Big data techniques for supplementing statistical business registers

Tomasz Klimanek, Marcin Szymkowiak[1]

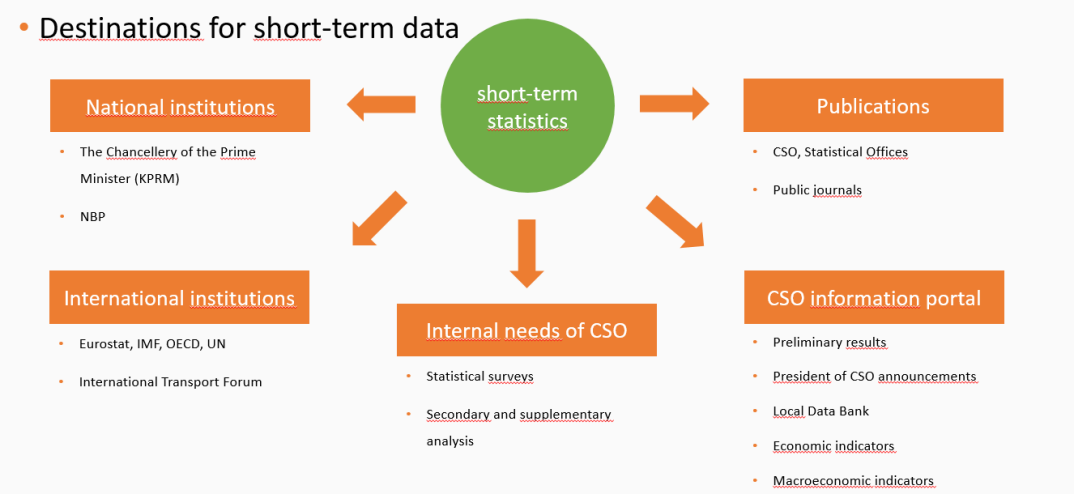
Introduction – DG1 survey description and it's challanges

The aim of the DG1 survey is to create a system of quickly accessible information about basic measures of economic activity of enterprises and provides data used to derive most short-term indicators about activity of enterprises and compile announcements made by the CSO President about the average monthly wage in the enterprise sector.

Survey results are used for purposes of:

– preparing current information about the socio-economic situation of the country and its provinces,
– quarterly national accounts,
– central and local government agencies and other institutions,
– National Bank of Poland,
– Local Data Bank
– EU, OECD, IMF and UN.

