ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

https://ec.europa.eu/eurostat/cros/content/essnet-big-data-0_en

Workpackage WPA

Coordination and Communication

Deliverable A5: Final Technical Report

Final version, 29 June 2021

Prepared by:

Marc Debusschere (WPA, SB, Belgium)
Martin van Sebille (WPA, CBS, the Netherlands)
Peter Struijs (WPA, CBS, the Netherlands)
Tomaž Špeh (WPB, SURS, Slovenia)
Galya Stateva (WPC, BNSI, Bulgaria)
Arko Kesküla (WPD, EE, Estonia)
Remco Paulussen (WPE, CBS, the Netherlands)
Monica Scannapieco (WPF, ISTAT, Italia)
Johan Fosen (WPG, SSB, Norway)
Marek Morze (WPH, GUS, Poland)
David Salgado (WPI, INE, Spain)
Marek Cierpiał-Wolan (WPJ, GUS, Poland)
Alexander Kowarik (WPK, STAT, Austria)
Natalie Rosenski (WPL, DESTATIS, Germany)

Workpackage Leader:
Peter Struijs (CBS, the Netherlands)
e-mail address: p.struijs@cbs.nl
telephone: +31 45 570 7441
mobile phone: +31 6 5248 7775

Contents

Executive Summary

This is deliverable A5, the Final Technical Report, of the grant agreement of the ESSnet Big Data II, which builds on the results of the ESSnet Big Data I (2016 – 2018). The ESSnet, which has 28 partners, covers the period from November 2018 to June 2021.

The ESSnet has organised its work in twelve workpackages, of which the first (WPA) concerns coordination and dissemination and the other produce statistical results. These eleven workpackages are organised in three tracks. The first comprises implementation projects (WPB to WPE) and cross-cutting implementation issues regarding the statistical process and architecture (WPF). The second consists of new pilot projects (WPG to WPJ) and cross-cutting issues regarding statistical methodology and quality. The third (WPL), which, as planned, finished earlier than the other tracks, investigated potential research areas for so-called trusted smart statistics.

The work of the ESSnet was concluded with two virtual events: One six-day dissemination event for a broad audience in November 2020, in which the results of the ESSnet were presented and discussed, and one three-day future-oriented event for a more limited audience in June 2021, in which eight themes related to the work of the ESSnet were discussed in groups, resulting in ideas to bring the use of big data for official statistics in the ESS to the next level of maturity.

These are the main results obtained in the ESSnet:

<u>WPB Online Job Vacancies</u>

The aims of implementing WPB on online job vacancies were to produce statistical estimates on the topic of online job adverts (OJA) and to identify statistical production processes and capabilities that may be affected at the national level and to define the conceptual production processes at the national level and at the level of the ESS. The implementation phase followed the work carried out during the pilot phase of the project ESSnet Big Data 2016-2018. The aim was also to establish the conditions under which OJA web scraping techniques can be used and to evaluate the quality of the data.

Methodological aspect related to the use of OJA data for producing statistical indicators has been addressed at the conceptual and practical level also by analysing the data produced by the European Centre for the Development of Vocational Training (CEDEFOP). Several use cases of using OJA data in official statistics have been developed that prove this data could be used successfully for the needs of official statistics for completing and estimating a number of characteristics of the job vacancies. These are not ready to be used as tools for decision making, and rather serve as proof-of-concepts illustrating some of the paths that are possible to follow in the future for creating statistical outputs based on OJA data. The quality issues still need to be addressed that would enable them to meet the standards expected of official statistics. On the other hand, OJA data provide many insights that official estimates cannot. Statistics based on OJA data can be published in a very timely and frequent manner, allowing for short-term tracking of labour market conditions and flash-estimates of labour demand. Because of the large data pool available, OJA data also allow for more granular analysis by subgroups or geographical regions. Additionally, OJA data might allow for the provision of completely new (to official statistics) labour market insights: for instance, indicators of labour market power by employers. There is, however, further need to address the challenges how OJA data should be interpreted and used together with official estimates. Their main advantages are their accessibility, timeliness,

comparatively low cost of producing statistics and lack of or minimal burden for the respondents. Most OJA contain information not only for the vacant position but also for the required qualifications and skills, including the so-called soft skills, for degree and kind of the graduate education and professional experience necessary for the potential applicants for taking it. Estimates could be made for labour demand by economy sectors, professions, specific skills and qualifications, and regions.

The analysis showed modest success in predicting job vacancy survey values using OJA data, so these data could be used for producing flash estimates of labour demand. It may also be possible to use these time series properties to produce more frequent estimates, or even reduce the frequency of the survey. CEDEFOP data seem to offer a promising basis for developing pertinent statistics and economic indicators. A preliminary analysis showed that existing relationships between OJAs, national and regional population sizes, and national economic activity expressed as GDP should be further explored. It is important to have in mind the significant differences between ESS countries in terms of the OJAs landscape. Thus, what is feasible in one country may not necessarily be reproducible in others. Secure long-term data acquisition is a crucial issue, since public statistics must offer grounds on which to compare changes over time using comparable data.

Integrating OJA in the statistical production may be a major goal of collecting OJA in the future. A good understanding of the "business models" underlying the development of job portals, as well as of the market of OJA is crucial. The online job portals have not been developed to produce statistics or indicators of the job market. Their goal is different from one portal to another, some of them are focused on human resources aspects, in finding the right person for a specific job, but other portals are focused only on making money from promoted adverts on their web pages. Understanding of their mechanisms is a key issue to understand the data collected and identify the best fit model to use them for statistical purposes, direct on job vacancies statistics or in correlation with other socio-economic indicators.

WPC Enterprise Characteristics

WPC, on Online Based Enterprise Characteristic (OBEC) is about understanding economic and business activity online from a national statistics perspective.

Statistical business register (SBR) enhancements with details on nationally listed businesses' online presence, such as websites, e-commerce or social media accounts, are a key output. The main aim of WPC to use web scraping, text mining and inference techniques for collecting and processing enterprise information, in order to improve or update existing information, such as Internet presence, kind of activity, address information, ownership structure, etc., in the national statistical business registers was achieved.

Within WPC, the methodology of the previous ESSnet Big Data I project was generalized and extended for use in any ESS country, taking into account the variety needed to support different use cases.

The WPC team has produced not just the experimental OBEC statistics, but also the Reference methodological framework and software Starter kit, so that NSIs can move ahead quickly with producing OBEC statistics within their own national contexts. As web scraping is a relatively new data sourcing method for NSIs, which requires due attention to data protection, the WPC team also

published an ESS web scraping Policy Template covering the key legal and ethical considerations and setting out a robust set of principles and practices for NSIs to follow.

WPD Smart Energy

WPD developed a production process to implement smart meter data for the following statistical products:

- electricity statistics of businesses, by sector
- electricity statistics of households
- identifying vacant or seasonally vacant dwellings by new estimation models

The production process includes setting up procedures and developing technical solutions to promote and support the collection, processing, and analysis of the data for statistical production. Additionally, it is shown that useful information can be extracted from smart meter data. With daily data, it is possible to extract businesses vs dwellings patterns. A Danish example shows how smart meter data can be used to explore the effects of the lockdown due to the COVID-19 pandemic on energy consumption.

WPE Tracking Ships

Maritime and inland waterways traffic has increased exponentially over the last decades. This required different solutions to ensure safety at sea, which resulted in the Automatic Identification System (AIS). Almost every ship broadcasts its location and status information by means of AIS, making it possible to detect other ships, wherever they are. Building on the developments of the completed work in the ESSnet Big Data I on the use of AIS data for maritime statistics, the aim of WPE was to develop functional production prototypes:

1. Improve statistics on Inland Waterway transport: information on some parts of the Dutch inland waters is missing, resulting in incomplete data on ships' journeys and with that, information on goods (un)loaded. Implementing information from AIS in the process, completes this missing information on ships' journeys. In turn, a methodology was developed to estimate the type and quantity of the commodities transported.
2. Develop statistics on the behaviour of fishing vessels: using AIS, the first results on the activity of fishing fleet (time when a fishing vessel is out of the port) and the traffic of fishing fleet (the number of fishing vessels in fishing areas) in Poland have been derived. Several visuals have been created: a map with the fishing ports and fishing areas, and the traffic intensity.
3. Improve the quality of statistics on port visits: a Port Visits Geo-Solution prototype was developed in PostgreSQL. AIS data is used to derive the port visits and port traffic of the Piraeus port in Greece. The prototype shows that we are able to generate the official so-called F2 table.
4. Improve statistics on air emissions and energy used for the environmental accounts: a method was developed to identify shipping vessels related to the Dutch economy using text mining techniques. Research was performed to investigate the use of distance travelled per ship as a proxy of fuel usage and emissions. As there was no direct opportunity to actually calculate fuel usage and emissions, due to e.g. insufficient availability of data sources, three theoretical methods were developed to implement AIS.

The products mentioned were created using several big data platforms: local platforms, the Big Data Test Infrastructure (BDTI) environment and its successor, the DataPlatform, both provided by the European Commission, and the United Nations Global Platform (UNGP) provided by the UN Global Working Group (GWG) on Big Data for Official Statistics. Each environment has its specific set of tools and uses different sources of AIS data with different coverages. AIS data is collected by several parties around the world, though the scope of the data may differ. Some parties only collect national data, some parties collect data on a European level and some parties collect all data around-the-world. Some parties collect maritime data and some parties collect data focused on inland waterways.

During the project, each time the best environment for developing and testing the product had to be selected. The choice depended mostly on the coverage of the AIS data.

The port visit product, which was developed by Greece, was executed to create the F2 tables for several Dutch and Polish ports. And the fishing fleet product, which was developed by Poland, was executed to create the first results on Dutch and Greek fishing fleet activities and behaviour. This shows that the products can be executed on other platforms than initially developed. And more importantly, it shows that the solutions are generic and can be used to produce the results for other countries.

WPF Process and Architecture

WPF, "Process and Architecture", has worked on defining a European reference architecture for Big Data to guide Big Data investments by NSIs and help the development of standardized solutions and services to be shared within the ESS and beyond.

The main outcome of WPF is what we called BREAL (Big Data REference Architecture and Layers), an architectural framework including several architectural artefacts: (i) in the Business Layer - a *List of principles*, a *Business functions model*, a *Life Cycle model*, a *Support functions model* and a *Stakeholder model*; (ii) in the Application Layer - a *Generic Application Architecture;* (iii) in the Information Layer - a *Generic Information Architecture; (iv)* in between the Application Layer and the Information Layer - an *Operational Model*.

The proposed Generic Application Architecture and Information Architecture are "generic" in the sense that they are not intended to be specific of a Big Data project or source. Conversely, the use of the services and data that are specific to the Big Data projects of the Implementation Track of the ESSnet Big Data II are described in a set of solution architectures, namely: (i) Solution architectures for *Online job vacancies* (WPB); (ii) Solution architectures for *Online based enterprise characteristics* (WPC); (iii) Solution architectures for *Smart energy* (WPD); (iv) Solution architecture for *Tracking ships* (WPE).

WPG Financial Transactions Data

The main aim of WPG was to provide an overview of the data infrastructure (metadata) of financial transactions data in the participating countries: to what extent are financial transactions data available and whether it is possible for NSIs to get metadata and ultimately to access the financial transactions data. Given the infrastructure, it has also been a main aim of the workpackage to assess the statistical potential of these data sources. WPG also aimed at studying the potential of financial transactions data in describing the sharing economy.

In the description of financial transactions data, a general overview was given of the payment system, payment instruments, the legal aspects, and the main actors. The financial transactions data situation in the WPG countries were described: Bulgaria, Germany, Italy, Norway, Portugal and Slovenia. An assessment was made on which promising statistics could be identified prior to accessing data, resulting in the following statistics: household budget survey, household expenditure of residents for tourism and balance of payments estimates, e-commerce turnover, retail trade index, early estimates of macroeconomic aggregates and indicators, business statistics indicators on flows between industries, and finally turnover of collaborative/sharing economy.

Case studies were performed to study more closely the potential of financial transactions data for official statistics. An Italian case study shows the gain of using time series of aggregated daily card transactions in forecasting and nowcasting of economic indicators. The Portuguese case study looked at how to improve the tourism statistics by addressing small lodgings, as well as how to measure new types of transport. The Norwegian case study looked at using financial transactions data for estimating the monthly total retail sales.

For sharing economy, different definitions were discussedbefore selecting an operational definition being useful for official statistics. Different approaches to assessing sharing economy were considered and it was decided to use financial transactions data for finding indicators of sharing economy. After an internal survey on the availability indicators of credit cards usage found in different WPG countries, sharing economy indicators were derived. It turned out that an empirical comparison between WPG countries was not possible in practice given the available data.

WPH Earth Observation

WPH aimed to support areal statistics with Earth Observation (EO) data. The project resulted in experimental statistics using remote sensing data. From the technological point of view, WPH used machine learning algorithms for statistical image analysis. The crucial goal of WPH was the usage of the EO data from different sources that contributed to building the geospatial framework to support the statistical registers. Within this project, the usefulness and practical usage of EO data in agriculture, built-up area, land cover, settlements, enumeration areas, and forestry thematic fields were proposed. The main objectives of WPH were implemented by the execution of nine case studies. Within the cases studies, big data sources were evaluated and possible statistical products from examined big data sources were defined. The expected products of WPH were grouped into three categories: statistical indicators, basic research and product supporting existing systems for statistical production. The application and information architecture according to the Big Data Reference Architecture and Layers (BREAL) was created. The quality and metadata framework with the protection of privacy and confidentiality and other legal issues were specified. All results were placed in the three deliverable documents.

WPI Mobile Networks Data

WPI on Mobile Network Data has focused on the incorporation of mobile network data into the production of official statistics in multiple statistical domains, with the concrete goal of developing and substantiating the so-called ESS Reference Methodological Framework proposed by Eurostat. Complementary, this workpackage has faced the stringent conditions to access this data source even for research. Reduced, limited, and intermittent access has been achieved only in some countries for this project. In this context, a modular approach to carry on the research has been followed, where, in

the one hand, efforts have been concentrated on the issues about the access retrieving direct information from the collaborating Mobile Network Operators (MNOs) about their viewpoints and also analyzing and proposing potential future collaboration scenarios. On the other hand, a modular end-to-end statistical process from the raw telco data to final statistical outputs (present population counts and origin-destination matrices in our selected business case) has been proposed and illustrated with the subsequent development of statistical methods and software tools. The statistical production architecture model (BREAL) proposed by WPF has been used to illustrate its usage with this data source and a glossary of terms has been produced to ease the communication between MNOs and the statistical community. To provide empirical illustrations, since access to data is blocked or severely restricted, WPI has developed a network event data simulator, which is a highly modular and configurable software providing synthetic telco data allowing us to test methods and software tools. These first steps will allow the community to develop and test proposals analyzing a range of details against a synthetic ground truth. The output of this WP has been articulated into 8 deliverables dealing respectively with the access, the simulator, the statistical methodology, the software tools, the glossary of terms, the quality and use of BREAL, the application to real data, and the visualization tools. Source code is openly shared in a public repository.

WPJ Innovative Tourism Statistics

The result of the inventory work carried out in individual partner countries in the initial stage of the WPJ pilot project is a list of tourism data sources (including big data). In this way, a list of 130 sources was prepared, of which 57.7% were sources managed by external entities. At the time of the project implementation, 52% of them were available, the remaining 48% were sources to which the partners had no access at all or the access was limited.

On the basis of the inventory, Flow Models were developed for each partner country presenting the directions of data merging. These compilations were the starting point for the development of a tool (visNetwork) for interactive graphical representation of relations and interrelationships between the above-mentioned sources, variables, domains, and countries and results of experimental statistics.

Another result of the WPJ is the creation of a new solution (Visual Modeler) to aid in the process of retrieving data from any websites using web scraping methods. This tool can also be used for data processing operations (analysis, calculations, etc.). It is one of the elements of the prepared Tourism Integration and Monitoring System (TIMS) prototype.

The TIMS prototype aims to support the process of creating statistics (including experimental ones) and integrating data sources identified in the field of tourism statistics in statistical databases (statistical research), data provided by external entities and data obtained from big data sources (e.g. web scraping websites, social media, etc.).

As part of the methodological work carried out by WPJ, a description of the methods of combining tourism data (mainly data on tourist accommodation establishments) was prepared in order to integrate statistical databases with data from web scraping of portals offering accommodation services and to improve the completeness, timeliness and consistency of the survey frame for tourist accommodation establishments. As a result, in all partner countries in total 3.369 new accommodation establishments not included in the statistical databases were identified, which constituted 6.7% of the total number of accommodation establishments in the survey frames.

On the other hand, the results obtained using the implemented data disaggregation methods significantly improved the quality of data on the capacity and use of tourist accommodation establishments.

The work of WPJ also produced flash estimates that responded to delays in publishing data on occupancy at tourist accommodation establishments. As part of the project work, the accuracy of flash estimates was also analysed in the face of the imbalance in the tourism market caused by the COVID-19 pandemic. Data sources from portals like Hotels.com and Google Trends have shown substantial potential and explanatory power. Therefore, taking into account all their advantages and disadvantages, modelling based on them could be further developed.

The final results of the WPJ work are the use of new data sources and the developed methodology to verify the estimates of the volume of tourist traffic by destination, means of transport and types of accommodation, to verify the estimated amount of travel expenses to improve the quality of the balance of payments and the tourism satellite account (TSA).

The aim of the work on the TSA in the context of the project was primarily to improve the quality and completeness of the currently collected data, as well as to identify and measure new, hitherto unknown phenomena in tourism, with a view to reduce the burden on respondents. The time constraints of the project allowed for the preparation of new estimates only for the sums relevant for determining the tourism demand in the TSA. Nevertheless, the proposed method of collecting data from websites offering tourist accommodation of second homes or passenger cars without a driver or the sale of conference services will allow to minimize the gaps between tourism statistics and national accounts data currently arising from the use of sample survey data. It will also make it possible to more accurately estimate the value of services purchased by tourists in the form of packages (accommodation, catering, transport, leisure, etc.), which according to the TSA methodology must be disaggregated into the corresponding items of tourism expenditure.

WPK Methodology and Quality

WPK consolidated the knowledge gained in this ESSnet in the area of methodology and quality for the usage of big data in the statistical production process and combine it with the insights from the previous ESSnet.

For quality the main outputs are quality guidelines and a quality report template. A report describing the methodological steps of using big data in official statistics is the main output in this work strand. A typification matrix was developed together with an evolution roadmap for big data projects. Finally, a literature overview has been prepared to list the important literature inputs to this workpackage and to the project as a whole.

WPL Preparing Smart Statistics

From November 2018 to October 2019 WPL examined the extended use of the Internet of Things (IoT) in order to prepare the ground for producing trusted smart statistics for the European Statistical System (ESS). The goal was to provide an overview of relevant topics for official statistics and to highlight topics that are promising and could be analysed further:

1. Smart Farming: The goal of this task was to give an overview of the data landscape, such as feeding machines, field robots and agricultural apps. Data produced by digital farming technologies are highly interesting for the agricultural statistical surveys. Due to a low standardization of interfaces and a low adoption rate of digital farming, the data might rather be a good source in the long term.
2. Smart Cities: On the one hand, the data generated by 12 exemplary Smart Cities was researched. On the other hand, two case studies explored new technological solutions and open data of smart cities. The data produced by smart cities is highly relevant for official statistics, e.g. sensor data, but the availability and access is generally rather poor at the moment. Nevertheless, in two case studies up-to-date and small-scale data about air pollution, traffic and noise as well as technological solutions such as Blockchains were successfully tested for official statistics.
3. Smart Devices: In this task, a comprehensive list of different types of smart devices that are categorized as follows was provided and analysed: smartphones, smart home devices, smart devices for health and fitness, smart devices for mobility, smart devices for travel and other smart devices. Public sector device data and community-generated device data are considered the most promising and recommended for further testing.
4. Smart Traffic: The use of inductive loop sensors for economic estimates and the possible improvement of transport statistics by using e.g. GPS-data or built-in sensor data was examined in this task. The use of traffic loop data for early economic estimates and of Remote Freight Management Systems for transport statistics has not only been proved useful, but is regularly published as experimental data already.

1. Introduction

1.1 Background

This is deliverable A5, the Final Technical Report, of the grant agreement of the ESSnet Big Data II. As indicated by its name, this ESSnet is the follow-up of the ESSnet Big Data I, which ended in May 2018. This report covers the whole period of the ESSnet Big Data II, that is, the period from November 2018 to June 2021. Originally, the agreement covered the period from November 2018 to December 2020, but the COVID-19 pandemic made an extension till June 2021 necessary, and an amendment was duly signed (see section 3.1 of this report). When reference is made in this report to the grant agreement, this is to be understood as the amended agreement.

The report describes the results of the action; it does not list the activities carried out by the workpackages that yielded those results, as that will be done in the Final Report on the Implementation of the Action, due 60 days following the closing date of the action. For the same reason, this report does not comprise an evaluation of the budget needed and used.

The consortium of the ESSnet consists of 28 partners, comprising 23 National Statistical Institutes (NSIs) and five other National Authorities. The overall objective of the ESSnet is to prepare the ESS for integration of big data sources into the production of official statistics. The consortium has organised its work in twelve workpackages (WPs), numbered A through L:

WPA	Coordination and Communication
WPB	Online Job Vacancies
WPC	Enterprise Characteristics
WPD	Smart Energy
WPE	Tracking Ships
WPF	Process and Architecture
WPG	Financial Transactions Data
WPH	Earth Observation
WPI	Mobile Networks Data
WPJ	Innovative Tourism Statistics
WPK	Methodology and Quality
WPL	Preparing Smart Statistics

The grant agreement specifies the agreed outputs of the workpackages, and its inputs, both in terms of number of days contributed by partner and workpackage and in terms of material costs. The total budget available from the grant is 2,465 million euro, but only 90% of costs, as a maximum, will be reimbursed. For each workpackage, Annex II of the grant agreement provides a description, specifying, among other things, the tasks to be carried out, the milestones and deliverables, and the number of days each partner contributes to the workpackage. A specification of the budget is given in Annex III of the agreement.

The remainder of this chapter describes the set-up of the workpackages (section 1.2) and the way the ESSnet has organised itself (section 1.3). The next chapter presents the results obtained for each of the eleven workpackages other than WPA, since coordination is covered in this introduction and administrative aspects will be elaborated in the Final Report on the Implementation of the Action. The communication and implementation results of WPA are described in Annex 1. The third chapter

describes the issues encountered in the action, starting with the impact of the pandemic on the ESSnet and other general issues and continuing with issues specific to the workpackages. At the end of the action a simple evaluation was held, the results of which are summarized in Annex II.

1.2 Approach to the workpackages

Apart from WPA, the workpackages listed in the previous section belong to three different strands or tracks:

Implementation projects:	WPB to WPF
New pilot projects:	WPG to WPK
Preparing for future trusted smart statistics:	WPL

The first track aims at the implementation of results from the ESSnet Big Data I concerning the use of web scraping aimed at online job vacancies and enterprise characteristics (WPB and WPC, respectively), the measurement of electricity consumption as registered by smart meters, identifying energy consumption patterns (WPD), and maritime and inland waterways statistics (WPE), based on AIS data (the Automatic Identification System for ships). Cross-cutting implementation issues are covered by WPF, which focuses on the design of the statistical process, including IT and architecture.

The second track adds new research areas to what has been investigated during the ESSnet Big Data I. Four domains were targeted: the potential of using financial transactions data for statistics (WPG), the application of earth observation (remote sensing) data to statistics (WPH), constructing a statistics production framework based on mobile phone data (WPI), which builds on research done during the previous ESSnet Big Data, and constructing a tourism information system using various sources (WPJ), also building on previous results. For the second track, cross-cutting issues are dealt with by WPK, which focuses on the methodological and quality questions that arise when using big data in the statistical production process.

Although WPF and WPK are assigned to different tracks, they provide frameworks that apply to all workpackages. Moreover, the architectural, methodological and quality framework form the starting point for a coherent approach to future statistics that are based on new data sources. These future statistics have been coined trusted smart statistics. Potential research areas for such trusted smart statistics were investigated in the first half of the ESSnet Big Data II by WPL.

At a practical level, this approach required to be facilitated in several respects. First of all, an organisational approach was needed to ensure that the agreed output would be produced with the resources foreseen. This is the subject of the next section, 1.3. Facilities were also needed for communication, in order to share and work on documents together and for virtual meetings, among other things. This is described in Annex I to this report.

1.3 Organisation

The organisation of the ESSnet has been carried out as foreseen in the grant agreement, with some adjustment due to the pandemic (see section 3.1). Each workpackage has a workpackage leader who

is in charge of organising the realisation of the milestones and deliverables of the workpackage. This includes the organisation of virtual and physical meetings of the workpackage members. The results of the workpackages are described in chapter 2.

At the level of the ESSnet as a whole, the main instrument for co-ordination is the bimonthly virtual meeting of the workpackage leaders. In addition, there are separate bimonthly virtual meetings of the workpackage leaders of the implementation projects and the new pilot projects, respectively. In total there have been 34 of these so-called co-ordination group (CG) meetings. The aim of the virtual CG meetings was to manage the project and stay in control of its realisation.

In addition, there have been two physical kick-off meetings for the combined workpackages of the implementation and pilots tracks, respectively, in December 2018 in Vienna, and two physical intermediate meetings for these tracks as well, in December 2019, also in Vienna.

For most full CG meetings the partners were asked to provide information on the realisation of the foreseen budget in the form of a spread sheet, which was consolidated by the secretariat (Martin van Sebille, WPA), thereby enabling the CG to link the progress in producing results to the resources actually spent. The meetings were also used, of course, to discuss cross-cutting issues. In addition to the workpackage leaders, the virtual CG meetings (full CG as well as track meetings) were also attended by the Eurostat project manager of the ESSnet, Albrecht Wirthmann.

In order to ensure the quality of the deliverables of the ESSnet, a Review Board was created at the beginning of the project, as was the case with the previous ESSnet Big Data. The members are Lilli Japec (chair, Statistics Sweden), Giulio Barcaroli (Italian expert), Florian Keusch (University of Mannheim) and Boriska Tóth (Statistics Norway). All workpackage leaders have arranged that their deliverables were reviewed by the Review Board. This has worked well: the reviews are considered to be of very high use and the comments have been taken into account in the final version of the deliverables. The members of the Review Board are also invited to the CG meetings.

The work of the ESSnet was concluded with two virtual events: One six-day dissemination event for a broad audience in November 2020, in which the results of the ESSnet were presented and discussed, and one three-day future-oriented event for a more limited audience in June 2021, in which eight themes related to the work of the ESSnet were discussed in groups, resulting in ideas to bring the use of big data for official statistics in the ESS to the next level of maturity.

The organisational arrangements of the ESSnet are considered to have been quite adequate. At the end of the project a simple internal evaluation was carried out, and feedback was collected from the two final events on dissemination and suggestions for the future, respectively. The conclusions were generally positive. They are summarised in Annex II.

2. Results of the workpackages

2.1 Online Job Vacancies *(full information on WPB available here on CROS)*

The aims of implementing WPB on Online job vacancies were to produce statistical estimates on the topic of online job adverts and to identify statistical production processes and capabilities that may be affected at the national level and to define the conceptual production processes at the national level and at the level of the ESS. Suitable techniques and specific methodologies were developed during the pilot phase of the project ESSnet Big Data 2016-2018 on Web Scraping for Job Vacancy Statistics. The implementation phase was based on work carried out to establish the conditions under which web scraping techniques can be used and to evaluate the quality of the scraped data.

1. Main findings

Using online job adverts (OJA) has a number of advantages. They can be used successfully for the needs of official statistics for completing and estimating a number of characteristics of the job vacancies. Their main advantages are their accessibility, timeliness, comparatively low cost of producing statistics and lack of or minimal burden for the respondents. Most OJA contain information not only for the vacant position but also for the required qualifications and skills, including the so-called soft skills, for degree and kind of the graduate education and professional experience necessary for the potential applicants for taking it. Often they also contain information on the employer (sector of the economy, size of the enterprise, kind of ownership, etc.). Estimates could be made for labour demand by economy sectors, professions, specific skills and qualifications, and regions. A comparison could also be made for which job vacancies similar education, qualification and specific skills are required. However, using OJA as an alternative information source for statistical surveys is always accompanied by challenges.

A good understanding of the “business models” underlying the development of job portals, as well as of the market of OJA is crucial. Understanding of their mechanisms is a key issue to understand the data collected and processed by CEDEFOP system and identify the best fit model to use them for statistical purposes, direct on JVS or in correlation with other socio-economic indicators. Beware of that simple aspect, the Online job portals have not been developed to produce statistics or indicators of the job market. Their goal is different from a portal to portal, some of them are focused on human resource aspects, in finding the right person for a specific job, but other portals are focused only in making money from promoted adverts on their web pages.

The quality issues are such that it is not clear if these data could be integrated in a way that would enable them to meet the standards expected of official statistics. On the other hand, OJA data can provide many insights that official estimates cannot. Statistics based on OJA data can be published in a very timely and frequent manner, allowing for short-term tracking of labour market conditions and flash-estimates of labour demand. Because of the large data pool available, OJA data also allow for more granular analysis by subgroups or geographical regions. Additionally, OJA data might allow for the provision of completely new (to official statistics) labour market insights: for instance, indicators of labour market power by employers. There is, however, further need to address the challenges how OJA data should be interpreted and used together with official estimates.

One area that shows some promise is to use the time series properties of OJA data to improve existing statistics. The analysis showed modest success in predicting job vacancy survey values using OJA data, so these data could be used for producing flash estimates of labour demand. It may also be possible to use these time series properties to produce more frequent estimates, or even reduce the frequency of the survey. An important limiting factor of current CEDEFOP data is that this system holds data from 2018 onwards, so it will take several years at least to collect a reasonable time series.

