Overview of Big-Data research in European Statistics Agencies

Loredana Di Consiglio, Martin Karlberg, Michail Skaliotis, Ioannis Xirouchakis[1]

Abstract

By adopting the Scheveningen Memorandum in 2013 and the big data action plan and roadmap in 2014, the European Statistical System (ESS) has committed itself to exploring the potential of big data for producing official statistics. As a result, an ambitious collaborative research programme has been launched in big data and official statistics. In parallel, several statistical agencies (as well as Eurostat) are engaged in further methodological research and experimentation in specific statistical areas (e.g. prices, enterprise characteristics, tourism accommodation statistics, culture, job vacancies, etc.).

These first experiences have highlighted the great potential, and the great challenges, of using big data for official statistics. The intensive use of 'web scraping' has emerged as a new tool for the collection of relevant statistical information, while change in the nature – and hence the complexity – of the new types of data has meant that we need to broaden the spectrum of methods in official statistics to include machine-learning and text-mining, for example. At the same time, the use of big data sources requires a (wider) consideration of ethical and privacy issues, which in turn may trigger methodological and technological investigations.

The aim of this overview is:

1. to provide a summary and assessment of recent initiatives by European statistical agencies in big data research of direct relevance to enterprise surveys, business registers and the EuroGroups Register (EGR); and
2. to raise awareness of research opportunities under the EU's Horizon 2020 programme.

Key Words: big data, official statistics, web scraping, business registers

1. Background

Official statistics agencies and other public bodies have recognised the need to undertake exploratory research into the potential of big data for policy. It took some time to get the agencies on board, but work to fill the gap has now picked up considerably.

In 2013, the Directors General of the National Statistical Institutes (DGINS) acknowledged that big data represented new opportunities and challenges for official statistics agencies and, with the adoption of the Scheveningen Memorandum (ESS Committee, 2013), they committed themselves to examining the potential of big data sources. In 2014, the ESS adopted an **action plan and roadmap** to implement the Memorandum (ESS Committee, 2014).

[1] Loredana Di Consiglio, Eurostat, European Commission, L-2920 Luxembourg, email: Loredana.DI-CONSIGLIO@ec.europa.eu, Martin Karlberg, Eurostat, email: Martin.KARLBERG@ec.europa.eu, Michail Skaliotis, Eurostat, email: Michail.SKALIOTIS@ec.europa.eu, Ioannis Xirouchakis, Eurostat, email: Ioannis.XIROUCHAKIS@ec.europa.eu.

The purpose of the roadmap is to enable NSIs gradually to integrate big data into the regular production processes of national and European statistics.

In the **long term** (beyond 2020), big data sources will be fully integrated in ESS official statistics production, i.e. legislation will have been adapted, business continuity guaranteed, skills made available and methods, tools and IT infrastructures reviewed and adjusted.

In the **medium term** (by 2020):

- big data strategies for official statistics should be integrated in official big data strategies at national and EU levels;
- the pilot experiments will be finalised;
- adequate IT infrastructures and methodological and quality frameworks will be available;
- data science skills will be an integral part of official statistics education;
- public-private partnerships will be in place on big data and official statistics; and
- ethical guidelines and a communication strategy will be established.

In the **short term** (by 2017), the roadmap envisages the development of an ESS big data strategy, assessment of several new data sources, a review of the adequacy of the European statistics code of practice and statistical legislation, the mainstreaming of data science courses into the European Statistical Training Programme (ESTP) and the development of joint research projects with private data holders.

The big data action plan and roadmap address nine broad areas of intervention, which are interconnected in highly formal, multi-dimensional and complex ways: policy, quality framework, skills, experience sharing, legislation, IT infrastructure, methods, ethics and communication, and pilot projects.
Accordingly, the ESS is already running an ambitious collaborative research programme on big data and official statistics.

2. The ESS Vision 2020 BIGD project

The inclusion of big data in the production of official statistics necessitates the adoption of new methods of data analysis and processing, and improvements to ESS members' IT infrastructure. In order to address the modernisation challenges, the ESS has launched a major initiative, Vision 2020 (http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020), to meet users' requirements, face the methodological and technological challenges, and update what it offers in terms of products and services.

The big data project is linked to a number of key areas of Vision 2020 and represents an immediate response to the call, under the Vision 2020 initiative, to harness new data sources in statistical production.

2.1 ESSnet

A first set of challenges identified in the action plan refers to cooperation and the exchange of best practice, methodology and transition to the 'real use' of data.

One way of tackling these issues is through a series of pilot projects. In November 2015, Eurostat signed a framework partnership agreement with 20 NSIs and two ministries (https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data), followed by a specific grant agreement (SGA), which sets out the pilot projects to be carried out in this first phase. The pilot projects will be an important pillar of big-data activity in the ESS in the coming years and should pave the way to data production driven by big data.
Basically, the pilots will explore the potential use of certain big data sources in the production of official statistics and investigate the application of results to specific statistical domains in response to the needs of users in particular policy areas.

