United Nations Global Working Group on Big Data for Official Statistics

7 November 2017, Bogotá, Colombia

The ESSnet Big Data and Its Results

Note for Information

Prepared by Peter Struijs[1]

Background

In the context of the European Statistical System (ESS), one of the initiatives concerning big data is the so-called ESSnet Big Data. The ESSnet is formed by 22 partners of 20 ESS countries, mainly National Statistical Institutes (NSIs). The ESSnet works on the basis of a Framework Partnership Agreement (FPA), stretching from January 2016 to May 2018. The work is funded by means of so-called Specific Grant Agreements (SGAs), of which there are two. The first one, SGA-1, covered the period from February 2016 to July 2017. The second one, SGA-2, covers the period of January 2017 till the end of the FPA. The two SGAs have an overlap in time. The size of each of the grants is one million euro, but only 90% of costs, as a maximum, is reimbursed.

The overall objective of the ESSnet is to prepare the ESS for integration of big data sources into the production of official statistics. For SGA-1 as well as SGA-2, the consortium has organised the core of its work around a number of work packages, each work package (WP) dealing with one pilot and a concrete output. In SGA-1 there were seven work packages, focused on specific sources or domains:

1. WP 1 Webscraping / Job Vacancies
2. WP 2 Webscraping / Enterprise Characteristics
3. WP 3 Smart Meters
4. WP 4 AIS Data
5. WP 5 Mobile Phone Data
6. WP 6 Early Estimates
7. WP 7 Multi Domains

A separate work package, WP 0, was created for the co-ordination of the ESSnet. For dissemination a separate work package was created as well, WP 9. That work package is also responsible for facilitating communication. Given the overall objective, the findings need to be generalised. This is done in SGA-2, for which a new work package was added, WP 8 (methodology, quality and IT).

The pilots have one thing in common: they cover the complete statistical process, from data acquisition to the production of statistical output. In addition, the pilots also consider future perspectives. Thus, all pilots recognise the following five phases:

[1] Co-ordinator of the ESSnet. The text is essentially an excerpt from one of the reports produced by the ESSnet.

1. Data access
2. Data handling
3. Methodology and technology
4. Statistical output
5. Future perspectives

All results are reviewed by a Review Board. For the ESSnet, a Mediawiki website has been created, which serves two functions:

- collaboration and communication platform for project participants: project information and backgrounds, resources and tools;
- extranet, presenting outputs in a well-structured way to anyone outside the project: backgrounds, documentation, public reports and deliverables.

In February 2017 a dissemination workshop was held in Sofia for a wider audience, in which the main results of the ESSnet achieved up till then were presented and discussed. A report of the workshop is available on the wiki of the ESSnet.

Results so far

These are the main results obtained for the seven pilots (see also the website):

WP 1 Webscraping / Job Vacancies
This work package focusses mainly on job portals. Since this pilot involves each country taking its own specific approach, there are a lot of country specific results. However, general selection criteria have been identified for targeting portals for scraping. Taking into account the distinction between job vacancy and job advertisement, a conceptual model is proposed of how on-line job advertisements correspond to the target population. In practical terms this may be defined as all vacancies that are available to be measured by existing job vacancy surveys. As well as providing a conceptual framework for understanding the coverage of job vacancies from on-line sources and how these relate to the measurement of all job vacancies, this approach may also provide the conceptual basis for an estimation framework, including an approach to data integration. Furthermore, the work package has identified an opportunity to work with the EU Centre for Vocational training, CEDEFOP.

WP2 Webscraping / Enterprise Characteristics
Six use cases have been identified in the pilot: (1) enterprise URLs inventory, (2) e-commerce in enterprises (about predicting whether or not an enterprise provides web sales facilities on its website), (3) job vacancies ads on enterprises' websites, (4) social media presence on enterprises webpages, (5) sustainability reporting on enterprises' websites (linked to the UN Sustainability Development Goals), and (6) relevant categories of enterprises' activity sector (NACE) aimed at checking or completing statistical business registers. A common use case template was developed and has been used. For the use cases, a total of sixteen pilots were performed and all of them were mapped to a general "logical architecture". Also, a report was produced on legal aspects related to web scraping of enterprise websites, aimed at showing the real possibilities for the NSIs to perform activities of web scraping. These appear to be generally favourable, although the situation differs from country to country.