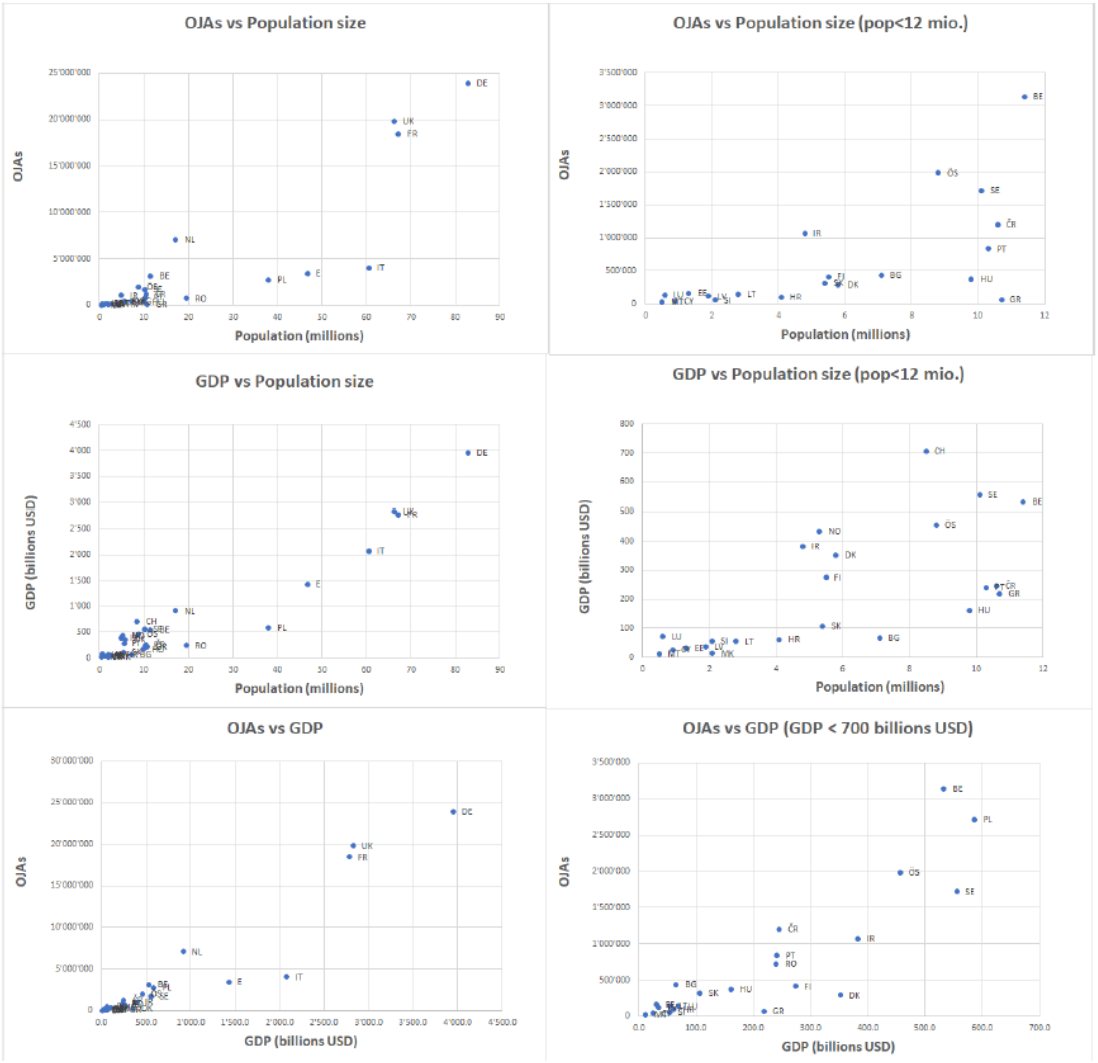


Figure 1: Scatterplots of relationships between OJA population size and GDP for V2 CEDEFOP datasets. Left: all countries; right : focus on smaller countries with populations size < 12 mio. or GDP < 700 billions USD)

CEDEFOP data demonstrate a strong relationship between OJAs and socio-economic variables such as national and regional population sizes and GDPs. These relationships can be modelled with promising outputs. Such models may potentially be used for predictive purposes. For instance, since OJAs are potentially instantly available, they might offer a data source for developing flash estimates of national and regional GDPs. The models and choice of variables should of course be studied in more detail and improved. The models also indicate that they offer the potential of predicting OJAs for countries without data for OJAs on the basis of their socio-economic variables. Such cases without CEDFOP OJAs data offer an interesting opportunity to test the models.

Integrating OJA in the production of JVS and complementing JVS by OJA may be major goals of collecting OJA in future years. Such an approach would significantly alleviate the survey burden on enterprises and would have important consequences on survey designs, with classical sampling surveys being replaced by censuses.

2. OJA use cases and experimental statistics

During the project execution several use cases on using OJA for producing statistical outputs of mainly experimental nature has been developed. These are not ready to be used as tools for decision making, and rather serve as proof-of-concepts illustrating some of the paths that are possible to follow in the future for creating statistical outputs based on OJA data.

Germany proposed and described the calculation of a labour market concentration index (LMC) based on OJAs. While market power due to monopolies and monopsonies in product markets is well documented and regulated by antitrust agencies, monopsony power and market concentration in labour markets is rarely in the spotlight. In part because empirical evidence of labour market concentration has historically been scarce due to lack of suitable data. In recent years however, public interest in labour market monopsony power has increased and several studies have shown that labour market concentration is widespread and has significant consequences. Concentrated labour markets allow employers to offer wages below worker's marginal productivity depress the wage distribution and generate rents just like in concentrated product markets. OJAs are uniquely suited for detecting and monitoring labour market concentration. Ideally, we would observe the universe of OJAs, in order to quantify market shares of all employers in each local labour market. Online OJA data collected by CEDEFOP comes reasonably close to this ideal. CEDEFOP OJA data also contains the necessary information about the location, the company and the occupation of the job. This data allows us to quantify the share of job offers for each employer by occupation and timeframe in local labour markets. From this data, one can calculate measures of market concentration like market shares or the Herfindahl-Hirschman-Index. Such indices of market concentration can be evaluated by the criteria used by antitrust agencies in order to classify markets as critically concentrated. LMC index can be published each quarter and can inform educational, regional and labour market policy. To measure LMC by regions, OJAs are broken down to Functional Urban Areas (FUA). These FUAs consists of city cores and their commuting zones. The LMC index can be published quarterly or yearly and can inform educational, regional and labour market policy.

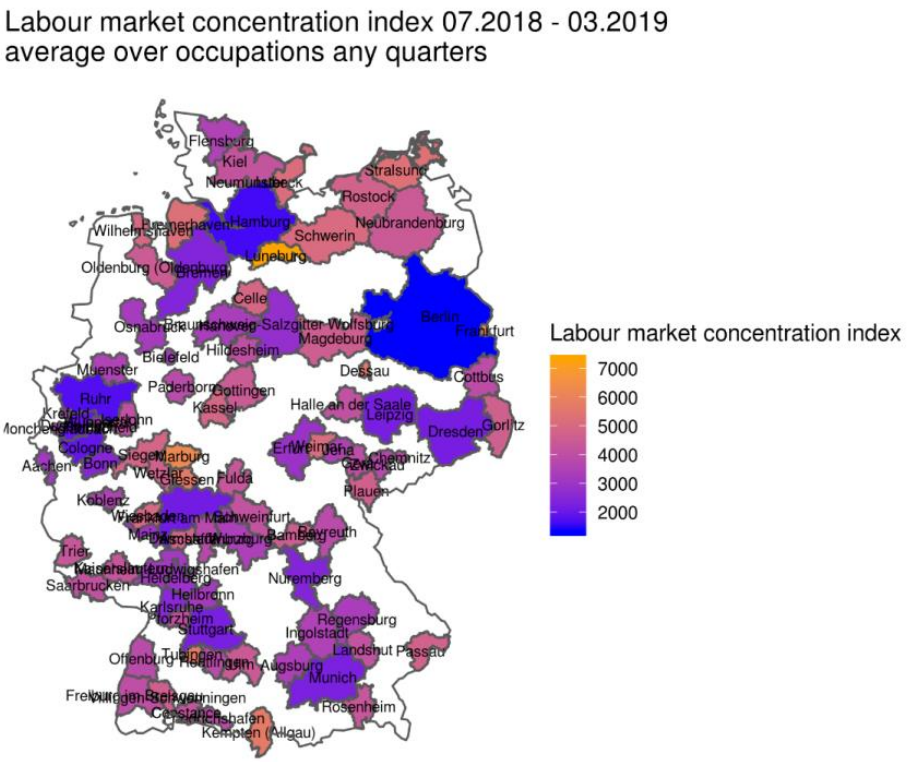


Figure 2: LMC index 072018 - 032019 average over occupations and quarters

ISTAT considered OJAs in several ways to complement the official JVS, in particular as a complementary source:

- as a new set of variables in business registers. Once JVs included in business registers, it would be possible to design surveys more specifically tuned to job vacancies, taking into account the actual distribution of JVs among enterprise sizes and possibly to reduce the burden on enterprises. This would of course imply a full coverage program of all enterprises
- to fill non-response units and for imputation purposes in general
- as auxiliary variables during JVS grossing up processes

The choice of websites for recording JVs and the quality of the linkage between OJAs and register units are a key issue in this context. The analysis carried out so far have shown some of the potential enrichments that OJAs could offer for the official job vacancy indicators, as required by EU Regulations.

Despite their coverage and representativeness limits, OJAs have the main advantage of providing information on the job positions announced, not only high detailed in terms of the characteristics of job positions, but also at a high frequency level, providing an information on a daily, weekly and monthly basis. Therefore, in addition to offer a potential enrichment for job vacancy official indicator from an informative point of view, OJAs can support the current production of JV official indicator from a methodological side.

Among the several ways in which OJAs can be used to complement the current official JVS production, three other uses are considered:

- to derive evidence on the vacancy flows during the reference quarters of the official surveys;

- to evaluate the representativeness of the vacancy stocks at the specific reference dates used in the official surveys;
- to indirectly calculate a monthly basis JV official indicators from the quarterly ones.

French NSI's Labour Department (Dares) and French National Employment Agency (NEA - Pôle emploi) has been working on a new version of the French labour market tightness indicator. Among other changes, a new tightness indicator includes online job advertisements. The initial purpose of including OJAs was to provide a more robust tightness indicator with a better coverage of job offers. It appears that most impacted groups are either groups with highly qualified occupations or, on the contrary, relatively lowly qualified occupations. Occupational tightness distribution broken down by qualification corroborates this finding. Overall, labour tightness in upper qualified occupations increases when adding OJAs (+0.28), whereas lower qualified occupations' tightness decreases (-0.25).

Slovenia used OJAs to estimate two statistics related to the national accounts field: the VAT from the turnover from the sale of services and the index of the turnover from the sale of services. The data does not yet include enough periods, to provide a reliable model, at least one more year of scraping would be needed. OJA data seems to lag too much to be useful for forecasting or nowcasting the chosen economic indicators, even with its timeliness. Even with these issues, however, it's reasonable to assume that both variables are linked. This implies a level of connectivity between OJA data and other economic indicators as well. In the future more analyses could determine a better and more advantageous use of OJA data for estimations in other fields. Furthermore, given the timeliness of OJA data, early estimations could be produced as early as the first week of every month, and subsequent estimations would improve the early values, giving us a dense time series of early estimations of diverse important indicators.

The Bulgaria use-case on web scraping of OJAs for producing experimental statistical started in the beginning of 2019 and the BNSI still continue to work on it, even after the end of the ESSnet project. The ambitions were to improve initial findings and to produce statistical indicators as such that can serve to enrich the job vacancy statistics or to measure the on-line labour by demand side. The OJAs indicators do not reflect the total number of job advertisements on the Bulgarian labor market as it does not include jobs advertised through other channels – enterprises websites, social media, newspapers or through informal communications such as word of mouth. In addition, it is possible that some vacant job positions are not advertised at all. The following experimental indicators for OJAs by weeks/months/quarters for the period were produced:

- Number of OJAs and change (flow)
- Number of OJAs and change by Educational_level (flow)
- Number of OJAs and change by Full_and_part_time_work (flow)
- Number of OJAs and change by Permanent and temporary job (flow)
- Number of OJAs and change by NACE 2.0 Level 1 (flow)
- Number of OJAs by NUTS 3 (flow)
- Number of OJAs by average salary (flow)

Bulgaria is doing OJAs experimental statistics in four main steps, which are performed in daily base and then the compilation is done for weekly, monthly and quarterly basis.

The results from OJAs data raise some important questions about the OJAs data and whether and how it should be used for official statistics. While results has shown that it is feasible to produce

experimental estimates of OJAs, there is still a conceptual difference between them and the official JV estimates. It is, therefore, clear that experimental OJAs data could not replace the official statistical indicators for JV. Unfortunately, OJAs experimental data give only a partial "picture" of overall labour market demand. It therefore seems that the role of OJAs data within official statistics is more likely to be as the basis for producing supplementary indicators. These could include indicators of local labour market demand and/or indicators of occupations and skills. It would be better not to measure only absolute numbers of OJAs, but also their change over time.

Sweden has explored possibilities to use online job vacancy data from different sources (including CEDEFOP) and came to a conclusion that, at this moment, the most reliable and the most useful source is the Swedish Employment Service (SES). In the short term, SCB has an intention to publish an experimental index based on the online job advertisements on monthly basis. This index is neither a replacement nor a monthly flash estimate for corresponding quarterly estimates but more like a complement in order to get a quick insight into the job market fluctuations. The index reveals a sharp downward movement from April 2020 to May 2020 with modest recovery in June 2020 due to Corona crisis. This kind of movement was impossible to identify with quarterly survey data so clearly.

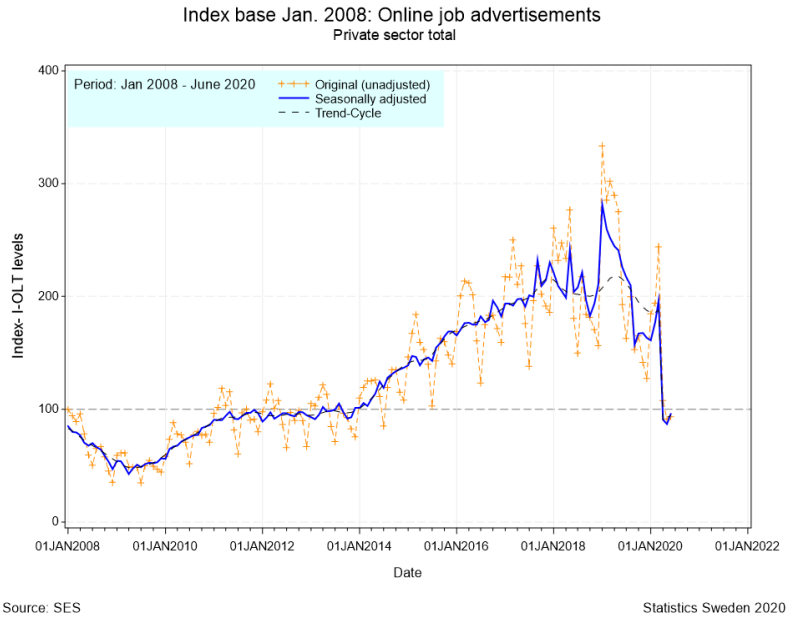


Figure 3: SCB:s Experimental Index Online Job Vacancies (I-OLT): Seasonally adjusted, trend- and unadjusted index values (Jan 2008= 100) ; Data from Jan 2008 to June 2020

3. Deliverables

Five deliverables were produced.

Deliverable B1: Methodological framework for processing online job adverts data for Official Statistics V.1.

Deliverable B2: Interim technical report

Deliverable B3: Methodological framework for processing online job adverts data for Official Statistics V.2.

Deliverable B4: Report on the statistical output, required quality and definition of the necessary metadata at European and national level

Deliverable B5: Technical report on the implementation requirements of prototypes in the relevant statistical production processes at European and national level

Interim technical report summarises the progress made during the first 11 months of the project (November 2018 to October 2019). It describes the main objectives and discusses work done to date and the issues identified.

Methodological framework for processing online job adverts data for Official Statistics contains three parts. The first part, traditional labour market indicators and concepts, describes traditional labour market concepts. The second part, methodological considerations around OJAs, deals with quantitative and qualitative description and analysis of the web data source and provides a few probabilistic approaches on how to integrate and link the web data source with in-house data, i.e. develop new statistics around web data, which can complement statistics from thejobs survey. The last part, architectural considerations around OJAs, contains a thorough description on a possible sustainable business process capable on integrating web data into official statisticsproduction systems.

Report on the statistical output, required quality and definition of the necessary metadata at European and national level contains three parts. The first part, methodological aspects related to OJA data, describes the methodological aspects and main findings of CEDEFOP data analysis, at the conceptual and practical level, regarding the use of data produced by CEDEFOP as an input for producing statistical indicators in the area of job vacancies. It addresses key conceptual, methodological, and technical aspects in using CEDEFOP data in OJA statistics. In the second part, use cases of using OJA data in official statistics, a few examples of using OJA for producing statistical outputs of mainly experimental nature are described. These are not ready to be used as tools for decision making, and rather serve as proof-of-concepts illustrating some of the paths that are possible to follow in the future for creating statistical outputs based on OJA data. The last part, conclusions and recommendations for future developments, describes major goals of collecting OJA in future years to alleviate the survey burden on enterprises and would have important consequences on survey designs, with classical sampling surveys being replaced by censuses. This section propose tentative sketches of integrating OJA in the process of JVS production, identify necessary requirements needed to assure standards of quality and identify pitfalls to avoid and issues to address.

ESS Standard for Quality Reports Structure quality template (ESQRS 2.0) for OJAs was developed. The template is derived from SIMS 2.0 for big data sources and comprises only concepts for the respective quality aspects for the experimental statistics on different big data sources in the case for OJAs experimental statistics. The definitions and guidelines are based on the recently published and updated version of the EHQMR (ESS handbook for quality and metadata reports). The new subconcepts for Big data are indicated by an "A" for "additional". The main idea is that each country, which is going to produce and publish experimental statistics for OJAs, fill in this ESQRS template, and disseminate together with experimental statistics.

Technical report on the implementation requirements of prototypes in the relevant statistical production processes at European and national level describes developed prototypes for collecting, processing and analysisof OJA data. The first part, OJA implementation requirements, outlines the architecture and processes for supporting OJA data integration and production of statistical outputs from OJA following Big Data Reference Architecture and Layers (BREAL) standard. The second part, prototypes and methods for processing OJA data, describes developed prototypes and their implementation requirements. Developed prototypes and their implementation requirements according to defined Application services can be found on the GitHub https://github.com/OnlineJobVacanciesESSnetBigData.

4. Data lab infrastructure

During the kick-off meeting in Vienna in December 2018 WPB participants and representatives from CEDEFOP and Eurostat participated in the discussion disussing the opportunities and interests of using the CEDEFOP data. WPB partners expressed interest in using CEDEFOP data. In April 2019 Big Data Task Force confirmed and proposed further exploration of potentials of the CEDEFOP data for producing official statisitcs. In the following months CEDEFOP, Eurostat and the WPB project coordinator tried to find a suitable solution for enabling access to and analysis of the CEDEFOP data. After several virtual meetings between CEDEFOP, Eurostat, ESSNet and CRISP it was decided that clean (pre-processed and classified) data will be available for the ESSNet using a Big Data Test Infrastructure (BDTI). The individual data processed by the CEDEFOP pan-EU system for collecting and analysing online job vacancies from 28 EU Member States were considered a basis for the work. CEDEFOP has developed first Data Lab allowing WPB partners to navigate and work with the data. In March 2020 Eurostat set up second data lab, EU Dataplatform, using Commision infrastructure and data provided by CEDEFOP.

5. Conclusions

Integrating OJA in the production of JVS and complementing JVS by OJA may be major goals of collecting OJA in the future. Such an approach would significantly alleviate the survey burden on enterprises and would have important consequences on survey designs, with classical sampling surveys being replaced by censuses.

Secure long-term data acquisition is a crucial issue, since public statistics must offer grounds on which to compare changes over time using comparable data. Since Job portals are usually privately developed and are not designed to produce reliable statistics, they are therefore susceptible to significant changes. One must therefore secure a long-term availability of the data, as well as control quality standards over time. Data matching between web data and statistical units is challenging issue to solve. Only a good matching will enable to move from classical data collection to online information in the production of JVS.

OJA comes from various job portals and enterprises' websites. Different portals often code professions with different coding systems. Therefore, it is important to use an occupation classifier to deduce for each job advertisement the standard occupation code that is necessary to calculate job vacancies by occupations. OJAs titles and offers' descriptions (if available) are often used to classify occupation.

Titles usually give a concise description of the jobs which makes them computationally less expensive to use than full description. However, sometimes titles may not contain enough information to build a satisfying classifier. Thus, the overall challenge is to find optimal algorithms considering available training datasets, which characteristics (such as size and features) may change from one country to another. Apart from a satisfying accuracy, such algorithms must also be not computationally costing and should be easy to maintain.

The CEDEFOP data collected as the European scale seem to offer a promising basis for developing pertinent statistics and economic indicators. A preliminary analysis shows that should be explored existing relationship between OJAs, national and regional population sizes, and national economic activity expressed as GDP. It is important to have in mind the significant differences between ESS countries in terms of the OJAs landscape. Thus, what is feasible in one country may not necessarily be reproducible in others.

2.2 Enterprise Characteristics *(full information on WPC available here on CROS)*

Work Package C, on Online Based Enterprise Characteristic (OBEC) is about understanding economic and business activity online from a national statistics perspective. Online business grows in importance every year, and this year, with a global COVID pandemic is more important than ever.

Statistical business register (SBR) enhancements with details on nationally listed businesses' online presence, such as websites or social media accounts, are a key output. This simple innovation is relevant to all ESS members, as every NSI maintains an SBR. It is powerful because SBRs are already linked to datasets underlying economic and business, meaning that integration is automatic, immediate, and seamless.

WPC team has produced not just the experimental OBEC statistics, but also the Reference methodological framework and software Starter kit, so that NSIs can move ahead quickly with producing OBEC statistics within their own national contexts. As web scraping is a relatively new data sourcing method for NSIs, which requires due attention to data protection, we've also published an ESS webscraping Policy Template covering the key legal and ethical considerations and setting out robust a set of principles and practices for NSIs to follow.

The main aim of WPC to use web scraping, text mining and inference techniques for collecting and processing enterprise information, in order to improve or update existing information, such as Internet presence, kind of activity, address information, ownership structure, etc., in the national statistical business registers was achieved. Within the WPC, the methodology of the previous ESSnet on BD I project (SGA-I and SGA-II) was generalized and extended for use in any ESS country, taking into account the variety needed to support different use cases.

The following sub-sections describe the all obtained results as reported in the WPC deliverables (see link on CROS portal:
https://ec.europa.eu/eurostat/cros/content/WPC_Milestones_and_deliverables_en).

1. Reference methodological framework and Reference architectures for OBEC data

Five use cases are defined in the ESSnet Big Data II WPC:

1. *URLs (Uniform Resource Locator) Inventory.* The URLs Inventory represents a database containing URLs for each enterprise in the target population, where there can be one, many or zero URLs for a given enterprise.

2. *Variables in the ICT usage in enterprise survey.* Use case 2 focuses on the extraction of information from an enterprise URL equivalent to variables from the ICT usage in enterprises survey.

3. *Data driven discovery of emergent enterprise classifications*.

4. *Experimental language statistics*. Both use cases 3 and 4 deal with classification of enterprise URLs based on the usage of language and text displayed in the URLs.

5. *Amazon seller data*. The goal of this use case is to scrape information about all Amazon sellers within one of Amazon's regional websites (e.g. amazon.de).

The first and second use case are set within the context of the ICT survey and can be implemented in the regular statistical production whereas use cases 3, 4 and 5 have more of an experimental character, not necessarily leading to new statistics.

The functional production prototypes have been produced by use cases. In the prototypes, the methods used, software used, procedure adopted and lessons learned have been described in details.

The process of web scraping for detecting enterprise characteristics naturally fits into the more general process of the big data processing life cycle. This generalized process provides a starting point for any ESS country that would like to derive online-based enterprise characteristics in their national context. To facilitate applicability in any ESS country, the WPC team sketched a high-level view on the process of determining enterprise characteristics as generically as possible (see Figure 1):

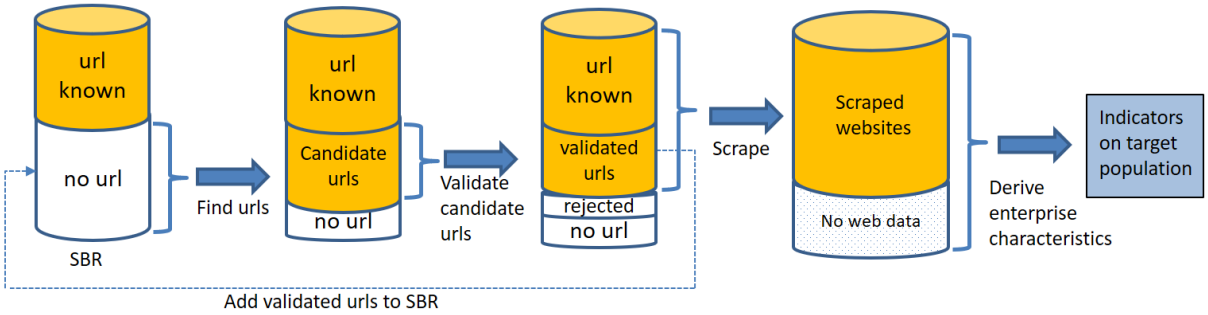


Figure 1: High level view on enterprise characteristics web scraping process

The phases of the process of determining enterprise characteristics are as follows: URL retrieval for businesses without assigned URLs; Validation of candidate URLs based on comparing enterprise known SBR data with data from search results; Scraping phase; Derive enterprise characteristics for the target population and Updating the SBR with validated URLs to the SBR. The most important process decisions to be taken in these phases, have been expressed in terms of GSBPM phases recognised in the big data lifecycle: Collect, Process, Analyse and Disseminate.

The description of generic Application solution architecture for OBECs' use cases 1&2: *URLs Inventory of enterprises* and *Variables in the ICT survey* has been done. The description is based on BREAL Application architecture (see Figure 2 and Figure 3).

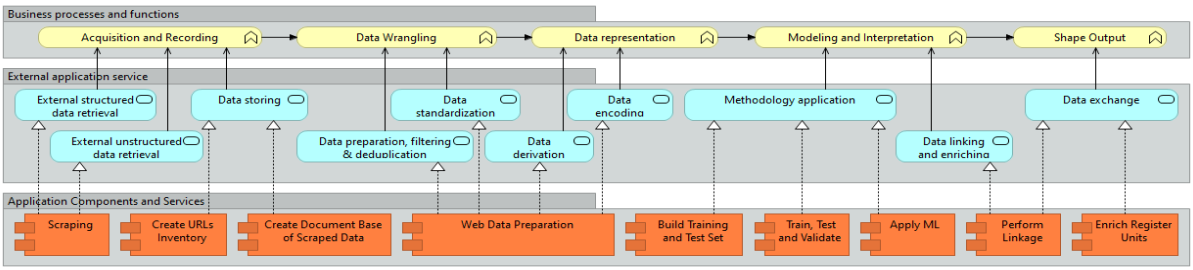


Figure 2: Generic graphical representation of Application architecture of OBECs' Use Case URLs Inventory of enterprises

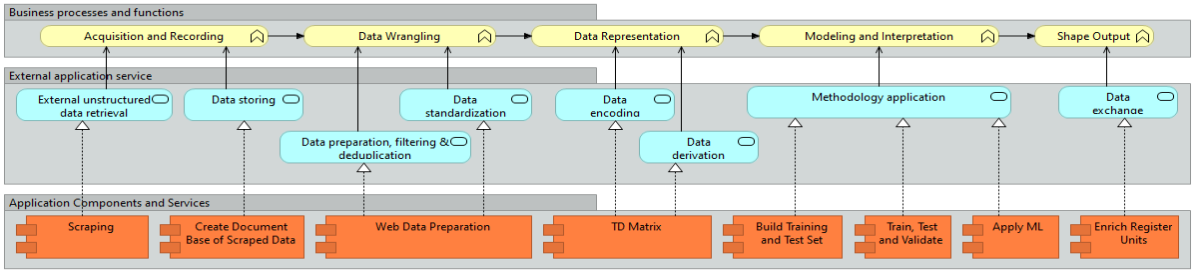


Figure 3 - Generic graphical representation of Application architecture of OBECs' Use Case Variables in the ICT usage in enterprise survey

The Business processes and functions layer describes the OBECs' business functions derived from BREAL model. Every business function triggers the next one. Every business function is served by one or more application services. Every application service is realized by one or more application components. The application components could be programs, modules, scripts, classes or functions. They could be stand-alone or part of a system. They could be written on different programing languages.

The description of generic Information solution architecture for OBECs' for OBECs' use cases 1&2: *URLs Inventory of enterprises* and *Variables in the ICT survey* (Figure 5 and Figure 6, respectively) is also based on BREAL Information architecture.

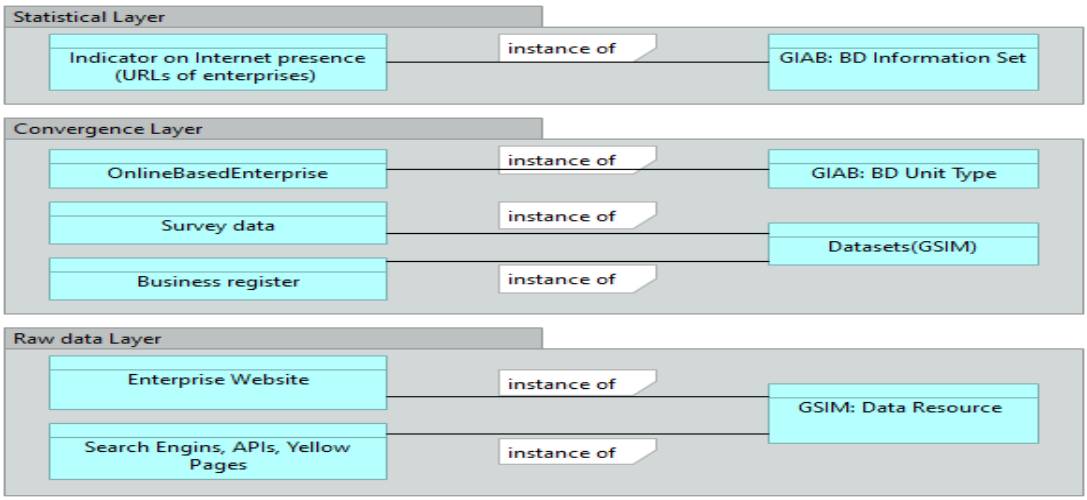


Figure 3 - Information solution architecture for OBECs' Use Cases URLs Inventory of enterprises

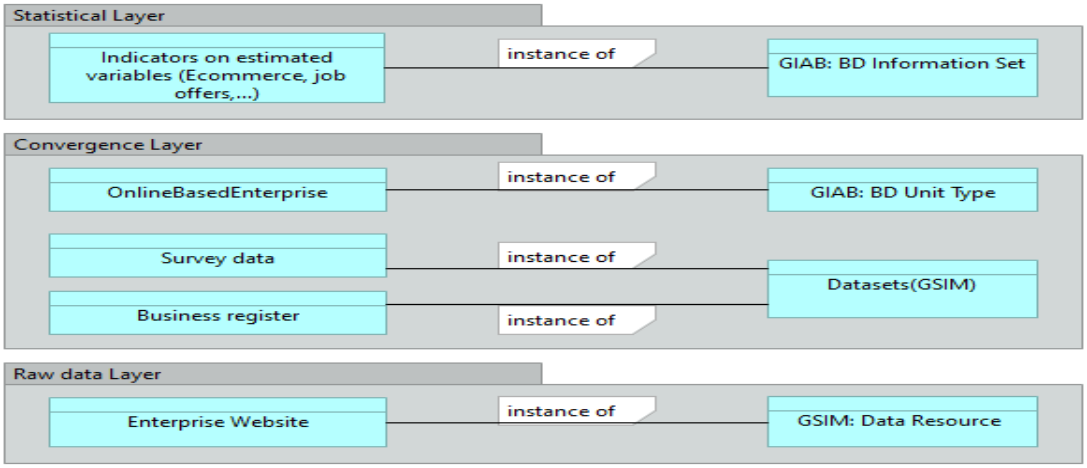


Figure 4 - Information solution architecture for OBECs' Use Cases Variables in the ICT usage in enterprise survey

The Raw data Layer describes the OBECs' initial data sources in terms of BREAL. It covers the Acquisition and Recording process of OBEC. The Convergence Layer describes the information entities derived from source data through Data Wrangling and Data Representation processes of OBEC. The Statistical Layer describes the information entities derived from Convergence Layer information sources trough Modelling and Interpretation and Shape Output processes of OBEC.