The first group of ESSnet pilots includes:

- web scraping (extracting information from websites) to compile statistics on job vacancies and enterprise characteristics;
- smart meters;
- automatic vessel identification systems; and
- mobile phone data.

Domain-oriented activities aimed at the integration of multiple sources are also under consideration.

All the pilots will consist of the following phases:

- investigation of data access, analysis of data availability and quality, analysis of relevant legal and privacy conditions, and establishing partnerships with private data holders;
- addressing the technical aspects of data-handling (e.g. implementation of web scraping, storage of content, location of processing, processing of raw data, micro vs aggregate data); and
- investigation of the methodology for output production (e.g. link to business registers or administrative data, text-mining, data-mining) and issues relating to statistical modelling, open algorithms, reproducibility, stability of the source, etc.

The plan is to draw cross-cutting conclusions as regards data access, data quality, methodology and IT architecture in the second phase of ESSnet (from early 2017).

2.1.1 Web scraping for job vacancies and enterprise characteristics

The pilot on job vacancies (led by the UK with DE, EL, IT, SE and SL as members) is exploring the potential of using web scraped data to compile statistics on job vacancies. It investigates methods focusing on job portals only. The first stage involves compiling a comprehensive inventory of job portals and their corresponding characteristics, with data being provided directly by job portal owners.
The second pilot on web scraping (led by IT with BG,NL,PL,SE and the UK as members) aims to enrich the business register with new information (e.g. on e-commerce activity) and reassess existing economic activity classifications. The work will involve exploring broader web scraping and machine-learning methods to identify relevant URLs.
It will be important to investigate the legal issues surrounding the use of scraped data and to assess the feasibility and quality aspects of the sources. The pilot is also expected to review and develop metadata frameworks.

2.1.2 Smart meters

Smart meters are electronic devices that record consumption at regular intervals and relay information back to the system operator for monitoring and billing purposes. A pilot on the use of smart meters (led by EE with AT, DK and SE as members) will seek to demonstrate, by means of concrete estimates, whether information from buildings equipped with smart meters can be used to produce (or provide supplementary information useful for producing) energy statistics and other statistics, e.g. housing census estimates and statistics on household costs, environmental impact and energy production.

The project will also assess methodological and IT issues and the potential and feasibility of linking to administrative registers.

2.1.3. Automatic identification systems

Data on the position of vessels obtained from the capture of signals from their automatic identification systems (AIS) are studied to estimate harbour visits and traffic intensities. It may also be possible to use AIS data for environmental statistics and a wide range of other purposes. AIS information is highly standardised, in order to ensure the overall integrity of the global AIS system and its components. The project (led by NL with DK, EL, NO and PL as members) analyses the feasibility of using the source as a reference frame by linking it to register-based data from port authorities. The added value of developing a uniform AIS-based reference frame in a multinational project is improved international comparability, timeliness and spatial granularity of estimates on transported goods and passengers.

2.1.4. Mobile phone network data

Mobile phone network data will be explored to produce daytime populations. In the initial phase, this project (led by ES with FI, FR, IT and RO as members) is exploring issues regarding access to data held by mobile network operators (MNOs), including current coverage and that expected in the near future. It is only in the second phase (2017) that the data will actually be analysed.

The above pilot projects examine the statistical potential of individual big data sources separately, rather than in combination.

2.1.5. Multiple sources / multi domains

ESSNet also includes a research activity which involves the analysis of methods needed when multiple sources are used to produce certain statistics.

A first cross-cutting pilot (led by SL with FI, NL and PL as members) investigates the use of a combination of multiple big data, administrative data, survey and other existing sources to produce early estimates for specific statistical purposes.

A second cross-cutting pilot (led by PL with IE, NL and the UK as members), on ‘multi domains’, investigates the potential of single big data sources for producing statistics on multiple topics such as population, border crossings and agriculture.

2.2 Eurostat internal research activities

In addition to managing and monitoring the ESSnet pilots, Eurostat has initiated various big data research projects either on its own or with a limited number of partners. Such

projects enable Eurostat staff to build internal analytical capabilities and technical expertise, and draw implications at strategic level for official statistics in general and the ESS in particular. Moreover, partnering with key stakeholders (e.g. MNOs, owners of digital platforms) helps to build trust between public and private bodies. Initial results from these projects are very encouraging as they demonstrate the capacity of novel data sources to provide policy insights that are not otherwise available.