WP 3 Smart Meters
This pilot has investigated data access and data handling of smart meters electricity data. It has carried out a literature study and a survey on access to smart meters data, which was sent to the NSIs of all EU member countries in the spring of 2016, with 18 responses. It appeared that only two countries currently have access to data: Denmark and Estonia. Several countries were aware of substantial legal barriers. Some countries such as Poland are in the process of drawing up legislation that will enable smart meters data use. For two countries, Estonia and Denmark, the pilot has defined and assessed the quality of smart meter electricity data, and a synthetic dataset was analysed as well, aimed at generating demo output and developing and testing statistics and algorithms for situations where linkage to enterprise or household characteristics is necessary.

WP 4 AIS Data
The work package investigates whether real-time measurement data of ship positions (measured by the so-called AIS-system) can be used to improve the quality and internal comparability of existing statistics and for new statistical products relevant for the ESS. Reports were produced on (1) the possibilities and pitfalls of creating a

database with AIS-data for official statistics, (2) deriving harbour visits and linking data from maritime statistics with AIS-data, and (3) sea traffic analyses using AIS-data. While the possibility of using AIS data from EMSA is still being investigated, AIS data from Dirkzwager was used, and the data quality analysed. Visualisations were made, showing the coverage of the ships by the data, and showing the path of a ship through time. A method to build a reference frame of maritime ships was developed. First results show that AIS data can be used as a backbone for maritime statistics. This is important, since the added value of running a pilot with AIS-data at European level is linked to the fact that the source data is generic worldwide and data may be obtained at European level.

WP 5 Mobile Phone Data

This work package has focused exclusively on data access during SGA-1, which was needed for SGA-2. A preliminary analysis of the issues regarding the access to mobile phone data was made, which was the basis for the design of a questionnaire surveying the status of this access across the ESS. Belgium, Finland, France, and Italy were found to have succeeded in their negotiations to have access to a concrete mobile phone data set that is used for SGA-2 (together with those of UK, Netherlands, and Germany, as new partners of WP 5for the second phase). Spain and Romania are still under contact with MNOs pursuing this goal. A workshop was held in Luxembourg to bring together mobile network operators, national statistical offices, Eurostat and other stakeholders, including some other international organizations (UN, OECD, ITU, DG Connect, DG Digit). Finally, with the technical assistance from the Estonian company Positium, which is an international expert in accessing and processing mobile phone data for statistical purposes, a set of guidelines for the access to these data has been produced with technical, business and practical recommendations for partners of the ESS.

WP 6 Early Estimates

The aim of this pilot is 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. A list was compiled of possible data sources and the statistical domains where they could be employed, and it was decided that the most promising and interesting ones concerned combining sources for early estimates on economic indicators. The economic indicators and possible sources were further specified. In this context two pilots were conducted, one by Statistics Finland and one by the Slovenian NSI. The relationship between GDP and Slovenian traffic sensor data was investigated, and Statistics Finland produced nowcasts of turnover indices. These results were used for a business case for the research currently carried out in SGA-2.

WP 7 Multi Domains

The aim of this pilot is to find out how a combination of big data sources, administrative data and statistical data may enrich current statistical output. Three statistical domains are being investigated: (1) population, (2) tourism/border crossings and (3) agriculture. For population, three areas are looked at: daily (life) satisfaction, the moods of population associated with public events (e.g., Brexit, voting), and morbidity areas (e.g., flu). For tourism/border crossings, a number of possible data sources have been identified and investigated, for instance with regard to traffic intensity information. For agriculture, the focus is on recognizing crop types based on satellite data.