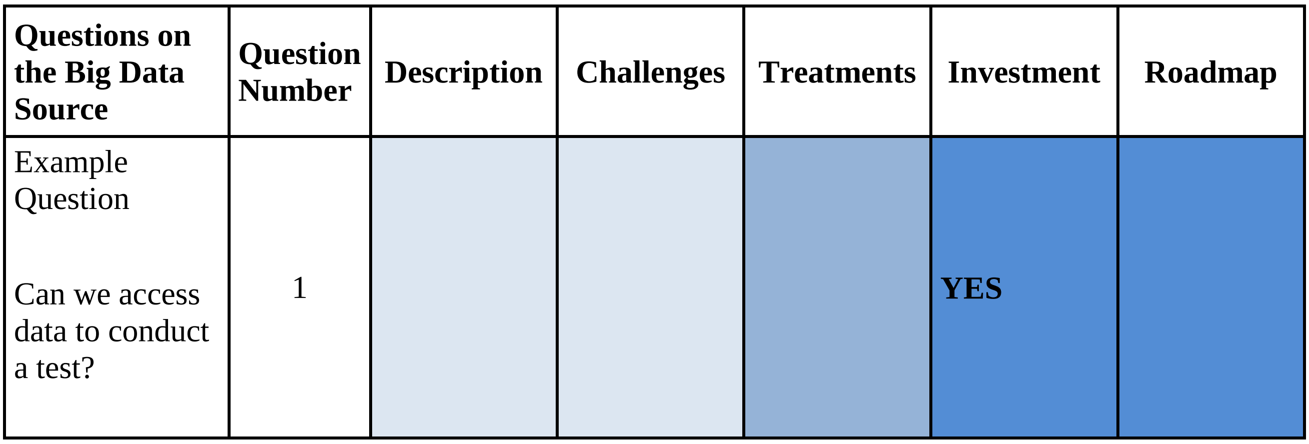
More information on any workpackage will be provided whenever needed in the next sections, particularly with some of the answers provided through TyM. The feedback from the work packages of the ESSnet on Big Data 2 are the basis for the assessments presented here, however an assessment of the present work packages is not made. Instead its the ability to answer to TyM questions that shows:

- how well one knows the data source,
- how advanced one is in it's exploration and
- how mature is the data source towards its exploitation for the production of official statistics.

The assessment proposed shows the 3 levels; exploratory, piloting and implementation mature source using a gradation of colour: light blue means able for exploration; medium blue apt for piloting and darker blue ready for implementation projects.

The starting point was an expectation, developed during the previous ESSnet on Big Data, that was called the basic assessment, where being able to describe and/or identify a source challenges was made to corresponds to the first maturity level, light blue. Knowing the treatments required was identified as necessary and enough for piloting, medium blue. Finally being able to forecast the necessary investment and/or tracing the roadmap to execute the plan would correspond to being ready for implementation, darker blue.

In some cases, it was found that it should be necessary not only to be able to provide an answer, but that the response be in a general direction. In this case, it is written in the cell of the summary table, as shown below.



Questions on the Big Data Source	**Question Number**	**Description**	**Challenges**	**Treatments**	**Investment**	**Roadmap**
Example Question Can we access data to conduct a test?	1				**YES**	

Figure 4 - Basic assessment pattern with required answers example

The basic assessment pattern stems from the practical examples collected during ESSnet on Big Data 1 and from the pragmatical distribution of the projects through the strands in the ESSnet on Big Data 2. However from the inability to provide answers, the weight of the business functions involved, the quality issues that arise or the lack of big data methodology this basic assessment pattern (BAP) was seldom found in the end. From the filled examples by the work packages and the subsequent discussions it was clear that the maturity assessment had to be relaxed or forced.

For example, although knowing the answer for a given question in terms of treatments to apply would generally place us in a medium (piloting) maturity level , it may in some cases be enough to place us already in an advanced (implementing) maturity level, thus the restrictions are in this case relaxed. On the contrary, it could be the case that even knowing the treatments is not considered enough to develop a pilot, in which case, instead of a medium (piloting) maturity level, a basic (exploratory) maturity level would be attributed, thus we would be forced to know more about the issue.