2.2.1 Mobile phone network data

Together with Statistics Belgium and Proximus (a major Belgian MNO), Eurostat is currently working with geolocation data from mobile phone networks to estimate current population and cross-border mobility. Proximus data have been analysed to gauge how much information is lost when the time resolution of the data is decreased, to test the conversion from mobile network cells polygons to a standard regular grid (which will be necessary when considering periods longer than one day as network cells change). Further goals are to:

- identify the appropriate space and time resolution;
- test the relationship between mobile network data and statistical data; and
- investigate whether downscaling using auxiliary variables (e.g. land use from CORINE, Digital Elevation Model, road networks, urban morphological zones) improves the results (see De Meersman et al (2016) for more details).

Test data from Orange France is being explored at Orange Labs premises (in cooperation with Orange Labs and INSEE) to analyse the impact of antenna density on statistical indicators and study uncertainty in home location algorithms. Machine-learning will be used to predict land use from mobile phone use.

2.2.2 Wikipedia

Eurostat is exploring the history of access to Wikipedia articles in different domains. Data on the number of page-views per hour for each article are made publicly available by the Wikimedia Foundation. The contributed content of the articles has also been used to a lesser extent.

Experiments are being run, on:

- insights regarding world heritage sites (culture statistics);
- the characterisation of cities (urban statistics); and
- tourism lead indicators and enhanced regional detail (tourism statistics).

The **first experiment** on culture statistics involves analysing page-views for articles on UNESCO World Heritage Sites (WHSs), of which there are around 1 000. There are one or more *Wikipedia* articles (in 31 different language versions) for each WHS and the total number of page-views is taken as a measure of the sites' popularity or of 'cultural consumption'. Possible analysis could involve comparing WHSs and tracking changes over time (see Di Consiglio *et al*., 2015 and Reis *et al*, 2016 for more details). The statistical figures on WHSs derived from this data source are relevant at least for culture and regional statistics.

Insights from *Wikipedia* relate to entirely new topics that are not covered by current data collection from traditional sources (e.g. WHSs).

Wikipedia has around 40 million articles in around 300 languages. The number of possible topics on which the source can provide insights is extremely large and the data

could therefore meet statistical needs in many domains. The timeliness of the data source may also make it very useful for nowcasting.
The **second experiment** involves expanding information on cities from relevant articles.
The **third experiment** seeks to evaluate the use of *Wikipedia* page-views as a source of information to identify factors that drive tourism to an area and whether it is possible to predict tourism flows using these data. The analysis will be performed at city level, considering all points of interest in the area (culture, heritage, sports, nature, leisure, etc.).

*2.2.3 **Origin/destination***

In partnership with DG MOVE, Eurostat plans to explore two potential sources of true origin/destination (O/D) data. Existing air transport statistics in the EU reflect on-flight origin/destination (OFOD), but where passengers change flight in a connecting airport the real airport of origin/destination is unknown. For this reason, Eurostat and DG MOVE decided to explore the use of ticket-booking data. The project is aimed at finding an appropriate source for true O/D data and, after a test phase, moving to production through regular data extractions from the commercial source.

2.2.4 Flash estimates

Eurostat is exploring the use of web activity data in order to produce flash estimates. The project has produced an overview of web activity sources that could be used in the production of statistics, with a particular focus on Google Trends (see Reis et al. 2014). It has also explored several forecasting models that take advantage of the high frequency and timeliness of this type of source and has entered a second phase where the focus is on the use of the data in the production of official statistics; in particular, a new indicator for EU unemployment is being tested.
Eurostat is cooperating with researchers from the Universities of Oviedo and Amsterdam to extend existing nowcasting models to all EU Member States. Also, cooperation has been set up with a researcher from the University of Paris on the use of machine-learning in nowcasting.

2.3 Study outsourced by Eurostat

As part of the action under the big data action plan and roadmap, Eurostat has outsourced studies on relevant issues not covered by ESSnet, as follows:

- ethics and the statistics code of practice;
- communication;
- legal environment;
- skills; and
- the organisation of a big data workshop for the purpose of experience-sharing (http://ec.europa.eu/eurostat/cros/content/ess-big-data-workshop-2016).

As a first step, work on the ethical review and development of guidelines associated with the use of big data in official statistics involves **assessing the European Statistics Code of Practice (CoP) from an ethical point of view**. In the context of big data, issues relating to data ownership, privacy and reputation are more complex than with traditional data, and public perceptions around these issues could have an impact on the image of official statistics.
As a second step, the review will produce **guidelines for (national and international) statistical offices**. It will be based on an analysis of ongoing projects relating to big data

and official statistics and should include consultation of various stakeholders. As well as ethical issues, the review will cover legal matters, with a comprehensive survey of relevant current and upcoming European and national legislation to identify cases where using big data sources for specific statistical purposes or official statistics in general would be inconsistent with relevant provisions and clauses that explicitly or implicitly allow access to, and use of, data sources. Due to resource constraints, the review focuses on four big data source types: telecom data, smart electricity meter data, textual data scraped from the internet and cash register payment data.

