# New forms of data for official statistics

**Niels Ploug**
**Statistics Denmark npl@dst.dk**

## Abstract

**Keywords**: administrative data, Big Data, data integration, meta data

## Introduction

The use of new forms of data in official statistics is both a challenge and a possible solution to more than one challenge facing official statistics worldwide. It is a challenge because official statistics traditionally is based on large survey samples where the content of data are carefully defined in order to serve a specific purpose for official statistics being e.g. the collection of data for the calculation of the labor market participation - or the unemployment – rate, while the use of new forms of data like e.g. data from administrative sources or Big Data not in the outset are data that are created to cater the specific and well defined purposes of specific official statistics.

At the same time is can be the solution to challenges facing official statistics for at least two reasons. It is well known that all kinds of survey based data collections have response rate problems being it survey for official statistics or research. The use of data from other sources like e.g. administrative records does not have response rate problems as the data are taken from a source where those who produce the data – the staff working in e.g. the educational system or the health care system – not only have an interest in but also an obligation to record the data. It is all well known that running a survey organization is a complicated and not least costly way of getting data. And as National Statistical Institutions (NSI's) in general are faced with budget cuts and the challenge to make ends meet the use of non-survey generated sources of data can be part of the solution to the budget challenge.

This paper deals with the possibilities and the challenges for NSI's in the use of administrative data for the production of official statistics and it deals with the potential future possibilities of developing new official statistics by using new data sources like Big Data.

## Big Data and official statistics

The use of Big Data has been on the agenda for official statistics for quite some time as an issue being presented and debated at international conferences and seminars world-wide. The ISI 2013 conference is no exception to this. On top of an UNECE/CES seminar in October 2012 a task team of national and international experts developed a condensed paper addressing the issue (UNECE, 2013). This paper classifies what is termed 'large data sources' into the following categories:

- Administrative data
- Commercial and transactional data
- Sensor data from e.g. satellites
- Tracking data from e.g. mobile telephones or GPS
- Behavioral e.g. on-line search
- Opinion e.g. comment on social media

Common features in relation to all kinds of Big Data are that they have a high frequency – contrary to survey data Big Data are generated every day, hour and minute – and they have a high degree of granularity. This is both and advantages and a challenge.

The advantages is that we are dealing with very rich data sets that are produced on a daily basis making it possible for official statistics to live up to the widespread user demand for more timely data. The challenges are conceptual, technical and strategic'. In order to be relevant Big Data should be able to support the production of statistics on some of the overall issues covered by official statistics like the development in the economy, in the business sector and social developments in relation to the citizens. The technical challenge has to do with the fact that Big Data are not only in relation to the amount of data but also in relation to an unstructured data model where data by no means are produced in order to cater official statistics. Even though this is a challenge it is not unknown to NSI's to deal with big and complicated sets of data and to transform raw data into data registers that can be used as the basis for the production of statistics.

The strategic challenge is related to the issue of relevance. NSI's are in many ways in a very strong position delivering high quality statistics for society – but this is a situation based on past efforts. In relation to the potential in the use of Big Data NSI's risk to lose their reputation as being the core institutions in the delivery of statistics on society and their relevance as Big Data has major advantages e.g. in relation to timeliness – if NSI's do not get on board and sets the agenda for the use of Big Data in official statistics.

NSI's has the advantage of being familiar with the organization, documentation and dissemination of data. In relation to Big Data it is especially important to focus on documentation and metadata in order to set a high quality standard for the use of Big Data. Not only are NSI's used to this kind of work – but at an international level NSI's work together through UN and regionally in e.g. ESS – The European Statistical System – to make sure that data are not only well documented but also comparable between countries – which is important also in relation to the future use of Big Data.

## Examples

There are already quite a lot of examples of the use of Big Data by NSI's. Administrative data has been used for statistical purposes for quite some time not least by the NSI's in the Nordic countries. At Statistics Denmark most social statistics has been based on administrative data for quite some time (cf: Eurostat & Statistics Denmark, 1995) and the use of administrative data has become an integrated part of the strategy for the development of statistics at European level (cf. Eurostat, 2013).

The use of administrative data is therefore not new in relation to the production of official statistics – but the digital revolution has created a huge pile of new opportunities in relation to the production and use of data on top of the use of survey and administrative data for official statistics.

The list of devices and actions that produce electronic footprints and by that micro level data is long:

- Credit card transactions
- Commodity (RFID) tracking
- Toll road recording
- Electronic tickets for travelling and entertainment
- Public services offered electronically
- Immigration control
- Mobile phone use
- Internet and social media use
- GPS tracking of traffic and transport

This creates the possibilities both for new data sources for existing statistics and for the creation of new kinds of statistics.

