Center for Big Data Statistics at Statistics Netherlands

Piet J.H. Daas (pjh.daas@cbs.nl)[1], Sofie De Broe[1] and Magchiel van Meeteren[2]

Keywords: Big Data, data science, modernizing official statistics, collaboration

1. INTRODUCTION

On September 27, 2016 Statistics Netherlands (CBS) launched its Center for Big Data Statistics (CBDS) during a Dutch economic mission to South Korea [1]. CBS has been using big data in its production of statistics for a number of years. One of the results was that in mid-2015, CBS became the first statistics bureau in the world to launch official traffic statistics produced with big data [2]. Main drivers for setting up the CBDS are the positive experiences with big data at CBS, the potential big data has to offer, the need for a single contact point for knowledge on this exciting and rapidly developing area and the choice to increase the cooperation with partners.

1.1. Objectives

The objectives of CBDS are threefold. The first is to enable the faster production of statistics, with real-time statistics as the ultimate goal. This will enhance the responses to the Dutch society's need to receive usable information more quickly. The second objective is for existing statistics to become available at a lower aggregation level, more in particular those with data on regional and urban areas. In addition, big data offers opportunities to make statistics production more flexible and to formulate new indicators. The third and final objective is based on the zero footprint concept; e.g. reducing the administrative burden at companies and for individuals further by deploying new sources. It is additionally expected that implementing big data will also lead to improvements in efficiency and quality.

2. CBDS IN MORE DETAIL

2.1. People

At CBDS, several groups of people will work together on the above mentioned objectives. Several roles are discerned in the center. These are: data scientist, data scout and domain expert. Data scientists are specialists with big data specific IT-skills that focus on the methodology of processing, checking, editing, and visualization of big data. They work for nearly all or for a major part of their time in the center. Here, they use their skills to analyse and combine data and extract information while developing big data specific methods. Data scouts are people that focus on finding interesting data sources and arranging access to them. They contact various companies and other organisations to discuss the possibilities of obtaining data by, for instance, starting joined projects or initiate other forms of cooperation. Data scouts work for a limited part of their time at CBDS. Domain experts are statistical specialists that work for the various statistical divisions of CBS. They join the CBDS when a project is initiated in which their particular expertise is required. These specialists are essential for producing reliable big data based statistics of high quality. Apart from the above mentioned CBS-employees, PhD-students, university students and experts associated with various

[1] Statistics Netherlands, Center for Big Data Statistics & Process Development and Methodology

[2] Statistics Netherlands, Center for Big Data Statistics & Data Collection, Production Services

national and international parties will work at CBDS when involved in various projects. Currently the possibility of involving Data engineers, big data IT specialists, are being discussed. The CBDS is led by a programme manager who is overall responsible for the work at the center, the daily guidance, and the work program. The scientific content is managed by a scientific director who leads the data scientists and assures the results obtained are methodologically sound and of the highest quality possible. A Steering Committee with representatives of the board of directors and the statistical divisions of CBS makes strategic decisions and support the programme manager.

2.2. Equipment

In the CBDS lots of IT equipment, such as workstations with lots of memory (≥ 32GB), a server on which e.g. web robots are run and a SPARK cluster, are available. On these machines large amounts of data are being collected, processed and analysed. The equipment is positioned within a secure environment. Physically, the CBDS is located in both offices of CBS; one in Heerlen and one in The Hague. Both locations are connected via a video connection. The main focus of CBDS is the office in Heerlen. In the center, predominantly open source software is used. The preferred programming languages are R, Python and Scala. In a separate room, big data training courses are given.