Almost all application services for both OBECs' use cases (URLs Inventory of enterprises and Variables in the ICT survey) have been designed to be flexible enough to be possibly interoperable, replicated or shared between NSIs.

For both OBECs' use cases the used data are European available on the Raw Data Layer due to the public nature of data contained in websites but local or European for the Convergence and Statistical Layers because of the presence of data coming from NSIs archives (e.g. Statistical Business register).

During the project all the services and all the data were locally managed, but in the future all the services (exception made for "Data linking and enriching" and "Data exchange" that use local data) could be shared by the NSIs.

Within the project activities, some generalized software solutions were implemented and are available at WPC GitHub.

2. Set of recommendations for NSIs (ESS level)

Based on the functional production prototypes and their testing for use-case 1 URLs Inventory and use-case 2 Variables in ICT survey, the set of recommendations were defined by WPC team to all NSIs within the ESS that would like to produce experimental OBEC data:

- *Local search engine strategy:* Since search engine results play an important role in the OBECs it is good to study the popularity of different – global or local – search engines in your country. Studies may be available from research institutes in your country. Based on that develop a strategy which search engine(s) to use in your OBEC process.

- *Local legislation on enterprise identification on the web:* There may be country-specific legislation describing how enterprises should identify on their 'about' web page. This is important for OBEC processes as one of the basic steps is to link web pages to units in your general business register. This is important for the development of a strategy to use (mapping of) variables from your general business register in this identification step.
- *Local practices for Ecommerce and / or social media*: The OBEC approach has been tested in a number of countries with comparable practices in the field of Ecommerce or social media practices. But even among these countries there are differences. Be aware of local practices on Ecommerce and social media that differ from the ones in the pilot countries. For example, local social media / Ecommerce strategies might use local providers that are popular in a specific country but completely unknown in another country.
- *Language specificities*: If there are multiple languages being used on the web in your country and websites provide access or web pages in multiple languages, you might use this for an even better identification, but also keep in mind that the information might also be contradictory between pages provided in different languages.
- *Other data sources:* Keep an eye on companies in your country that also scrape data. Although the purpose of their activities might have a different goal, it could be that the data they collect is still valuable for a NSIs, as a primary data sources (after careful quality control) or as an alternative data source for comparison.
- *Re-use software:* Within the ESSnet WPC project there have been a number of implementations of the OBEC process. These have been brought together in the starting kit which is available on github. These are open source, so you may either use them out of the box, or – if necessary – adapt them to your own needs.
- *Frequency for URLs Retrieval:* The process of URLs Retrieval needs to be repeated regularly for updating and maintaining the URLs Inventory and to check validity of URLs. The frequency could be at least once a year.
- *Internet stability:* The stability of Internet as a data source for official statistics need also consideration because the Internet policies of search engines may vary from country to country.
- *Internet restrictions:* Internet without any restrictions in accessing external websites (e.g., no firewalls filtering the websites);
- *Quality of webscraping algorithm:* For programming the own webscraping algorithm or adopting an existing one it is highly recommended to build minimal examples containing html codes used for unit testing. The examples should reflect various issues and specifics which are encountered while scraping enterprise homepages and exhaustive tests should be written to check the capabilities of the webscraping algorithm.
- *Main scraping languages:* In general, Python, Java, PHP, Node and R are the main web scraping languages. The experienced gained shows that the most convenient programming languages, used for the URLs Retrieval, depends on the in-house competence.
- *Allowable response time for a websites:* It is also very important to set the maximum allowable response time for a websites, as they may not always return a response in a timely manner.

- *URLs for the websites:* If there is no URL for the websites the first recommended step is to extract the domain from the contact list of the person for selected enterprise and survey, excluding most common e-mail domains, such as gmail.com, hotmail.com, yahoo.com;
- *Big Data environment*: One simple solution for big data environment could be Python and NoSQL database. A database is not necessarily needed for storing the scraped data or indicators derived from scraped data but it is still recommended to use one. In general, the big data environment depends on the in-house preference and competence.
- *Programming skills:* The basic programming skills are required to adopt and change existing solutions or rebuild a scraper using existing software libraries.
- *Webscraping Policy:* Applying the rules from the Policy of Webscraping (e.g. delays between requests etc.).

3. Starter Kit for NSIs

The WPC team has produced the Starter kit for NSIs, intended as an introduction to web scraping for OBEC, to support producers of official statistics with implementing their own web scraping routines. However, the methods and functions in this Starter Kit most likely need to be adapted to the individual needs and particularities in the respective countries. The Starter kit for NSIs covers all the needs required to perform the two main tasks, i.e. the search for the URLs of the websites of the enterprises in the population of interest, and the scraping of the contents of these websites. The third module gives an example of the use that can be made of the scraped content, by selecting from each website the links of the social media contained in it. The notebook format allows to understand how to perform each step in the task, giving indications on the required input and permitting to analyze the output. The Starter Kit has three parts and each part of the Starter Kit has a Jupyter Notebook that serves as a manual on how to use the functions and methods. These manuals are intended for statisticians with little to no programming background. The source code can be consulted by users with a background in programming. All other details and explanations could be find when start to reading the Starter Kit for NSIs, V.2.

4. Experimental statistics and reference metadata

WPC produced experimental OBEC statistics for 2019 and 2020. All experimental results for OBEC are available on the CROS portal, experimental statistics section together with short methodological notes and reference metadata (quality reports in ESQRS format (as deliverable C5) and ESMS metadata mainly intended for users). The methodological notes are reflect the process and methods used for production of OBEC experimental data.

The calculated indicators were obtained according to the described methods in *Reference methodological Framework* document.

Austria, Bulgaria, Netherlands and Poland have published OBEC statistics for the *use-case 1 URLs Inventory* and *use-case 2 Variables in ICT survey* for 2019 and 2020, as well as Ireland – for 2019. In addition, Italy and the UK have published results for 2019 about use-cases 3 and 4.

2.3 Smart Energy *(full information on WPD available here on CROS)*

The Smart Energy work package D (WPD) is one of the work packages in the ESSnet Big Data II project and aims to implement smart meter data for production of official statistics. In addition, information and suggestions about smart meters metadata, data delivery, storage, validation, preparation and linking to administrative sources was investigated. The work package describes the methodology on how to find the electricity consumption of businesses and households, and how to identify empty dwellings from smart meters data.

1. Production process

Smart meters provide a possibility to innovate existing statistics and produce new statistics. Development of smart meter hubs provides a possibility to ease transfer of the data by getting them from a centralised place

Work of WPD was devoted to establishing official statistics production based on smart meter data. Below, a high-level outline of the recommended production process is provided:

- Before receiving the data, explore metadata and conceptualise the data you need.
- Come to a detailed agreement with data owners, specifying a data exchange protocol, file structure, and file content.
    - Files containing Smart meter data will be large and some information (background information) does not change often. To optimise the receiving process it is beneficial to split the data into several files: consumption/production data, background information on the end user and on smart meters.
    - Aggregation level of consumption/production data by time period should be considered based on needs. Hourly data lead to increased size of the data files and are not always available and not always necessary. At the same time, monthly aggregated data might not be enough to satisfy all needs. As an example one can find business and household statistics for monthly consumption which can be obtained based on monthly aggregated data. On the other hand, daily data give better possibilities for identifying vacant dwellings. One could also consider averages for weekends and weekdays. The decision will be based on conclusions from the metadata exploration.
- Control the quality of the received data. Validate that data were received as agreed, that they do not contain duplicates, and make other controls as described in the section "Data validation".
- Choice of tools for data storage is strongly dependent on data size and the needs and requirements of the National Statistics Institute (NSI).
- The raw data should be prepared and ready for production of statistics.
    - Smart meter data are sensitive data since they contain names and national identification numbers, which should be replaced with correspondent pseudonymous values
    - Usage of geocoding is beneficial as it will make address codes unaffected by address changes.

- Linking smart meter data with registers allows for extracting additional information, however, might be challenging.

2. Business consumption statistics

The expected outputs for business statistics are final energy consumption statistics of businesses by economic activity (monthly, quarterly and annual estimates). The overall process of finding electricity consumption of businesses can be divided into four larger phases.

1. **Data reception and preparation** - NSIs receive the data from the data- hub/provider. This phase includes preliminary data validation, anonymization of the data (personal codes, customer codes, etc.) and geocoding of the data. Generally a data architect is responsible for these operations with the data.
2. **Data storage and aggregation** - The data can either be stored in a classical relational database management system (RDBMS) or a more big data oriented solution like Hadoop HDFS. To find the electricity consumption of businesses the data granularity does not have to be very fine. Daily or even monthly aggregates will save computing power and time.
3. **Data linking** - linking the electricity data to the business register. This step will enhance the data quality and give the possibility to calculate the electricity consumption statistics by economic activity sector. It is necessary to find a link between statistical units: business entities and the observed units which are metering points. Linking is done only for metering points with a valid grid agreement. We also exclude metering points related to apartment associations, open suppliers and other network companies due to the fact that they are not the end consumers.
4. **Data modelling and interpretation** - In this stage we identify the actual end user (business), and the amount of electricity the business has consumed or produced. To compute the annual and quarterly electricity consumption statistics by economic activity sector, the NACE code is used to group each business into an economic sector and the consumption data is aggregated on groups. To validate the business consumption outputs, survey data, where companies have declared their consumption, can be used.

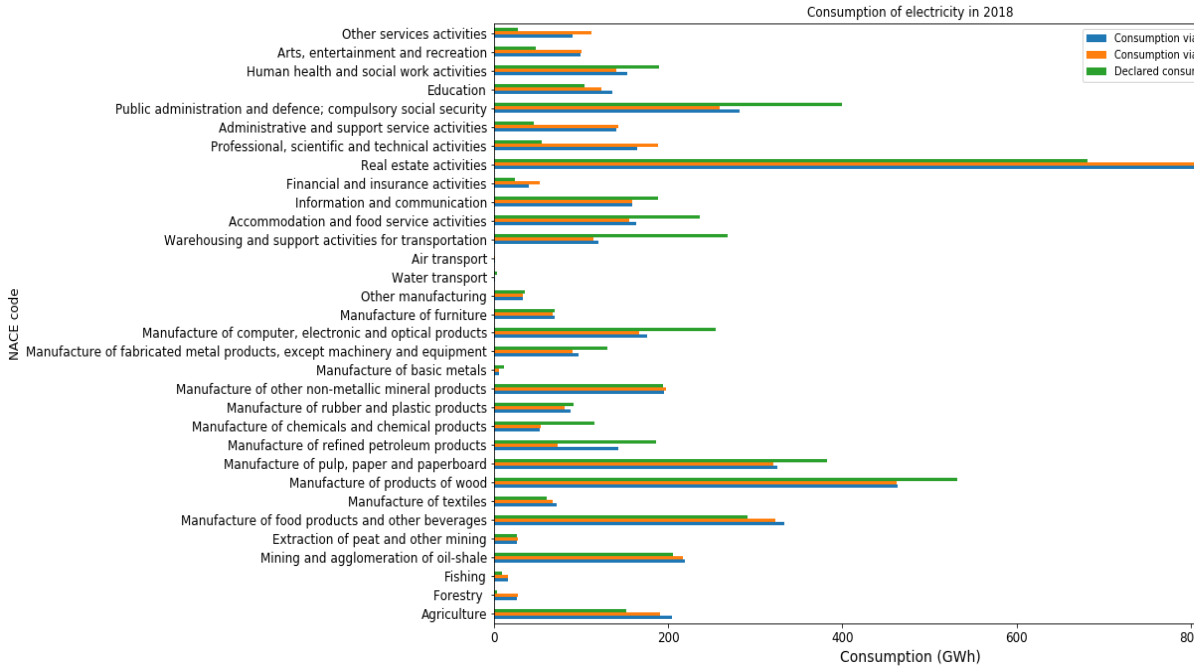


Figure 1: Electricity consumption of Estonian businesses by economic sector

Estonia's main goal was to find the total business electricity consumption by economic activity (Figure 1). The main differences between the declared consumption via survey and smart meters estimates are due to undercoverage. One explanation is that the end user's own produced electricity for their own consumption is not recorded in the data hub. We know that there are over 2700 businesses that have produced electricity for the network, so we can assume that they also produce electricity for their own consumption. We found no major issues with overcoverage.

3. Household statistics

To find electricity consumption statistics of households the smart meter data undergo a similar process as it did with business statistics. Data is received, stored, aggregated and linked to the national building and dwelling register, which contains usecodes for all buildings, units and dwellings. Usecodes selected were household, summerhouse or apartment. Further, the building- and unitcodes are linked to the population register in order to find only the units where people were registered with an official address. Addresses are selected for the statistics if more than zero and less than ten people were registered on the address in 2019.

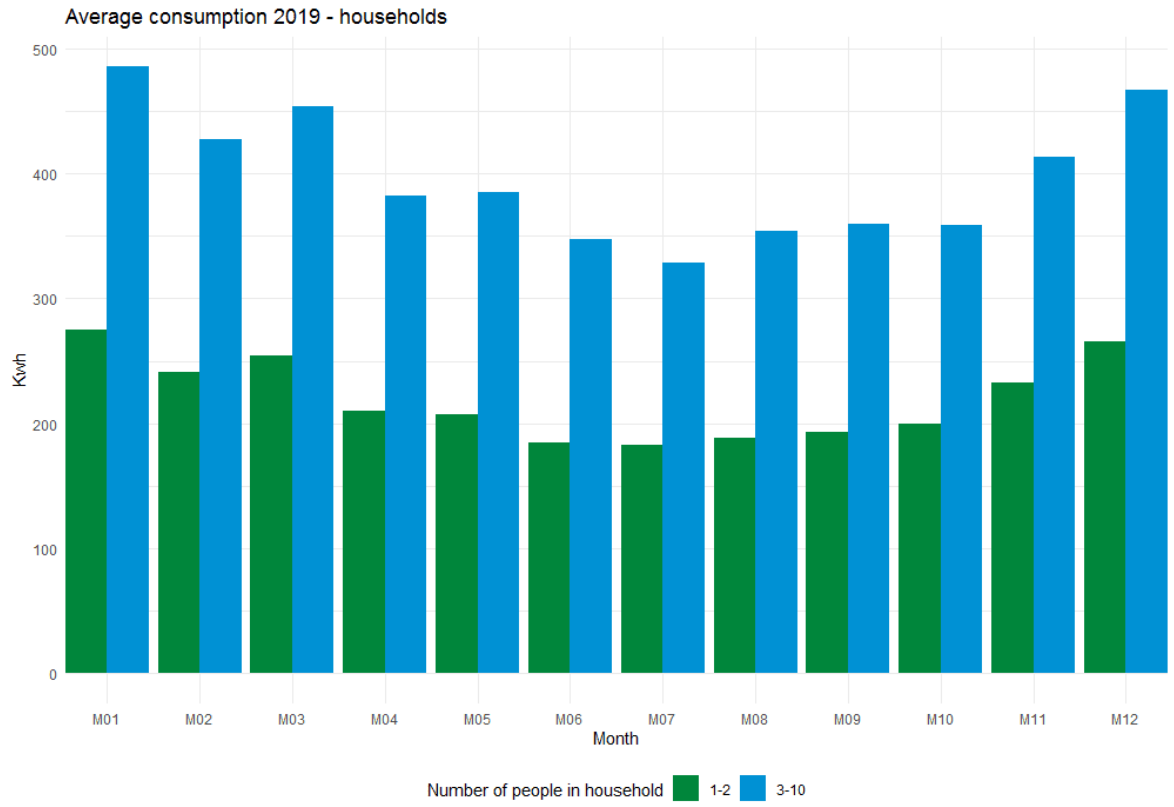


Figure 2: Average monthly consumption of Danish households

4. Vacant dwellings

Identifying vacant or seasonally empty living spaces from electricity consumption data can be relevant for housing statistics as it provides information where people actually live. The results can be used in the population and housing census. It would be desirable to separate dwellings into the categories occupied dwelling, dwellings for seasonal or secondary use and vacant dwelling in accordance with the recommendations on "Occupancy status of conventional dwellings" by United Nations Economic Commission for Europe.

WPD has a special emphasis on vacant dwellings. One way to find out which of the houses are empty (vacant), seasonal (summer), and occupied houses is to use clustering. Cluster analysis gives valuable insights from our data by grouping the data points with a clustering algorithm. Since smart meter data is unlabelled, the only way to cluster the data is to apply unsupervised machine learning. There are alot of clustering algorithms. We chose K-means as it is the most prevalent algorithm used for clustering smart meter electricity consumption data (Tureczek & Nielsen, 2018) and performed 2 case studies for clustering monthly and daily aggregated data.

4.1. Case study with monthly data (Norwegian example)

From this case study, clustering of smart meters into the three occupancy groups (occupied, seasonally/secondary use, vacant) by only using information on monthly aggregated energy

consumption is not possible. It is neither possible to adequately separate all the vacant dwellings from the non-vacant dwellings: some smart meters have an energy consumption pattern over months that might correspond to both a vacant and a non-vacant dwelling. Smart meters with permanently low energy consumption were defined as vacant, together with those smart meters with an end user change and suddenly reduced energy consumption by the time of the reference period.

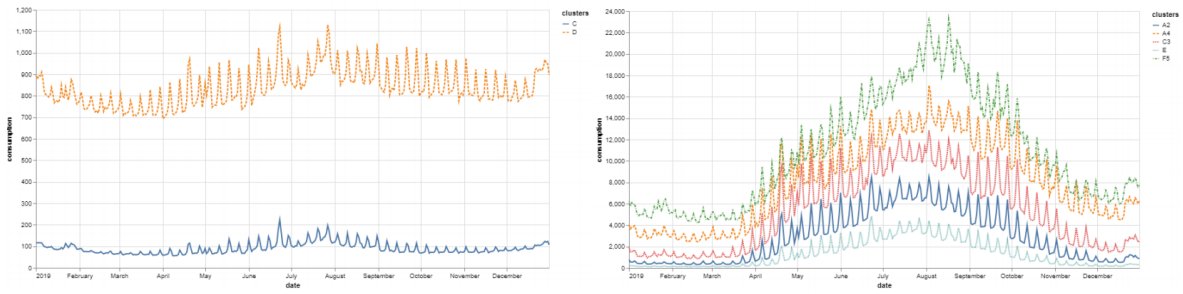
Since smart meters are not installed in all utility units, e.g. some users rejected installation, and since the occupancy groups are poorly identified by monthly aggregated energy consumption, it should be considered to combine electricity data with other data sources containing information on human activity by location, e.g. mobile phone data.

4.2. Case study with daily data (Estonian example)

The Estonian consumption data set contains hourly data for each smart meter. For ease of analysis, the hourly granularity has been aggregated to days, resulting in a data set where each smart meter has 365 instances of consumption. The clustering process can be divided into five sections.

1. Split the data set into eight subsets based on consumption threshold. Splitting the data set into smaller subsets gives more control over the process and makes it easier to identify different clusters.
2. Run the elbow method separately for the eight subsets and look for probable cluster numbers.
3. Perform k-means clustering in those eight subsets with the number of clusters decided from the previous step.
4. Analysis of the result.
5. Validate the clusters and the result.

After clustering there were 11 distinct groups that shared similar electricity consumption patterns (Table 1). There were 105 783 metering points where the daily average consumption of the cluster was lower than 1 kWh (Figure 3a). Estonia's largest energy provider has defined a 250 kWh yearly consumption threshold for vacant living spaces, hence this cluster can be classified as vacant. We observe 20 288 unique metering points where the consumption increases during the summer months and there are high consumption peaks during the weekends (Figure 3b). This cluster was classified as summer houses. It is necessary to point out that there were no summer houses that exceeded yearly consumption of 4500 kWh.



(a) vacant cluster (b) summer_houses cluster

Figure 3: Average consumption patterns a) vacant dwellings; b) summer houses

Cluster	Metering points
vacant	105 783
occupied_low	110 414
summer_house	20 288
occupied_1	94 948
occupied_2	262 988
occupied_spring	15 588
occupied_autumn	15 039
industry_1	66
industry_2	95 752
industry_3	1
outlier	4

Table 1: Results of clustering and number of metering points in each cluster

5. Choice of tools

WPD has evaluated two solutions for data storage and processing:

- Apache Hadoop framework, both as a data storage backend (HDFS) and data querying and processing platform (Hive and Spark).
- PostgreSQL database together with widely used data science tools available for R and Python languages.

5.1. Apache Hadoop

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than relying on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Hadoop Distributed File System (HDFS) is a distributed file system that provides high-throughput access to application data.

Hive is a data warehouse infrastructure that provides data summarization and ad hoc querying.

Spark is a fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.

5.2. PostgreSQL, R and Python

PostgreSQL is a powerful, open source object-relational database system that uses and extends the SQL language combined with many features that safely store and scale the most complicated data workloads.

R is a programming language and free software environment for statistical computing and graphics. R has grown from a statistics package to a popular data science toolset with a thriving community, lots of add-on packages and visualisation options.

Python is a high-level, general-purpose programming language. One of the most popular programming languages in the world, Python has an extensive set of data science packages.

5.3. Hadoop or PostgreSQL

Probably the most important aspect while choosing the tools is to look at the dataset size. In official statistics hourly and even daily data is rarely needed, so to optimise further processing aggregates for every day, month and year are suggested. Using the aggregate datasets means that raw data almost never need to be accessed and resource requirements for memory and processing power are considerably smaller.

Another thing to consider is the complexity of installing and managing Apache Hadoop cluster - this requires specific skills that might not be readily available in every organisation. Also with small scale setup more hardware resources are needed for Hadoop compared to the use of simple relational database. Apache Hadoop should be considered in cases when:

- data size is very large (more than ten million metering points with hourly data)
- highly parallelizable algorithms are used that can benefit from the distributed processing

6. Conclusions

The work of WPD provides examples of statistics on household and energy consumption. The statistics were obtained from smart meter data and are already implemented in production. In addition, it is shown that useful information can be extracted from smart meter data. With daily data, it is possible to extract businesses vs dwellings patterns. A Danish example shows how smart meter data have been used to explore the effect of the lockdown due to the Covid-19 pandemic on energy consumption. The results on identification of vacant dwelling based on daily data will be input to the Estonian population census 2021.

2.4 Tracking Ships *(full information on WPE available here on CROS)*

Maritime and inland waterways traffic has increased exponentially over the last decades. This required different solutions to ensure safety at sea, which resulted in the Automatic Identification System (AIS). Almost every ship broadcasts its location and status information by means of AIS, making it possible to detect other ships, wherever they were. Building on the developments of the completed work in the ESSnet big data programme I on the use of AIS data (Automatic Identification System) for maritime statistics, the aim of this Work Package (WP) tracking ships (WPE) is to develop functional production prototypes. Results of the WP in the first program showed the potential wealth of AIS data to improve current statistics and to generate new statistical products. Although some important elements of current maritime statistics such as type and quantity of goods loaded or unloaded at the port are not part of AIS, AIS is useful to improve other aspects of statistics and provide new products. In addition, AIS data is also useful for statistics on inland waterways, emissions and can provide other new products.

For each of the four main products identified in the first deliverable, a first version of the prototype using AIS data has been developed in the current program (all code and manuals are published on GitHub):

(1) Improve statistics on Inland Waterway transport: statistics on Dutch Inland Waterways (IWW) are based on incomplete information. Information on some parts of the Dutch inland waters is missing, resulting in incomplete data on ships' journeys and with that, information on goods (un)loaded. Implementing information from AIS in the process, completes this missing information on ships' journeys. This was done by developing a production infrastructure that can deal with big data. Furthermore, algorithms were developed to link AIS data to traditional source data. In turn, a methodology was developed to estimate the type and quantity of the commodities transported.

(2) Develop statistics on the behaviour of fishing vessels: using AIS, the first results on the activity of fishing fleet (time when a fishing vessel is out of the port) and the traffic of fishing fleet (the number of fishing vessels in fishing areas) in Poland have been derived. Several visuals have been created: a map with the fishing ports and fishing areas, and the traffic intensity. The fishing fleet was derived using the EU fishing fleet register, applying the rules of the Common Fisheries Policy. The prototype was also tested for fishing ports in the Netherlands and Greece.

(3) Improve quality of statistics on port visits: a Port Visits Geo-Solution prototype was developed in PostgreSQL. AIS data is used to derive the port visits and port traffic of the Piraeus port in Greece. The prototype shows that we are able to generate the official so-called F2-table. The prototype was also tested for ports in Poland and the Netherlands. Two manuals have been developed providing analytical guidelines, technical instructions for this prototype for full-fledged implementation in the ESS.

(4) Improve statistics on air emissions and energy used for the environmental accounts: a method was developed to identify shipping vessels related to the Dutch economy using text mining techniques. Research was performed to investigate the use of distance travelled per ship as a proxy of fuel usage and emissions. As there was no direct opportunity to actually calculate fuel usage and emissions, e.g. due to insufficient availability of data sources, three theoretical methods were developed to implement AIS.

Additionally, during this workpackage, a prototype is being created to develop an early indicator for trend changes in the real economy.

Concerns, signalled during the previous ESSnet big data programme I, like the need for suitable hardware and software tools for the exploration and exploitation of the huge amount of AIS data, still applied. The products mentioned were created using several big data platforms: local platforms, the Big Data Test Infrastructure (BDTI) environment and its successor, the Dataplatform, both provided by European Commission, and the United Nations Global Platform (UNGP) provided by the UN Global Working Group (GWG) on Big Data for Official Statistics. Each environment has its specific set of tools and uses other sources of AIS data with different coverages. AIS data is collected by several parties around the world, though the scope of the data may differ. Some parties only collect national data, some parties collect data on a European level and some parties collect all data around-the-world. Some parties collect maritime data and some parties collect data focused on inland waterways.

During the project, each time the best environment for developing and testing the product had to be selected. The choice depended mostly on the coverage of the AIS data. The inland waterways product requires AIS data on inland waterways, which only is available nationally because of privacy reasons. Both the products on port visits and on fishing fleet were developed locally and later implemented and tested on the Dataplatform. Almost 7 months after the start of the project of the project, the BDTI environment was provided by the European Commission. It was available for 5 months for an interrupted period. Then, the environment was deprecated and in month 18 of the project, it was replaced by the Dataplatform. Setting up the Dataplatform and loading the AIS data required a lot of valuable project resources. Finally, for the product on air emissions and energy used for the environmental accounts, worldwide coverage is required, which was only available on the UN Global Platform.

Even though the Dataplatform was provided late during the project, it was used to test two of our products for other countries. The port visit product, which was developed by Greece, was executed to create the F2 tables for several Dutch and Polish ports. And the fishing fleet product, which was developed by Poland, was executed to create the first results on Dutch and Greek fishing fleet activities and behaviour. This shows that the products can be executed on other platforms than initially developed. And more importantly, it shows that the solutions are generic and can be used to produce the results for other countries.

1. Outcome of the prototypes

Data sources

Depending on the prototype, different data sources were used for AIS. In addition, different sources of complementary information are used. The reason different AIS sources were used, is that not all sources provided the data needed for a particular prototype. Here, an overview is provided of the different sources used. Table 1 shows the different data sources for AIS and Table 2 the complementary data sources. In general, the AIS messages 1, 2, 3, 4, and 5 were used. These contain static and dynamic information (i.e. information on ships' characteristics and information such as location and speed, respectively). For inland waterways, message 8 is relevant as it contains static information on inland

waterway vessels. Note that not all sources were obtained specifically for this work package WPE, but for example also for the previous ESSnet I (WP4), or local projects such as improving statistics on inland waterways.

AIS data						
	AIS EMSA	**AIS Dirkzwager**	**AIS Polish data**	**AIS Greek data**	**AIS Orbcomm**	**AIS Inland waterways**
Source	EMSA (not open source)	Dirkzwager (not open source)	Polish Maritime Authority	Hellenic Coast Guard	UN Global Platform	Rijkswaterstaat
Coverage obtained	European coastline	European coastline	Poland	Greece	Worldwide	Dutch Inland Waterways
Time period available	2017-2018	March 2015 - March 2016	yearly	yearly	October 2018 - present	2017
Filtering	3-minute filter	unclear	none	none	unclear	none
Specifics	Does not contain fishing fleet	Bad coverage of Greece				Message 8 is included, specific for IWW
Applied for	Port visits	Fishing fleet	Fishing fleet	Port visits	Energy used and air emissions	Inland Waterways
More information	WPE Deliverable E3 [3] and current deliverable	ESSnet I: WP 4 Deliverable 4.3 [4]	ESSnet I: WP 4 Deliverable 4.3 [4]	ESSnet I: WP 4 Deliverable 4.3 [4]	WPE Deliverable E3 [3] and current deliverable	WPE Deliverable E3 [3] and current deliverable

Table 1: Different data sources for AIS

Complementary information					
	RIS (IVS)	**Fleet register data**	**Lloyd's Register of Ships**	**General Business Register**	**EU Monitoring, Reporting and Verification of CO2 from marine transport (EU-MRV)**
Type of information	Lock information (not open source)	Fishing fleet (open)	Ship information (not open source)	Business information of Dutch businesses	Energy use and air emissions of ships
Source	Rijkswaterstaat	DG Mare	IHS Markit	Statistics Netherlands	European commission / EMSA
Coverage obtained	Dutch inland waterways	European fleet	Ships entering Greek ports, and Dutch ports	The Netherlands	Only for large ships voyaging to and from EU ports, between EU ports and at berth
Time period available	2017	yearly	yearly	yearly	yearly (from 2018)
Filtering	n.a.	n.a.	n.a.	n.a.	n.a.
Specifics	Contains all ships Dutch inland waterways that pass information points	Missing/incorrect MMSI and IMO-numbers	n.a.	n.a.	n.a.
Applied for	RWS (Dutch water infrastructure Authority)	Fishing fleet	Greek and Dutch Maritime Statistics	Energy used and air emissions	Energy use and air emissions
More information	WPE Deliverable E3 [3] and Current deliverable	WPE Deliverable E3 [3] and Current deliverable	Current Deliverable	WPE Deliverable E3 [3] and Current deliverable	Current Deliverable

Table 2: Data sources for complementary information

2. Inland waterways

In statistics on the transport of goods via Inland Waterways (IWW), the amount and type of goods transported via inland waterway are described for the national territory. In the Netherlands, these statistics are based on information that has to be provided when inland waterway ships pass by a lock. At the locks, so-called Information and Following system for Shipping (IVS-information) is gathered on the origin and destination of a ship, the goods the ship carries and some ship's characteristics. However, this information is known to be incomplete as not all shipping routes pass by a lock, resulting in incomplete information for these journeys. AIS provides information on all ships and therefore is useful in gaining a complete image of IWW shipping. This project focuses on supplementing the information missing from IVS using AIS data. The national source of AIS is used because of the sensitive

aspect of AIS of IWW ships, as the home address of inland skippers is usually their ship. Furthermore, EMSA data does not contain information on IWW ships.