On the basis of the review, a **strategy** will be outlined to allow Eurostat or any NSI to communicate their values and commitments to respecting privacy and data-protection legislation. The overall goal is to ensure an adequate level of public awareness regarding the use of big data for the purposes of official statistics and how this is consistent with our core values. In line with a risk analysis based on factors relating to potential public-relations crises, several plans should be outlined so as to ensure preparedness for various eventualities. Mitigating measures will be proposed to cover the various risks.

Finally, a **training** strategy will be drawn up that identifies training objectives in the field of big data and the means by which these can be achieved.

2.4 Preliminary investigation on selectivity issues

Big data offers great advantages for official statistics, but it also poses particular challenges, e.g. big data sources often suffer from self-selectivity (Buelens *et al.*, 2014). Eurostat is therefore conducting a preliminary methodological investigation on this topic (in view of a possible in-depth follow-up study in the near future).

The main objective of the investigation is to identify existing methods that can be used to address selectivity in big data sources, in order to be able to make unbiased inferences for populations of interest in official statistics. After analysing the possible causes of selectivity in some relevant big data sources, e.g. mobile phone data and social media, the investigation will identify the potential relevant literature for adjusting that selectivity, considering at the same time the implications of the sources' features for the application of the methods (e.g. auxiliary variables, uncertainty in profiling, etc.). A first presentation of the results of the preliminary investigation is due to be given at the 2016 ESS Big Data Workshop (http://ec.europa.eu/eurostat/cros/content/ess-big-data-workshop-2016).

3. Other experiments by ESS NSIs

NSIs in the ESS are involved in numerous activities focusing on big data and many have shared information on these through the UNECE big data inventory (BDI), in which eight of the 32 ESS member states are currently represented (see UNECE, 2015), or the UNSD/UNECE big data survey (BDS), in which 21 member states participated (see UNSD/UNECE, 2015).

Owing to the incomplete ESS coverage of the BDI and the fact that the BDS was conducted approximately a year ago, we asked a small subset of respondents (six NSIs) about the completeness of the information and to list any additional activities, in particular those with a potential use in business statistics. In total, the consolidated dataset (from BDS, BDI and the answers to the follow-up) covers 66 big data activities by 22 ESS member states.

3.1 Scanner data and web scraping – for CPI and more

There is a clear shift towards using web questionnaires, scanner data and online data as consumer price index (CPI) sources and virtually all the NSIs (20 out of 22) report on

projects for modernising CPI production by compiling scanner data or by means of web scraping; many report projects using a combination of both approaches. Normally, techniques for using scanner data are far more mature (some NSIs do not even report them as big data projects, as they already use them in regular production processes). Given their relative maturity, scanner data are sometimes used to test/benchmark/validate web scraping results (see Nygaard, 2015, for an example).

While scanner data are used virtually exclusively for price statistics, web scraping is also being explored for a variety of other purposes. Most notably, a number of NSIs are using it to collect job vacancy data from enterprises, in many cases under the Enterprise Websites project (see Vaccari, 2015). There are two projects aimed at applying web scraping for statistics on ICT use in enterprises.

Another type of web scraping application is sampling frame construction. One project in this area aims to create a list of training institutions (for subsequent use as a sampling frame). Another (using web scraping and satellite imagery) aims to identify 'difficult to enumerate' dwelling types in preparation for the next population census. A third, on agritourism farm statistics, also lists enumeration/register issues among the challenges to be tackled.

3.2. Mobile phone data

The second most prevalent data source in the consolidated dataset is mobile phone data, with 10 NSIs reporting on activities based on this source. As is to be expected, the applications deal with movement (transport, mobility, tourism, migration) as well as location (general geolocalisation as well as population statistics).

3.3. Social media

Although social media constitute a big data source, current experimentation with social media data in the ESS appears to be limited. The consolidated dataset contains information on only four projects, of which two deal with Twitter data for measuring consumer confidence (see Daas and Puts, 2014, for a more detailed description of one of these projects) and one uses geolocated Twitter data to infer residence and mobility (see Swier *et al.*, 2015).

One interesting social media application is the ONS (UK) 'Deriving characteristics and potential estimation methods from social media' project, which uses Twitter data and a bespoke survey of Twitter use. It is precisely this type of 'do you tweet?' survey that was suggested by Groves (2013) as a means of allowing extrapolation from big data to the general population and it is encouraging to see experimentation being conducted along these lines.

A number of NSIs and Eurostat are currently in contact with LinkedIn with regard to possible joint research projects in the fields of workforce skills and mobility.