2.3. Work program

The program of the CBDS consist of various themes; shown as horizontal bars in Figure 1. These times are also called flagships and cover social, economy and mobility statistics. Two special areas of interest are identified, which are Sustainable Development Goals and Smart Cities. The themes are product oriented and contain a combination of potentially interesting statistics which may have a short or long term objective. The first are usually part of case studies that are either the result of short sprints (following a Scrum approach) or are produced with the assistance of students or partners. Examples of some of the short term topics currently investigated are: a social media based real-time unsafety monitor, a traffic oriented quarterly indicator for the economy, a dot-map based visualization and a so-called ‘pepernut’ index; the latter focuses on food product sales related to the feast of Saint Nicholas. Statistics with a long term objective are either initiated by a statistical division of CBS or (part of) an externally financed research project. Examples of projects with a long term goal are: Big Data based Sustainable Development Goals Indicators, statistics based on data from the Automatic Identification System of ships and modernization of the Consumer price Index.

The vertical bars in figure 1 are composed of one of four activities essential for the CBDS and the statistical products produced. These are i) Data scouting and data access, ii) Ethics and privacy, ii) Methodology and data integration and vi) Big data in official statistics related. The first, data scouting, assures that the data needs are identified and made available for the products intended to be produced. Ethics and privacy focuses on assuring that measures are taken to protect the privacy of the ‘units’ in the sources and that only the part of the data that is needed is collected, stored and processed; more on this in [3]. The methodology part focuses on applying big data methodology (see below). The last vertical activity shown in figure 1 focuses on making sure regular production of big data based official statistics is possible. Experimental products are first published on the innovation portal of CBS [4] on which feedback is requested.

2.4. Methodology

This work in this area is performed in close cooperation with the methodologist involved

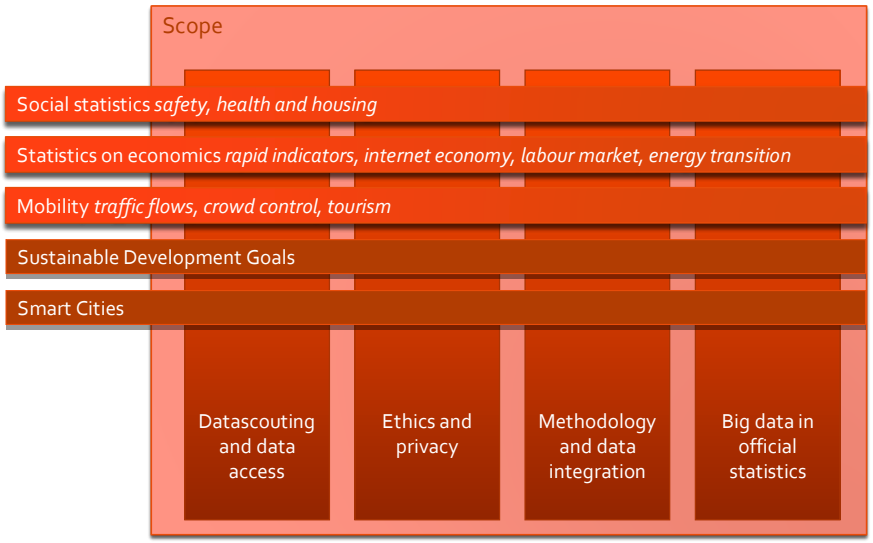

Figure 1. General overview of the working program of CBDS and the interaction of the various themes (horizontal bars) and activities (vertical bars).

in the big data research theme of CBS. This work lays the foundation on which reliable big data based statistics can be produced. Three ways can be discerned by which the big data can be used for the production of official statistics. These are:

1. Survey based with Big Data as an additional source
Here the results obtained from survey or admin data essentially lay the foundation. Big Data is used as an additional source of information in two different ways. The first is as an additional data collection mode. The use of scanner data and data collected by web robots as input for the Consumer Price Index illustrate this. The second way is using big data as auxiliary information in a model-based inference approach. Examples in which this approach has been applied successfully are: the study of the effect of social media sentiment on the Consumer Confidence Survey and the use of mobility data for the study of well-being and other socioeconomic phenomena. From a Big Data perspective, the methods used to extract the information from Big Data are of interest which are also relevant for the approaches described below.