Process, concepts & methodology

The process, which is shown below in Figure 4, is described in detail in the deliverables. The main steps and methodology used are:

- Pre-processing: AIS messages 1, 2, 3, 5, and 8 are selected, decoded, filtered (within Dutch territory and with a speed over 0.2 knots) and merged.
- Method to derive journeys: a journey is a movement between two locations where goods are loaded and/or unloaded or the point where a ship leaves Dutch waters.
- Method to match AIS & IVS journeys: based on Probabilistic Record Linkage
- Method to estimate type and quantity of goods

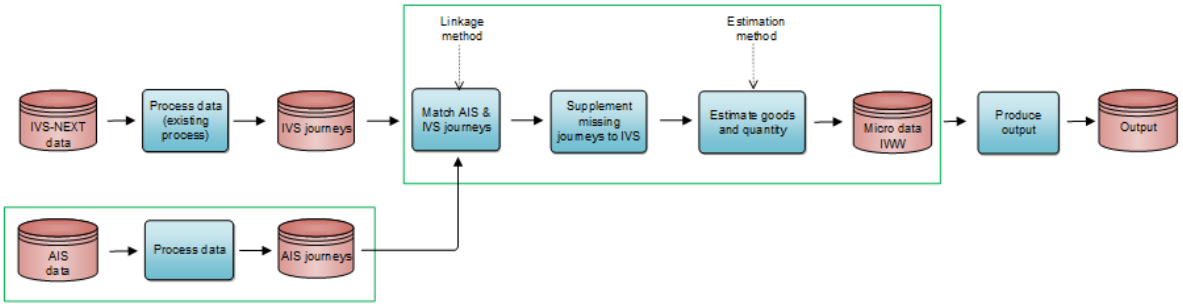


Figure 4: New IWW process with AIS as an additional source

Infrastructure and software

The original setup proved to be too slow to process a week of data. At Statistics Netherlands a proof-of-concept (PoC) has been performed with Greenplum as a high-performance computing (HPC) environment based on PostgreSQL. In using AIS for the process of statistics on IWW, the use of an HPC environment is necessary to parallelize the processing of AIS data to decrypt and process AIS.

The experience with the platform is very good. The data processing performance scales (almost) linear with the number of server segments. It has resulted in increasingly faster runtimes while keeping the flexibility of adapting the code. We were able to convert most of the steps in the process of transferring AIS data to AIS journeys, such that they were usable in the environment.

3. Fishing fleet

Behaviour of fishing fleet (activity and traffic) has not been a widely researched topic yet. However, it is becoming a topic of great interest to the central and local government administration in the coastal areas. For a large part, this is a result of the sustainable development policy (see Goal 14: Conserve and sustainably use the oceans, seas and marine resources).

The analysis performed for this WP focuses on the behaviour of the fishing fleet and its activity within fishing areas, which shall be of novelty and allow production of experimental statistics on fishing fleet behaviour. The work concerning Polish fleet and selected Polish seaports should result in an easy-to-implement software with universal methods supporting the process of collecting, processing and analysing AIS data for the purposes of statistics. The developed methods were also tested on a single fishing port in the Netherlands (Yerseke) and in Greece (Kavala).

An important aspect of the work is the use of available sources, namely AIS and Fishing Fleet Register data, and the fact that the developed model shall be so universal that it can be replicated for the ESS and other international organizations.

Process, concepts & methodology

The process, which is shown below in Figure 5, is described in detail in the deliverables. The main steps and methodology used are:

- Create reference frame of Polish fishing fleets: by correcting and completing the EU fleet register using Python web scraping methods
- Cleansing data: remove outliers, technically faulty messages and duplicate messages
- Method to validate data: based on the Haversine method
- Create register of fishing ports with its geographical coordinates: usually each fishing boat is "assigned" to one fishing port.
- Method to measure and analyse activity and traffic of fishing ships: deriving many indicators by combining ports' coordinates and ships' AIS messages together.
- Method to differentiate between itineraries typical for fishing vessels: routes involving fishing activity take the shape of zigzags, while the routes travelled to reach a fishery are generally straight lines

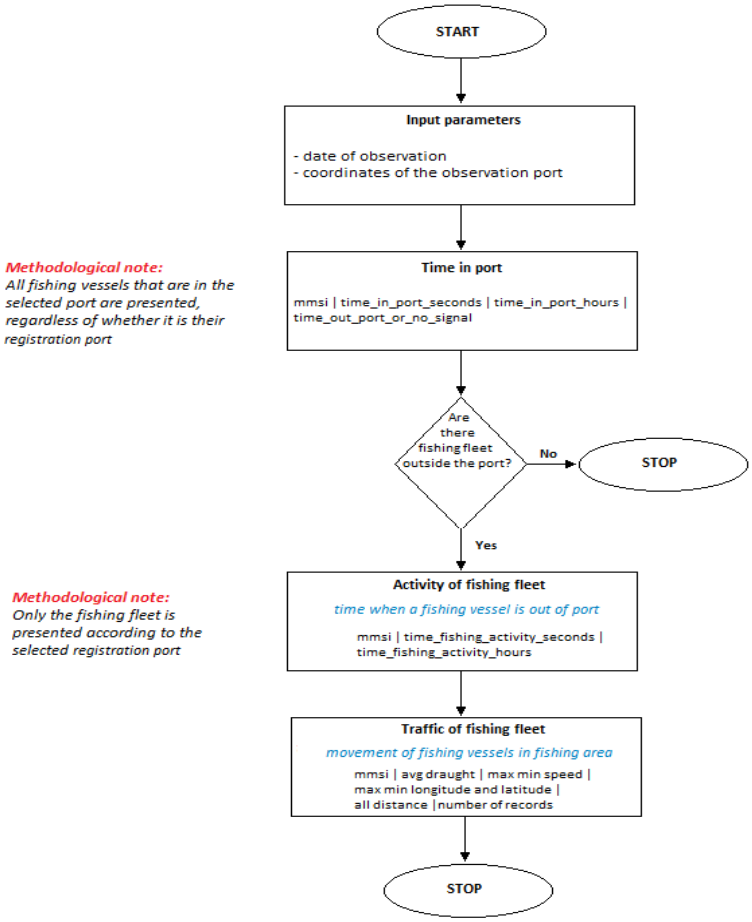


Figure 5: Schematic diagram – General flow of the algorithm for data processing

Results

In the deliverables, the many analyses performed are described and resulting visualisations are shown. It shows that AIS data is a good source for analyses of fleet activity.

The investigation also shows that the codes and software worked correctly for all three countries (Poland, Greece and The Netherlands). The prototype developed is universal and can be implemented for other countries interested.

4. Port Visits Geo-Solution

The initial aim of this prototype was to use AIS data for the compilation and verification of existing Maritime Transport Statistics (MTS) under the Directive 2009/42/EC.

The Port Visits Geo-Solution prototype provides a general tool for monitoring ships movements using AIS data in selected area(s) defined by a polygon(s). It is a data driven prototype, not tailor-made for a specific Port, adaptable to the specificities of any Port. The solution has been implemented in a powerful, open-source object-relational database system, PostgreSQL with PostGIS.

In this way, an AIS spatiotemporal database (DB) of ships movements was built. The added value of this solution is that it can be adapted to any area(s), defined by polygon(s), giving the opportunity to

the user to compile interactively spatiotemporal queries. The results will be visualized in the map, using the geometry viewer, without using any extra visualization tool.

Process, concepts & methodology

The process is described in detail in the deliverables. The main steps and methodology used are:

- Developing the AIS spatiotemporal database of ships movements
- Cleansing data: filter port area, deduplication of AIS vessel characteristics information, filter ship types based on terminals
- Method to link AIS data to Ship Register
- Compilation of the F2-table

Results

The prototype has a very good performance of compiling the F2 table, as shown by the quality indicators in Table 3. The overall accuracy, which measures how well the prototype correctly identifies or excludes the port calls, is 88.8%. The overall sensitivity, which measures the proportion of positives that are correctly identified by the prototype, is 93%. The overall precision, which measures how close the prototype's and official results are to each other, is 93%.

Types of Ships	**Quality Indicators**		
	Accuracy	***Sensitivity***	***Precision***
	(TP+TN)/(TP+TN+FP+FN)	***TP/(TP+FN)***	***TP/(TP+FP)***
31 - Container	*88,9%*	*96,4%*	*91,9%*
32- Specialised cargo / 34-Dry cargo barge	*71,3%*	*95,7%*	*73,6%*
33 - General cargo, non-specialised	*88,1%*	*93,8%*	*93,6%*
35 - Passenger	*95,1%*	*96,7%*	*98,3%*
36 - Cruise Passenger	*100,0%*	*100,0%*	*100,0%*
Total	*88,8%*	*93,0%*	*93,0%*

Table 3: Quality Indicators

The prototype has also been tested at the Dataplatform for both Poland and the Netherlands (port of Amsterdam). For testing purposes, only two days of AIS data (March 6th and 7th, 2017) was loaded. For Poland, the F2 table with the number of arrivals per ship (MMSI) for the port of Świnoujście was created. For The Netherlands, it was decided to use the prototype to find all oil tankers that arrive on March 7th, 2017, at port of Amsterdam, which is one of the busiest in Europe. The outcome was a perfect match to the official maritime statistics on port visits by ship type of the Netherland's Maritime Statistics Department.

5. Energy used and air emissions of shipping vessels for the environmental accounts

The aim of this feasibility study was to investigate whether AIS data can be used for compiling energy use (for the PEFA) and air emissions (for the AEA) by shipping vessels for the environmental accounts. The research was divided in three parts:

1. Developing a method to identify shipping vessels related to the Dutch economy

2. Investigating whether it is possible to retrieve and implement worldwide AIS data for Dutch shipping vessels to calculate travel distances
3. Analysing how to compile energy use (for the PEFA) and air emissions (for the AEA)

The study not only focused on Dutch residents' ships, but also on ships sailing under the Dutch flag. This latter selection is the reference group for some Dutch marine policies. Also, for some EU countries this selection is the reference group for compiling air emissions.

Process, concepts & methodology

The process is described in detail in the deliverables. The main steps and methodology used are:

- Method to combine the annual inventory list of ships, provided by the Dutch department of Information and Transport, with the General Business Register and the Production Statistics (see Figure 6)
- Shown that the operator characteristic of the ship represents the connection of a ship with the national economy, as the operators are related to the residents' shipping companies: average matching percentage of 87 percent for the period 2014 to 2017
- Method to calculate distance travelled per ship for a certain time period

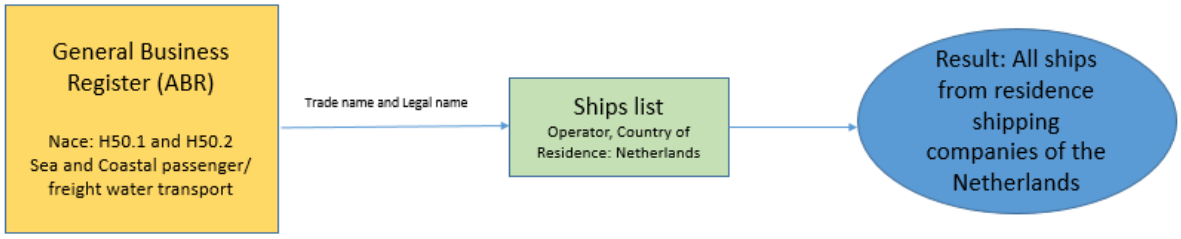


Figure 6: Identifying shipping vessels related to the Dutch economy using the general business register

The main challenge in this project was the insufficient availability of data sources. Not only could we not obtain longer times series of worldwide AIS data, but we also did not get access to a ship database to retrieve technical information from ships. As a result of the research on the definition and identification of Dutch residents' ships, three theoretical methods were developed to implement AIS data for the purpose of estimating energy used and emissions of the ships: 1) distance as a proxy, 2) national actual emissions plus distance, and 3) emissions per ship. With the first method being on the most aggregated level and easiest to perform, to the third method being a bottom-up approach and most difficult to execute. All three methods can be applied to either a nationalities residents' ships or ships sailing under nationalities flag.

6. Future work

During this work package, we have developed several functional production prototypes, showing the possibilities of AIS data in statistical production. These prototypes prove that AIS data is a valuable source to generate statistics in a production environment. By implementing some of these prototypes in the Dataplatform environment and executing these prototypes to create partial results for several countries (Greece, Poland and The Netherlands), we have shown that these solutions are generic and can be used to generate these statistics from one central environment.

In order to use AIS as a source for official statistics and fully benefit from it, some prerequisites need to be met:

1. AIS maritime data with worldwide coverage is crucial

   For certain statistics, like air emissions and economic indicators, AIS data with worldwide coverage is crucial, but very costly. EMSA collects AIS data with European coverage, and if agreed by the member states, it could share this data. EMSA also buys additional satellite and global data from commercial parties. EMSA cannot share this data under the current contracts. At the UN Global Platform however, AIS data with a worldwide coverage is currently available and shared.

2. AIS inland waterways data is desired

   Different member states could benefit from using AIS to improve statistics on inland waterways. If that is the case, AIS data with coverage of European IWW could be desirable to provide to the statistical offices. Note that, under current legislation, AIS data related to inland waterways ships is sensitive information, as the ship is also the home location of the skipper, and therefore considered confidential data. This confidentiality issue would first have to be investigated.

3. Ship register with worldwide coverage is required

   A worldwide ship register, both maritime and inland waterways, including ship characteristics and ownership/flag is required. The latter is required for air emissions from the environmental accounts. A ship register with worldwide coverage is Lloyd's register from IHS Markit's. A good example of an initiative to create such a register by the European Commission, is the European fishing ship register (see https://webgate.ec.europa.eu/fleet-europa/search_en). Unfortunately, this does not include the MMSI number, which makes it difficult to connect it directly to AIS data.

4. Port register with European coverage is needed

   A port register with at least European coverage is needed. This register should at least contain all European ports with their geographical boundaries, preferably including more detailed information on the terminals, the type of cargo supported (containers, bulks or even people) and their geographical boundaries.

5. Global IT platform and infrastructure is required

   In order to develop, test and implement solutions based on AIS data, a well-suited IT platform and infrastructure, including access to the AIS data, is mandatory. Setting up this environment requires special skills that are currently not always available. Note that the UN Global Platform provides such a platform. Eurostat is in the process of developing such a platform, but extra attention should be geared towards sustainability, user-friendly access and use.

In the first ESSnet on AIS, we have described a number of other possible new statistical products. Unfortunately, due to time restraints on the availability of the platform and the situation surrounding COVID-19, we did not have extra time to work on these extra products. We still think it is worth pursuing some of the following topics:

- Intra-port distances
- Fluvio-maritime transport: AIS data could help getting insight in the relationship between maritime and IWW transport

- Port-to-port distance matrix based on model estimating shipping routes: AIS provides the opportunity to follow ships' routes and thus calculate the distance travelled. In addition, it can also provide more insight into patterns and factors determining these travel patterns
- Insight in effects of disruptions or regulations: AIS could be used to detect for example, a shift in traffic patterns caused by port closure, or speed regulations in emissions zones.
- Visualisations: although this is already an important aspect of our prototype for fishery, visualisation can provide many more options of gaining insight in maritime transport.

2.5 Process and Architecture *(full information on WPF available here on CROS)*

WPF, "Process and Architecture", has worked for defining a European reference architecture for Big Data to (i) guide Big Data investments by NSIs and (ii) help the development of standardized solutions and services to be shared within the ESS and beyond.

The main outcome of WPF is what we called BREAL (Big Data REference Architecture and Layers), namely a reference architecture defined as a set of artefacts to be used by:

- NSIs that aim to introduce the use of Big Data in their production processes.
- Public and private organizations that would like to follow a defined and controlled way of producing Big Data-based statistics guided by the Official Statistics expertise.

From a practical point of view, BREAL can be used as follows:

- As a reference framework to be used at national and ESS-level by enterprise architects to align business and IT needs.
- As a language for IT/solution architects to describe information systems projects that make use of Big Data sources.
- As an instrument for NSIs' top management to plan national investments related to Big Data projects, taking into account the economies of scale that are offered by European infrastructures and services for Big Data.

The following sections describe the principal obtained results as reported in the deliverables F1 "BREAL- Big data Reference Architecture and Layers- Business Layer" and F2-"BREAL-Big data Reference Architecture and Layers- Application Layer and Information Layer" (see https://ec.europa.eu/eurostat/cros/content/WPF_Milestones_and_deliverables_en).

1. Overview of BREAL

BREAL is an architectural framework including several architectural artefacts. The specific BREAL artefacts are represented in Figure 1, where they are placed on the specific architectural layer they belong to.

In the Business Layer, the BREAL artefacts are:

- *List of principles*, intended as guidelines and general rules that support (statistical) organizations implementing statistics based on Big Data sources.
- *Business functions model*, related to the development, production and deployment process necessary for a Big Data-based production line.
- *Life Cycle model*, sequencing BREAL business functions in a Big Data-based production line.
- *Support functions model*, including the functions that are thought to support the production process, like, e.g. human resource management.

- *Stakeholder model*, with the stakeholders playing relevant roles in a production system intended to manage Big Data sources.

The Application Layer consists of a:

- *Generic Application Architecture*, with a set of generic application services, proposed to show how the identified business functions can be implemented.

The Information Layer includes a:

- *Generic Information Architecture*, consisting of the three layers described by the 'hourglass model' proposed for Trusted Smart Statistics, namely raw data, convergence, and statistical layer.

In addition, BREAL includes an Operational Model, describing how data and services can be deployed in a Big Data solution. This can be considered as part of both the Application and Information layers.

The proposed Generic Application Architecture and Information Architecture are "generic" in the sense that they are not intended to be specific of a Big Data project or source. The use of the services and data that are specific to the Big Data projects of the Implementation Track of the ESSnet Big Data Pilots II are described in a set of solution architectures, namely (see Figure 1): (i) Solution architectures for *Online job vacancies* (WPB); (ii) Solution architectures for *Online based enterprise characteristics* (WPC); (iii) Solution architectures for *Smart energy* (WPD); (iv) Solution architecture for *Tracking ships* (WPE).

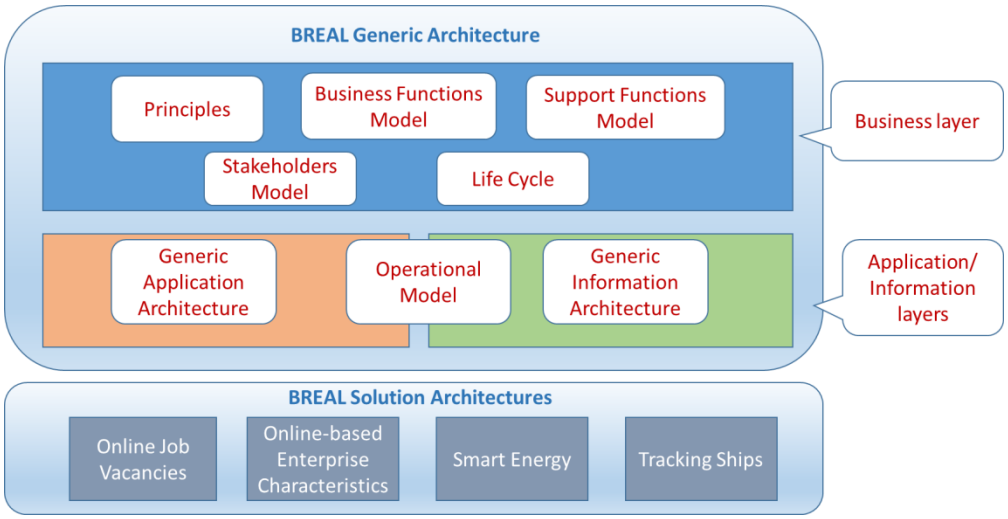


Figure 7: Overview of BREAL

2. BREAL Artefacts at a Glance

In this section, among the main artefacts of BREAL, we focus on shortly describing: BREAL Life Cycle (Section 2.1), BREAL Application Architecture (Section 2.2) and BREAL Information Architecture (Section 2.3)

BREAL Life Cycle

Using existing reference architectures (in particular EARF1, and GSBPM2), we designed the BREAL business functions model, including already defined business functions and several new business functions. These business functions were used to build the BREAL Big data Life Cycle, shown in Figure 2. It encompasses three major areas:

- Development and Information Discovery – where the exploration of the Big Data source, its integration with other data and the discovery of information take place
- Production – actually creating statistical products through the use of Big Data sources
- Continuous Improvement – monitoring and assessing the Big Data source usage with a focus on the population coverage issues and the validity of the models used.

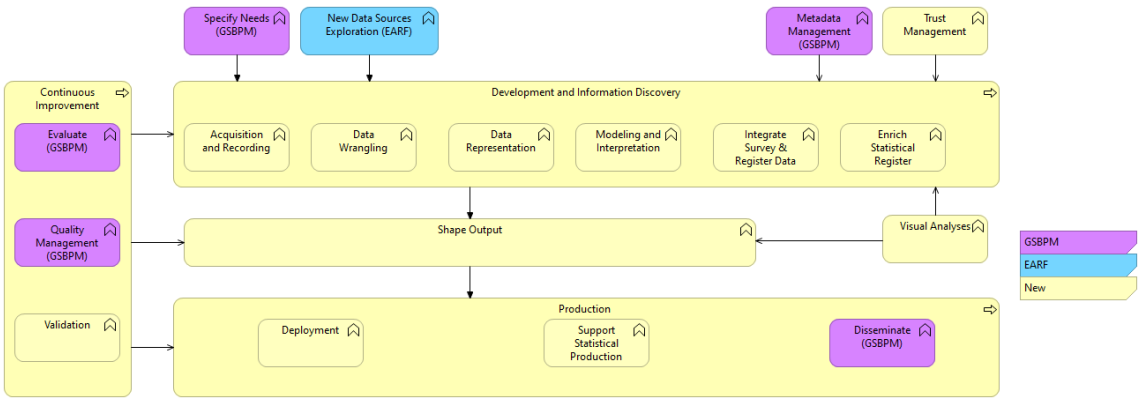


Figure 8: BREAL life Cycle

Generic Application Architecture

For most of the business functions, identified within the Business Layer of BREAL, it is relevant to identify some Application Services, i.e. logical components belonging to a software application layer that implement the specific functions.

The BREAL functions for which related application services have been specified are shown in Figure 9, with the business functions in yellow and the application services in blue. The application services have been formulated by researching the solutions for WPB through WPE. Even though we tried to provide a complete list, it might need be extended with future developments.

[1] Eurostat: ESS Enterprise Architecture Reference Framework: https://ec.europa.eu/eurostat/cros/content/ess-enterprise-architecture-reference-framework_en
[2] GSBPM vs 5.1: https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

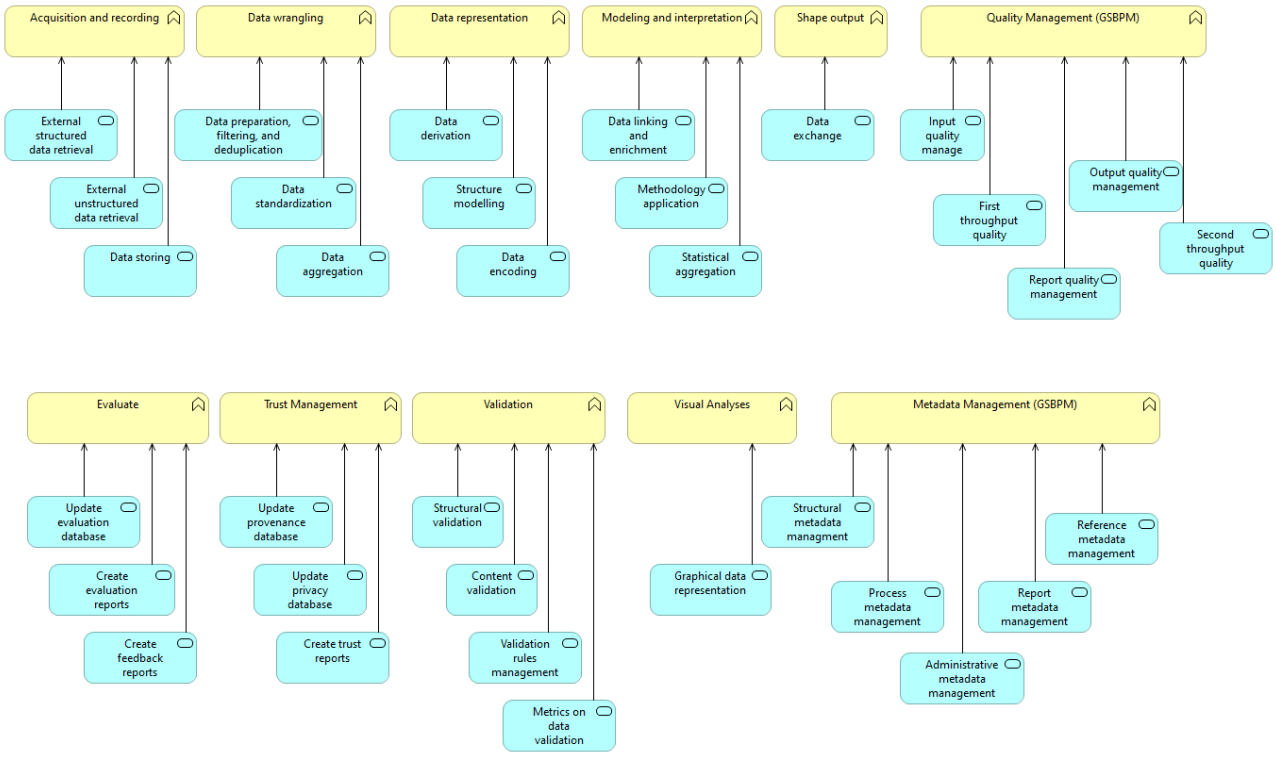


Figure 9: BREAL Application services for Development, Production and Deployment business function

In addition to business functions related to the production life cycle, a set of *support business functions* have been identified as relevant to big data management. In Figure 10, the related application services are represented as blue boxes for each of the business functions (yellow boxes).

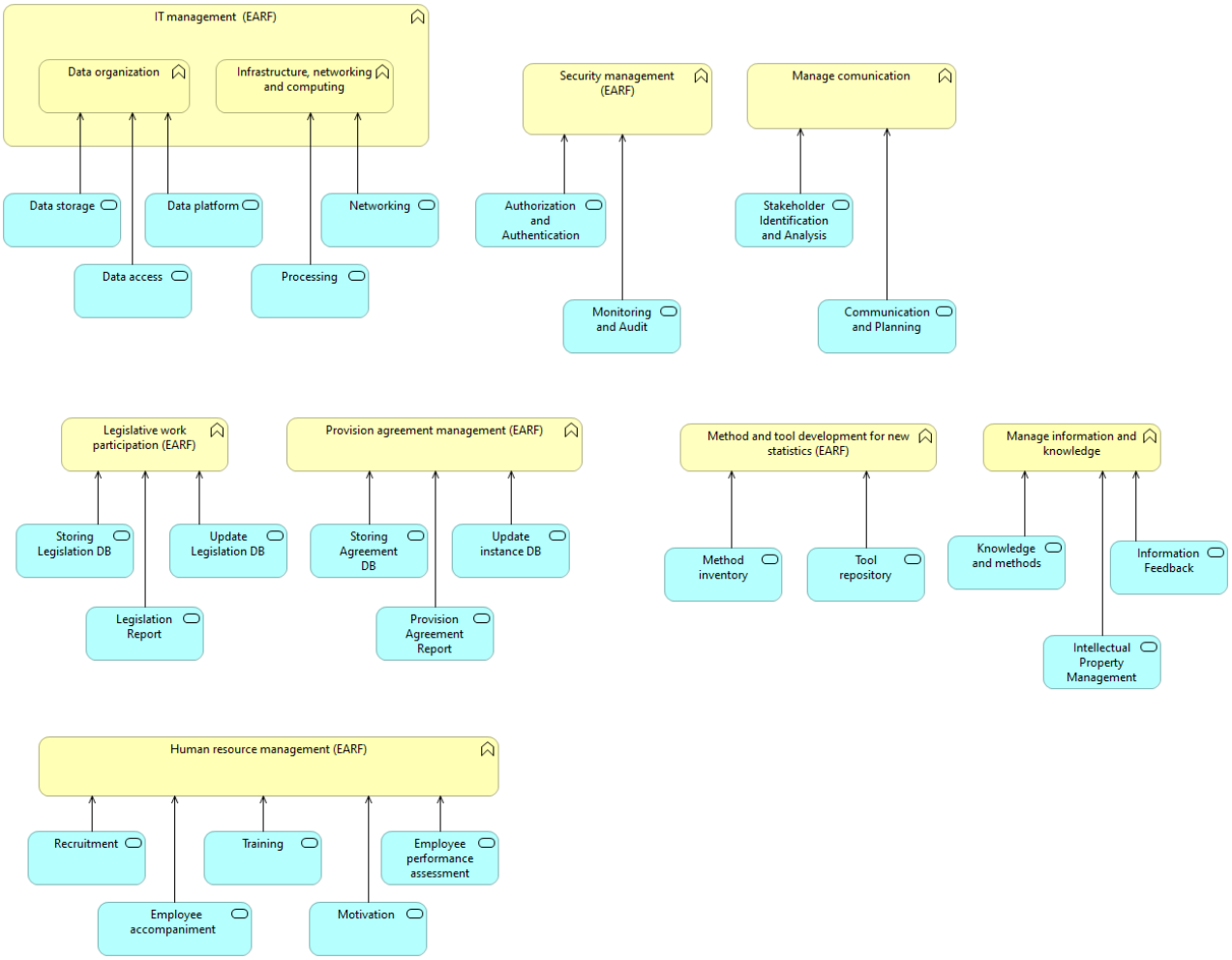


Figure 10: BREAL Application Services for Support Functions

Generic Information Architecture

The Information Layer of BREAL consists of three sublayers, as described by the 'hourglass model' proposed for Trusted Smart Statistics, namely a raw data layer, a convergence layer, and a statistical layer.

- The *Raw Data Sublayer* includes data that are acquired and stored by the BREAL 'Acquisition and Recording' business function. At this stage, we just identify the concepts and no detail is provided on formats or other technical specifications that can be useful for raw data acquisition and storage.

- The *Convergence Sublayer* contains data represented as units of interest for the analyses. These data are produced as results of the BREAL life cycle functions of 'Data Wrangling' and 'Data Representation'.

- The *Statistical Sublayer* includes those concepts that are the targets of the analysis. These data are produced by 'Modelling and Interpretation', 'Integrate Survey and Register Data', 'Enrich Statistical Registers', and 'Shape Output'.

The intention of this model and the implied solution architecture is to hide the real-world complexity (the raw data sublayer) from the statistician and the statistical complexity (the statistical sublayer) from the data provider. In the ideal situation, the convergence layer contains the data that is actually shared between the parties.

In addition to the data concepts, some metadata concepts are introduced for each of the three layers. In particular, one specific category of metadata has been selected as very specific of Big Data, namely Provenance metadata. These metadata have been specified for each of the three layers mentioned above.

Moreover, in order to emphasize the integration of the Information Layer with respect to existing standards for data modelling in Official Statistics, we included some key concepts from GSIM (Generic Statistical Information Model)[3].

The resulting composition of the Information Layer of BREAL is shown in Figure 11. Each sublayer consists of: (i) specific BD data entities (blue color); (ii) GSIM entities (pink color) and (iii) specific provenance metadata entities (yellow color).

