The ESSnet Big Data: Highlights of Results and Outlook for Future Work

Peter STRUIJS

Coordinator ESSnet Big Data, CBS, Netherlands

Abstract. One of the main actions resulting from the Big Data Action plan and Roadmap (BDAR) is the ESSnet Big Data. In fact, there are two consecutive ESSnets, the first one ending in May 2018 and the next one planned to start in November 2018. In the first ESSnet, which was launched in February 2016, 22 partners from 20 countries of the ESS collaborated in exploring the possibilities of using big data as a source for official statistics. The research comprised seven big data pilot projects and covered aspects such as data access, statistical processes and methods, and combining different data sources for specific domains. Cross-cutting features such as methodology, quality and IT were also looked at. A call for a second ESSnet was published in May 2018. That ESSnet is planned to continue till the end of 2020. In total 27 partners from 22 countries have expressed interest in joining the new ESSnet. Apart from carrying out new pilot projects, this ESSnet aims at implementing results from the first ESSnet in a limited number of countries. In addition, the ESSnet will explore the possibilities of developing trusted smart statistics, thereby laying the foundation for future actions in this field.

One of the main actions resulting from the Big Data Action plan and Roadmap (BDAR) is the ESSnet Big Data. The overall objective of the ESSnet is to prepare the ESS for integration of big data sources into the production of official statistics. At the beginning of 2016, the ESSnet was launched. It ended in May 2018, but a new ESSnet Big Data is already being planned, in response to a call published in that same month. The approach to the first ESSnet and the main results so far are described in the first two chapters. The third chapter presents an outlook for the second ESSnet.

1. The ESSnet Big Data I: Approach

The first ESSnet Big Data was based on a Framework Partnership Agreement (FPA), covering the period from January 2016 to May 2018. The FPA was founded on a consortium of 22 partners, consisting of 20 National Statistical Institutes (NSIs) and two Statistical Authorities. In the context of the FPA, two Specific Grant Agreements (SGA-1 and SGA-2) were signed, the first one starting in February 2016 and the second one ending in May 2018. The two SGAs had an overlap in time.

The budget for each of the SGAs was one million euro, with a maximum reimbursement of 90%. The ESSnet organised the core of its work around a number of workpackages, each workpackage dealing with one pilot and a concrete output. There were eight workpackages, of which seven were focused on specific sources or domains and the eighth on overarching aspects:

- WP 1 Webscraping / Job Vacancies
- WP 2 Webscraping / Enterprise Characteristics
- WP 3 Smart Meters
- WP 4 AIS Data
- WP 5 Mobile Phone Data
- WP 6 Early Estimates
- WP 7 Multi Domains
- WP 8 Methodology, Quality and IT

A separate workpackage, WP 0, was created for the co-ordination of the ESSnet. For dissemination a separate workpackage was created as well, WP 9. That workpackage was also responsible for facilitating communication.

The pilots had one thing in common: they covered the complete statistical process, from data acquisition to the production of statistical output. In addition the pilots also considered future perspectives. Thus, all pilots recognised the following five phases:

- Data access
- Data handling
- Methodology and technology
- Statistical output
- Future perspectives

At a practical level, this approach required to be facilitated in several respects. First of all, an organisational approach was chosen to ensure that the agreed output would be produced with the resources foreseen. In order to ensure the quality of the results of the ESSnet, a Review Board was created with three members: Lilli Japec (chair, SE), Anders Holmberg (NO) and Faiz Alsuhail (FI). They reviewed all deliverables of all workpackages. In order for the partners of the ESSnet to be able to process big data, some IT facilities were considered necessary. Therefore the ESSnet subscribed to the so-called Sandbox in Ireland, and a GitHub repository was used. Facilities were also needed for communication, in order to share and work on documents together and for virtual meetings, among other things. A Mediawiki-based wiki was set up in which all project participants can edit and which anyone else may freely consult. Among many other things it contains all deliverables of the ESSnet, including the results described in the next chapter: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata

A dissemination conference, called BDES 2018 (Big Data for European Statistics) was held in Sofia (BG) in May 2018 for a wider audience, in which the main results of the ESSnet were presented and discussed. Speakers included the Director-Generals of Eurostat, Mariana Kotzeva, of Statistics Bulgaria, Sergey Tsvetarsky, and of Statistics Poland, Dominik Rozkrut. The audience included representatives from academia, ESS and non-ESS statistical institutes, and international organisations.