The representation of this is lightning or darkening a colour. When a cell presents a darker shade than what is expected, following BAP, we say that it’s relaxed. On the contrary if it shows a lighter shade than anticipated, it was forced. These situations will be explained along with the Assessment Pattern found to work for TyM.

The next sections will go in detail into the Source, Metadata and Data layers of TyM.

TyM Source Layer

The first questions when addressing a data source relate naturally to access, ownership and legal as well as ethical issues that may prevent access to the data. TyM asks 5 questions that cover these 3 areas:

1. Who owns the data? Public administration, one company, several companies?
2. Is it possible to get regular access? Does it have to be paid?
3. Are there legal or ethical problems to access the data?
4. 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?
5. Is there a possibility to access the data to study its relevance?

Their organization is explained further in the next subsection on business architecture BREAL.

Source - Business Functions

Source and Access is the first section of TyM and it focus on the BREAL business functions that precede the big data life cycle, as introduced in figure 2, while covering the access, ownership, legal and ethical issues.

The questions of access must distinguish between a single or one time only access, for a study or proof of concept and continuous access, as is required for the regular official statistical production. This distinction is obvious when we think in terms of the business functions (BF) involved. The question "*Is there a possibility to access the data to study its relevance*" adresses the "**New Data Sources Exploration**" BF while questions such as "*Is it possible to get regular access? Does it have to be paid?*" or "*Are there legal or ethical problems to access the data?*" target the "**Acquisition and Recording**" BF.

The legal and ethical problems that may limit the usage of a source as well as the costs involved in the process have to be accounted for by the "**Specify Needs**" BF and are covered by questions "*Are there legal or ethical problems to access the data?*" and "*Are there limitations? What is the nature?Which costsare involved? How can the costs be covered?*".

At last, but not least, the ownership of the data must be adressed. The access to data from a big data source may differ considerably from survey or administrative data, specially in the case of privately owned data where completely new ways of cooperation have to be developed. The question "*Who owns the data? Public administration, one company, several companies?*" targets the "**Trust Management**" BF.

Although the information is collected through five simple questions, the columns on challenges, treatments, investment and roadmap identified are able to differentiate the distinct levels of maturity regarding source knowledge as we shall present in the examples in the next subsection.

Source - Examples description

As expected most work packages identified few problems accessing the data, because they resorted to data owned by the public administration or webscraped. Among the webscraping projects (WP B, C and J) several concerns were expressed; questions on copyright protection, privacy rights and ethical principles on the one hand, and questions on site selection, search engine usage, universal scraping tools and methods on the other.

For all the work packages, the questions of access required establishing a cooperation between the institution that intends to use the data and the data provider. This was regarded as indispensable however in several cases it didn’t exist or didn’t prevent problems with data formats, on time transmission, etc. The dates of data transmission or access have to be established in order to allow the timeliness of the statistical production. In the same way the formats must be maintained and agreed upon between the involved parties, as they may compromise the production processes.

The work packages that scraped the data from the web may have their timeliness compromised due to changes on the pages they are scraping.

Source - Maturity Level Assessment

The first question that targets the Trust Management BF shows the basic levels of maturity, with an enforcement on investment. Once one is able to describe the source (provided one has access) the conditions for the exploratory level are met. Being able to devise the necessary treatment places the source on the piloting level and finally the roadmap is the key to the implementation level. In this case knowing the investments and being able to apply them is not enough to go into implementation because a path to access the data has to be established with the data provider, thus the cell was forced. When a protocol or cooperation with the data provider is guaranteed, its considered that the roadmap exists, thus the conditions are fulfilled.

The second question presents the basic assessment pattern, with a relaxation on challenges. However this result may emanate from the fact that all cases, except for WP I, consider public administration or webscraped data, thus with free access to data.

The third and fourth questions relate to the specify needs BF, which may be the explanation for the relaxation on the column challenges. This seems to suggest that giving the need to access the data, i.e., supported by a strong use case, the legal and financial obstacles can be overcome. However in both cases the relaxation is dependent on the privacy concerns and copyright questions being addressed and solved.

Finally the fifth question shows an extreme relaxation. However being related with the New Data Sources Exploration, it should be noted that in this case the implementation level means the implementation of case studies, and not the production of regular official statistics.