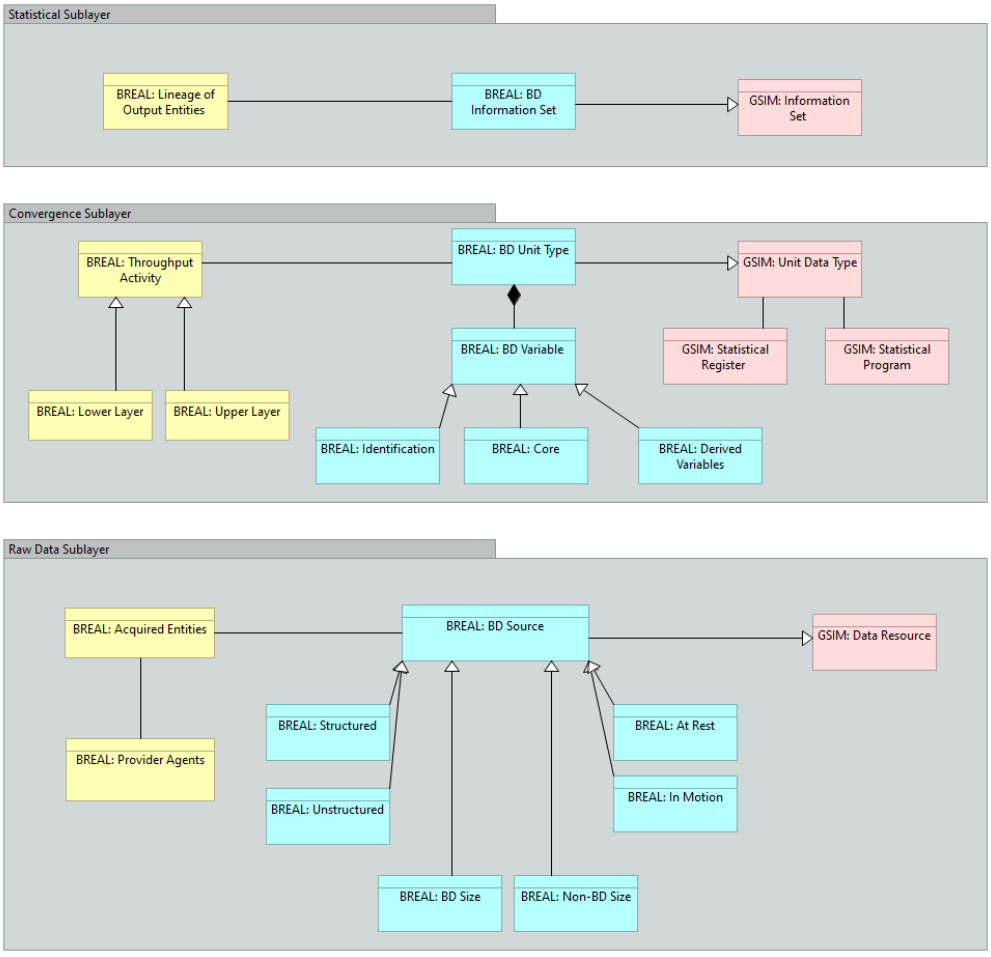


Figure 11: BREAL Information Architecture

[3] GSIM vs 1.2: https://statswiki.unece.org/display/gsim

3. Outlook

BREAL is a product of the ESSnet Big Data II resulting from a process involving two strands of work: a *bottom-up one* where the concrete solutions emerging from the vertical lines of work inside the WPs were abstracted and taken into account and a *top-down one*, where the design of BREAL was figured out by WPF participants.

BREAL work takes explicitly into account different types of sources (Web data and Sensor data principally) but aims at being general enough to represent also other kind of Big Data sources. The explicit reference to existing standards, like GSBPM and GSIM, is also very much important, especially for the complementary role that BREAL can have with respect to them.

However, though BREAL is quite mature, for sure it can be enriched and possibly revised in the future by practical applications.

Some concrete applications of BREAL have already started in a 'spontaneous' way, and include:

- Within the ESSnet Big Data Pilots II, the application of BREAL has been experimented within the Pilot Track workpackages, such as for instance by WPI and WPH.
- Outside the ESSnet Big Data Pilots II, and more specifically, within the ESSnet Smart Survey, BREAL has been using for modelling the architecture of a European level platform to support smart surveys.

2.6 Financial Transactions Data *(full information on WPG available here on CROS)*

1. Overview

The main aim of WPG has been to get an overview of the sources and the data infrastructure (metadata) of financial transactions data in the participating countries. The objective has been to describe to what extent financial transactions data are available as well as whether it is possible for the NSIs to access them. Given the infrastructure, it has also been a main aim of WPG to assess the statistical potential of these data sources. The statistical potential could be the quality improvement of official statistics as well as a new statistical product. Case studies should be performed to assess the statistical potential in practice.

In addition to the financial transaction part described above, WPG has also aimed at utilising financial transactions data to describe the sharing economy, after finding a definition of sharing economy suitable for official statistics.

The WPG results are described in the final deliverablewhich contains the cumulative results of WPG (see https://ec.europa.eu/eurostat/cros/system/files/wpg_deliverable_g4_final_report_09_04_2021.pdf).

Part I of WPG was about describing financial transactions data without analysing such data. First the concept is described followed by an overview of some EU initiatives. Then the main characteristics of payment system, payment instruments and actors are described followed by the legal aspects in general and in particular for each of the participants in WPG, the *WPG countries*: Bulgaria, Germany, Italy, Norway, Portugal and Slovenia. Based on the work described above, an assessment was made on which official statistics that could be improved by financial transactions data. The potential improvement could be either by using financial transaction data as additional data or by replacing the current data source completely, the latter could then reduce both response burden and cost. By having the assessment before financial transaction data were accessed and analysed, the results could help in selecting which financial transaction to access and analyse later in the project. The promising official statistics indentified: household budget survey, household expenditure of residents for tourism and balance of payments estimates, e-commerce turnover, retail trade index, early estimates of macroeconomic aggregates and indicators, business statistics indicators on flow between industries, and finally turnover of collaborative/sharing economy.

Part II of WPG was application of financial transactions data. Case studies was an important part of WPG. Before the project started, the number of case studies was uncertain due to the uncertainty concerning data access. The selection of case studies could not only be based on which statistics were promising, but also on which data could be accessible. In the end, Italy, Norway and Portugal were able to access data and perform case studies. The Italian case studies look at the gain of using time series of daily card transactions by aggregated merchant category code groups in forecasting and nowcasting of economic indicators.

A tourism survey will typically not cover very small lodgings. A topic in the Portuguese case study, is the improvement of the tourism statistics using financial transaction data covering these lodgings. A second topic is how to measure transport made by the new alternatives to taxi. The Norwegian case study looks at to what extent debit card transactions at legal unit level can be used to estimate the

monthly total retail sales by the ten industry classification groups that is being used in the Retail sales index statistics.

As the final application, sharing economy was studied. Both the theoretical defintion and an operational definition are discussed, since sharing economy is defined in several ways in the literature. The project started with a "pure" theoretical definition that limited sharing economy to the temporarily sharing of physical goods between persons. The operational definition finally adopted, was the definition of collaborative or platform economy, since this is easier to measure and also more relevant for economic statistics. As an example, the chosen operational definition covers small lodgings owned by businesses, and these are useful for travel accommodation statistics that would otherwise typically only cover hotels and regular guesthouses.

Different approaches to assessing the sharing economy was discussed. The initial intention was to study, from the literature, which data is being collected and maintained by the sharing economy platforms, but this was not possible due to lack of available information in the literature. The selected approach was to see how financial transactions data can be used for deriving indicators on sharing economy. Based on an internal survey among the WPG countries, a methodological study was performed to derive indicators. It was attempted to find indicators that could be compared omong the WPG countries, but this turned out impossible to do given the available data. The end of the sharing economy part is a discussion of the underlying problems connected to measuring sharing economy (in the operational definition of collaborative economy) compared to measuring the wider e-commerce. The collaborative economy part of e-commerce is difficult to separate from the rest of e-commerce. Another point is that financial transactions data on e-commerce is easier to integrate with other sources such as administrative data, and there are also freely available statistics on the amount of internet activity on commercial pages.

2. Conclusions

An inventory of the financial transactions data infrastructure was been performed for the different countries of WPG, and case studies were made to assess the potential of financial transactions data in official statistics. Also, the possibility of extracting information on sharing economy from the financial transactions data, has been investigated.

The infrastructure of financial transactions data is different in different countries although there are some similarities. In Portugal and Slovenia, the NSI can only access financial transactions data through the tax authorities. Both in Italy and Norway, banks, card companies and service providers are potential sources, but in Italy this access need to go through the National Bank. The legislation grants the The National Bank of Italy access to microdata, and they can produce aggregated data for Istat.

Due to the legislations, only in Norway and Portugal are micro-financial transactions data going to be available for the NSI. A good system for pseudonymisation of any identificatory variables is a requirement before accessing such data. Further, financial transactions data are widely regarded as sensitive and risk assessments are required by GDPR before such data are accessed. As an example, the society's need for official statistics must be balanced against the perceived burden of the persons whose transactions are a part of the financial transactions dataset.

A practical point is that Micro-financial transactions data from the complete population are very large data, a consequently the NSI must have sufficient server capacity both to store and process the data.

Since most countries can only access aggregated data, an important process is for the NSI to ensure that the data provider aggregates the data to a level that is useful for the NSI's further processing within the production process of the official statistics. This interaction between the data provider and the NSI is only one part of the the cooperation needed to agree on the data delivery to the NSI. A good cooperation is necessary also when the data access is granted by the legislation. The data providers must use resources on preparing data for the NSI, and for many countries the NSI cannot pay for the data. Thus, a positive cooperation between the NSI and the data providers is an essential part of the process to access data. Even when the cooperation works excellently, the access to financial transactions data takes some time. The EU NSI's Directors Expert Group on privately held data on how to overcome the critical issues related to the accessibility of financial transactions data, concludes that the NSIs should try to contact potential data providers directly instead of through intermediaries. The direct contact can substantially improve the level of co-operation and also improve the quality of the data being accessed.

The case studies in the application part of WPG show some of the potential of financial transactions data for official statistics. The Italian case studies use time series of total daily card transactions by aggregated mcc-categories for forecasting and nowcasting of economic indicators. This results in an improvement both of the quality and production time. The Portuguese case study looks at in which ways the tourism statistics can be improved by using financial transactions data to find information on lodgings being too small to be covered by the current tourism survey. The Portuguese case study also investigates how transport by the new alternatives to taxi can be measured. One challenge is how the data are organised at the Tax Authorities as well as how the Tax Authoritites pre-process the data before they are accessed by the NSI: it is not possible to identify those invoices that corresponds to cancelled purchases, since the credit note cannot be linked to corresponding the invoice, and therefore it is not possible to identify which invoices that are associated with a cancelled purchase. The Norwegian case study investigates to what extent debit card transactions at legal unit level can be used to estimate the monthly total retail sales by the ten industry classification groups used for the Retail sales index statistics. This case study had to use financial transactions data for only one day since it was impossible to access monthly data in time for the project, and this illustrates that it often is a complicated process to receive financial transactions data, and that often a long time has passed before the data are accessed and available for analysis within the NSI.

A result of the case studies is that the financial transactions data indicators or variables can be used as predictors in models for the near future, but the incorporation of financial transactions data into the production of official statistics needs to be better explored. For testing the models, larger databases being more complete in terms of enclosed activities (firms, groups etc.) are needed.

The timeliness of financial transactions data was relevant during the Covid-19 crisis for improving the forecast of some selected macroeconomic indicators (e-commerce, final consumption, retail trade index), especially when financial transactions data was broken down by Merchant category code. The reason for this is that different payment flows have registered different trends in the economic terms during the crisis.

The sharing economy is an increasing part of the economy. A framework has been made for extracting information on sharing economy from the financial transactions data through indicators on the sharing economy activity. Here, sharing economy was defined broadly as collaborative economy since a narrower definition is both less relevant for official statistics as well as more difficult to measure. The existence of relevant input data within each of the WPG countries for producing indicators, showed that it was not possible to produce indicators that could be compared between countries. In the project in ultimately became clear that that sharing economy (collaborative economy) often cannot be measured separately from e-commerce, and this suggests that future work on sharing economy indicators should concentrate on the whole e-commerce instead of narrowing down to collaborative economy.

3. Recommendations

Most NSIs have long experience in receiving administrative registers from the governmental authorities. The financial transactions are however mostly handled by private companies, and thus the potential data providers are then also private companies such as banks or private service providers. When the data provider is a private company, the challenges are different from the administrative data situation. Below are some recommendations for the NSI that wants to get access to financial transactions data

- The data provider is concerned about whether their data will be handled securely, especially since financial transactions data are usually considered as sensitive. Informing the data provider of the secure handling of data being performed within the NSI is an important part of the NSI's contact with the data provider. It could also be useful to inform the data provider of the role of the NSI. Governmental organisations such as the Tax Authorities, the National Insurancy Agency and the police are concerned about accessing data for auditing the economic activity of a single business or a single person. In contrast, the NSI is not interested in the contents of the transactions for the single units, but only to aggregate data for statistical purposes. The purpose of the NSI is to deliver aggregated numbers. The potential data provider may not be aware of this distinction between the NSI and other governmental institutions.
- The access should be authorised by a legal act, e.g. the Statistical act. If a legal act is not relevant, typically if the data are sufficiently aggregated, there should be be a formal agreement between the NSI and the potential data provider. An institutional agreement between the NSI and the National Bank typically is appropriate, since the National Bank in many countries has access to more data than the NSI**.**
- A good cooperation through a good relationship is essential for obtaining financial transactions data, both in the case of autorisation by a legal act and in the case with a formal agreement. It is always more tempting for a potential data provider to use resources on data access for the NSI if there is a positive relationship with the contact persons in the NSI. This cooperation is essential since the data provider's need to use resources in preparing and transferring data for the NSI as well as in explaining the metadata. The NSI usually cannot pay for these expenses and should acknowledge the burden that has been put on the data provider. The NSI should tell the data provider about the data provider's contribution to the improved official statistics for the society as well as about any reduced response burden from surveys that can be the consequence of the new data source. As an example, the data provider itself or its associated businesses may face reduced response burden. In addition, the NSI might be able to help the

data provider in its metadata description, development of data infrastructure, or in data anlysis which all might be useful also for the data provider's own usage of the data in question. One solution to increase the NSI's ability to help the data provider, is to let some NSI staff work on e.g. metadata description for a period at the data provider's premises without cost for the data provider. This leads to a useful knowledge transfer both from the NSI to the data provider and vice versa.

- A thorough study of metadata in close dialogue with the data provider is an important task to perform as early as possible in the process. The variable descriptions should be studied, but the nature of the unit should be identified as well as coverage and known quality issues. Based on the metdata studies, the data may turn out not to be relevant for the official statistics after all, either because of undercoverage or overcoverage of the dataset, lack of important variables, missing values on important variables for many units, or that the data are aggregated in a way that destroys the usability for the targeted official statistics. An example of the latter is the data from the clearing and settlement system which for some countries means that many transactions are aggregated to the net transfer between the banks, and then the information on the type of payer and receiver is missing. Since the process of getting data usually takes a lot of calendar time as well as personnel resources within the NSI, a thorough study of metadata can save a lot of time, e.g. by termating the process if it turns out that the data will be not relevant after all.
- The appropriate aggregation of data, when made outside the NSI's premises, should be studied carefully through metadata studies, since otherwise the data may lose some of their usability, c.f. the previous bullet point. For the same reason, the level of aggregation should also be carefully agreed upon with the the institution performing the aggregation.
- The chosen aggregation of the variables in the dataset is of central value when it comes to the possibilities of calculating statistics for the relevant groups instead of for less relevant groups. One example is the use of merchant category codes (MCC) for indicators on sharing economy. Some merchant category codes are very detailed, but they do not fully match the industry classification codes used by the NSIs. For international comparisons of sharing indicators, it is important that the mcc codes are aggregated in a manner that allows for such comparisons.

Role of ESS in enhancing the use of financial transactions data

- The definition of a new regulatory and methodological context for the NSIs' use of privately-held data is necessary for an effective use of financial transactions data for statistical purposes. The ESS should try to obtain a framework agreement between the NSIs of ESS and the most important companies managing financial transactions in Europe. A relevant starting point could be to target the network of financial institutions currently negotiating an agreement on an EU-wide debit card[4].

Financial transactions data should be a part of the priority areas that Eurostat has identified for exploring potential options for public-private cooperation[5]. A role for financial transactions data should be found also in the definition of the European Data Spaces designed by the European Commission to

[4] https://www.ecb.europa.eu/press/pr/date/2020/html/ecb.pr200702~214c52c76b.en.html

[5] So far four priority areas have been identified for action: Mobile network operators, Digital platforms (mostly related to tourism statistics), Trusted Smart surveys and Smart traffic.

enhance the access to privately-held data via industrial data platforms[6]. On such platforms, the financial sector businesses could find a secure and trusted environment to share their data based on voluntary agreements. Concerning data processing, a solution is that the data providers perform the data processing using a programming code that has been developed by the NSI. This can be benefitial both for the data provider that can use less resources, as well as for the NSI that can ensure a data processing that is

[6] Planned activities include data spaces in the domains of industrial (manufacturing), Green Deal, mobility, health, financial, energy, agriculture, public administrations, skills and open science.

2.7 Earth Observation *(full information on WPH available here on CROS)*

This WPH pilot project aimed to support areal statistics with Earth Observation (EO) data. Project results in experimental statistics using remote sensing data. From the technological point of view, the WPH uses new methods like machine learning algorithms for image analysis. The crucial goal of the WPH is the usage of the EO data from different sources that will contribute to building the geospatial framework to support the statistical registers. Within this project, the usefulness and practical usage of EO data to fill the gap between statistical and geographical information named "geospatial breakdown" are proposed. The main objectives of WPH are implemented by the execution of nine case studies (CS) divided into thematic fields: agriculture, build-up area, land cover, settlements, enumeration areas, and forestry.

The expected products of WPH can be grouped into three categories: statistical indicators, basic research and product supporting existing systems for statistical production (table 1**Fout! Verwijzingsbron niet gevonden.**).

Table 4. Classification of expected products within WPH in reference to each case study.

Thematic field	Case study	Main specification of final product		
		Statistical indicator	Basic research (experimental and theoretical research) = analysis	Support/improvement of existing systems for statistical production
Agriculture	CS1	+		
	CS2	+		
	CS3	+		
Build-up area	CS4	+		
	CS5	+		
	CS6		+	
Land cover	CS7	+		+
	CS8	+		+
Settlements, Enumeration Areas and Forestry	CS9	+		+

Statistical indicators can be defined as the area under cultivation and cropped area (CS1, CS3), the proportion of agricultural area under vegetation cover in winter – SDG indicator 2.4.1 (CS2), the average share of the build-up area of cities that is open space for public use for all, by sex, age and persons with disabilities – SDG indicator 11.7.1 (CS4), Weighted Urban proliferation index (CS5), land change (CS7), indicators of the evolution of the area occupied by the Eucalyptus and a related indicator concerning the susceptibility of the forest to fire (CS9). Basic research is the output of the CS6. Through the combination of different data sources, a statistical analysis of how the urban quality of life differs between socio-economic groups can be conducted. The collected geographic data in CS6 is output as well, which can be used as a basis for further research and statistical production. In the case of the CS8 the research goal is to design and develop an automatic Land Cover (LC) estimation system. Such a

system should be able to take as input a satellite image depicting a portion of territory and to return as output a table of LC statistics.

Three deliverables were produced during the project:

- Interim Technical Report (Deliverable H1) concerning the description of data sources, state of the art, statistical product definition, test site with collected data, methods of data preparation adequate to each case study
- Report on results and activities executed by case studies (Deliverable H2) concerning the detailed description of research including data processing, data analysis, and conclusions.
- Final Technical Report (Deliverable H3) concerning the evaluation of big data sources and definition of possible statistical products from examined big data sources; business architecture suitable for big data processing; methodological framework; quality framework with the protection of privacy and confidentiality and other legal issues; IT infrastructure.

All reports are available on the website:

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

1. Results in the agriculture thematic field

The agriculture thematic field was executed by three case studies. First involved crop recognition, mapping, and monitoring using radar and optical satellite data in Northern Europe conditions (Case study 1). The second involved monitoring the off-season vegetation cover of agricultural soils in high-latitude agricultural systems using radar and optical satellite data (Case study 2). The third was about crop recognition with very high-resolution aerial data (Case study 3).

Results and conclusions of crop recognition, mapping, and monitoring:

- Crop identification using time series of radar Sentinel-1data, supported by optical Sentinel-2 data in crop recognition and mapping with high accuracy was achieved.
- The highest accuracy of single crops classification was achieved for Artificial Neural Network (ANN) 0.89, Random Forest (RF) 0.88, and Supported Vector Machine (SVM) algorithms 0.88. A slightly smaller overall accuracy is for Decision Tree (DT) 0.81 and K-Nearest Neighbour (KNN) 0.85. Considering classification based on aggregated crop types the overall accuracy increased 4% for ANN, RF, SVM, 5% for KNN, and 6 % for SVM.
- Random Forest or Artificial Neural Network is recommended for pilot production. In this case study, RF was selected due to the relatively smaller sample dataset requirements than for ANN algorithm. Since different types of cereals aren't satisfactory distinguishable the use of aggregated classes was chosen. Based on the classification result, the area of crops was estimated.
- The presented case study shows a huge potential of using EO data in agricultural statistical production due to easy data access, global coverage, and reliable results.

- One of the limitations can be complex processing and 10 m spatial resolution causes small parcels elimination. Despite this global estimation of crop groups area is possible.

- The distinction between some crops is hard due to the similarity in a plant structure (i.e cereals) taking into account the method of obtaining data (i.e satellite radar).

- The presented methodology can support agricultural statistics in further projects including crop yields and growth models.

Results and conclusions of monitoring the off-season vegetation cover:

- In all classification scenarios using only Sentinel-2 data provides almost as good performance as a fusion of Sentinel-1, Sentinel-2, and soil data. However, including Sentinel-2 and soil data does improve performance. Thus, in the best scenario, the authors would use all three data sets for classification. In reality, due to cloudy conditions, not all parcels of interest have any Sentinel-2 data in a narrow time window in spring. In these cases, we need to settle for Sentinel-1 and soil data only. Note that the performance with Sentinel-1 and soil data only in bare soil classification is still quite close to the best performing 'all data' scenario.

- The classifier performs nearly equally well for both classes; 86% correctly for bare soil, and 82% correctly covered soil.

- Integrated Administration and Control System (IACS) data can stand as a sole source to produce statistics on off-season vegetation cover. IACS data becomes available at the end of June each year. The uptake of remote sensing data to produce near real-time statistics on off-season vegetation cover seems very much feasible. The prediction model performance is quite high, especially when detecting bare soil conditions.

- The whole processing stack from data preprocessing to the statistical product can be automatized and it is highly manageable in the current IT environment. Authors estimated that the new indicator can be published within 7 days from the latest Sentinel image acquisition.

Results and conclusions of crop recognition with very high-resolution aerial data:

- Satellite data of intermediate resolution like a Sentinel are freely available at the required high frequency, while high-resolution aerial photography data can only be provided annually (25 cm - in winter, moreover), tri-annually (40 cm) or 10-yearly (10 cm), a temporal resolution unsuitable for detailed crop recognition.

- A possible alternative to accessing freely available high-resolution aerial images of sufficient frequency might be to produce these via aeroplane or drone (UAV, Unmanned Aerial Vehicle), possibly partnering with other public agencies or research institutes to lower the cost.

- Aerial photography datasets seem too expensive to be used exclusively but they could be integrated into a three-tier mixed-mode statistical production consisting of satellite data analysed via AI algorithms, trained and supplemented by survey and administrative data, as the 'normal' and usual way to recognise crops, aerial photography data for the limited and

clearly defined areas with small parcels and a wide range of high-value-added crops, and surveying in those areas where the use of drones is not allowed.

- Because the first subtask, assessing the data situation, concluded that high-resolution aerial photography data are not readily available with the required frequency, the next foreseen stage of testing machine learning methods to analyse the data could not be commenced. It was, for the time being, limited to a first look into the data science capabilities of the various partners to address the research question (Statbel, Statistics Flanders, VITO, Geocounter).

2. Results in the build-up area thematic field

The build-up area thematic field was executed by three case studies. First aimed to implement the UN methodology to compute the SDG indicator 11.7.1 (Case study 4). Second case study concerned a method of UN SDG indicator 11.3.1: Urban Sprawl estimation by using Earth Observation data(Case study 5). Third one aimed to investigate the benefit of combining earth observation data with official statistics (Case study 6).

Results and conclusions of implementing SDG indicator 11.7.1:

- The cities boundaries were obtained by pixel clustering for 31 urban units in France with more than 200 000 inhabitants.
- Open public spaces and streets delineations within cities are more tricky as no sources can provide accurately all the open public spaces. For this reason, the area types of interest should be clearly stated and data from various sources should be combined.
- Three indicators quality scores for open public spaces were calculated and normalised by the highest score reached among all cities.
- There is a small positive correlation between the indicator and the normalized quality score: cities with the lowest value of the indicator has also the lowest values for the quality score.
- Some cities with a low-quality score have a high value of the indicator like Le Mans or Le Havre. Moreover, most cities have a low-quality score which means that there is great uncertainty over the final results and the need for deeper analysis in the future.

Results and conclusions of urban sprawl across urban areas in Europe:

- Overall accuracy was 0.89 for the best model on the test set, an overall accuracy of 0.83 on the blind test area. Recall, precision, and F1-score were also measured per category. While overall performance was satisfactory, the variations for these measures per land cover category were quite large.
- The model especially had difficulties identifying forest/semi natural areas and water bodies. However, the main goal of the model was distinguishing between build-up area and other areas, which the model seems to be able to do sufficiently.

- A method for estimating urban sprawl using the weighted urban proliferation index (WUP); a metric evaluating the build-up area, its dispersion, and the uptake of built-up area per inhabitant was done, but further validation using domain experts is still needed.

Results and conclusions of a combination of official statistics and Earth Observation data to determine the quality of life:

- Due to its high spatial and temporal resolution, earth observation data can be used to measure certain aspects of environmental quality of life such as air quality, urban heat islands and urban green.
- Different aspects concerning the environmental quality of life were identified through reviewing the quality of life initiatives, discussions with earth observation experts and a literature review on determining the quality of life using remote sensing data.
- A literature review of the identified topics (environmental surroundings indicators like temperature and vegetation) was conducted. Based on the identified aspects, possible data sources have been investigated, specifically for the identified topics together with further data sets that could be linked to a geocoded survey.
- This case study shows that even if the concrete correlation results do not exactly match the theoretical hypothesis from literature, there is a great potential for new and more granular data analysis regarding the aspect of spatial resolution. Further data on the environmental living conditions (e.g. air quality and noise pollution) would enhance this kind of analysis for an evidence-based democratic decision-making process concerning diverse socio-economic groups of the population. This can be achieved through the combination of earth observation data with official statistics.

3. Results in the land cover thematic field

This thematic field was executed by two case studies. First focused on comparing «in-situ» and «remote-sensing» collection mode for land cover data (Case study 7). Second involved land cover maps at a very detailed scale (Case study 8).

Results and conclusions of comparing «in-situ» and «remote-sensing» collection mode for land cover data:

- The method designed for detecting land cover changes combines two levels of observation: pixel and buffer. It allows adapting the change detection strategy according to the land cover class studied and the level of accuracy desired. As the values of the OSO maps are extracted at the pixel and buffer scale for each point of the Teruti grid, it is indeed possible to switch between the 2 levels and compare them. In the same way, as the land cover classes confidence and the calculated heterogeneity indexes are available for each point, it is possible to adjust the change detection criteria and make them more specific in order to improve the accuracy of the results.

- Over the 2017-2019 period of time, approximately 90% of the Teruti points selected by the land cover change detection method are not surveyed in the field but imputed by external geographic databases. Over the 10 % remaining points, the land cover transitions breakdown is : 26% deforestation, 21% agricultural intensification, 19% agricultural abandonment, 13% revegetalization, 7% reforestation, 5% disartficialization, 4% de-vegetation, 4% artificialization, 1 % waterflood and 0.5% dewatering. NUTS2 regions with the highest number of changing points per km² were Corsica and Bretagne, with 0.25 and 0.21 points/km², respectively. And the NUTS2 regions with the least number of points detected per km² are Île-de-France, 0.07 points/km², and Hauts-de-France, 0.08 points/km².

- The statistical land cover areas breakdowns are comparable both nationally and regionally. Gaps are often very localized and easily identifiable. Out of the 2 majority classes, Cropland and Woodland, OSO seems to underestimate the areas compared to Teruti. On the contrary, OSO provides larger estimates than Teruti for agricultural grassland, shrubland and Urban areas.

- In 2020, Teruti's investigators will return to the previously surveyed points three years ago in 2017. The 2020 collected data will thus constitute the ground truth for land cover changes between 2017 and 2020. These results will be used to adjust the parameters of the change detection method designed from the OSO map. Confidence levels in the OSO land cover classes and the combination of pixel and buffer approaches will be adjusted to improve the relevance of detected land cover changes.

- For the Teruti 2021 and subsequent surveys, the changes in land cover detected from the OSO map will allow better targeting of field observations. They should eventually form a new stratum of the Teruti survey design.

Results and conclusions of land cover maps at a very detailed scale:

- Istat automated LC estimates had remarkably good accuracy for most LC classes, except for a systematic upward bias affecting narrow linear structures like rivers and highways. This overestimation issue turned out to be inextricably linked to the adoption of EuroSAT as a training set and to Istat tile-based classify-and-count approach.

- The new integrated architecture (CNN + U-Net) works very well for all LC classes 1) The U-Net takes care of LC classes “River” and “Highway” 2) The CNN copes with all the other LC classes 3) Partial LC maps produced by 1) and 2) are merged to yield a final complete LC map

- Results seem very promising. Output LC maps are detailed and accurate. Output LC statistics are sound

4. Results in the settlements, enumeration areas and forestry thematic field

The settlements, enumeration areas and forestry thematic field were executed by one case study (Case study 9) divided into two topics. Topic 1 aimed to update the Settlements and Enumeration Areas integrated into the INSPIRE Theme Statistical Units dataset by redefining the geography of the statistical subsections or census tracts to the next 2021 Census. Topic 2 involved exploration the possibility of studying the forest and the eucalyptus plantation and its impact in preventing forest fire.

Results and conclusions of update the INSPIRE Theme Statistical Units dataset, namely the Settlements and Enumeration Areas:

- High-resolution layers produced by satellite imagery and methodologies of processing earth observation data can be included in the production of official statistics as well as in the national statistical production framework since these data sources and related methods present scientific support, simple data access, data availability and regular periodicity and global scope, therefore supporting traditional statistical procedures and reducing the gap between spatial and statistical data.

- The Global Human Settlement Layer (GHSL) data based on built-up area and other Copernicus high-resolution datasets can produce statistical indicators with multi-level dissemination which subsequently can support and monitor the INSPIRE Theme Statistical Units dataset, as well to build the geospatial framework to support 2021 Census.