2. The ESSnet Big Data I: Results

2.1 WP 1 Webscraping / Job Vacancies

This pilot explored the potential of on-line job vacancy (OJV) advertisements as a data source for job vacancy statistics. OJV data contain more variables and detail than existing job vacancy statistics. Various approaches to accessing data were explored including web-scraping of job portals and enterprise web-sites as well as arranged access with owners of on on-line job vacancy data. A key outcome of this pilot was the establishment of a relationship with the European Centre for the Development of Vocational Training (CEDEFOP), who are undertaking a major project to collect OJV data for all EU member states. This is expected to be the main source of OJV data for use within the ESS. A range of data processing, quality and methodological issues were explored including deduplication, classification of data from unstructured text, data linking, validation and time series analysis. However, quality issues around coverage, representativity, definitions and structural differences in on-line labour markets between countries, mean that it cannot replace existing job vacancy statistics. Future challenges will centre on how to produce and present new statistics based on OJV data together with existing statistics.

2.2 WP 2 Webscraping / Enterprise Characteristics

The purpose of this workpackage was to investigate whether web-scraping, text mining and inference techniques can be used to collect, process and improve general information about enterprises. In particular, the aim was twofold: (i) to demonstrate whether business registers can be improved by using web scraping techniques and by applying model-based approaches in order to predict for each enterprise the values of some key variables; (ii) to verify the possibility to produce statistical outputs with more predictive power combined or not with other sources of data, like the

"ICT use by enterprises" survey. Six use cases were investigated: URLs retrieval, e-commerce/web sales, social media detection, job advertisement detection, NACE detection and Sustainable Development Goals (SDGs) detection. Final methodological and technological conclusions based on all the work done within the ESSnet were drawn and a set of output indicators were published on the ESSnet wiki as experimental statistics.

2.3 WP 3 Smart Meters

The main goal of this workpackage was to demonstrate the potential use of data from smart electricity meters for production of official statistics. The pilot had three goals with regard to expected outputs. First, to assess whether current survey based business statistics can be replaced by statistics produced from electricity smart meter data, second, to produce new household statistics and third, to identify vacant or seasonally vacant dwellings. Furthermore, the team looked beyond the goals set to identify potential statistical products in the domain of energy consumption or in other statistical domains by relaying on the data produced early on. The workpackage also studied classification and provided different use cases of using the smart meter data. The smart meter data was used to produce regional electricity statistics and it was evaluated whether the data can be used to produce tourism statistics. In addition, from the theoretical viewpoint potential uses for different kinds of smart meters (e.g. natural gas, water) were proposed. The impact of aggregation for producing different statistical products from the smart meter data was also evaluated. Finally, a compact overview of lessons learned during the project produced and recommendations given to other countries which start using smart meter data.

2.4 WP 4 AIS Data

The workpackage investigated whether real-time measurement data of ship positions (measured by the so-called AIS-system) can be used for improving the quality and internal comparability of existing statistics and for new statistical products relevant to the ESS. Reports were produced on the use of AIS data to determine emissions and on its use to produce possible new statistical output. Scenarios were developed for production of European and national statistics based on one single European data source and a cost-benefit analysis of using AIS-data for official statistics presented. The original plan assumed getting access to data from the European Maritime Safety Agency (EMSA). However, access to AIS data from EMSA was not granted[1]. Therefore European AIS data from Dirkzwager was used and the quality of this source was compared to satellite data from Luxspace and national data from Greece. Visualisations were made of transshipment by container giants, and of the average and maximum speed of ships in European seas and oceans. Results from

[1] Since the end of the ESSnet the prospects of obtaining data from EMSA for statistical purposes have improved.

this workpackage show 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 still is useful to improve other aspects of maritime statistics and provide new products.

2.5 WP 5 Mobile Phone Data

This workpackage has focused on the development of a methodological framework, the analysis of the IT infrastructure and software tools, and the assessment of quality issues regarding the use of mobile phone data in the production of official statistics. Concerning methodology, the workpackage investigated the whole process from data collection to statistical output. Concrete methodological proposals were provided for different elements of this process. Linked to this framework, IT platforms for data access and processing were described and two R packages developed. The first one aims at implementing the Bayesian approach to geolocate network events based on the signal strength and the second one at implementing the statistical model developed to estimate population counts. Concerning quality, the workpackage has focused on two aspects. On the one hand, an analysis was made on how the European Statistics Code of Practice is going to be affected according to the preceding proposals. On the other hand, proposals have been made to deal with the accuracy dimension of quality in the context of the new inference model for the production of official statistics using mobile network data. Finally, the workpackage has made recommendations for future research.