The summary colored TyM is presented below.

Questions on the Big Data Source	Question Number	Description	Challenges	Treatments	Investment	Roadmap
Source & Access						
Who owns the data? Public administration, one company, several companies?	1	ACCESS				
Is it possible to get access? Does it have to be paid?	2		YES FREE			
Are there legal or ethical problems to access the data?	3			ADRESSED SOLVED		
Are there limitations … a. What is the nature … b. Which costs are involved? c. How can the costs be covered?	4			ADRESSED SOLVED		
Is there a possibility to access the data to study its relevance?	5	YES				

Figure 5 - Maturity Assessment of TyM's Source Layer

Source - Evolution Roadmap

The most pressing question in the source layer is that the data owner/holder can prevent the adoption of a very promising source, hence the cooperation with the data provider is essential. To advance the maturity level of the source some kind of protocol must be agreed on. This provision agreement must stipulate the requirements of the data being provided: timeliness, confidentiality, quality, transmission protocol, authorship and pre-processing characteristics. For more information see the Provision Agreement Management Services of (BREAL, 2020b) on Support Services Description.

Quality guidelines on this issue, are also presented. To improve the applicability of the guidelines these are arranged according to the ways the National Statistical Institutes (NSIs) access the data from a new data source and the level of cooperation with the data provider. Some of the suggestions provided like establishing a "New Data Sources Apointee" relate with the new actors and roles required by Big Data Architecture. For more on the "New Data Source Apointee" refer to (Quality Guidelines, 2020) chapter *Input Phase – data source*. For more on the actors and roles refer to (BREAL, 2020a) chapter *BREAL Business Layer – Actors and relationships*.

Regarding the particular case of webscraping projects, there are specific quality guidelines under the subchapter *No cooperation with the data owner needed to gain access to raw data* (Quality Guidelines,2020) which adress the column challenges of TyM's Source Layer. The subsequent column, Treatments, when webscraping may be found on the (Methodological Report, 2020) chapter 2 on *Text and Text Mining*.

The next section focuses on TyM Metadata layer.

TyM Metadata Layer

Probably the most important questions regarding statistical production that we may pose to a source are metadata related. For this reason it was necessary to include 8 questions in this layer of TyM. The Metadata layer collects knowledge about the units and populations definition, identifiers, linkage and auxiliary information. These are the questions:

1. Is the definition of the population known? If not do you already have a method to address this issue?
2. Is the base unit of the data set known?
3. Do the units have an identifier?
4. Do you have the necessary variables to reach the relevant granularity level for the statistical unit?
5. 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?
6. Is there other information to make the data set useful with auxiliary data (NSI or other sources)?
7. Are there known quality issues with any of the variables?
8. Does the data contain sensitive variables? (Meaning legal or ethical issues related to its use)

Their organization is explained further in the next subsection on business architecture BREAL.

Metadata - Business Functions

The Metadata section of TyM picks up in terms of business functions where the Source Layer stopped. It covers the only remaining preceding life-cycle BREAL business function that wasn't covered by the Sources, i.e. the Metadata Management, as introduced in figure 2. The questions "*Is the base unit of the data set known?*" and "*Do the units have an identifier?*" adress the "**Metadata Management**".

Besides this point the other questions focus on the BREAL big data life-cycle Development and Continous Improvement processes.

In the current Metadata TyM layer the first question *"Is the definition of the population known? If not do you already have a method to address this issue?"* targets the "**Data Wrangling**" BF and depending on the degree of knowledge collected in the answers may also be relevant for the "**Methods and Tools Development**" support services. These questions together with "*Do you have the necessary variables to reach the relevant granularity level for the statistical unit?*" are connected with the **Modelling and Interpretation** BF, in particular directly to BREAL application architecture of **Statistical Aggregates** as presented in BREAL subchapter on Development Services and Components, chapter on BREAL Application Architecture (BREAL, 2020b).

With big data, as with any data source there is the need to examine data to try to identify potential problems, errors and discrepancies (such as outliers, item non-response and miscoding). But with big data one needs to go further than the mere input data validation, particularly when there is the need to relate it with other data. The question "*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?*" targets the **Review and Validate** BF.