- The use of Google Earth Engine (GEE) to create indicators showed that with relatively little work and requirements for IT architecture it is possible to use publically available data indicative for urban growth for the creation of indicators showing where new urban growth has developed. GEE offer much quicker times to create zonal statistics than using conventional GIS, for example, the Zonal Statistics function in ArcGIS takes a long time to execute and needs to download data.

- The statistical products developed constitute a very important auxiliary data to support the delineation of Enumeration Areas and Settlements for the 2021 Census. Although these data do not have a spatial or geometrical methodological feature to create the new geography based on the earth observation data regarding the human presence, they will be integrated into the updating process of the delineation of the Enumeration Areas for the next 2021 Census.

Results and conclusions of exploring the possibility of studying the forest and the eucalyptus plantation and its impact in preventing forest fire:

- The methodology used by this work can help monitor the evolution of this species in our country since it can provide data faster than the methodologies that use photo interpretation as the Land Cover chart from the Directorate-General of the Territory.

- Training a model to do image classification saves time and costs in scanning and data collection. It also allows you to expand your analysis using the model with different data sets and in different locations, using images with the same characteristics as those used in modelling.

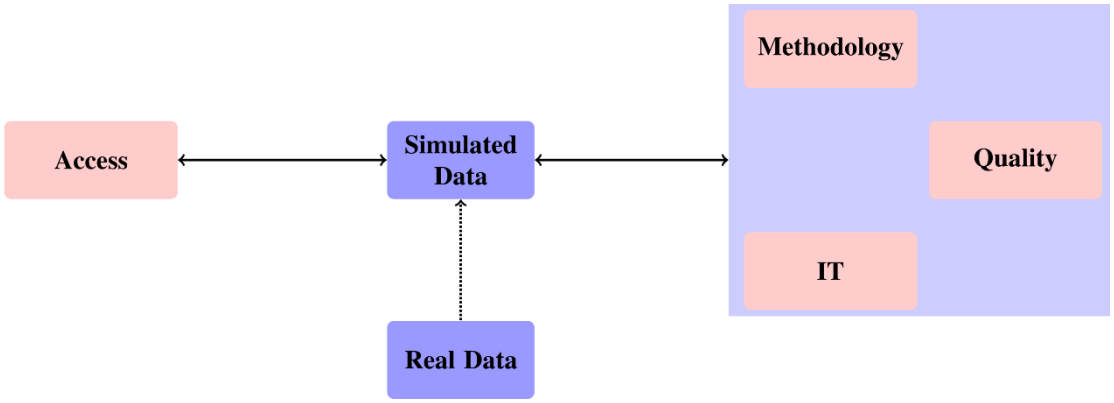
- These models require a large number of training samples, without these training samples they do not produce quality inferences, which may hinder their implementation.

2.8 Mobile Networks Data *(full information on WPI available here on CROS)*

1. General Objectives and Approach

WPI on Mobile Network Data is devoted to the incorporation of mobile network data, i.e. the digital trace of human mobility in mobile telecommunication networks into the production of official statistics. Two main circumstances have strongly conditioned the development of this work package. Firstly, access to this data source is an intricately complex issue and is not granted even for research purposes in normal conditions (hence the new inclusion as a pilot in this second edition of the present ESSnet). Secondly, the specific goal in the project call is "[d]evelopment of an adequate Reference Methodological Framework […] inspired by the work on the Reference Methodological Framework for processing Mobile Network Operators data for official statistics" proposed by Eurostat.

In this line, the ESSnet on Big Data followed the natural course of action of getting access to data, developing the statistical methodology, implementing this methodology into software tools, assessing the quality, and foreseeing extensions to more products in multiple statistical domains. This linear approach, however, was immediately blocked due to the lack of access to data. To mitigate this paralysing situation, we have followed a modular approach detaching the issue about the access to data from the rest of aspects such as methodology, IT tools, and quality. To achieve this we decided to develop a data simulator producing synthetic network event data, so that the research could progress in this limited, but hopefully useful, scenario so that when finally access to real data will be granted, then the production framework will be partially ready and developed (see figure).



At the same time, since this is the second time that this data source is object of research in the ESSnet on Big Data, we committed ourselves to come as close as possible to the results of the Implementation Track, and especially to the development of the production architecture produced by the Work Package F, namely the Big Data Reference Architecture and Layers (BREAL).

All in all, WPI has produced eight deliverables dealing with different complementary aspects of the use of mobile network data for the production of official statistics:

i. Deliverable I1 on Access, focusing on key aspects to have access to mobile network data from Mobile Network Operators (MNOs).
ii. Deliverable I2 on the Data Simulator, focusing on the development of a network event data simulator.

iii. Deliverable I.3 on Methodology, focusing on different aspects of the process such as geolocation of mobile devices or inference to the target population proposing statistical methods to deal with them.
iv. Deliverable I.4 on Software Tools, focusing on prototyping software implementations of the aforementioned statistical methods.
v. Deliverable I.5 on Structural Metadata, focusing on the construction of a glossary of terms introducing the most relevant concepts related to the underlying telecommunication technology.
vi. Deliverable I.6 on Quality, focusing on a first application of the BREAL to an end-to-end production process instance from to raw telco data to final statistical estimates.
vii. Deliverable I.7 on Experiences with Real Data, gathering very relevant national experiences with some limited data sets.
viii. Deliverable I.8 on Visualization Tools, focusing on the development of visualization tools specific for this data source.

Next, we provide an overview of the main outcome from each deliverable.

2. Main Outputs

Deliverable I.1 on Access

This deliverable focuses on the intricate issue about the access to mobile network data. In general, statistical offices of the Member States of the ESS do not have access to this data source for long-term production of official statistics and the collaboration with MNOs is reduced to research analyses under restricted and limited conditions in the optimal cases. This situation was indeed made explicit already during the ESSnet on Big Data I and the parallel work by the ESS Task Force on Big Data/Trusted Smart Statistics. For the ESSnet on Big Data II we have concentrated on collecting the direct feedback from MNOs regarding the issue about the access in the different country members of this WP and proposing some potential collaboration scenarios between NSIs and MNOs to be explored.

To get the feedback from the collaborating MNOs we produced guidelines to hold a direct dialogue with experts and different staff from these companies. The motivation was to clearly and explicitly identify the obstacles to reach a sustainable access to data to produce official statistics in a routinely basis. These guidelines focus on the more potentially controversial aspects to have access to data. The feedback varies greatly from country to country and even from one MNO to another. By and large, a revision of the legal framework is apparently needed, an alignment of public and private interests needs further investigation, and compensation costs and investment return also need further consideration.

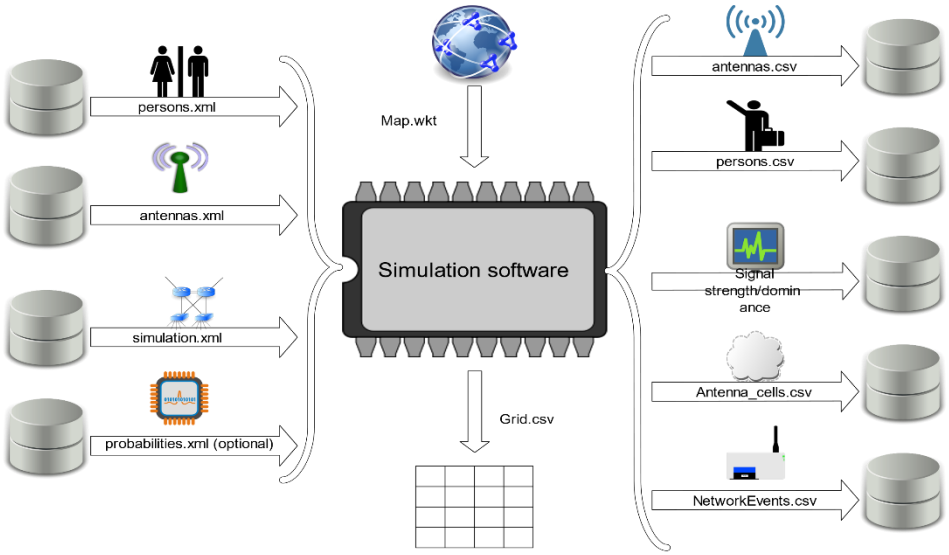
We propose three potential scenarios. Firstly, we assess the possibility of imposing a mandatory access by Law either to individual data at the device level or to aggregated data processed by the MNOs. Secondly, we consider the option of establishing commercial relationships, thus paying through a commercial contract for the aggregated data. Finally, we consider building a win-win partnership between NSIs and MNOs. We argue that a collaboration will foster the development of a promising market in the benefit of both producers and users, and society in general.

In summary, further work within the ESS in close collaboration among statistical offices and MNOs is needed. We argue that the role of NSIs cannot be reduced to acquiring aggregated data produced by MNOs for commercial purposes with an undisclosed methodology. A two-way feedback between methodological developments and access conditions are in our view advisable to reach an optimal collaboration scenario. The European dimension to achieve standard solutions and common approaches is also underlined.

Deliverable I.2 on the Network Event Data Simulator

Following the strategy introduced above, to mitigate the lack of access even for research we have developed a data simulator producing synthetic network event data, thus allowing us to test methodological proposals and develop prototyping software solutions.

The simulator is designed and built in a highly modular fashion so that the user can specify the geographical territory under analysis (as a map), the target population together with its movement pattern, the population of mobile devices, and the telecommunication network and its configuration. The software outputs not only the simulated network event data but also the synthetic ground truth (real positions of each individual) so that the analyst can assess the performance of the different statistical methods.



It is important to underline the instrumental character of this software. Firstly, from a methodological point of view, it is expected that any statistical method devised for real production conditions should also work properly under the simplified conditions in a synthetic scenario like that provided by the simulator. In this sense, it is advisable to make the software grow with more diverse and realistic features (like more event network variables, more complex movement patterns, the introduction of usual environments for individuals, etc.). Secondly, the use of synthetic data makes us free about the restricting constraints about privacy and confidentiality of real data. Furthermore, it allows us to undertake analysis of reidentification of individuals even for aggregate data in different scenarios.

Finally, from a strategic point of view, the analyses potentially undertaken relating accuracy of final estimates and access conditions (e.g. which network event variables to use?) will be helpful in unravelling some aspects about the access to real data.

Deliverable I.3 on the Statistical Methodology: the Reference Methodological Framework

This deliverable embraces the statistical methods proposed to substantiate the ESS Reference Methodological Framework. We focus on building an end-to-end statistical production process from the raw telco data to the final estimates about the target population. To be more specific, in our proposal we are considering the estimation of present population counts and associated origin-destination matrices. We use instrumentally the data simulator to illustrate and test the proposals.

It is absolutely essential to underline that we adopt a modular approach for the design of the statistical process. The different modules deal with different aspects of the complex process going from the geolocation of mobile devices to the final inference step connecting the dataset with the target population. The modularity will allow us to cope with the complexity of the process and to reuse the modules for multiple statistical domains according to the ESS Reference Methodological Framework.

Each module is characterised by its input and output data and the throughput process. As an important design decision we take probability distributions of different quantities along the process as the input/output information objects of each module. The use of probability distributions is motivated because they provide the most adequate mathematical treatment of inference, they allow us to rigorously deal with uncertainty and propose point estimates and accuracy indicators in a natural way, and make it possible to integrate external data through the use of priors, posteriors, and hierarchical models in a seamless way. Furthermore, using these probability distributions, we can integrate the sequence of modules as a production chain using the total probability theorem:

$$\mathbb{P}(z_{out}|z_{in}) = \int dz_1 \int dz_2 \cdots \int dz_N \, \mathbb{P}(z_{out}|z_N) \cdots \mathbb{P}(z_2|z_1)\mathbb{P}(z_1|z_{in})$$

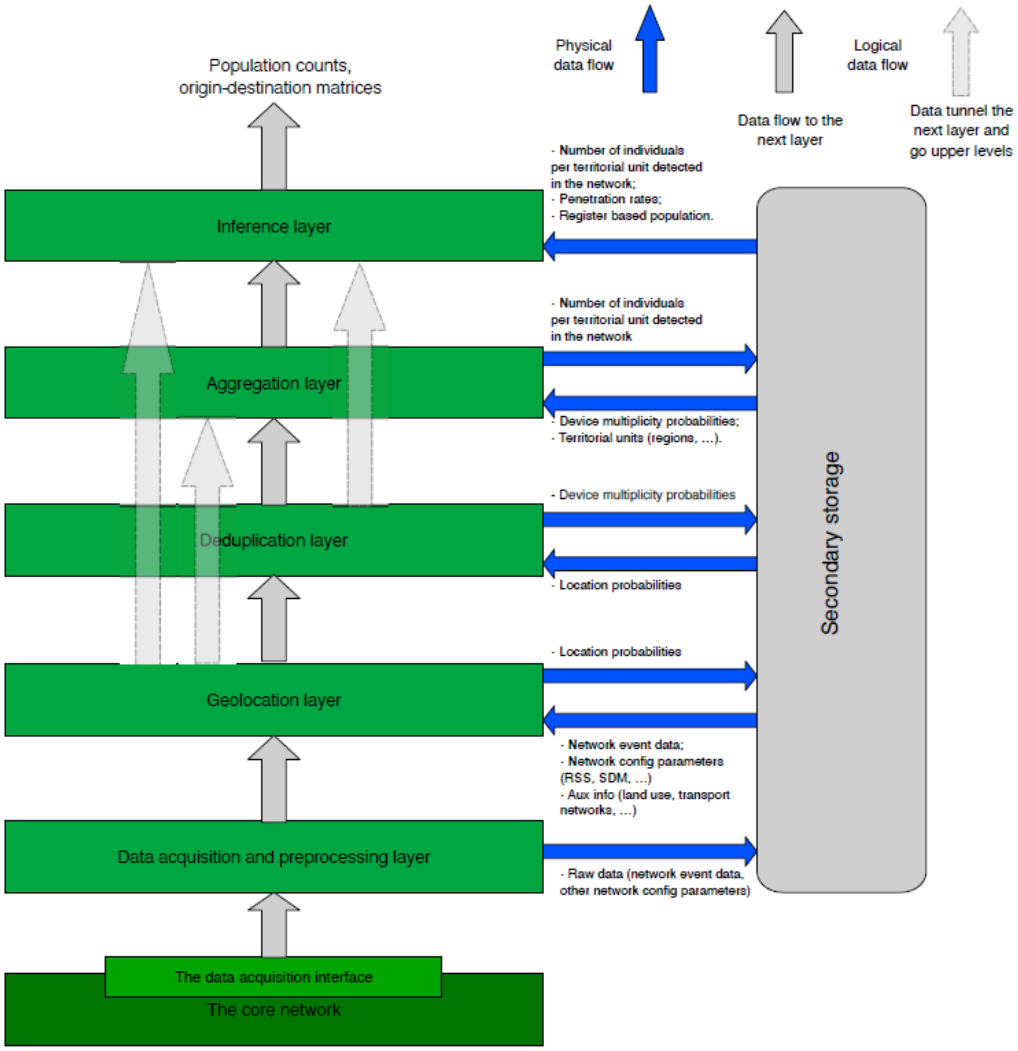
The modules proposed so far are:

- Geolocation of mobile devices, focused on providing location probabilities of each device over a grid of reference at each time instant.
- Deduplication, focused on providing duplicity probabilities for each device to belong to an individual either with two or one devices.
- Statistical filtering, focused on constructing algorithms to identify devices belonging to the target population of analysis. We have only provided a general approach for this module, since we need more complex movement patterns in the simulator. For our present population business case, it is not needed.
- Aggregation, focused on computing probability distributions of aggregate population counts for the population detected by the network.
- Inference, focused on computing probability distributions of aggregate population counts for the whole target population over the reference grid at each time instant.

Deliverable I.4 on Software Tools

This deliverable provides a detailed description of the software stack developed to implement the preceding statistical methods. Each production module is implemented as an independent R package with the corresponding interface (input/output). This deliverable is naturally complemented with the corresponding source code for each R package, which is fully shared in an open-access public repository[7].

All packages follow the principles of modularity and abstraction, have vignettes and reference manuals, and are illustrated using end-to-end examples with synthetic data generated from the simulator (thus, compared to the simulated ground truth). They are computationally intensive and use parallel computing techniques.



[7] https://github.com/MobilePhoneESSnetBigData.

Each of the preceding layers (see figure) has been implemented in the following R packages:

- geolocation, implementing the dynamical approach to the geolocation of mobile devices using Hidden Markov Models (HMMs). The package makes use of the Rcpp package for those functions with higher computational burden.
- deduplication, implementing three proposals to identify pairs of devices belonging to a same individual.
- aggregation, implementing the method to compute the probability distribution of the number of individuals detected by the network aggregating the location probabilities of the mobile devices.
- Inference, implementing the method to compute the probability distribution of the number of individuals of the target population from that of those detected by the network and auxiliary information.

Scalability issues have not been investigated for real data volume, but the design includes consideration to further explore these issues in a Big Data platform.

Deliverable I.5 on Metadata

This digital data source is characterized, among other things, by the remarkable technological complexity arising from the telecommunication industry. In order to set up a standardised statistical process for multiple statistical domains, structural metadata must be provided to clearly understand the different concepts with accessible definitions so that their roles are optimally assigned in the process. This will favour the communication between statistical offices and MNOs.

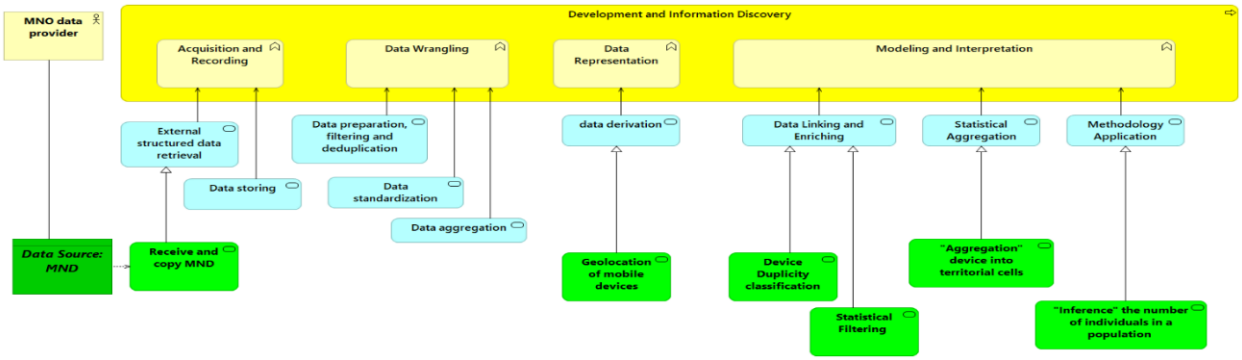
In this line, we have created a glossary of terms following the modular structure of the designed end-to-end process and, in particular, completely aligned with the ESS Reference Methodological Framework. More specifically, terms and concepts have been identified and defined for input data, for intermediate data, and for output data, thus providing their corresponding metadata. Following the international standards (GSIM, in particular) the glossary is built providing the name/lemma, acronym, and definition of each concept.

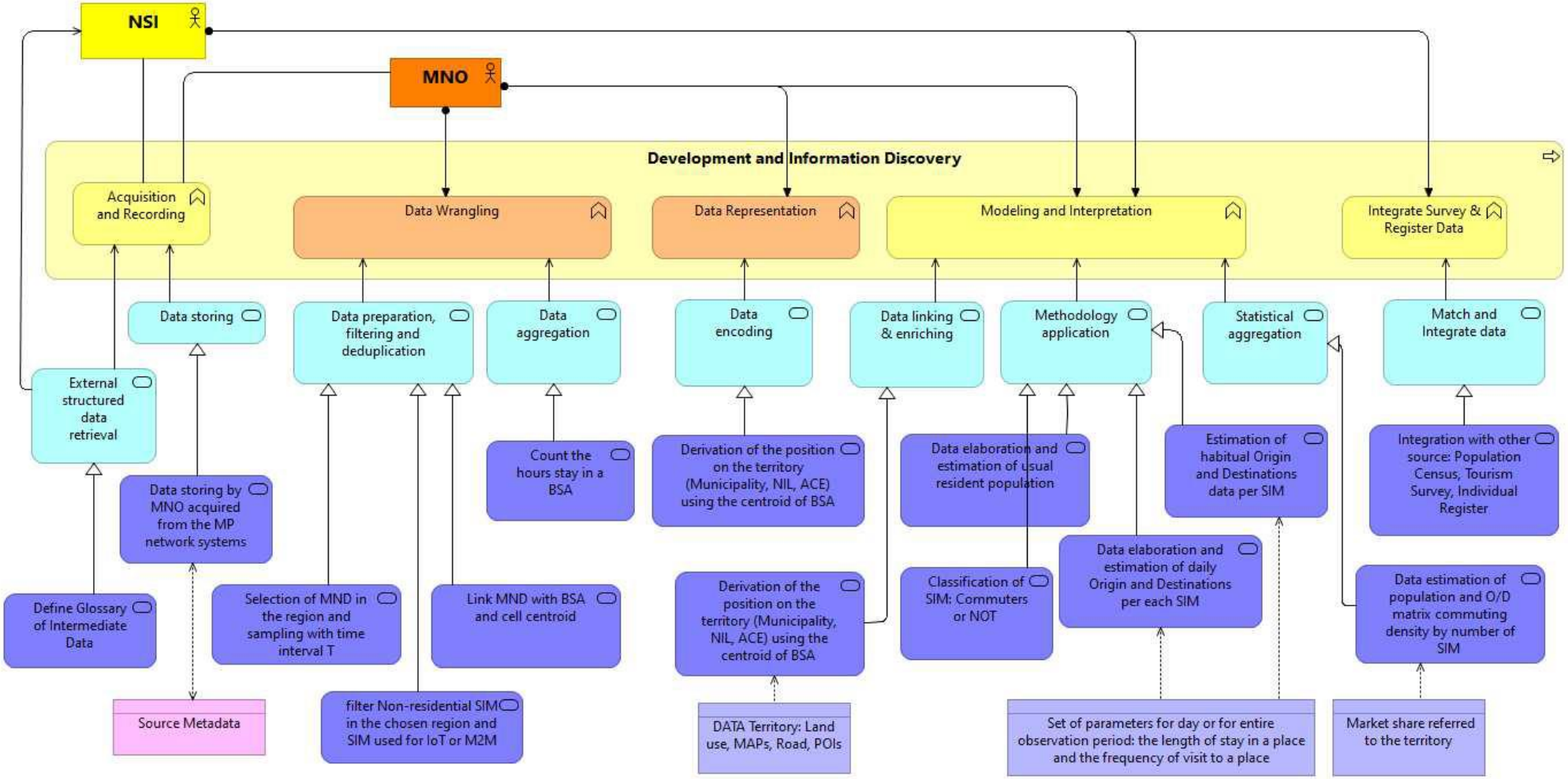
The deliverable provides both a detailed explanation of the structure of the glossary and the whole glossary as an appendix. An online version has also been created.

Deliverable I.6 on Quality

This document is focused on some quality issues (more specifically included in WPK on Methodology and Quality, where this data source has been used as an assisting element) and, more concretely, on the use of the BREAL architecture developed by WPF to describe a specific instance of the statistical process with mobile network data to compute origin-destination matrices.

In this deliverable, the BREAL model has been proved to be easily adapted to describe the different methodological steps of the whole process both identifying business functions and application services (see figures below).


MNO data provider
Development and Information Discovery
Acquisition and Recording
Data Wrangling
Data Representation
Modeling and Interpretation
External structured data retrieval
Data storing
Data preparation, filtering and deduplication
Data standardization
Data aggregation
data derivation
Data Linking and Enriching
Statistical Aggregation
Methodology Application
Data Source: MND
Receive and copy MND
Geolocation of mobile devices
Device Duplicity classification
Statistical Filtering
"Aggregation" device into territorial cells
"Inference" the number of individuals in a population


NSI
MNO
Development and Information Discovery
Acquisition and Recording
Data Wrangling
Data Representation
Modeling and Interpretation
Integrate Survey & Register Data
Data storing
Data preparation, filtering and deduplication
Data aggregation
Data encoding
Data linking & enriching
Methodology application
Statistical aggregation
Match and Integrate data
External structured data retrieval
Data storing by MNO acquired from the MP network systems
Count the hours stay in a BSA
Derivation of the position on the territory (Municipality, NIL, ACE) using the centroid of BSA
Data elaboration and estimation of usual resident population
Estimation of habitual Origin and Destinations data per SIM
Integration with other source: Population Census, Tourism Survey, Individual Register
Define Glossary of Intermediate Data
Selection of MND in the region and sampling with time interval T
Link MND with BSA and cell centroid
Derivation of the position on the territory (Municipality, NIL, ACE) using the centroid of BSA
Classification of SIM: Commuters or NOT
Data elaboration and estimation of daily Origin and Destinations per each SIM
Data estimation of population and O/D matrix commuting density by number of SIM
Source Metadata
filter Non-residential SIM in the chosen region and SIM used for IoT or M2M
DATA Territory: Land use, MAPs, Road, POIs
Set of parameters for day or for entire observation period: the length of stay in a place and the frequency of visit to a place
Market share referred to the territory

The description of the process according to BREAL is conducted for a real process jointly designed and implementation at Istat (the Italian National Statistics Office) by statistical officers and MNO experts. This description is fully aligned with the rest of deliverables and proposals and, more specifically, with the ESS Reference Methodological Framework, thus providing empirical evidence of this approach to build an adaptable and versatile statistical process using mobile network data.

Deliverable I.7 on Real Data

The ultimate goal of the design and implementation of a statistical process is to produce outputs using real data. In this sense, this document contains the current experience of four partners of the WPI with real data, namely that of Destatis (Germany), Istat (Italy), CBS (The Netherlands), and INSEE (France). Regarding access not only is the situation different in each country, but also within the same country every MNO poses specificities and even along time the situation with a given MNO can change.

However, in this deliverable the focus is placed on the statistical outputs in real production conditions. In no case access for long-term production has been achieved and feasibility studies and research analyses have been at most conducted.

In the case of Destatis, only aggregated data have been analysed and processed complemented with auxiliary information from official data. Applications in population statistics, commuter statistics, and tourism statistics have been explored. The main conclusion is that the restriction to have only aggregated mobile network data for the research available limits the number and the possibilities of the feasibility studies noticeably. Hence, one of the biggest challenges in all the feasibility studies is and will remain the definition of a suitable mobile network data set for each use case, as Destatis has no unrestricted access to individual data. Especially in view of the limited adjustment possibilities and the almost non-existent influence on data processing, the determination of a suitable data set is extremely difficult. Without the urgently needed knowledge of the detailed methodological procedures at the MNOs, it is not possible to exploit the full potential of the data for each research question.

In the case of Istat, a detailed account of the mobility pattern in the province of Pisa using CDRs at the level of municipalities is included in the deliverable. The goal is the computation of origin-destination matrices, which is explained. At present, the localization based on the position of the antenna shows weaknesses that compromise the reliability of population estimates, as well as all the other statistics connected with this concept. This is ameliorated with close collaboration with the MNO using information about the Best Service Areas. Considerations for further precision in the localization are agreed for the future. A detailed account of the composition of the Data Protection Impact Assessment is discussed, which needs close agreement and approval by the National Data Protection Agency. The need for further research on input privacy is underlined.

In the case of CBS, a pioneering end-to-end statistical process from raw telco data to final present population estimates is described in detail. It is fully based on the static approach making use of Bayes' theorem, as explained in deliverable I.3 on methodology. A detailed account of every element of the methodology is included (grid tile, radio cell, auxiliary data, prior probability, event location, posterior probability,…) and how they are processed. A visualization dashboard has been developed to share the final outputs. More research is needed in the calibration step together with the use of auxiliary data sources.

In the case of INSEE; the first experimental present population statistics for France is presented with a rich and interesting discussion of the many aspects involved in the process (geolocation, calibration, anchor points identification, etc.). Access is recognised to be of utmost importance to develop a fully-fledged methodology, thus inviting to develop a firmer legal basis for it as well as close cooperation with MNOs. In particular, access to some form of longitudinal access is identified as key for the quality of the final statistical output. Collaboration from MNOs in the form of data management (network topology modelization and cell register management, M2M and IoT filtering, reports on data collection failure, etc.) have been identified empirically as necessary. Combination of different sources both from MNOs and statistical offices arises as very promising to achieve high-quality statistics, calling for cooperation on privacy-preserving transfer of information and transparent sharing of computations (not yet possible).

Deliverable I.8 on Visualization

This last deliverable contains the visualization techniques for different elements of the whole statistical process. As all the preceding deliverables, it contains the first elements towards a more general visualization framework. Currently, an R package called mobvis has been developed and shared as open-source code[8].

The tools allow the user to visualize radio cell configurations, event location probabilities and posterior location probabilities across the reference grid. Animated visualizations can be produced showing the dynamical aspects of the evolution of devices.

The package can be used both for real data and for synthetic data.

[8] https://github.com/MobilePhoneESSnetBigData/mobvis.

2.9 Innovative Tourism Statistics *(full information on WPJ available here on CROS)*

Innovative Tourism Statistics (WPJ) was a pilot project undertaken in the frame of the ESSnet Big Data II. In its implementation were involved employees from eight European statistical institutes:

- Statistics Poland (GUS) represented by the Statistical Office in Rzeszów (leader of the WPJ),
- National Statistical Institute of the Republic of Bulgaria (BNSI),
- Hellenic Statistical Authority (ELSTAT),
- Hesse Statistical Office (HESSE, Germany),
- Italian National Institute of Statistics (ISTAT),
- Statistics Netherlands (CBS),
- Statistics Portugal (INE),
- and Statistical Office of the Slovak Republic (SOSR).

The pilot project attempted to meet the following challenges:

- preparing an inventory of data sources related to tourism statistics (including big data sources) in individual partner countries along with their description and classification,
- developing a scalable solution for downloading data using web scraping techniques from selected web portals offering the possibility of booking accommodation,
- implementing methods of combining and matching data on tourist accommodation establishments in order to integrate statistical databases with data from web scraping for the purpose of improving the completeness of the survey population of tourist accommodation establishments,
- spatial-temporal disaggregation of the use of tourist accommodation establishments,
- preparing flash estimates of occupancy of tourist accommodation establishments to shorten the time of data publication for external recipients,
- developing a methodology for estimating the volume of tourist traffic and tourist expenses with the use of various data sources (statistical and non-statistical).

All the activities that were carried out allowed for a preliminary assessment of the impact of the obtained results on the improvement of the data presented in the Tourism Satellite Account (TSA). In addition, the key issue addressed by WPJ was the preparation of the Tourism Integration and Monitoring System (TIMS) prototype along with the micro-services dedicated to the above-mentioned areas that would support statistical production in the field of tourism statistics and assist in monitoring changes in the tourism sector.

During the pilot study five tasks were carried out. Detailed information related to each of the issues under consideration by the project participants can be found in the five reports available on the CROS Portal platform referring to the ESSnet Big Data II project, in the part dedicated to the WPJ package[9].

Task 1: Inventory of big data sources related to tourism statistics

[9] https://ec.europa.eu/eurostat/cros/content/wpj-milestones-and-deliverables_en [accessed: 29.01.2021]

a. Inventory of data sources

As part of the work carried out in this task the general list of data sources (including big data) identified by individual WPJ partner countries during the inventory process was established. The catalogue is divided according to different subdirectories, e.g. types of sources and frequency, availability, usefulness in estimating the demand and supply side of tourism.