3.4. Road and administration data

A good example of official statistics based on big data is the Statistics Netherlands project on road sensors for traffic intensity statistics (see Puts *et al*., 2014). The consolidated dataset lists a couple of other applications based on road-use data: Statistics Finland (see Piela, 2014) uses road-sensor data as a source of commuting statistics and Statistics Austria uses road-pricing data to enhance road freight statistics.

3.5. Smart meters

Three NSIs report projects using smart meters (for electricity, gas or water). These involve two categories of use: energy (environment) or (as described, for example, in

Williams, 2015) social (population, household occupancy or household consumption) statistics.

3.6. Statistics from web services

Apart from web scraping, there are instances of using information from open web services (such as *Wikipedia* for culture statistics; see Di Consiglio *et al*., 2015) or proprietary web services (such as Google maps) being used, often for the purpose of geolocalisation.

One or two instances of Google Trends being used (for such varied purposes as unemployment and language usage/skills) could also be put into this category (although the amount of 'black box' pre-processing is considerable).

4. European system of interoperable statistical business registers (ESBRs)

4.1 The project and data collection

The European System of interoperable statistical Business Registers (ESBRs) project is part of the ESS's Vision 2020 implementation portfolio and aims to improve the quality of business statistics and, in particular, statistics on globalisation. The ESBRs is a network of statistical business registers in Eurostat and the 32 participating members of the ESS. The project has already delivered important results, including the EuroGroups Register (EGR) 2.0 and the Interactive Profiling Tool (IPT) prototype.

EGR 2.0 is the statistical business register of multinational enterprise groups (so-called 'Global Enterprise Groups') in Europe and contains structural economic information on these groups, their constituent legal units and corresponding enterprises, having at least one legal unit located in an EU or EFTA country. In practice, EGR 2.0 is the central node of the ESBRs and is located in Eurostat.

The IPT prototype has been developed by ESBRs to implement European profiling of the groups (analysing their economic and operational structure, irrespective of borders). It takes into account the groups' global dimension and reflects their activities more accurately.

Statistical business register officers in the 32 countries typically use administrative data to populate their registers. Part of this information reaches EGR 2.0, where it is consolidated and released back to national statistical authorities for national and European statistical purposes. In the context of national and European profiling, statisticians and profilers use various sources, including the internet, to gain access to recent and current information on the groups that they analyse.

4.2 ESBRs inquiry to profilers on the use of big data

A short questionnaire was sent to profilers using the IPT prototype for their European profiling activities in 2016. Replies were received from profilers in 17 EU and EFTA countries.

Profilers use the internet massively, searching for information on groups' structure, operational segments, activities, turnover and employment. All profilers look for consolidated annual reports or even individual legal unit accounts or other financial statements.

All profilers use the internet to retrieve information on the groups, chiefly their websites and the websites of the constituent legal units, the numbers of which can run into several thousands. Most profilers use various other sources on the internet, including business/commercial/trade registers, national and European authorities, commercial data providers and other private databases, digital articles and other web references,

Wikipedia, etc. Most profilers use internet search engines (such as Google) to locate information concerning the groups.
However, they do not use big data approaches, which results in costly and cumbersome 'manual web scraping'.

4.3 Is there potential to use web scraping for profiling?

As seen in section 4.2, profilers use the internet massively, but they do not use big data approaches, which results in costly and cumbersome 'manual web scraping'. There is thus significant untapped potential, as automatic web scraping to achieve profiling would lead to considerable efficiency gains — and the ESS is acquiring the general skills for web scraping in the course of various experiments. Applying ESS expertise on web scraping to the profiling of multinational enterprise groups does therefore seem to be an attractive solution.
However, the devil is in the detail: will techniques that have proven effective for certain web scraping applications be applicable to profiling? Some answers to this might be provided in ESSnet work package 2 on web scraping enterprise characteristics (see section 2.1).

5. Experimentation at the JRC with official statistics potential

The Joint Research Centre (JRC) is the Commission's in-house research service. It has many lines of activity dealing with big data: remote sensing, environmental monitoring, satellite imagery, text- and data-mining, genomics, bioinformatics, migration, transport, internet traffic, analysis of financial transactions, energy modelling, smart networks, etc.
Many of these activities have great potential for statistical analysis. In the following section, we describe a small subset of JRC activities in specific EU policy areas.

5.1 Container tracking

ConTraffic is a project that JRC started over 10 years ago in cooperation with OLAF and DG TAXUD with the aim of supporting customs authorities dealing with the control of containerised cargo.
Information on container routes is not readily available, but it can be compiled from individual Container Status Messages (CSMs). Ocean carriers transporting the cargo containers collect, store and transmit their own CSMs.
These whole-data records describe the overall movement and status of the containers and represent an independent source of information to complement that available to customs and other authorities.
The key idea of ConTraffic is that CSM data can be used efficiently to reconstruct container routes, conduct route-based risk analysis and support ongoing investigations. This allows detailed analysis of the origin of arriving goods, transhipments, where and when they occur, and time stamps. Various web services have been designed on the basis of the source:

- Track and Trace, Container Surveillance is an online service that tracks the movements of specific containers in near-real time;
- Port2Port is an application that shows the results of pre-computed statistical analysis on the logistics routes used by carriers to transport containers between particular departure and destination ports.