Validating is different than evaluating. When Big Data sources are used, evaluation plays an important role. Most of the specificities of Big Data are related to its quick pace of change, in terms of the population covered and of their behaviour. Both the data set and its potential connections with other data must be closely monitored. The question *"Is there other information to make the data set useful with auxiliary data (NSI or other sources)?"* addresses the **Evaluate** BF.

The continuos improvement processes are tackled by two important BFs Evaluate, and Quality Management. All evaluations result in feedback, which should be used to improve the relevant process, phase or sub-process, creating a quality loop. The question "*Are there known quality issues with any of the variables?"* targets the **Quality Management** BF.

Probably the major causes of concern, when dealing with big data are the ethic and legal aspects of the data usage. For this reason it was included the question "*Does the data contain sensitive variables? (Meaning legal or ethical issues related to its use*)" which maps to the **Legislative Work Participation** support service as presented in BREAL subchapter on Support Services and Components, chapter on BREAL Application Architecture (BREAL, 2020b).

In the next subsection we discuss the examples gathered in the ESSnet regarding the Metadata layer of TyM.

Metadata – Examples description

The information collected through the work packages of the ESSnet on Big Data 2, encompasses several types of big data sources, but 3 of the projects webscrape their data and in particular WP B on online job vacancies and WP J on Tourism share the same preoccupations. These are extensive to any data source and do not come from the data representation questions, posed by unstructured or semi structured data, that may rise from the usage of web scraped data. New methods and tools development is required and the statisticians and researchers are only now getting used to the new units present in their data sets that are not particularly suited to their statistical needs.

Besides it's virtually impossible to choose representative samples and the matching across multiple sources is not easily solvable requiring further investigation.

While WPC also relies on web scraped data on enterprises variables, it is one of the serendipitiuos cases where the data sets unit matches the target unit. For comparison purposes, the enterprise as the unit for retrieving OBECs is the same as defined in Council Regulation 7 (EEC) No 696 /93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community. This makes the example of WPC closer to WPG on financial transaction where the units also don't pose any problem, whenever the micro data is available, i.e. every transaction by itself. Even if this is the case with WPC the difference in the structure and content of the websites still depends to a large extent on the size of the enterprises, the complexity of their organization, and the economic activities they carry out, thus requiring some sort of stratification.

This again focuses the completeness problem extensively to all the web scraping projects; which enterprises, which job offers, which accomodations information is captured, only the ones with online presence. Furthermore the classifications gathered from the internet, for the accomodations for example, do not match the statistical nomenclatures. There is also no standardization on how the adresses are presented, be they addresses of companies or of accomodation locations, and although in many cases the geographic coordinates are available, it is not simple to assign them to a specific unit, i.e. reduce the unit to a single point.

This difficulty is shared by WPH on the usage of satellite data. The question in this case is to interpret the pixels' values and connect them with physical objects on the ground. Here some of the problems are generic to the usage of satellite data, like cloud cover and spatial resolution, while others reside in the ability to define meaningful units that may serve the statistical purpose at hand. The holy grail when using big data sources is to find this correspondence between the data set units and the statistical units. Even in the cases where all the data (usually sensor generated data) could potentially be available, like in the cases of WPD, WPE and WPI (smart meters, AIS and mobile phone data respectively) it doesn't exempt us from the need of developing rigorous statitistical treatments that may connect the datasets with the target populations under analysis. Examples for this last three work packages are the connection between a smart meter and the end user for WPD, where several consumers share a metering point; the need to decide on coordinates for the end of a journey on WPE, since the destination filled by the shipper is not always accurate; and in WPI the transformations needed to go from the base units that are network events (like connecting the phone or sending a message) to the intermediate units that are the mobile device, and finally to the statistical units, the population.

Last but not least it's refered that not only the legal landscape but also the public perception are in constant change, which renders the usage of big data sources even more sensitive. The changes in these answers introduced in the expected basic assessment pattern BAP are discussed in the following subsection.

Metadata – Maturity Level Assessment

The first question regarding the population definition shows a relaxation on treatment and investment, which is natural given its very practical nature of data wrangling; which involves the data preparation, cleaning and enrichment specific when dealing with big data but that doesn't entail other statistical methodological difficulties.