During inventory work, project partners have inventoried a total of 130 sources of information. Just over half of them (57.7%) were external sources, while the remaining 42.3% were internal sources. External data sources, i.e. whose administrators are outside from official statistics, include those that partner countries do not yet have access to as well as those with limited availability. Taking into account the criterion of the availability of external data sources in all partner countries, it was found that 52% of them were available.

Table 1. Number of data sources identified broken down by type

Type of source	Country participating in the grant								Total
	BG	**Hesse**	**EL**	**IT**	**NL**	**PL**	**PT**	**SK**	
	Number of identified sources								
Total	**6**	**12**	**11**	**14**	**44**	**16**	**7**	**20**	**130**
Internal sources	4	6	2	9	22	3	1	8	55
External sources	**2**	**6**	**9**	**5**	**22**	**13**	**6**	**12**	**75**
including:									
Supply side	-	-	2	1	10	4	2	3	**22**
Demand side	2	6	7	4	20	12	4	11	**66**
Sources not available (temporarily or permanently)	-	4	7	3	9	2	1	10	**36**

On the basis of the data sources inventory, the project partners have developed Flow Models for their countries according to the accepted scheme.

Each model presents proposed directions for combining data collected from external sources and web scraping with data from official statistics. The models developed by the project partners have been adapted to the number and types of inventoried sources in each country and to the areas where they will be applied in the short term. They also served as a starting point for the implementation of the task 1c entitled "Source characteristics".

b. Web scraping

During the project, work was carried out on identifying tourist portals that will provide the most complete information on accommodation, transport or food, among other things. Based on the data from https://www.similarweb.com, an analysis of the most popular websites for the countries

participating in the project was carried out. Based on the results, it was determined that the most popular websites in Europe are: Booking.com, Airbnb.com and Trivago.com.

Widely used methods of extracting data from Internet portals often require a lot of specialised knowledge and workload devoted to programming, data processing and analysis. The WPJ team proposed a new web scraping solution (Visual Modeler) that was checked during the project and it will also be part of the Tourism Integration and Monitoring System (TIMS) prototype supporting official statistical production in the field of tourism.

The Visual Modeler has a graphical interface and can be used to download data from any website. In addition, this tool can be used for data processing operations (analysing, conducting calculations, etc.). The use of the application does not require advanced programming knowledge from the user. The tool allows to create a web scraping process directly in the application, load data from external APIs or scripts created in other languages (JavaScript, R).

c. Source characteristics

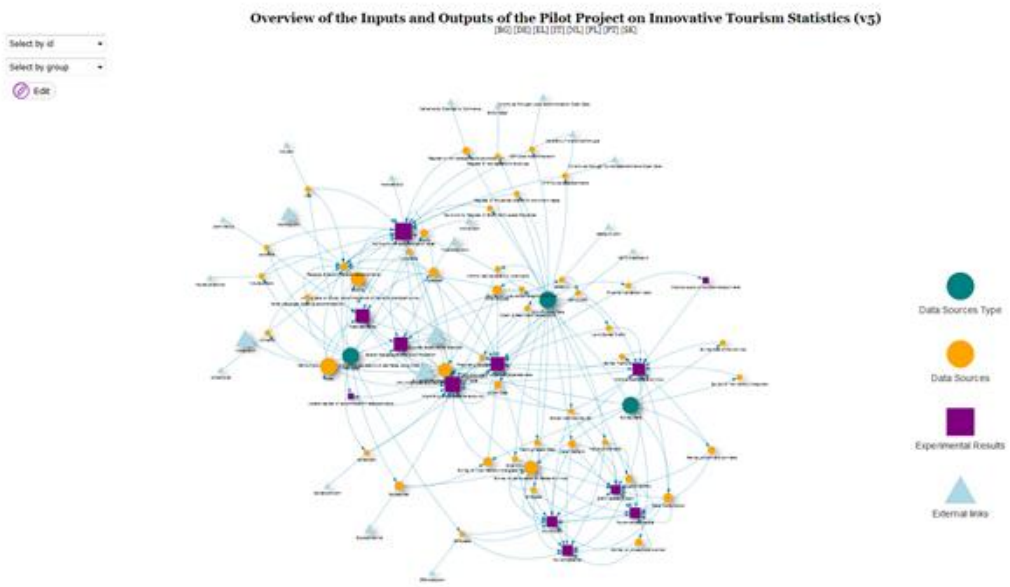
The objective of this task was to provide an interactive graphical representation of the relationships and interconnections between data sources (multi-purpose data sources, survey data and web scraped data), variables and domains as well as countries and experimental results.

This approach taken was inspired by network analysis which can be considered as map of relationships. These relationships are composed by nodes (for example, statistical domains or data sources) and edges that represent the connections between nodes. In this map of relationships not only different types of nodes but also different types or intensities of connections (edges) can be represented.

Network visualization is valuable alternative to represent a complex and otherwise static set of heavily interconnected data. Therefore the graphic Flow Models prepared by all country partners were “translated” into two input data objects describing nodes and their connections (edges) in order to create an interactive network for WPJ. The package has highly customizable options such as shapes, styles, colours, sizes, images but, most importantly, interactive controls (highlight, collapsed nodes, selection, zoom, physics, movement of noes, tooltip and events). Additionally it uses HTML canvas for rendering. It is based on html widgets, so it is compatible with shiny, R Markdown documents and RStudio viewer. R’s ability to read and write in a multitude of formats makes it very appealing to this sort of task.

The **Fout! Verwijzingsbron niet gevonden.**1 presents a general (static) view of the visNetwork.

Figure 1. Overview of the Inputs and Outputs of the Pilot Project on Innovative Tourism Statistics



A small video on how to use and make a visNetwork is available on https://youtu.be/cYETq0rIT9k as extension of this work.

Task 2: Examining availability, legal aspects and the quality of the new identified data sources used in the project

Big data as well as administrative data are not produced for the purpose of official statistics and, therefore, before integrating them in the production system of official statistics, potential quality, methodological, legal, privacy and other issues should be identified and addressed.

Legality of web scraping and netiquette

The work of the package included extracting data from web portals related to tourism. In this context, the legal aspects of web scraping should be taken into account [Stateva G. et al. 2017].

During the web scraping process, as well as after gathering the necessary information, WPJ followed the obligation to maintain statistical confidentiality as well as the provisions regarding the protection of personal data and the protection of copyrights. Good web scraping practices were also preserved.

Detailed characteristics regarding the legal aspects of web scraping, based on the experience of the WPJ partners gained during the work in ESSnet Big Data I is described in Deliverable J1 - ESSnet Methods for web scraping, data processing and analyses.

Due to differences in the legal aspects of web scraping in different countries, the legality of web scraping had been examined before the process started. Web scraping of websites should be non-invasive and used on the basis of good practices while maintaining statistical confidentiality and comply with data protection and copyright laws.

Task 3: Developing a methodology for combining and disaggregation data from various sources

A very crucial and at the same time difficult to implement element of workpackage J (WPJ) work, was to prepare a methodology for combining and disaggregation of tourism data. The work aimed at improving the completeness, timeliness and consistency of the survey frame of the tourism accommodation establishments.

a. Combining of tourism data

The successful linkage of data from web scraping with data from administrative registers and statistical survey frames was largely dependent on an appropriate matching strategy between these sources. Two main approaches to this task have been applied by WPJ partners: deterministic (or rule-based) record linkage and probabilistic (or fuzzy) record linkage. The first one was based on comparing text strings formed from addresses and postal codes of accommodation establishments, the second one was based on geographical coordinates and distances between accommodation establishments. Both approaches were implemented and reviewed separately and jointly. The results showed that applying both methods produced better quality of results.

Quality of matching was checked by reviewing names and addresses of paired establishments and deriving the number of correctly and incorrectly matched and not-matched establishments. Results of matching can be summarized in the form of a confusion matrix, which is often used when evaluating the performance of a classification model.

Table 2. Confusion matrix for data linkage results

		Actual	
	Total Population (all establishments in the portals and survey frames)	Match (establishment is present in portals and survey frame)	Non-match (establishment is present in portals only)
Predicted	Match	true matches True Positives (TP)	false matches False Positives (FP)
	Non-match	false non-matches False Negatives (FN)	true non-matches True Negatives (TN)

It must be noted that increasing the threshold (i.e. the distance between a pair of coordinates) generates a higher number of True Positives and False Positives and, at the same time, a lower number of True Negatives and False Negatives.

Following table presents the number of new establishments in each partner country. Number of newly found establishments vary from country to country. In the framework of WPJ we found 3 369 new establishments in project countries i.e. 6.7% of number of accommodation establishments in survey frames.

Table 3. Results for combining of data

Country partner	DE	EL	IT	NL	PL	PT	SK
Area	Hesse	Attica	Emilia-Romagna				
	Number of accommodation establishment						
Survey frame	3 483	876	12 877	8 843	16 561	1 474	5 930
Web scraping	731	508	6 419	1 937	3 742	1 856	1 134
Matches	490	343	4 270	1 540	3 591	1 591	933
New establishments	**241**	**165**	**2 149**	**197**	**151**	**265**	**201**

b. Disaggregation of tourism data

In the majority of European Union countries, the survey of the occupancy of tourist accommodation establishments is a monthly survey. The structure of the form does not allow to answer the following questions:

- How many tourists stayed overnight during the last weekend of the month?
- How many tourists stayed at the accommodation establishment on each day of the month?

Two data sources were used for temporal disaggregation: low-frequency data, e.g. monthly data that are a subject to disaggregation as well as high-frequency data, e.g. daily data containing auxiliary variables. In the case of disaggregation of data on the occupancy of accommodation establishments.

The first step in temporal disaggregation was regression. Using auxiliary variables, a regression model was built and preliminary estimates of the variable prepared. The simplest approach selected one variable and estimated the model using the classical method of least squares, or selects multiple variables and estimates the model using the generalized least squares method.

Typically, daily aggregates do not add up to a monthly value. Hence, the second stage of temporal disaggregation was benchmarking. The difference between the monthly value and the daily aggregates were estimated for each day.

The results obtained using the implemented methods significantly improved the quality of data on the capacity and occupancy of tourist accommodation establishments as well as trips of tourists and their incurred expenses. Thanks to the high frequency of data from web scraping of accommodation portals, it is possible to disaggregate monthly data on the occupancy of accommodation establishments into daily data.

Task 4: Flash estimates in the field of tourism

The scope of information about tourist accommodation establishments provided to Eurostat can be and in fact is different between countries. In many countries establishments having less than ten bed places are not covered by the surveys, additionally different thresholds are used in specific countries to collect information on establishments. The frequency of data collection is not very diverse and

usually data are collected on a monthly basis (in the case of project partners, all countries collect information on a monthly basis), while differences often occur in the observed population of establishments as well as methods of data collection.

Data on the occupancy of tourist accommodation establishments are often published with more than a month delay. This is due to the allotted time the entities have to submit data on the occupancy of tourist accommodation establishments and the time needed to process data by services of official statistics. The use of non-statistical data, e.g. from web scraping, is an attempt to speed up the process of producing statistical data. The advantage of this data sources is that complete data sets are available immediately after the end of the month which allows for rapid analysis.

Data from the web scraping of accommodation booking portals contain information about the accommodation establishment, i.e. its location and data on the offer of this entity, the most important of which are the price and type of the accommodation establishment. Data from a statistical survey - data from a monthly survey on the occupancy of tourist accommodation establishments, which covers entities of the national economy conducting activities classified according to NACE to the groups: 55.1, 55.2 and 55.3.

Analyses has been carried out on the accuracy of the forecasts in the face of the imbalance in the tourism market caused by the pandemic. The quality of the models has decreased after the introduction of restrictions related to COVID-19 pandemic. This is due to very large nominal drops in the tested values. The observed declines significantly reduce the precision of forecasts. As a simplified picture of reality, the model will only be able to reflect this state of affairs after some time. It should be remembered that the proposed solution is based on statistics on the prices of accommodation establishments, and even in the case of such large falls in nights spent, the prices of accommodation establishments have not changed significantly. This may have been due to the fact that, in a situation of uncertainty as to the duration of the restrictions, the owners could not or did not want to make decisions which would have had a very significant impact on prices.

From tested data sources that could serve as instrumental for modelling of flash estimates, Hotels.com and Google Trends data exhibited some potential and explanatory power. It is hard to say if the model based on Google Trends data outperformed the one based on Hotels.com data, they both showed some pros and cons:

Model based on Hotels.com data

- \+ rather stable data source with direct connection to domestic tourism
- \+ price variables strongly correlate with the tourism indicators
- \+ provided sound predictions for even pandemic lockdown (but only for number of nights spent, which could be due to a discrepancy between survey indicators)
- − extensive daily scraping necessary, which can crash at any time
- − not specifically robust in unexpected scenarios (accommodation owners advertising even if there are some restrictions, distribution of prices with low variance)
- − still too short time series

Model based on Google Trends data

+ easily accessible real time data on user's searching activity, long time series available
+ significant correlation with tourism indicators
+ provided promising predictions for peaks and troughs
+ robust in unexpected scenarios as the user behaviour on the Internet is in line with the current circumstances
− indirect connection to domestic tourism as it is mixed with foreign tourism
− impossible to calculate regional estimates

Both of the approaches can be further developed. In case of unusual situation in the field of tourism, which was encountered, e.g. in 2020, they can provide vital auxiliary information for the models. On the other hand, in case of stable progress, time series forecasting without external explanatory variables could yield more feasible flash estimates.

Task 5: Use of big data sources and developed methodology to improve the quality of data in various statistical areas

As part of this task, the following works were carried out:

- verification of estimations of the size of tourist traffic according to travel directions, means of transport, types of accommodation
- verification of estimation of the amount of expenses connected with trips to improve the quality of balance of payments
- use of developed methodology to improve the quality of the Tourism Satellite Accounts

a. Verification of estimations of the size of tourist traffic according to travel directions, means of transport, types of accommodation

Estimating a size of tourist traffic is based on simple linkage of several data sources using unique identifiers and calculations of conditionals distribution. Data on flights usually covers name and location of an airport, IATA and ICAO codes, type of aircraft, date of arrival and departure etc. Country name, type of aircraft, IATA and ICAO codes are the keys to join several data sources into one database.

Distribution of trips with respect to air traffic can be estimated with following procedure: collect all origin and destination airports with a use of an online flight connection search engine. use flights schedules to derive a distribution of flights for each origin airport, attach a country where the airports are located with airport code lists (IATA, ICAO, FAA), attach a capacity of aircraft (seats) with data on technical information on aircrafts, calculate a distribution of flights (measured with seats) for each airport available from national airports. The destination airports identified in this step will be called the hub airports, calculate a distribution of flights for each hub airport. Repeat this step until all destination airports are reached. Remove all routes that are irrelevant with respect to time or cost efficiency, time and cost efficiency does not provide an unambiguous rule. Nevertheless, the more obviously irrelevant routes are removed, the better the results, use a data from civil aviation office to

benchmark distribution of flights from a given origin airports to the known total, for each hub airport calculate the number of passengers that travel further using a relevant statistic from the sample survey (share of tourist using airports from a given country as a hub), sum up passengers from all routes: travelling directly from origin country, travelling with a use of one hub airport, etc.

Combination of administrative data, survey data and big data pertaining to air traffic enabled to estimate trips to a larger number of countries than it results only from the survey of the participation of Polish residents in trips. In the survey of trips of Poles, there are about 80 countries each year, and when estimating trips using big data over 140 destinations are obtained. Due to the very large number of countries for which it would be necessary to collect data through web scraping, the project work was limited to the countries of South America with 13 countries only. In 2019, the number of South American countries that appeared in the survey of the participation of Polish residents in trips was between 4 and 6, depending on the quarter. In the same period, the number of countries generated on the basis of big data was between 8 and 10. Finally, the number of countries with estimated trips in 2019, using the James-Stein estimator, was between 9 and 10.

b. Verification of estimation of the amount of expenses connected with trips to improve the quality of balance of payments

In the presented approach, there are two groups of data sources used to build the expenditure database: the sample survey of trips and data from various portals obtained through web scraping.

The basic data sources that can be linked to air trip expenditure include websites offering the booking and sale of accommodation, airline tickets, and websites with catering establishments. Through web scraping, one can collect data on, among other things: prices of accommodation by type of accommodation establishment, ticket prices by airline and route, average prices of meals in catering establishments (on websites, in the restaurant description, price ranges for standard dishes are often given and not prices for specific meals).

These data can be divided into two groups: data that should be collected frequently due to rapid changes, e.g. flight tickets prices, accommodation prices, and data that can be collected with low frequency, e.g. food and beverages prices, local transportation costs. The data from the sample survey on trips should cover several consecutive years to include a wide range of trips with respect to their descriptions (e.g. destination country, purpose of trip, type of accommodation, means of transportation) and their expenditure.

As a result of comparative analyses of journeys with the same characteristics to different South American countries, it can be seen that the model differentiated expenditure, taking into account the variability of data coming from big data. Based on the model, expenditure per person for a trip with 16 overnight stays in a hotel in Brazil was estimated at EUR 1 445, and expenditure for a trip to Paraguay, under the same assumptions, was estimated at EUR 1 603. The model expenditure showed less volatility than the expenditure from the survey of trips, which improved the stability of the results for individual countries and quarters.

The final result was influenced both by the new distribution of trips with new countries and by the estimation of expenses including big data sources. In order to determine how these two elements

affected the final result, tourism expenditure was estimated in the following options, taking into account:

1. the average expenditure directly from the sample survey of trips and the distribution of trips from the model,
2. the distribution of trips from the survey and the average expenditure from the model.

For 2019, total expenditure increased after modelling from EUR 22 335 000 by EUR 3 978 000 to EUR 26 313 000. The amount of EUR 3 978 000 can be broken down into the effect of modelling the expenditure itself, the change in the distribution of trips itself and the remaining effect of the total change.

c. Use of developed methodology to improve the quality of the Tourism Satellite Accounts

The time constraints of the project have now allowed for the preparation of new estimates only for the totals relevant for the determination of tourism demand in TSA and no significant changes in the total value of tourist expenditure are expected, even with the continuation and improvement of the proposed methods. The data collected with the use of the methods and tools used so far are reliable, and the aim is, above all, to improve their quality and completeness, as well as to identify and measure new, hitherto unknown phenomena in tourism. One should also not forget about measures to reduce the response burden.

In the case of this project, however, the project partners emphasize that its development it is important not only to improve the basic data on tourism statistics as input to the TSA tables, but in many cases even more important than the total value is the expected improvement in the quality of estimation of some items related to specific expenditure, difficult to estimate on the basis of data provided by sample surveys. As an example, Italy gives the underestimation on home rental expenditure for domestic tourism, but the problem of second homes is increasingly also affecting inbound tourism. This issue relates not only to the occasional provision of such services in the circle of family or friends, but also, in an increasingly serious dimension, to services provided under the so-called collaborative economy. In the near future, it will be possible to receive, via Eurostat, data on short-term accommodation from four large international platforms, which should significantly improve the estimates of this phenomenon. Nevertheless, the method proposed in the project for web scraping on other sites devoted to renting such houses will allow, among other things, to minimize gaps between tourism statistics and national accounts data currently resulting from the use of sample survey data. The scope of current web scraping of websites related to tourism covers more and more issues and will also allow to determine the sizes of other phenomena difficult to estimate but very relevant in the development of TSA, such as the phenomenon of renting passenger cars without a driver by tourist or the sales volume of conference services. Above all, however, it will enable a more accurate estimation of the value of services purchased by tourists in the form of packages (accommodation, catering, transport, recreational, etc.), which, in accordance with the TSA methodology, must be disaggregated by the relevant items of tourist expenditure.

Tourism Integration and Monitoring System (TIMS) prototype

Across the European Union statistical organizations conducting research in the field of tourism undertake similar activities, although each country uses different processes to describe individual phenomena. Processes associated with the acquisition, processing and publication of the results, however, may vary depending on the country and institutions. Rapidly changing and available new technologies require statistical offices to adapt their IT systems to collect and process data, develop them continuously, as well as to improve the consistency and comparability of results obtained on the basis of data gathered. All this hinders communication, exchange of tools and methods in the framework of cooperation between statistical organizations and comparison of obtained results. In addition, the use of big data sources in this process is an additional challenge in the area of methodology and applications of modern technologies.

The work carried during the pilot project of WPJ helped to provide a solution in form of a prototype (IT-system tool) to monitor changes in the Tourism sector. The potentiality and corresponding feasibility of this prototype (set of system components) are based on the experiences gathered and tested across the eight countries participating in the WPJ, and more importantly, it relies greatly on the assistance of an IT- platform and its well-performing components.

The TIMS prototype is designed using currently used modern information technologies. Its construction will use programming languages and libraries available as free/open software. The entire system infrastructure will be prepared on the basis of a set of rules that define communication between computer programs (API) in the REST architecture (client-server). Adoption of such a solution will ensure a high versatility of the tool. In addition, all queries sent by a client will be authenticated, which will protect the system against unauthorized access to data.

The solution fulfils the task of integrating data sources identified in the field of tourism statistics in the system databases. In addition, it aims to support the production process of statistics (including experimental ones), as well as data from both statistical databases (statistical surveys), data made available by external providers and those obtained from big data sources (e.g. web scraping, social media). The integration of data sources will be supported by their analysis and documentation in a metadata system that allows describing the variables and relating them to each other.

Thanks to the implemented internal mechanisms, the designed TIMS, will allow for pre-processing of data, including cleaning, unification of their structure and saving to databases. At a later stage, it will be possible to analyse the collected data and process it to obtain the final results.

The system will operate based on a number of individual modules, whose number and scope can be successively expanded depending on the needs of users. Dedicated modules (e.g. queries, integration, calculation, etc.) with implemented mechanisms, functionalities and methods of data estimation, will allow for data to be filtered, supplemented (imputation of missing data) as well as calculated and presented in various formats. The output data will be disseminated using existing official statistics systems, as well as via a dedicated API.

2.10 Methodology and Quality *(full information on WPK available here on CROS)*

1. General Objectives

The aim of this workpackage was to consolidate knowledge gained in this ESSnet in the area of methodology and quality for the usage of big data in the statistical production process and combine it with the insights from the previous ESSnet. Several deliverables from WP8 of the previous ESSnet were used as base of deliverables of this workpackage. Also deliverables of other workpackages from the previous ESSNet were used for input.

Since the process and architecture were dealt with in WPF and these issues are closely connected with methodology and quality, there was a need for coordination between WPK and WPF – especially to take the developed architectural model BREAL into account whenever possible.

Experiences obtained in the pilots of WPB to WPE, WPG to WPJ and WPL were used as input to this WP. To achieve a high degree of acceptance, feedback from relevant ESS bodies was collected for key deliverables (especially the quality guidelines and the quality reporting template).

2. Covered Topics

The following topics included in key deliverables of the previous ESSnet were extended and enhanced.

Literature

This task started by updating the literature overview of the previous ESSnet. Apart from the most important publications in the area of big data, the continously updated reports of the pilots of this ESSnet were included in this work package. A new focus of the literature overview was to identify useful quality indicators which could be used to quantify some quality dimension specific when using big data sources in the production of statistics.

The final literature overview can be found on the CROS-portal as a PDF document and is also available as searchable online tool via http://sa.ug.edu.pl/lr/ .

K12 Revised Literature Overview

Quality

When producing official statistics assessing quality is a crucial point of utmost importance. Beside the previous ESSnet, the quality framework developed by the UNECE and other external sources served as input for developing quality guidelines.

Based on the report on quality delivered by the previous ESSnet Big data, the quality framework for big data was further enhanced. There were three goals of this work. Firstly the information from the current pilots were used to update the report. Secondly quality guidelines for the usage of big data in official statistics were written based on the know-how from within the project an outside sources, e.g. the UNECE big data quality framework. The third goal was to develop a template for a quality report

when using big data in the production of statistics. Just as the methodological part, the quality report takes into account the diverse nature of big data sources as well as the access to the data sources.

The produced guidelines (K3) help users and potential future users of new data sources in the following relevant questions:

- What are the key quality issues with respect to the data access?
- What quality dimensions are relevant while processing the new data?
- What are the key quality issues with respect to the usage of new data in the statistical production process?

The structure of the guidelines follows the phases of the statistical production process of official statistics including big data sources. These phases – input, throughput (1 & 2) and output – are described in detail and with examples for different data sources in the introduction of the guidelines.

The structure of the quality report template (K6) was taken from the widely known SIMS (Single Integrated Metadata Structure). The definitions and guidelines are based on the recently updated version of the EHQMR (ESS handbook for quality and metadata reports). All (sub)concept are checked if they are fit for new data sources.

Some of them are adapted to better reflect the challenge posed by new data sources. A major challenge in using SIMS was that SIMS is generally output-oriented, but not all projects with big data produces directly output.

The quality template was tested by the WP members of the pilot track. It was used as basis for a questionnaire about quality issues for the pilot track intermediate meeting from 11th to 12th of December 2019 in Vienna.

The report describing quality aspects (K11) has two main parts,

- Quality related issues and
- Connection to Quality guidelines for multisource statistics in the ESS.

The aim of the first part is to put quality issues of the pilots to record in a summarized way. The second one is a discussion about the harmonisation of the “Guidelines for the Acquisition and Usage of Big Data” (QGBD) with the already existing “Quality Guidelines for Multisource statistics” (QGMSS, Komuso guidelines), which have been developed within different frameworks, but have poins in common.

K3 Quality Guidelines for the Acquisition and Usage of Big Data

K6 Quality Report Template

K11 Report Describing the Quality Aspects of the Different Pilots and the Way Forward

Methodology

The methodological report of the previous ESSnet was updated with new findings from the literature review and the outcome of the pilots. Based upon the methodological report and the general stepwise

approach proposed in the previous ESSnet methodological report, a more advanced framework was developed. Important future research questions were also identified.

The methodological report (K9) provides an overview of the current state of art of Big Data methodology. As we have found that the range of what one considers Big Data methodology differs tremendously on the experience and area of expertise of a particular statistical researcher, we decided to apply a very strict (narrow) focus on what is included in this report. Essential starting point of any method described in this report is that it:

i) has been used by researchers involved in the ESSnet Big Data I and II projects,

ii) is included in one of the Big Data based official statistics currently being published,

iii) is used in one of the Big Data based statistics that is in the process of becoming officially published.

Any other method, however interesting it may seem, is not included in this report. The methods included in this report are divided in various (sub)sections that relate to the various phases of the statistical process in which they are used. The same phases, input, throughput (1 & 2) and output, were used in the quality guidelines report (K3) and described in detail there.

The report describing the methodological steps (K10) combines the processes developed for 13 examples with the personal experiences of the authors to a generic process by which Big data can be used in official statistics production.

K9 Methodological Report

K10 Report describing the methodological steps of using big data in official statistics with a section on the most important research questions for the future including guidelines

Typification

Although the insight about a possible typification of Big Data projects has emerged from the previous ESSnet, it had not yet been explicitly expressed and it was a task during this project to explore and describe the elements that will classify a possible big data in a more comprehensive and exhaustive way.

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.

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.

The goal of the Typification Matrix (K7) is to identify the questions that will make it possible to assess the maturity level of a big data source. This assists the future evaluation of big data sources/projects.

The typficiation matrix was tested by the WP members of the pilot track. It was filled out for the pilot track intermediate meeting from 11th to 12th of December 2019 in Vienna.

Building on the Typification Matrix, the second document on typification (K8) aims to provide an evolution roadmap based on the Typification Matrix for big data projects. A major outcome is the assement of maturity level of different big data projects. The assessment into different levels of maturity was based on feedback from the work packages and 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. Therefore, the assessment may change over time. Since such assessments rarely are really objective, it represents the authors' subjective view on this issue.

K7 Typification Matrix for Big Data Projects

K8 Evolution Roadmap Based on the Typification for Big Data Projects

Meeting Organisation for the Pilot Track

Two meetings for two days each were organised, bringing together members of the WPK with members of all pilot workpackages (WPG-WPJ), a kick-off meeting and an intermediate meeting. The objective of these meetings was to coordinate and organise the work of the whole pilots track together and to communicate the progress of each workpackage within the project. The meetings were organised at Statistics Austria in Vienna. The agenda and meeting minutes were prepared by WPK.

2.11 Preparing Smart Statistics *(full information on WPL available here on CROS)*

1. Goal of WPL

The aim of WPL "Preparing Smart Statistics" is to examine the extended use of the Internet of Things (IoT) in order to produce trusted smart statistics for the European Statistical System (ESS). As the range of topics regarding the subject IoT is huge, the goal is to provide an overview of relevant topics for official statistics, to show their variety and to highlight topics that are promising and could be analysed further. Therefore, the availability and accessibility of the different data sources are also checked.

Furthermore, the goal is to explore how the digital footprints of daily life created by human wearables, city and vehicles sensors and other smart systems could change the way to produce trusted smart statistics taking advantage of societies' datafication.

The development of trusted smart statistics aims for data that might be pre-processed by the data providers or data sources ready to use for official statistics, which leads to a combination of privately held data with official survey and/or administrative data.

WPL was designed to prepare the ground for future actions on trusted smart statistics by concentrating on the IoT data sources and devices with potential relevance to official statistics, resulting in an overview of the data landscape. The results of WPL should enable future actions to result in proof of concepts and experimental statistics.

All the results of this WP keep in mind that the overall goal is the development of generic solutions and harmonized as well as standardized approaches and recommendations for the ESS.

There are four main topics and therefore tasks conducted in WPL: 1. Smart Farming, 2. Smart Cities, 3. Smart Devices and 4. Smart Traffic.

As the objective of WPL was to give an overview of the 4 topics and to give recommendations for possible follow-up studies, the duration was only 12 months, conducted from November 2018 to October 2019.

There were 12 partners from 11 countries involved in WPL (leader in bold):

- Statistics Austria (STAT, AT)
- The Bulgarian National Statistical Institute (BNSI, BG)
- The State Statistical Office of Berlin-Brandenburg (SSO BB, DE)
- **The Federal Statistical Office of Germany (Destatis, DE)**
- Statistics Finland (FI)
- The National Statistical Insitute of France (INSEE, FR)
- The National Statistical Institute of Italy (ISTAT, IT)
- Statistics Poland (PL)
- Statistics Netherlands (CBS, NL)
- Statistics Norway (NO)
- Office for National Statistics (ONS, UK)
- Statistics Portugal (INE, PT)

Also involved as a subcontractor of BNSI was the company Mimirium Ltd, in the work of task 2, case study 1.

The milestones and deliverables of WPL are available via the following link:

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WPL_Milestones_and_Deliverables

2. Content of the four tasks

Details about the four tasks can be found in the four deliverables of this WP.

Task 1 – Smart Farming

Task 1 was conducted by STAT, Statistics Poland and Destatis. The aim of this task is to explore the possible use of data coming from precision farming in agricultural statistics and to identify potential data sources. In this regard, task 1 provides an overview of the data landscape, that is of smart technologies and agricultural technology providers, and of the data generated by these technologies. Furthermore, it explores the use of smart technologies by farmers on the example of Germany. Finally, it presents a promising use case and gives recommendations for possible follow-up studies.