2. Census based with Big Data as the single or main source
In this case focus is on the coverage of the target population in the source. When the latter is completely included, one can produce a so-called census-based Big Data statistic. Examples of this are: traffic intensity statistics based on road sensor data, a consumer price index solely based on product prices collected from the web, land-use statistics based on satellite images of a country and waterway transport statistics based on inland Automatic Identification System transponder data from ships. Of these examples, only the first two have actually been produced. Coverage is key in all these applications. Most important methodological issues here are efficiently processing huge amounts of data, dealing with errors in large data sets, removing units not belonging to the target population and comparing the variable(s) available with the information needed. The latter is an area familiar for statisticians using admin data.

3. Big Data based with the target population not completely included
When Big Data is used as the single or main source and the target population is not completely covered, finding ways to deal with the selectivity of Big Data becomes the most important question. Examples of studies in which this is the case are: a social media based unsafety monitor, a social media based sentiment statistics, day time population statistics based on mobile phone data and tourism statistics based on mobile phone data.

In the second and third case attempts were made to correct for the missing part of the population. However, even though the findings are obviously biased, they also reveal the potential of using Big Data in these specific areas. Additional challenges when dealing with Big Data in these cases are the fact that most Big Data sources are event oriented and many of these sources hardly contain background characteristics of the units included. This stimulated the need to found ways to derive background characteristics which has been demonstrated to be possible.

2.5. Partners

A long list of national and international organisations has endorsed the CBDS. These organisations will provide data, resources, infrastructure, expertise and knowledge which are needed to successfully perform joint projects. Associated partners are: APG, Brightlands Smart Services Campus, CapGemini Nederland, Centre for Innovation Leiden University, CGI, De Nederlandsche Bank / Dutch Central Bank, Dell-EMC, Deloitte, Dutch Open University, Eurostat, Fontys University, HumanityX, IBM, Jheronimus Academy of Data Science (JADS), KPN, Leiden University, Maastricht University, Maryland University, Michigan University, Microsoft, Ministry of Economic Affairs, Province of Limburg, SURFSara, TNO, University of Amsterdam, University of Twente, World Bank and Zuyd University. Associated national statistical offices are the statistical offices of Italy (ISTAT), Slovenia (SORS), Estonia (ES), Finland, Korea (KoStat), Sweden (SCB) and the United Kingdom (ONS). The most recent version of this list can be found at [5].

3. CONCLUSIONS

The initiative to launch the CBDS at CBS has resulted in a focus of all Big Data initiatives at the office. This has brought together all disciplines in the area of Big Data, speed up efforts and created a clear identity inside and outside the office. As a result of this, an increasing number of Big Data projects are being initiated at the office, some of which will undoubtedly form the basis for new, quicker available and/or statistics with a zero footprint.

REFERENCES

[1] CBS, CBS starts unique initiative for Big Data Research, September 27 (2016), Located at: https://www.cbs.nl/en-gb/news/2016/39/cbs-starts-unique-initiative-for-big-data-research

[2] CBS, A first for Statistics Netherlands: first Big Data based statics launched, (2015), 1-3, Located at: https://www.cbs.nl/NR/rdonlyres/4E3C7500-03EB-4C54-8A0A-753C017165F2/0/afirstforlaunchingstatisticsbasedonbigdata.pdf

[3] ESSnet Big Data, Work Package 7, Milestone 7.5: List of available Big Data sources in the domain(s), Legal aspects, The Netherlands, section 3.1.3, (2016), 15-19, Located at: document https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/9/92/WP7_Milestone_7_5.docx

[4] CBS, Innovation page, October 25, (2016), Located at: https://www.cbs.nl/en-gb/our-services/innovation

[5] CBS, List of CBDS partners, October 25, (2016). Located at: https://www.cbs.nl/nl-nl/onze-diensten/unieke-samenwerking-voor-big-data-onderzoek/internationaal-partnernetwerk-van-cbds (in Dutch)