Traffic and transport statistics is one area where it will be obvious to use Big Data. Both toll road recording and electronic tickets for public transportation can among other things be used for statistics on internal mobility e.g. mobility between an economic (city) center and the periphery Based on Big Data this kind of statistics can be used to a general mobility statistics covering the number of vehicles and persons moving in and out of the economic center. A more detailed statistics breaking the movements down into time slots could be useful for planning purposes, and a statistics on rise or fall in the movements of big vehicles (lorries and trucks) in and out of the economic center could be an early economic indicator.

Consumer statistics is another area where Big Data could be useful. The use of data from RFID tracking could be an alternative to the traditionally way of collecting price data for a major part of the consumer goods included in price statistics. In many countries consumer goods a sold by specialized chains of e.g. supermarkets, stores for electronics or clothing. Access to scanner data from cash registers from these chains could be an important and cost efficient input to price statistics. Data from electronic tickets in relation to entertainment could be another input in relation to consumer statistics – and a potential early economic indicator as money spend on entertainment potentially are an important part of the more cyclical fraction of consumption. Data from internet based reservation

systems for hotels, summer houses and flight tickets could also be an important input to an 'early economic indicators' statistics.

Internet traffic is a new area for statistics which can give important knowledge on the use of internet in different areas of a country – as an indicator of the countries it literacy – which is important knowledge in relation to e.g. plans for digitalization of the communication between citizens and the public sector.

On top of that the potential digitalization of the public sectors communication with citizens can in itself create important and interesting Big Data. Statistics on the development in applications for social benefits, the development in permissions to build or expand an existing house or a factory could also be important early economic indicators.

All in all Big Data creates a lot of new possibilities for official statistics.

## Issues to be addressed in relation to the use of Big Data in official statistics

The use of Big Data being it data from administrative sources or data from electronic footprints is not only a possibility for official statistics it is also a challenge.

The challenge lies in how to use the data for the production of statistics and in the documentation of the quality of the statistics. Metadata are important in general and very important in relation to the use of Big Data.

Contrary to statistics based on surveys directed for the production of statistics in a specific area like e.g. the labor force survey or the time use survey – statistics based on Big Data has to use several sources of data. This calls for a system of data integration cf. Zhang, 2012. At Statistics Denmark data integration from different administrative sources is performed at a daily basis to check the quality of information in e.g. the employment and/or unemployment statistics. It is among other things checked whether persons who show up in the tax registers as being employed during a specific period of time e.g. a month has received any kind of income transfers during that specific period of time – and vice versa if persons who have received e.g. unemployment benefit for a specific period of time receives any kind of earned income during the same period. This is done through the integration of information from the tax registers with information from other e.g. the register on recipients of sickness benefits, parental benefits etc.

In general there is a need to develop a coherent system of quality assurance for the use of Big Data including the use of administrative data for the production of official statistics. This system should include at least the five following points[1]:

[1] All based on a presentation by Li-Chun Zhang 'On quality of statistics based on administrative register' at Statistics Denmark……

- Technical checks – is it technical possible to use the source file and data in the file.
- Accuracy – are data correct, reliable and certified e.g. to what extent exist erroneous and/or untrustworthy objects in the source.
- Completeness – the degree to which a data source includes data describing the corresponding set of real world objects and variables e.g. under-coverage (absence of target objects) or over-coverage (presence of non-target objects).
- Time-related dimension – indicators that are time and/or stability related; e.g. time lag between the end of the reference period in the source and the moment of receipt.
- Integrability – the extent to which the data source is capable of undergoing integration or of being integrated in the statistical system e.g. the linking ability by use of common identifies in the data source and the statistical system.

There is a huge need for the creation of a quality assurance system for Big Data in order to enable the users of data to check data quality in relation to the above mentioned points. This need certainly exists for the use of administrative data but it becomes even more important for the use of Big Data like electronic footprints.

## Conclusions

It is very easy to become very enthusiastic about the potential that lies in the use of Big Data for the production of new, interesting and timely statistics. But Big Data also pose important challenges to not least to official statistics. Ahead is hard work in the codification and production of a meta data system to support the use of Big Data.

In using data from administrative sources and Big Data like electronic footprints official statistics moves away from survey sources that is totally in the control of the statisticians into using data from sources like administrative registers and electronic devises that in the outset are not there for the production of data for statistics. There a lots and lots of advantages in relation to cost and timeliness but also challenges in relation to usability, quality and documentation.

The 20$^{th}$ century was the century of the birth and maturing of survey sampling. The 21th century should be the century of data integration.

## References

Eurostat (2013): European Statistical Programme - Annual Statistical Work Programme 2013

Eurostat and Statistics denmark, “Statistics on Persons in Denmark, A register-based statistical system” (1995).

UNECE (2013): What does ‘Big Data’ mean for official statistics?

Zhang (2012): Two-phase life-cycle model for integrated statistical data, Statistica Neerlandica.