2.6 WP 6 Early Estimates

The aim of the workpackage was to investigate how a combination of multiple big data sources and existing official statistical data can be used in order to create existing or new early estimates for statistics. Several pilots were carried out. The most promising estimator was GDP, but the pilots were not limited to GDP due to the fact that results of analysing data sources suggested the calculation of estimates of other economic indicators. The main outcome was the calculation of a testing set of early estimates of concrete economic indicators (GDP, TIO, IPI, …) together with the defined methodology and process needed for this purpose. It was shown that incomplete early micro-level sources and a real-time big data source such as the traffic loop data can be used to produce early estimates of economic indicators. NSIs can shorten the publication lag in a straightforward way without compromising the quality of results. Concrete advice on how to embrace the opportunities of nowcasting was provided. NSIs can address timeliness by using a range of micro-level data sources accumulated in the registers well before the official release is made, by employing large dimensional econometric models, to form an initial quick estimate of the target indicator. This does not necessarily lead to too large revisions, but adds significantly to the quality of official statistics through the timeliness dimension.

2.7 WP 7 Multi Domains

This workpackage prepared and tested 6 intra-domain pilots in three different domains: three in Population, two in Tourism, one in Agriculture. The best tested pilot and most promising in the Population domain was Life Satisfaction – it used a machine learning algorithm to produce the results of life satisfaction according to the classification from the EU-SILC survey (happy, neutral, calm, upset, depressed, discouraged), based on Twitter data. The other two pilots in the Population domain were related to the selected health status of population and to Peoples opinion/interest by topics based on websites by ONS (UK) by Facebook. In Tourism there were two different pilots: Tourism accommodation establishments and Internal EU Border Crossing, including data sources by Air Traffic – Flight Movement web scraping and Traffic Loops data. The agriculture domain had one pilot prepared by two different methodological approaches: the first was prepared by Statistics Poland and the second by CSO Ireland. For combining data two different approaches were carried out – intra-domain data combining (all domains) and inter-domain data combining (agriculture-tourism). The results of the pilots conducted show that the greatest potential is in the agriculture domain – to identify crop types. The case is ready for use with the open data that can be accessed on the Internet.

2.8 WP 8 Methodology

The aim of this workpackage was to lay down a general foundation in the areas of methodology, quality and IT infrastructure when using big data for statistics produced within the ESS. The workpackage 8 therefore started with a workshop in which the most important topics in the area of IT, quality and methodology were identified. Next an overview of important papers, project results, presentations and webpages relevant to the application of big data for official statistics was created. This overview was linked with the findings of the pilots in the first phase (SGA-1) and the input available from the second one (SGA-2) of the ESSnet Big Data. The main outcome of the workpackage was an overview of the methodological, quality and IT findings when using big data for official statistics. Experiences obtained both in- and outside the ESSnet Big Data were used as input for the workpackage.

3. The ESSnet Big Data II

While the ESSnet Big Data I was still ongoing, the need for a successor ESSnet was already recognised. In February 2018 the ESSC adopted the Business Case Smart Statistics & Big Data, which explained and specified what would be needed after finishing the ongoing ESSnet. In April 2018, a sprint session with members of the ESS Big Data Task Force was held in the Netherlands to explore and identify relevant issues related to trusted smart statistics. In May 2018, a call for proposals for a new ESSnet on Big Data was published, building on the Business Case and the smart statistics sprint session. As the first ESSnet Big Data was generally considered a success, the approach of the call is similar. The available call budget for the multibeneficial grant is estimated at 2.465.000 euro, with, again, a maximum co-financing rate of 90%. The deadline for submission is 20 September 2018 (17:00 CET)[2]. The ESSnet is expected to start before the end of 2018 and finish by the end of 2020.

Whereas the ESSnet Big Data I was all about pilot projects, the call has three different tracks. The first one aims at implementing results from the previous ESSnet, that is, functional production prototypes will be developed building on four of the seven pilot domains of the ESSnet Big Data I. These are the following:

- Online job vacancies
- Enterprise characteristics
- Measuring electricity consumption, identifying energy consumption patterns
- Maritime and inland waterways statistics, environmental statistics

The second track concerns five new pilot projects, of which at least three have to be carried out. These five are the following:

- Use of financial transactions data
- Use of remotely sensed data
- Use of online platforms such as social media and sharing economy platforms
- Use of mobile network operator data
- Use of innovative sources and methods for tourism statistics

The third track pertains to the domain of trusted smart statistics, focusing on the extended ecosystem of the Internet of Things (IoT). The third track can be seen as preparing the ground for future actions aimed at developing trusted smart statistics.

The new ESSnet is, once again, expected to also look at horizontal and cross-cutting issues.

[2] This is after the submission of this paper for the DGINS 2018 conference.

Preparations for a proposal in reply to the call are underway. In total, 27 members of the ESS (22 NSIs and five other national authorities) have expressed interest in participation. This is a substantial increase compared with the previous ESSnet. The activities will also cover more subjects and be more diverse than the ESSnet Big Data I.

Bringing the new ESSnet to a good end will thus be a unique challenge to all involved. The ambitions are high, and rightly so, given the strategic importance of preparing the ESS for the integration of big data sources into the production of official statistics and the development of trusted smart statistics. It will be worth the effort.