JRC has a web scraping project to analyse the electronic messages (EDIs) exchanged by the logistics companies. The task of reconstructing events is challenging due to the

quality of the information: there are non-codified values for locations, events and vessels, incomplete histories (i.e. missing records), messages do not always arrive in chronological order, contain errors, duplicates and obsolete records, and register future events.
Container trip information (CTI) provides core data such as locations, time periods and vessels, and vessel stop information (VSI) describes a (container) vessel's stopover at a terminal, where loading/unloading takes place.
Information is extracted from raw CSMs by applying artificial intelligence techniques to sequences of CSMs for each container to derive its CTI and reconstruct the vessel stops from the aggregated analysis of the CSMs of all the containers on the vessel (using load/discharge/transhipment events).
More information can be found on https://contraffic.jrc.ec.europa.eu/

5.2 AIS, LRIT or VMS

Vessel data consists of self-reporting positioning (often referred to as 'cooperative') data, e.g. AIS, Long range identification and tracking (LRIT) or Vessel Monitoring System (VMS). The fusion of different data is a necessary step to determining the position of vessels at sea.
The vessel positions are submitted via LRIT every 6 hours, and are used to identify traffic routes. This can be used for vessel position prediction and to detect divergence from usual ship routes for safety purposes. Using data-mining and other track-processing techniques, JRC decomposed maritime traffic densities into a set of routes identified by properties such as distribution of speed along the route, travel time and ship type, size, draught, etc. Additional information can be found in Pallotta *et al.* (2013), Mazzarella *et al.* (2015) and on the JRC project website (https://bluehub.jrc.ec.europa.eu/).
AIS can also be used to analyse fisheries activities. In the JRC project, AIS messages were classified as relating either to fishing or to steaming using a classification algorithm based on the analysis of individual vessels' speed profiles. This classification approach proved to be sufficiently robust in the case of trawlers, which represent the majority of vessels over 15 m long. JRC recorded the frequency of AIS to produce a map of fishing intensity and examine the dependencies of coastal communities on fishing grounds.
For each port, additional information is available on the total number of vessels registered in the fleet register, estimated employment and gross value added (GVA) and coverage in terms of the number of fishing vessels for which AIS data were available. (see https://bluehub.jrc.ec.europa.eu/webgis_fish/).

As described in section 2, this source is also being investigated by ESSnet.

5.3 Smart meters and smart grids

JRC also conducts scientific research and supports EU policymaking on the conversion to smarter, interoperable electricity systems. Intelligent electricity networks (smart grids) are the key component in the EU's energy strategy.
JRC manages and regularly updates a comprehensive inventory of power systems/networks and smart grid projects in Europe.
The power grid model is used to run static and dynamic analyses of the European transmission network via advanced power simulation platforms.
JRC's models include:

- EUPowerDispatch, a European electricity transmission network model to analyse the impacts of the future increase of variable renewable energy on European cross-border transmission capacity needs and future plans. The model

covers a time-frame of one year on an hourly basis. The main model outputs are generation levels for each energy source at each node, marginal variable electricity production costs at each node and cross-border flows. In addition, CO_2 emissions and renewable energy curtailment needs are calculated;

- resLoadSim, a tool for residential electric load simulation and predicting residential loads, which uses a probabilistic approach to predict the electric load profiles of individual households by calculating their total appliance consumption.

On the basis of its experience of analysing and modelling energy supply, JRC is also studying smart-meter data to predict energy usage and match demand and supply.
The data can also be used to identify 'energy profiles' in order to redesign energy policy models.

As described in section 2, ESSnet is also investigating this source. Eurostat has set up a network of cooperation between ESSnet partners and JRC researchers in order to share experience in this field (and others).

JRC's activities are described here: http://ses.jrc.ec.europa.eu/smart-grids-observatory.

5.4 Text analytics

The European Media Monitor (EMM) tool (https://ec.europa.eu/jrc/en/scientific-tool/europe-media-monitor-newsbrief) is an online project run by JRC. A number of research projects are investigating the possibilities for extracting value based on text analysis of scraped data. One of these is about disambiguating company names in text. One major potential application is exploiting company introduction paragraphs in job vacancy notices, with potential benefit for the activities of the ESSnet work package on web scraping (see section 2). The application areas of text-mining are numerous and JRC has already started to develop tools and prototypes that can be extended in several fields, e.g. JRC's work on technology innovation monitoring (TIM) http://www.technologymonitoring.eu.