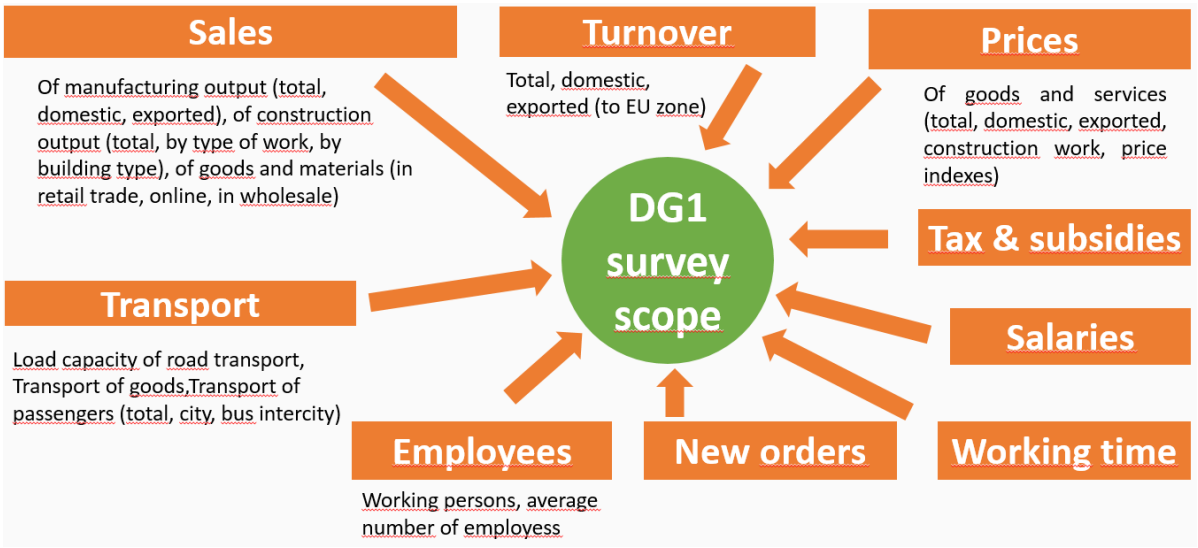
Fig. 1. Importance of DG1 survey



[1] Tomasz Klimanek - Statistical Office in Poznań, t.klimanek@stat.gov.pl, Marcin Szymkowiak, Statistical Office in Poznań, m.szymkowiak@stat.gov.pl, Poznań University of Economics and Business, m.szymkowiak@ue.poznan.pl

Target population in the survey is defined as legal persons and organizational units without legal personality and natural persons employing at least 10 persons and conducting economic activity classified into sections B through J, K, M (with the exception of divisions 72 and 75), N R and divisions 02, 95, 96 and class 03.11.

Fig. 2. Collected information



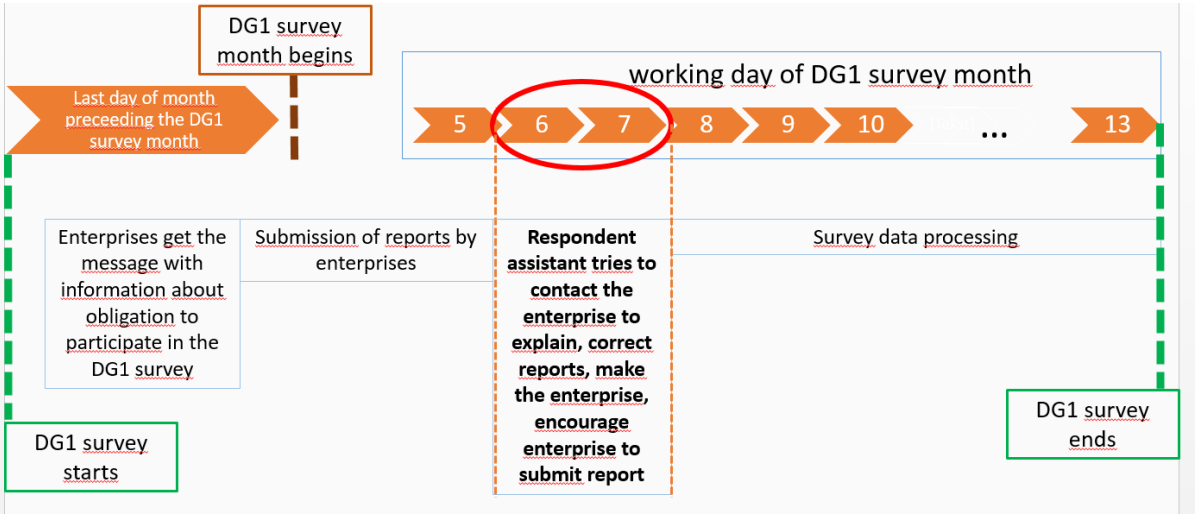
The measurement of current economic activities of enterprises (monthly DG-1 survey) is an element of short-term enterprise surveys. Data collected in this survey are used to calculate short-term indicators and meet the requirements of European statistics. These short-term indicators represent basic measures of economic activity in different sectors of the economy (manufacturing, construction, retail and wholesale trade, accommodation and food service activities, transportation and storage, information and communication, professional, scientific and technical activities, administrative and support service activities): sales, domestic and non-domestic turnover, new domestic and non-domestic orders in selected sections of economic activities, employment and average wages and hours worked. In addition, the DG-1 survey is a source of data about the structure of sales revenue, which serves as a system of weights for calculating producer prices indexes.

The DG-1 survey is administered at the level of particular provinces. Data are collected, processed and generalized at the level of 16 provinces by designated teams of statisticians assigned to each province. The teams consist of employees of the Statistical Office in Poznań and the survey is conducted by the Poznań office and its five branch offices.