In comparison the questions regarding the base units, data linking and quality issues (questions 2, 3, 6 and 7) with the variables follow the expected BAP. This correspondence shows that there is no unexpected easy way out in these areas, but also no surprising difficulties.

On the other hand, the questions on data granularity and the transformation of base units to statistical units (questions 4 and 5), as we have seen through the examples provided, show that more knowledge and experience is required before progressing to the piloting phase. Thus they both are forced on treatment. Displaying the same forcing on treatment is the last question on the sensitive variables, reflecting the legal and ethical concerns mentioned on the examples.

The summary colored TyM is presented below.

Questions on the Big Data Source	**Question Number**	**Description**	**Challenges**	**Treatments**	**Investment**	**Roadmap**
Metadata						
Is the definition of the population known? If not do you already have a method to address this issue?	1	░	░	▓	▓	▓
Is the base unit of the data set known?	2	░	░	▒	▓	▓
Do the units have an identifier?	3	░	░	▒	▓	▓
Do you have the necessary variables to reach the relevant granularity level for the statistical unit?	4	░	░	░	▒	▓
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?	5	░	░	░	▒	▓
Is there other information to make the data set useful with auxiliary data (NSI or other sources)?	6	░	░	▒	▓	▓
Are there known quality issues with any of the variables?	7	░	░	▒	▓	▓
Does the data contain sensitive variable? (Meaning legal or ethical issues related to its use)	8	░	░	░	▒	▓

Figure 6 - Maturity Assessment of TyM's Metadata Layer

Metadata – Evolution Roadmap

From the information gathered through the work packages testing of TyM regarding Metadata, two main obstacles emerge on the maturity evolution path of any given big data source.

One of them has to do with the legal and ethical issues of using the big data source, or some of its data, and the other with establishing a connection between base units and statistical units.

We will start by focusing on the first problem where some resolution paths may be suggested by the General Data Protection Regulation that was adopted and became enforceable shortly before the second ESSnet on Big Data started. In the past the individuals supplied their information directly to NSIs answering to their surveys. Nowadays many of the activities carried out by the people leave a digital footprint that can help the citizens to answer accurately to these surveys or, in some cases, even to substitute them, reducing their burden. In fact, every individual is entitled to have their personal information protected (GDPR, 2018), used in a fair and legal way, and made available to them when they ask for a copy. Theoretically then, when a citizen must fulfill a survey one can choose to retrieve their data from a relevant big data source and provide it to the NSI.

This solution may be impractical, probably encompassing the retrieval of the data of thousands of citizens. However in some circumstances, one may be entitled to obtain his/hers personal data from a data controller in a format that makes it easier to reuse the information in another context, and to transmit this data to another data controller of one's choosing without hindrance. This right to data portability applies where processing of personal data (supplied by the data subject) is carried out by automated means, and where one has either consented to processing, or where processing is conducted on the basis of a contract between the individual and the data controller. Still according to the GDPR, an organisation can process personal data to carry out a task that is in the public interest, such as official statistics production.

Models of collaboration with data controllers can be devised that protect the citizen privacy rights transferring only aggregated and statistical data instead of individual and raw data. A deeper explanation on the different characteristics of statistical data versus raw data as well as examples and quality guidelines are provided in the chapter "Throughput phase I: Deriving statistical data from raw data of a big data source" (Quality Guidelines, 2020).

The second difficulty mentioned, establishing a connection between base units and statistical units, is also known as the inference problem. The usage of big data sources does not imply that new ways of drawing inference must be devised. Instead what is necessary is a high awareness to the questions that the usage of big data intrinsically entails like completeness, coverage and missing or low quality data, that must be accounted for and compensated. Statistical inference is the focus of (Methodological Report, 2020) chapter 6.

Finally, in terms of architecture, the path to increase the big data source maturity level is naturally the Modelling and Interpretation BF, since virtually in every case a model will ultimately be used to infer from the data. Moreover, when big data is used for statistical production, conceptual differences in what is measured and the fast pace of change will require very constant continuos improvement processes loops as described in (BREAL, 2020a) Evaluate and Quality Management BFs.