Generally, data generated by smart farming technologies are very promising for official statistics, but - mainly due to the restricted use of digital technologies by the farmers - it might be more interesting for official statistics in the long run. Nevertheless, data sources explored in the European project "Big Data Grapes" might also be interesting for official statistics and could be examined further in a follow-up study.

Task 2 – Smart Cities

Task 2 was conducted by BNSI, INSEE, ISTAT, SSO BB and Destatis. This task aims at investigating the potential of IoT tools (such as smart sensors) in order to produce smart statistics related to the topic Smart Cities.

In case study 1, BNSI explored if new technological solutions of Smart Cities, such as Blockchains or LoRaWAN (a media access control protocol for wide area networks), can be used for the purposes of official statistics. In this regard, information about problem areas for disabled people was collected via an app, that was developed for this case study, and via GPS and additionally installed devices in the city. The case study shows that the used tools allow to collect data that previously couldn't be collected and at the same time secure the privacy and confidentiality of the respondents. This was achieved by storing the gathered data only locally in an encrypted model on the users' devices and by only transmitting anonymized and aggregated data to the National Statistical Institutes (NSIs).

It is highly recommended to establish not only an ESS interest group that builds knowledge and skills about the chances of the technologies used here, but also to generate more use cases based on the technologies explored here and to test the use of these solutions for all member states.

In case study 2, ISTAT, SSO BB and Destatis researched the outcomes of the 12 projects that were funded by the Horizon 2020 Programme of the European Commission. Furthermore, the cities Berlin and Darmstadt were explored regarding the smart technologies they implemented and how the

generated data might be of use for official statistics. The aim of this case study is to give a first overview of the different 'smart solutions' that are partly still developed and partly already implemented in cities across Europe and to highlight the most promising solutions for official statistics.

Although there seem to be a lot of valuable data generated by Smart Cities, the availability and access to it is generally rather poor at the moment. Nevertheless, for a follow-up study it is recommended to further explore the city Darmstadt, especially as there is already some data available as open data that was collected via sensors on traffic lights. Another use case could explore the data platforms that combine all data that is generated by different technological solutions in Smart Cities. The third use case could further investigate so-called smart lamp posts that capture different kind of data, such as data about vehicle traffic, air pollution and water level.

In case study 3, INSEE and ISTAT analysed the correlation between socioeconomic characteristics of people and the level of pollution in the cities Nice and Rome. Especially in Nice, a successful cooperation with partners that provide and use raw open data was established. In Nice, people of higher socio-economic status are significantly less exposed to air pollution than people of medium and low socio-economic status. This was probably the most advanced case study of task 2, combining open source data, used by a Smart City, with official data.

Possible follow-up studies could include on the one hand an incitation to other countries to follow the example of INSEE and ISTAT and on the other hand methodological developments in order to improve the quality of the estimation. Also, a deeper collaboration with smart cities is foreseen, in order to better understand the reasons for pollution by analyzing traffic data collected by the city.

Task 3 – Smart Devices

This task was conducted by CBS, INE and ISTAT. The aim of this task was to conduct a first study on smart devices from the viewpoint of official statistics. It was meant to be preparatory for future projects on trusted smart statistics that will be able to go deeper into the most promising use cases identified in this task. Task 3 provides a comprehensive list of different types of smart devices that are categorized as follows: smartphones, smart home devices, smart devices for health and fitness, smart devices for mobility, smart devices for travel and other smart devices. The devices are described regarding their purpose for the users of the specific device, their distribution and also concerning their potential use for official statistics.

Generally speaking, data from any smart device or citizen science project is very interesting for official statistics. However, preconditions are the type of data access (if accessible at all), the kind of data it concerns, the maturity of the devices and the transferability to the ESS. In this regard, three use cases were identified: 1. The realisation of a citizen science project in order to examine voluntary public participation for statistical purposes, 2. the development of an app to deduct people's mobility pattern using machine learning techniques and 3. the analysis of travel patterns based on smart travel cards.

Task 4 – Smart Traffic

Task 4 was conducted by Statistics Finland, Statistics Norway and ONS. It contains two subtasks exploring new data sources that can be used to analyse traffic for official statistics.

In subtask 4.1, Statistics Finland and ONS explored the use of traffic inductive loop sensors for monitoring traffic in order to use it for economic estimates. Although there are some data quality issues, traffic loop data has many advantages, especially the availability of time series data often going back to times prior to the 2008 financial crisis. ONS opts for creating a set of indicators that allow early identification of large economic changes in trade, whereas Statistics Finland derives truck traffic growth rates to produce early estimates of GDP. These new products based on traffic loop data are able to give policymakers early warning of changes in the economy, although care must be taken when using them and it's important not to conflate them with regular official estimates of GDP and trade. Finland and the UK published their research on their websites already and will continue to improve the products. They highly recommend that other NSIs follow their example.

In subtask 4.2, Statistics Norway explored promising new data sources for producing statistics on road freight transport. From all the data collected nowadays by modern trucks, three data sources were identified as particularly promising: Build-in sensor data, GPS-data and tachograph data. If the data would be used by official statistics, the response burden in the freight transport survey would decrease significantly. In order to use the data, either a uniform interface for data delivery needs to be built or alternatively third-party GPS-data could be used or integrated smart tachograph data could be utilized as position data.

The data source is evaluated as interesting for the whole ESS, as information requirements, quality issues and technologies on sensor data in trucks apparently are very alike throughout European countries.

3. Conclusion

In WPL, a wide variety of topics concerning the subject IoT have been examined and many promising data sources have been identified. The results show that it might be rewarding to dig deeper in follow-up studies. However, generally the data landscape is still in a development state. Data access and availabilty are often an issue. Even though merging of sensor data with geolocalized official data has been tested succesfully and some data sources are already used for estimates and have been published as research papers, the data format and the transferability to the ESS still has to be examined in many cases. Also, data protection is an important subject to take care of by official statistics, but there are quite promising trends leading to open data, which is usually anonymized and aggregated. Also, the confidentiality of data could be protected by using new technologies and procedures such as cryptography technologies or blockchains. Certainly, these technologies have to be further investigated. In addition, the quality of data gathered by smart systems and their representativity should be examined carefully as well.

Altogether the use of data generated by IoT tools offers many possibilities to enrich offical statistics, but both using new data sources and adopting/changing the statistical production processes of the NSIs to involve those have to be explored in further studies.

3. Issues encountered

3.1. General issues

The main issue encountered will not surprise: the COVID-19 pandemic, which manifested itself in ESS countries from about ten months before the foreseen end of the project (i.e., end of December 2020), and which started to have an impact on the project from April 2020 onwards. Obviously, all meetings planned during the pandemic had to be held virtually. This affected all mid-year meetings that were planned for 2020. The partners of the ESSnet adapted well to the circumstances and the ESSnet succeeded in producing all 58 deliverables of WPB to WPL at least in draft before the end of 2020, with the exception of two deliverables of WPI, which were produced in draft early 2021.

These two deliverables of WPI critically depended on the collaboration of MNOs (Mobile Network Operators) with individual NSIs, as one concerned access to MNO data and the other the processing of such data. The collaboration with NSIs for the ESSnet was not and could not be given priority by the MNOs during the pandemic, as their data was suddenly requested for urgent other needs, which demanded much effort and attention. Despite all efforts, this made it impossible for the ESSnet to deliver the two deliverables in time. However, collaboration with NSIs picked up in time for the deliverables to be produced within the adjusted planning.

Another deliverable, the final conference of the ESSnet, BDES 2020, was affected as well. This conference could not be held physically, which meant that interaction would be reduced, in particular informal interaction. The ESSnet had intended at BDES 2020 to not only present its results, but also to discuss the possible use of the results of the ESSnet in the future, the conditions to be fulfilled for such future use, and the related broader strategic orientation needed for the ESS after the end of the project, but a virtual event was considered less suited for that purpose. The ESSnet took the decision to hold BDES 2020 in a virtual format in November 2020, but to limit its scope to the dissemination of the results of the ESSnet. And the decision was taken to request an amendment to the grant contract to allow for a second BDES conference, in June 2021, aimed at discussing the future oriented issues. This second conference, BDES 2021, would be held in a physical format if circumstances allowed. These decisions had to be taken at the end of the summer of 2020 already because of the preparation time needed.

The request for an amendment to extend the project till the end of June 2021 was submitted and granted, including both the extension of the deadlines of the two WPI deliverables and the possibility to organize BDES 2021 in addition to BDES 2020. (Of course, it also covered the extension of related reporting obligations, mainly affecting WPA.) In the end, it turned out that circumstances did not allow for BDES 2021 to take place in a physical format, but the split of the original conference in two parts, each with its own objectives and set-up, may still be considered the best way to get the best conference results under the circumstances of the pandemic.

The extension of the grant contract till June 2021 proved to be a blessing to the ESSnet for other reasons as well. Apart from the two WPI deliverables, a bottleneck in the planning had emerged in the second half of 2020 which had to do with delays in the review process of deliverables. There had been changes in the composition of the Review Board, which together with personal circumstances led to a backlog of reviews in the last months of 2020. Thanks to special efforts of the members of the Review

Board and the workpackages concerned, the delays could be kept to a minimum and the last deliverables could be formally submitted by mid-April 2021.

3.2. Issues at the level of the workpackages

WPB Online Job Vacancies

Staffing issues

Some of the issues encountered within this workpackage are country specific. United Kingdom has experienced difficulties in retaining staff, in particular data science specialists, who have not been replaced in the second part of project execution. Nevertheless, they fulfilled the envisaged tasks. Ireland has withdrawn from active participation due to their internal policy on the use of scraped data. Due to limited resources, Lithuania did not actively participate in the second half of the project execution.

Data quality and data access issues

The EU DataPlatform contains multiple sets of CEDEFOP data. These datasets differ dramatically in their size, both absolute and relative to the covered timeframe. Each new version changed all previous data retroactively. Because the actual collection of job ads can not change retroactively, the different versions of the data was caused by changes in the processing pipeline which has not been explained. During data analysis data quality issues have been identified, which have not been documented or clarified as well. During the implementation of the project it turned out that CEDEFOP data processing system is very effective but also very complex. Despite great efforts of the administrators of the Data lab and EU DataPlatform, resources were limitited and in particular cases insufficient for working directly with such a large datasets. As a result, certain products could not be published as experimental statistics.

WPC Enterprise Characteristics

For WPC no significant issues occurred.

WPD Smart Energy

The main issues that WPD encountered during this project where data related. Norway had access to monthly aggregated data and Sweden only had a test dataset to work with.

WPE Tracking Ships

Concerns, signalled during the previous ESSnet Big Data project, like the need for suitable hardware and software tools for the exploration and exploitation of the huge amount of AIS data, still applied. The products mentioned were created using several big data platforms: local platforms, the Big Data Test Infrastructure (BDTI) environment and its successor, the DataPlatform, both provided by European Commission, and the United Nations Global Platform (UNGP) provided by the UN Global Working Group (GWG) on Big Data for Official Statistics. Each environment has its specific set of tools and uses other sources of AIS data with different coverages. AIS data is collected by several parties

around the world, though the scope of the data may differ. Some parties only collect national data, some parties collect data on a European level and some parties collect all data around-the-world. Some parties collect maritime data and some parties collect data focused on inland waterways.

During the project, each time the best environment for developing and testing the product had to be selected. The choice depended mostly on the coverage of the AIS data. The inland waterways product requires AIS data on inland waterways, which only is available nationally because of privacy reasons. For the product on air emissions and energy used for the environmental accounts, worldwide coverage is required, which was only available on the UN Global Platform.

Both the products on port visits and on fishing fleet were developed locally and later implemented and tested on the DataPlatform. Almost 7 months after the start of the project, the BDTI environment was provided by the European Commission. It was available for 5 months, with interruptions. Then, the environment was deprecated and in month 18 of the project, it was replaced by the DataPlatform. Setting up the DataPlatform and loading the AIS data required a lot of valuable project resources.

Much was learned about what data is needed. In order to use AIS as a source for statistics and fully benefit from it, some prerequisites need to be met:

1. AIS maritime data with worldwide coverage is crucial

   For certain statistics, like air emissions and economic indicators, AIS data with worldwide coverage is crucial, but very costly. EMSA collects AIS data with European coverage, and if agreed by the member states, it could share this data. EMSA also buys additional satellite and global data from commercial parties. EMSA cannot share this data under the current contracts. At the UN Global Platform however, AIS data with a worldwide coverage is currently available and shared.

2. AIS inland waterways data is desired

   Different member states could benefit from using AIS to improve statistics on inland waterways. If that is the case, AIS data with coverage of European IWW could be desirable to provide to the statistical offices. Note that, under current legislation, AIS data related to inland waterways ships is sensitive information, as the ship is also the home location of the skipper, and therefore considered confidential data. This confidentiality issue would first have to be investigated.

3. Ship register with worldwide coverage is required

   A worldwide ship register, both maritime and inland waterways, including ship characteristics and ownership/flag is required. The latter is required for air emissions from the environmental accounts. A ship register with worldwide coverage is Lloyd's register from IHS Markit's. A good example of an initiative to create such a register by the European Commission, is the European fishing ship register (see https://webgate.ec.europa.eu/fleet-europa/search_en). Unfortunately, this does not include the MMSI number, which makes it difficult to connect it directly to AIS data. For the product on fishing fleet, we resolved this problem by scraping this missing data.

4. Port register with European coverage is needed

   A port register with at least European coverage is needed. This register should at least contain all European ports with their geographical boundaries, preferably including more detailed information on the terminals, the type of cargo supported (containers, bulks or even people) and their geographical boundaries.

5. Global IT platform and infrastructure is required

   In order to develop, test and implement solutions based on AIS data, a well-suited IT platform and infrastructure, including access to the AIS data, is mandatory. Setting up this environment requires special skills that are currently not always available. Note that the UN Global Platform provides such a platform. Eurostat is in the process of developing such a platform, but extra attention should be geared towards sustainability, user-friendly access and use.

WPF Process and Architecture

For WPF no significant issues occurred.

WPG Financial Transactions Data

Change of contents in Sharing economy part

During the first half year of the WPG work, it turned out that the original Sharing economy part of WPG (Task 5-8) could not be performed according to the contract. The contract stated that it should be a literature study of the accessibility and metadata of relevant data within the sharing economy platforms. During the work, it was revealed that hardly any literature existed on the topic. As a result, in the summer of 2019 there was an agreement with Eurostat to rather study to what extent financial transactions data could be used to derive indicators on sharing economy, where sharing economy was defined broader as "platform economy", since the latter definition made it less complicated to identify the sharing economy part of financial transactions.

Change of budget

Because of the change of contents described above, it was agreed between all the WPG members to re-allocate some resources since the new contents put more burden on Italy. The agreed decision was that Italy increased the original budget with about 25 days, whereas Bulgaria reduced their budget with 13 percent and Germany reduced their budget with four percent. The rest of the countries (Norway, Portugal and Slovenia) reduced their budgets with eight percent. This resulted in no change in the total budget for WPG.

WPH Earth Observation

Staffing issues

During the project, some staff changes have taken place (data scientist from Destatis/Germany), two project participants left and they have not been replaced (data scientists from Insee/France and CBS/Nederlands). Nevertheless, the envisaged tasks were fulfilled.

Data access

The main issue encountered within WPH concerns data access. In most cases, intermediate resolution satellite data were used. These data are freely available and can be gathered with a high frequency of a few days. The data access issue refers to cases where high-resolution data are required, i.e Case study 3 - Crop recognition with very high-resolution aerial data. Publicly available aerial images cannot be used for detailed crop recognition via machine learning despite the adequate resolution, because

they are recorded at an insufficient frequency (annually and tri-annually, and even 10-yearly for the highest-resolution data) to constitute the time series needed for analysis which should probably be at least weekly in the critical periods. Furthermore, some datasets cannot be used for crop recognition at all due to a recording time unfit for this purpose (during winter). One possible alternative might be for statistical institutes themselves to produce the required datasets via aeroplane or drone (UAV, Unmanned Aerial Vehicle). The cost of this could be lowered by partnering with public agencies or research institutes interested in aerial photography datasets for their own testing or operational objectives in the domains of agriculture or GIS. The second alternative may be buying detailed satellite data at a sufficient resolution from private providers.

WPI Mobile Networks Data

The main issue encountered regarding WPI on mobile network data is still the question about the access to real data. This is hampering and slowing down the development of the production framework for official statistics and two coordinated courses of action are identified, namely, (i) the development of an appropriate legal framework and (ii) a close collaboration and partnership agreement with MNOs. The foregoing activities clearly recommend a European dimension in the initiatives both at the strategic and methodological levels. In this line, the proliferation of isolated business cases potentially arise as an obstacle to finding standard and common solutions, not only in terms of statistical methods but also to gain access to data for production in the long term.

Finally, in the efforts to come closer to real production conditions, scalability of all developed proposals, especially those derived from the generation of synthetic data, must be investigated. Adequate hardware platforms and software tools need to be researched to implement the ideas presented.

WPJ Innovative Tourism Statistics

The main issue encountered during the work of the WPJ was the drastic quantitative and qualitative decline in tourism-related web scraping data due to restrictions on the activity of accommodation establishments during the COVID-19 pandemic. This required the adaptation of models used and a re-inventory of the available data sources in a short time at an advanced stage of the project.

Some countries (mainly Germany (Hesse) and Portugal), due to the internal policy of using web scraped data or blockage of the web scraping processes, were not able to maintain continuity in data collection. This resulted in too short time series and it was not possible to use the created models to generate data of appropriate quality.

Another issue encountered was related to accessing data from non-statistical sources, including administrative data that was identified during the inventory of tourism-related data sources. In many cases, obtaining information from external administrators required time-consuming negotiations, including at the local administration level. Due to constraints resulting from the short term of the eligible grant agreement, not all negotiations started were completed within the project life span and therefore data were not used in the development of the tourism related models.

WPK Methodology and Quality

Issues encountered in WPK were more of a content-related nature than of an organizational one: As WPK was not dependend from data providers directly, no issues encountered here. Also, WPK concludes adequate time and budget calculations.

However, it has become clear that the access, as well as the processing and the usage of new data sources include very source and data-specific processes. Due to the diverse nature of the new data as well as the new data sources, it was a huge challenge to formulate generally applicable quality guidelines with practical relevance which are more than a general reassertion of the very abstract principles.

To overcome this issue, additionally source-specific guidelines were formulated within the framework of quality guidelines. The same procedure was followed in other deliverables produced in WPK. This modular approach of WPK gives a good overview on quality and methodology for different new data sources. But it also underlines the importance of follow-up source-specific projects -- as the “ESS Web Intelligence Network” for web data – to deepen the research on source-specific issues on quality and methodological issues.

WPL Preparing Smart Statistics

In contrast to the other workpackages, the duration of WPL was only 12 months, from November 2018 to October 2019. Therefore, it was not designed to conduct feasibility studies, but only to find out if there are relevant and promising topics for official statistics, which are worth to be researched further in a follow-up study. For some tasks a longer duration might have been helpful, e.g. to establish cooperations with external partners. Due to this fact, an exchange with the other workpackages was not possible.

It is already important to keep in mind that this workpackage was finished before the COVID-19-crisis and might have had other goals and objectives if it was conducted during the pandemic.

Annex I: Communication and Dissemination

1. Objectives and instruments

The success of a large project with as many as 28 partners from 23 different countries and an estimated 150 active participants depends to a large extent on high-quality **internal communication**: a free, easy and frictionless exchange of reflections, opinions, methods and data within and between tracks and workpackages. Additionally, efficient **external communication** is required, as the ESSnet has the explicit objective to have an impact on the ground, and to promote the adoption of big data and smart statistics for the statistics production of the European Statistical System and to disseminate results beyond as widely as possible.

Both internal exchanges and external dissemination were served simultaneously by a combination workplatform/extranet **Mediawiki wiki website** open for viewing to all, but limited for editing to project participants, supplemented by a restricted Confluence website for storing personal, financial and confidential project information.

To further disseminate the project's results to stakeholders and to discuss future directions, an international conference **BDES 2020** was planned to be held in Warsaw (PL) on 24-25 November 2020, aiming at an attendance of approximately 150 persons.

Unfortunately, two external factors negatively impacted the full implementation of this communication planning and made extensive rethinking and reorganisation necessary. The COVID-19 pandemic and the ensuing contact limitations and travel bans led to a change of plans regarding the final ESSnet conference: Two events were organized, 6 days of **BDES 2020 Online Sessions**, from 23 to 30 November 2020, to disseminate results, and the **virtual BDES 2021** conference, from 16 to 18 June 2021, to focus on the future (see section 3.1 for further details).

A second setback was the decision by EC Digit to eliminate the project's Mediawiki collaboration platform and extranet website by 10 December 2020. Fortunately, its content could be salvaged to a more permanent location at Eurostat's CROS portal (see https://ec.europa.eu/eurostat/cros/content/essnet-big-data-0_en), thus ensuring continuity but providing considerably less editing and site management functionality.

2. Websites

2.1 The original setup

One of the first and most urgent requirements at the start of ESSnet Big Data II in November 2018 was building the foreseen suite of 3 websites:

- A public **Mediawiki wiki/intranet/extranet website**, open for viewing to all but limited as to editing to project participants. Mediawiki has a sophisticated site management for handling different levels of access, for keeping track of changes, for discussion between contributors and

for quickly and efficiently getting content online under controlled conditions. The Mediawiki site combines multiple functions:

  - Project intranet with meeting calendar and minutes, planning, timetables, project documents, etc.Work platform for collaboratively sharing data, methods and reflections and for co-creating deliverables.
  - Showcase to the public at large for results: milestones and deliverables, experimental statistics.
  - Toolbox with big data event calendar, documentation section containing relevant articles, links page, access to satellite GitHub repositories, etc.

- A restricted **Confluence wiki website**, accessible only to project participants, for storing personal, financial and otherwise confidential information such as data from partners or non-final results.

- A **mirror site on the CROS portal** duplicating the upper layers of the Mediawiki site but without content, thus providing a second and more 'official' access to the content uniquely stored in the Mediawiki site.

Because ESSnet Big Data II was a continuation of ESSnet Big Data I, with a similar setup, this structure was available within one month after the project's start, hosted by Eurostat with the technical support and maintenance of EC Digit.

2.2 Switch to CROS portal

In the middle of 2020 Eurostat communicated that EC Digit was no longer competent to maintain the security of the Mediawiki site and that it would be closed down by September 2020, thus scrapping an accepted project deliverable before the project's end, removing an essential tool for collaboration and dissemination, and imperilling the organisation of the BDES 2020 and BDES 2021 events.

Fortunately, some workaround solutions were found. The shutdown was postponed to December 2020, making it possible to still efficiently organise the BDES 2020 Online Sessions. And Eurostat invested in moving the content to the CROS portal (https://ec.europa.eu/eurostat/cros/content/essnet-big-data-0_en), with the considerable additional advantage that results and tools will continue to be available after the end of ESSnet Big Data II. The downside of markedly less editing and site management functionality is somewhat offset by the fact that most deliverables and documents had been put online before the switch. However, this did not make the organisation of BDES 2021 any easier.

3. Conferences: BDES 2020 and BDES 2021

3.1 Original planning and COVID-19 impact

The project planning foresaw a conference, co-organised by Statbel (Statistics Belgium) and GUS (Statistics Poland) and with the dual purpose of disseminating results and discussing the next steps, towards the end of the project period. At the end of 2019 the location, date and format had been decided: an international 2-day conference BDES (Big Data for European Statistics) 2020 in Warsaw (PL) on 24 and 25 November 2020, with a projected attendance of about 150 persons. But then, at the beginning of 2020, came COVID-19 …

In spite of misgivings from the beginning of the pandemic, a final decision on the November conference was postponed until the last possible moment in June 2020, and then it was decided to abandon the idea of a physical conference and to replace it with two events: on the one hand a series of online sessions to disseminate results in November 2020, under the original name of BDES 2020; and on the other hand a symposium-like limited gathering to discuss the future with a select attendance of 50-55 people, physically in Warsaw (PL) if at all possible, in May or June 2021, under the name of BDES 2021.

3.2 BDES 2020 Online Sessions

The BDES 2020 Online Sessions were held on six consecutive days from 23 to 30 November 2020. They were very well-attended, with over 400 registered users, considerably more than could have been accommodated in a physical meeting, and no major technical problems occurred, not a given with so many live room presenters.

BDES 2020 consisted of 12 half-day sessions, an introductory one and 11 sessions presenting the results of the 11 ESSnet Big Data content workpackages. Each of the latter sessions started off with a presentation of the main results of the workpackage, followed by a wrap-up by a discussant segueing into a question & answer session with the active participation of the audience.

All information about the BDES 2020 Online Sessions, including presentations, presenters' bios and full recordings of all sessions, can be found at https://ec.europa.eu/eurostat/cros/content/bdes-2020_en.

A short survey of participants after the event returned very favourable comments about content and format. One remark consistently made was that attendance would not even have been possible if the event had not been online, due to time and budget restrictions. A major side benefit of the online format which allows recording is also that viewing the sessions is even now possible, and that they can be used at any time for training purposes in national statistical institutes and other organisations, within the European Statistical System or globally.

3.3 BDES 2021 Roundtable on future directions

The second replacement event, the BDES 2021 Roundtable on future directions, took place on three consecutive days, from 16 to 18 June 2021. Due to the Covid-19 situation still not permitting physical international meetings, the roundtable was organised online, using Zoom. The first two days were dedicated to intense discussion in small groups, 16 in total, followed by a plenary day to report on these discussions, to collect takeaway points and to draw conclusions. Attendance was intentionally limited and on invitation, to allow free and optimal exchange of ideas. In total 47 persons from 28 organisations participated.

During the discussion rounds on the first two days, 16 and 17 June 2021, 8 themes, prepared and documented beforehand, were treated in-depth, each being discussed independently by two different groups:

1. From exploration to exploitation
2. Leaner and meaner statistics
3. Access modalities
4. International and ESS cooperation and division of labour
5. Fostering culture change
6. Smart statistics

7. Business architecture and methodological frameworks: next steps
8. Further research priorities

Discussion groups were composed in line with participants' preferences, expressed beforehand, and ranged in size from 5 to 13 participants. Each was assisted by a moderator facilitating the exchange of opinions and by a reporter summarising the discussion's conclusions. In order to guarantee free and open speech, discussions proceeded under the Chatham House Rule stating that participants do not speak for an institution or in any formal role, but as individuals with personal ideas and views. To further ensure free expression, no recording and no detailed verbatim reporting was done.

On the final day, 18 June 2021, themes were taken up again in plenary sessions and final conclusions were drawn. For each of the 8 themes two reports were presented by the reporters of the discussion groups were the theme had been treated, and a further exchange of ideas in the whole group arrived at final conclusions and takeway points.

Discussions, both in the small groups and in the plenary sessions, were lively and active, resulting in a list of concrete ideas and points for future action in the field of big data, new data sources and smart statistics. The online roundtable approach with well-prepared and ably assisted discussion in small groups worked exceptionally well, and might in itself be an innovative takeaway point for future similar purposes.

All information about the BDES 2021 Roundtable can be found at https://ec.europa.eu/eurostat/cros/content/bdes-2021_en.

Annex II: Evaluation of the ESSnet

It is a good custom in project management at the end of a project to look back in order to identify instances from which one can learn, in particular things to do or not to do. In that way others can profit from the experiences of the project. Based on feedback from members of the Coordination Group and on the replies to a request to participants of the BDES 2020 conference to provide feedback, a number of points were identified. They are grouped into the following categories: general points, project organization, the BDES 2020 conference, the BDES 2021 roundtable, and other points. The two BDES events get special attention, since they had an experimental character, as the originally planned more traditional physical conference could not take place because of the pandemic. Specific issues that were encountered are not included, as these are listed in section 3.2.

General points

- The initial expectations and project objectives have been largely met. The ESSnet project form is a very effective way to share and obtain knowledge, especially from more advanced partners.
- The achieved results are highly promising for statistical production, in particular for the WPs of the implementation track. However, it should be recognized that in many cases much remains to be done for the obtained results to be ready to be incorporated in real statistical production.

Project organization

- The project organization was effective in the sense that all contents deliverables were produced in time (with only a few in the extra time provided by the amendment), after having been reviewed by the Review Board, which can be considered as a quality mark. Having a Review Board for such projects can be considered a good practice.
- Success factors were the much appreciated, responsive administrative and financial management and the fact that all WP leaders and participants were fully committed and assumed their responsibilities. Regular meetings made the project alive and active. The commitment of the partners was generally high, although there were exceptions due to local circumstances. However, the coordination burden for WP leaders was considered high.
- The communication through the mediawiki was quite good and well organized, but the unforeseen and forced transition to CROS, though well managed with good support from Eurostat, was not considered an improvement. An issue was also the sharing of draft documents and interim results within WPs. Some missed editorial standards for the contents of documents/reports.
- The end date for most WPs was chosen to be two months before the (original) project end date, because the previous ESSnet Big Data had experienced a bottleneck at the end, in particular concerning the high number of reviews that had to be carried out by the Review Board in the last months. This “slack” has proven to be very useful.

BDES 2020

- Participants to BDES 2020, albeit with a low response rate, gave the conference very high marks, from an organisational, technical as well as substantive perspective. The organizers

themselves were also quite satisfied as the conference clearly reached its dissemination goal and went generally speaking (very) smoothly.

- With more than 400 participants in total and an average of more than 100 per session, attendance was much higher than what had been possible with the originally planned physical conference. The threshold for participation was low (no traveling, no budget needed, no limitation to delegations); for instance some staff of regional offices indicated that they could only attend because the event was virtual.

BDES 2021

- The roundtable went technically and organizationally smoothly. The use of Zoom with breakout rooms worked very well, as the size of the groups was small by design. It was good that the use of more complex tools was avoided.
- The format, which was quite innovative, worked out extremely well. An essential element was the application of the Chatham House Rule, which made the participants feel free to express their own views. Other success factors were the fact that all participants had an active role, the discussion sessions had a moderator and a reporter (for whom instructions and guidance had been prepared), and for each discussion theme a briefing sheet had been prepared. The involvement of the participants showed in their staying till the end.
- Despite its virtual character, the roundtable cemented the community that had been emerging from the ESSnet Big Data. Participants appreciate being part of a network of ESS experts.
- Some expressed the opinion that the format chosen for BDES 2021 may be worth considering for certain future occasions as well, for instance task forces charged with proposing the way forward.

Other points

- Concerning the preparation of the proposal for the grant, the active participation of partners was in some cases hampered by its timing around the summer. Also, during the preparation partners had to deal with uncertainties, such as concerning data access and the time-line of usability (e.g. Cedefop data). As they may necessitate adjustments of the project during execution, such uncertainties have to be better assessed in advance (e.g. by using SWOT analyses).
- For some aspects, the ESSnet was dependent on Eurostat or other Commission services. This was in particular the case for some data sources (AIS data, Cedefop) and infrastructure (Datalab, termination of mediawiki). Where policy or other changes occurred, this forced the ESSnet to adjust, for instance by abandoning the mediawiki or having to use the UN Global Platform for AIS data.
- Also, the effectiveness of the results of the ESSnet depends, among other things, on their follow-up by Eurostat. In the case of the products of WPL this follow-up was not considered entirely satisfactory. To a certain extent such situations cannot be avoided, but awareness of such dependencies and addressing them will improve the effectiveness of ESSnets.