6. Experimentation in the European System of Central Banks (ESCB)

Like the NSIs, central banks have started to see a need to cooperate and build a roadmap to facilitate the use of big data sources. A few projects have already started: central banks are now collecting data from various sources (e.g. administrative records, search engines, reporting banks' and financial institutions' databases, online news and supermarket records), in order, for example, to nowcast unemployment, conduct research on price dynamics, improve understanding of credit risk, market risk and financial operations, and analyse micro data to support their policymaking (Irving Fisher Committee on Central Banks Statistics, 2015).
Pilot projects have just been launched to assess the use of big data from four source groups: administrative data, internet data, commercial data and financial market data. The two main potential projects using **administrative data** relate to trade and tax income. The pilots on **internet data** (queries, social media, internet media text and portals, internet price information) are intended to measure confidence, sentiment and attitudes to economic activities and entities. The nowcasting of industry and housing production, retail sales and unemployment are other important goals. Internet price information will be assessed to study the dynamics of consumer prices and house prices.

Pilots on **mobile phone data** will address their usefulness in measuring tourism and estimating balance of payments.

7. Horizon 2020 research opportunities

Horizon 2020 (https://ec.europa.eu/programmes/horizon2020/) is the biggest EU research and innovation programme ever, with nearly €80 billion of funding available over seven years (2014 to 2020). The current Horizon 2020 work programme, for 2016-2017, was published in October 2015. Among the very large number of calls relating to big data, a few could apply to official statistics strategies.

7.1 Call on data-driven policymaking

The call on 'Policy development in the age of big data: data-driven policymaking, policy modelling and policy implementation' (H2020 CO-CREATION-06-2017) is aimed at exploring and experimenting with ways of exploiting data, focusing on the development of methods for using big data in policy development, the assessment of economic, political, ethical and legal issues, and the implications of big data practices. The project will develop:

- methods and tools for compiling, analysing and visualising data;
- methods for metadata schemes, data-linking or the reconciliation of multiple datasets for coherence; and
- data-mining methods for policy modelling and simulation.

The deadline for the call is 2 February 2017.

7.2 Other Horizon 2020 big data calls

The following domain-specific calls in the Horizon 2020 *societal challenges* strand may also have an official statistics impact:

- 'Big data supporting Public Health policies' (SC1-PM-18-2016) on the acquisition, management, sharing (this includes security and privacy aspects), modelling, processing and exploitation of data in support of public authorities; and
- 'Big data in Transport: Research opportunities, challenges and limitation' (MG-8.2-2017), which is aimed at identifying methodological issues and developing the necessary tools to enable the effective mining and exploitation of big-data sources. With a view to reducing limitations, projects will examine the institutional and governmental issues and barriers as regards the use of big data in transport and provide policy recommendations on 'data openness' and sharing.

8. The way forward for big data research in the ESS

As discussed at length in this paper, several ongoing experimentation and research initiatives in Europe are studying the use of big data in official statistics. There is considerable fragmentation, as many of these being carried out locally and independently of each other (see section 3 for examples). The main exception to this pattern is the ESS Vision 2020 BIGD project (see section 2), within which ESSnet is a major collaborative

undertaking and activities carried out (directly or outsourced) by Eurostat play a complementary role in implementing the ESS Big Data action plan and roadmap 1.0 (ESS Committee, 2014), where it is noted that:

> 'Embarking on the integration of big data into official statistics is a nontrivial activity, taking place in a dynamic environment. External events, as well as findings made along the way during the implementation of the Action Plan will most likely trigger the inclusion of new actions and the refocusing of existing ones. For this reason, the planning of the project should be circular and follow a stepwise approach. After each step there should be an evaluation of the results which should feed a review of respective parts of the action plan and the roadmap.'

In the next such review, an update to a version 1.1 or even 2.0 of the action plan might be needed to achieve the medium-to-long-term objectives of the roadmap. This could include a dynamic research strategy, possibly with (but certainly not limited to) the following strands of action:

- the use of Horizon 2020 research lines (see Section 7);
- the promotion of (and development of enabling frameworks for) large-scale public-private partnerships; these are necessary not only for reasons of funding and/or requirements for multi-disciplinary teams, but also in order to address issues of data access. Very often, research projects fail to ensure adequate access to privately held data; and
- earmarking part of the research funding for *Smart Statistics*, a project to analyse the technological foundations of a future statistical information system operating in a context of wider digitisation in society and greater interactions between the web, the web of data and a multitude of smart environments based on the 'internet of things', such as smart cities and Industry 4.0.