The next section describes the Data Layer.

TyM Data Layer

The final TyM Layer focuses on the data and on the type of issues that may arise because of it being big data instead of the traditional data that is traditionally used for statistical production. This layer covers the type, size, format and structure of the data on these 3 questions:

1. What is the type/format of the data?
    a. Human-produced records
    b. Machine-produced records
    c. Satellite Images
    d. Web scraped text
    e. Video files
    f. Audio files
    g. Other. Which?
2. Do you know the size of the dataset? Will it be a problem…
3. Do you know the structure of the dataset? Are many different files considered a 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?...

The aim and importance of each question will be introduced in the next subsection with the BFs of BREAL.

Data - Business Functions

When data is unstructured or semi-structured, as for example with webscraped data, it is necessary to establish a data structure to represent the data. The first question of the TyM Data Layer "*What is the type/format of the data?"* aims to assert the needs regarding the **Data Representation** BF of BREAL.

Closely connected to the type and format of the data are the questions of volume and the capabilities needed to manage tasks and decisions surrounding the ingestion, annotation, cataloguing, storage, retrieval and distribution of IT assets. The second question, although intentionally open "*Do you know the size of the dataset? Will it be a problem…*" refers to the **IT Management** BF.

Finally, the third question continues exploring the **Modelling and Interpretation** BF in the particular aspects of :

- the capability to enrich the statistical register(s) with the information retrieved from the big data source with question "*Do you have to relate several files to have the entire dataset?*" – **Enrich Statistical Register** BF
- the ability to reuse and integrate other data like survey & register data in order to enhance the quality or improve the value of the results derived so far with question "*Are the variables that enable linking of the data already known*" – **Integrate Survey and Register Data** BF.

The examples collected from all the work packages will further illustrate these aspects in the next subsection.

Data – Examples description

Unsurprisingly after the difficulties enumerated with the base units connection with the statistical units, and of the big data sets to auxiliary data, the linking with other data that would permit the integration with surveys or the enrichment of statistical registers remains an obstacle and methodological analysis and assumptions are needed, as refered by the work packages in their feedback.

All work packages without exception mention as a complication the intensive processing required, be it because of text mining treatments (when webscraping), the process of complete sets of data at once (AIS), satellite images processing requirements, bayesian hierarchical models or other reasons. This difficulties were to be expected since we are dealing with big data but somehow they are stronger than anticipated by all the work packages in progress.

The repercussion of this result will be shown in the assessment subsection but before we close the answers presentation it is worth mentioning that in connection with the size and structure of the datasets it was mentioned by work packages already in the implementation phase how much they are dependent on the collaboration and good will of the data providers that change structures without previous agreement, due to the lack of well established provision agreements. This issue was discussed in the source layer of TyM. For more information see the Provision Agreement Management Services of (BREAL, 2020b) on Support Services Description.

Data – Maturity Level Assessment

Based on the information gathered through the examples the data representation seems not to be a problem and even projects that are identifying the challenges of data representation are ready for piloting, thus we show a relaxation on the challenges column for the first question.

However the questions related with using and processing big data seem more pressing than anticipated, which explains the enforcement on the Investment column of the second question, related to the need to ensure the IT support before going into implementation.

Finally the most complex issues are the ones associated with unit linking, therefore the third question shows two columns forced, treatments and investment. Usually exploring treatments is considered suitable for pilots, while it was found advisable in linking units case to experiment with those still on the exploratory phase. Similarly working out the investment is generally appropriate for the implementation phase, although in this case seems preferable to stay on the piloting stage.

The summary colored TyM is presented below.

Questions on the Big Data Source	Question Number	Description	Challenges	Treatments	Investment	Roadmap
Data Type						
Choose below the type/format of the data? a. Human-produced records b. Machine-produced records c. Satellite Images d. Web scraped text e. Video files f. Audio files g. Other. Which?	1	░	▒	▒	▓	▓
Do you know the size of the dataset? Will it be a problem…	2	░	░	▒	▒	▓
Do you know the structure of the dataset? Are many different files considered a 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?...	3	░	░	░	▒	▓

Figure 7 - Maturity Assessment of TyM's Data Layer

Data – Evolution Roadmap

The difficulties reported on processing the big data, either because it would be desirable to process it all at once (AIS data – WP E) or because the algorithms necessary are processing intensive (MNOs data – WP I) may be adressed on the architecture level. Namely working jointly with other NSIs and combining efforts to share infrastructure platforms, application services or even data sources. This is described in (BREAL, 2020b) where an operational model is proposed to deploy such solutions.