Data are mainly collected via the reporting portal and are either provided directly by enterprises or transferred to the system by statisticians who have received the data by phone or in paper questionnaires by mail.

Since data completeness is crucial for the survey, the recommended approach in contacts with respondents is one characterized by a high degree of flexibility, in an effort to obtain data regardless of how they are provided.

Fig. 3. Timeline of DG1 survey



One of the main challenge in the survey is pointed in the Fig.3. – 6th and 7th working day of DG1 survey in the given survey month. This is the period when respondent assistants try to get in touch with the enterprise to explain data in reports (structural breaks), correct reports in case of errors, make or encourage enterprise to submit the report in case of the nonresponse. However this is very often a difficult part of the survey as there is no obligation to submit phone, e-mail address, website address to Business Register (see Fig. 4. and Fig. 5.).

Fig. 4. DG1 survey – January 2016 edition – BJS contact information

numer of all units	numer of units obliged to take part in DG1 survey	with phone number	no phone number	with at least 1 e-mail address	no e-mail address at all
COUNTRY LEVEL					
100722	33295	32768	527	33096	199
WIELKOPOLSKA REGION LEVEL					
11536	3438	3380	58	3420	18

Fig. 5. DG1 survey – January 2016 realization

	Respondent contact				Changes in BJS		Not updated		Web search	
	Phone		E-mail							
	Explanations/ Errors	Reminders	Explanations/ Errors	Reminders	Phones	E-mails	Phones	E-mails	Phones	E-mails
POLAND	4087	3595	735	3000	102	67	327	202	592	352
Wielkopolska Region	261	520	43	319	64	14	70	20	87	29

Framework of Web Scraping for Official Statistics – Law Issues in Poland[2]

The law issued on 29 June 1995 on official statistics in article 5 p. 1 gives a general permission to collect and store the data from all data sources specified in this legal act as well as in other regulations or in different legal acts. Therefore it is a legal act that is used by official statistics to confirm formally the basics of data processing.

Big Data analysis for experimental purposes is possible because of the obligation for official statistics to conduct research work as well as research and development work, including methodology or mathematical methods and applying IT methods in statistics. This work must be done bearing in mind the protection of personal data.

Current law or provisions in agreements/rules of websites concerns only commercial use of data. It is important to update the current regulations and to introduce a new law for non-commercial use of the data and further processing of this data with Big Data tools.

Based on art. 5 p. 1 written in the Law issued on 29 June 1995 on official statistics, it can be concluded that other data sources and data tools can be used to gather statistical data by official statistics. However it is necessary to formulate those additional forms in a legal act.

Testing 'rvest' R package

The growth of data availability from the WWW is getting more and more easy. Therefore, it presents the opportunity for data collection also for statistical purposes. Only valuable data should be extracted from the web pages of interest. There are a lot of irrelevant data, such as advertisements, which should be excluded by data extraction process. Data

[2] Based on the unpublished paper prepared for WP2 ESSnet on BigData project (J.Maślankowski, 2017)

extraction, or data scraping, is the problem of extracting target information from web pages to produce structured data that is ready for post-processing. It could be used by the use different software and application. In the presentation we show the initial results of using pattern matching on extracted data from webpages to find some basic contact information which is missing in statistical official business register (eg. telephone number, e-mail address, tax id number, company registration number, etc.). For that purpose R software environment and rvest package were used for statistical analysis, data processing, and web mining. Rvest package however do not provide basic crawling, because it can only parse and extract contents from URLs. They must be collected and provided manually. This will be the aim of our future work.

References

Klimanek T., Potocki K., Big Data techniques supporting a short-term statistics survey, presentation prepared for ICES-V The Fifth International Conference on Establishment Surveys Geneva, Switzerland, June 20-23, 2016.

Maślankowski J., 2017 Framework of Web Scraping for Official Statistics – Law Issues in Poland version 0.5, unpublished paper prepared for WP2 ESSnet on BigData project.

Rvest package description https://cran.r-project.org/web/packages/rvest/rvest.pdf