Eurostat is committed to raising awareness of the need for the ESS and the global statistical system to undertake (even mainstream) research on *smart statistics*. The question is not whether smart statistics will become a reality, but whether the official statistics community will actively engage in the development of such (largely) automated production processes sufficiently early to ensure that NSIs remain relevant in such a data ecosystem. If we fail to rise to this enormous challenge now, others will doubtless take over.

Acknowledgements

The authors would like to thank the colleagues at the ESS NSIs that took their time to reply to the UNSD/UNECE survey on organisational context and individual projects of Big Data, contribute to the UNECE big data inventory and to reply to the survey to the profilers on Big Data practices in their profiling activities. Without their detailed input, our overview of big data and profiling activities in the ESS would not have been possible.

References

B. Buelens, P. Daas, J. Burger, M. Puts and J. van den Brakel (2014), Selectivity of big data; http://www.pietdaas.nl/beta/pubs/pubs/Selectivity_Buelens.pdf

P.J.H. Daas and M.J.H. Puts, (2014). Social Media Sentiment and Consumer Confidence. European Central Bank Statistics Paper Series No. 5.

F. De Merseen, G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis, H. I. Reuter (2016) Assessing the Quality of Mobile Phone Data as a Source of Statistics, Quality Conference 2016

L. Di Consiglio, C. Donovan, B. Kovachev, A. Murray and F. Reis (2015), Wikistats project report; http://www1.unece.org/stat/platform/display/bigdata/Report%3A+Wikistats

ESS Committee (2013), Scheveningen Memorandum on 'Big Data and Official Statistics', http://www.cros-portal.eu/content/scheveningen-memorandum

ESS Committee (2014), ESS big data action plan and roadmap 1.0 http://ec.europa.eu/eurostat/cros/content/ess-big-data-action-plan-and-roadmap-10

R. Nygaard (2015), The use of online prices in the Norwegian Consumer Price Index. Paper presented at the 14th meeting of the Ottawa Group; http://www.stat.go.jp/english/info/meetings/og2015/pdf/t1s2p5_pap.pdf

R.M. Groves (2013), Official Statistics and "Big Data". Keynote address at the NTTS 2013 conference http://ec.europa.eu/eurostat/cros/content/keynote-address-robert-m-groves-slides.

I. Jansson, M. Limbek, B. Nikic, O. Nyqvist, K. Potocki, M. Puts and D. Wu (2015), Report:Enterprise Web sites; http://www1.unece.org/stat/platform/display/bigdata/Report%3A+Enterprise+Web+sites.

Irving Fisher Committee on Central Bank Statistics (2015), Central banks' use of and interest in "big data", http://www.bis.org/ifc/publ/ifc-report-bigdata.pdf

F. Mazzarella, M. Vespe and C. Santamaria, (2015) SAR Ship Detection and Self-Reporting Data Fusion based on Traffic Knowledge, IEEE Geoscience and Remote Sensing Letters, vol. 12.

G. Pallotta, M. Vespe and K. Bryan (2013) Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction, Entropy, vol. 15,

P. Piela (2014). Commuting time for every employed: combining traffic sensors and many other data sources for population statistics. Paper presented at the European Forum for Geography andStatistics; https://geo.stat.gov.pl/documents/20182/25356/4_EFGS+2014+paper+Piela.doc

M.J.H. Puts, M. Tennekes and P.J.H. Daas (2014) Using Road Sensor Data for Official Statistics: Towards a Big Data Methodology. Paper presented at Strata+Hadoop World; http://conferences.oreilly.com/strata/strataeu2014/public/schedule/detail/37462.

F. Reis, P. Ferreira and V. Perduca (2014) The use of web activity evidence to increase the timeliness of official statistics indicators, presented at the IAOS 2014 Conference on Official Statistics

F. Reis, L. Di Consiglio, L., B. Kovachev (2016), Big data in official statistics - Insights about world heritage from the analysis of Wikipedia use, International Symposium on the Measurement of Digitized Cultural Products

N. Swier, B. Komarniczky and B. Clapperton (2015). Using geolocated Twitter traces to infer residence and mobility. GSS Methodology Series No 41; http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/gss-methodology-series--41--using-geolocated-twitter-traces-to-infer-residence-and-mobility.pdf.

UNECE (2015). Big Data Inventory; http://www1.unece.org/stat/platform/display/BDI

UNSD/UNECE (2015). Results of the UNSD/UNECE Survey on organizational context and individual projects of Big Data; http://unstats.un.org/unsd/statcom/doc15/BG-BigData.pdf.

C. Vaccari (2015). Report: Enterprise Web sites; http://www1.unece.org/stat/platform/display/bigdata/Report%3A+Enterprise+Web+sites

S. Williams (2015). Modelling sample data from smart-type meter electricity usage. paper presented at the NTTS 2015 conference; http://ec.europa.eu/eurostat/cros/content/ntts-2015-proceedings .