Regarding the linking units issues (Quality Guidelines, 2020) provides relevant guidelines for each of the big data classes that match partially the WPs of these ESSnet and (Methodological Report, 2020) devotes chapter 4 to linking units.

Naturally the linking of units presents more difficulties and harder challenges when we combine several data sources. TyM doesn't deal with more than one data source at a time but we can also recommend (Komuso, 2018) that particularly describes processing errors that do not play a role for unique data sources.

Conclusions

The current attempt of maturity assessment of a big data source builds on the work of the Typification Matrix (TyM), covering the most relevant elements to characterize a big data source maturity. To propose an evolution roadmap for the big data sources heuristic connections were established with:

- (Quality Guidelines, 2020) through TyM's column Challenges
- (Methodological Report, 2020) through TyM's column Treatments
- (BREAL, 2020a) and whenever possible with (BREAL, 2020b) through the questions. Annex A presents the correspondence of each question with BREAL processes and business functions following the big data life cycle in official statistics production.

The evolution roadmap process described was used with all the types/kinds of big data sources available in the projects of the ESSnet 2 on Big Data, without showing problems for any specific type of data source. The three levelled score (exploratory, piloting and implementation) seemed appropriate and the need to distinguish further maturity levels did not arise from the feedback received from the work packages. However it represents the current knowledge in the ESSnet Big Data II about the different data sources, so the assessment may change over time. Since such assessments rarely are really objective, it represents the authors' subjective view on this issue.

Although the Big Data sources available and relevant for the production of official statistics are expected to change, the data types used may evolve but will remain in nature similar to the ones addressed. More esoteric types of data will be increasingly common like Genomic Data (DNA genomes), High Dimensionality Data (facial recognition technologies) or Translytic Data (simultaneously transactional and analytic) but as is the case in any very specialized field this data tends to be highly structured. Less structured and disperse will be the personal information gathered through IOT (internet of things). Even so personal information follows the dimensions of human life: space and time, therefore posing mainly the same challenges of sensors or mobile phone data.

During the discussion of TyM's data layer the concern with errors and issues that may come up when combining sources was mentioned. As previously stated, TyM does not deal with several big data sources simultaneously and neither can the present work, based as it is on TyM. Despite this limitation it is still a scalable solution. Each data source must be evaluated on a matrix and those can be collected in a cube where each source appraisal corresponds to a slice of the cube. If one chooses to do that, the questions on linkage can be changed to consider specifically a pair of sources of the Typification cube. Both for fill and visualization purposes, the number of TyM slices must equal the number of sources being evaluated. Since this question is related to the evolution of big data projects, as using more than one source is in many cases necessary, or advisable, for the production of official statistics this could be the focus of future work.

References

BREAL, 2020a - Big data REference Architecture Level - Essnet on Big Data 2, Workpackage F Deliverable F1 (2020) https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WPF_Deliverable_F1_BREAL_Big_Data_REference_Architecture_and_Layers_v.03012020.pdf (accessed 16th June, 2020)

BREAL, 2020b – Data and Application Architecture - Essnet on Big Data 2, Workpackage F Deliverable F2 (2020) https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WPF_Milestones_and_deliverables (accessed 26th July, 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)

GDPR, 2018 - Regulation (EU) 2016/679 (General Data Protection Regulation) in the current version of the OJ L 119, 04.05.2016; cor. OJ L 127, 23.5.2018 https://gdpr-info.eu/ (accessed 20th July, 2020)

KOMUSO, 2019 - ESSnet KOMUSO, Quality Guidelines for Multisource Statistics (QGMSS), Version 0.81 https://ec.europa.eu/eurostat/cros/system/files/wp1_guidelines_-_v0_8_1.pdf (accessed 20th July, 2020)

Methodological Report, 2020 – Methodological Report - Essnet on Big Data 2, Workpackage K Deliverable K5 (2020). https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/d/df/WPK_Deliverable_K5_First_draft_of_the_methodological_report_2020_06_17_final.pdf (accessed 16th June, 2020)

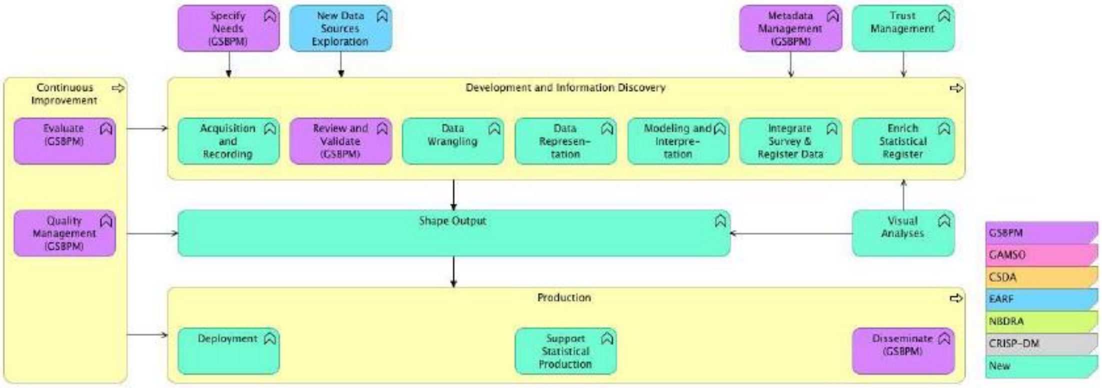
Quality Guidelines, 2020 – Quality Guidelines for the Acquisition and Usage of Big Data - Essnet on Big Data 2, Workpackage K Deliverable K3 (2020). https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/f/f8/WP3_Deliverable_K3_Revised_Version_of_the_Quality_Guidelines_for_the_Acquisition_and_Usage_of_Big_Data_Final_version.pdf (accessed 16th June, 2020)

Ricciato et al, 2019 - Ricciato, F., Wirthmann, A., Hahn, M. (2019), Integrating alternative data sources into official statistics: a system-design approach.

TyM, 2020 – Typification matrix for big data projects – Essnet on Big Data 2, Workpackage K Deliverable K7 (2020). https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/e0/WPK_Deliverable_K7_Typification_matrix_for_big_data_projects_2020_04_30.pdf (accessed 16th June, 2020)

Annex A

For convenience the BREAL Big Data Life-Cycle (LC) figure 2 is reproduced again, preceding the organization of TyM questions by BREAL processes and business functions.


Specify Needs (GSBPM)
New Data Sources Exploration
Metadata Management (GSBPM)
Trust Management
Continuous Improvement
Evaluate (GSBPM)
Quality Management (GSBPM)
Development and Information Discovery
Acquisition and Recording
Review and Validate (GSBPM)
Data Wrangling
Data Represen-tation
Modeling and Interpre-tation
Integrate Survey & Register Data
Enrich Statistical Register
Shape Output
Visual Analyses
Production
Deployment
Support Statistical Production
Disseminate (GSBPM)
GSBPM
GAMSO
CSDA
EARF
NBDRA
CRISP-DM
New

BREAL processes	BREAL BF	TyM Layer	Questions
Preceding LC	Specify Needs	Source	3, 4
Preceding LC	New Data Sources Exploration	Source	5
Preceding LC	Metadata Management	Metadata	2, 3
Preceding LC	Trust Management	Source	1
Development LC processes	Acquisition and Recording	Source	2, 3
Development LC processes	Review and Validate	Metadata	5
Development LC processes	Data Wrangling	Metadata	1
Development LC processes	Data Representation	Data	1
Development LC processes	Modelling and Interpretation	Metadata	4
Development LC processes	Integrate Survey & Register	Data	3 (a)
Development LC processes	Enrich Statistical Register	Data	3 (b)
Continuos Improvement Processes	Quality Management	Metadata	7
Continuos Improvement Processes	Evaluate	Metadata	6
Supporting Services	Legislative Participation	Metadata	8
Supporting Services	IT Management	Data	2