Strategies in producing statistics with privately held data at SURS

Tomaž Špeh
Head of IT production and infrastructure, SURS

Karmen Hren
Deputy Director-General, SURS

The paper aims to present potential ways of addressing the challenges related to the inclusion of privately held data sources into official statistics, taking into account legal requirements, business incentives as well as enhancing internal management and organizational capacity suitable for smaller offices with limited resources dedicated to research capabilities; for example, project-oriented organization, collaboration with academia, commitment to innovations, open-collaborative development and introduction of related standards and frameworks. The paper is based on experiences acquired in the recent implementation of projects related to the use of privately held data for the production of official statistics at SURS.

1. Introduction

New data sources, commonly referred to as “big data”, have attracted growing interest of national statistical offices in recent years. They range from diverse business transactions to social media, high-resolution sensors, and the Internet of Things. The statistical community has recognized that these data have the potential to complement existing approaches for producing official statistics in different ways, for example by improving cost effectiveness, timeliness, granularity, accuracy and relevance, and by reducing the reporting burden.

However, in order to make optimal use of new data sources, a number of challenges have to be addressed. The most pressing is the access to data. They are owned or held by private companies and collected for purposes other than producing official statistics. Private companies are reluctant to share data with national statistical offices not only because of a concern over giving up commercially sensitive data, but also because of concerns related to privacy infringement risks, to the costs and to the risks of anonymization techniques. In addition to the data access issue, there are several other challenges to be dealt with (legal requirements, enhancing internal management and organizational capacity, developing methodologies, business incentives, competitive risks, privacy and ethics).

The Statistical Office of Slovenia (SURS) has recently started several projects in order to discover the possibilities for the usage of new data sources. Developing and implementing new approaches for collecting prices and compiling price indices started (scanned data, scrapped data). Mobile positioning data were examined to study their adequacy for statistical purposes (e.g. statistics about population, time and location distribution, population mobility). Scraped job vacancies data started to be gathered and their usability studied. Data coming from traffic counters were analysed as a possible source for different statistics. This paper focuses on experiences that SURS acquired in these projects.

2. SURS's case studies

This section provides a brief description of three cases of working with privately held data at SURS.

Scanner data for the consumer price index

The main aim of the project was to modernize the data collection and other methods for compiling the harmonised index of consumer prices (HICP), namely by introducing price scanner data.

Due to the commercial sensitivity of the scanner data (detailed information at item code level - detailed item description, turnover and quantities sold) and treatment of these data as a business secret, quite an effort was necessary to convince retailers to participate in the project. At numerous meetings at all levels, management and operational, the mission of official statistics, its activity and importance, the data protection policy and the reasons for modernising the processes were presented. Especially pointed out was that these data will be particularly protected. The negotiation process lasted three years but finally all retailers agreed to cooperate. Special contracts were signed with all retailers as an additional sign of SURS's commitment to data protection even though from a purely legal view they are not needed.

After the data were obtained, they were integrated in the regular production and are now part of the monthly HICP compilation.

Mobile phone data

The main aim of the project was to study the adequacy of the mobile data for statistical purposes, to recognise SURS's potential deficiencies regarding the data infrastructures or human resources, and to study questions regarding the safety of handling these sensitive data.

The main challenge addressed was the data access. The reaction of the mobile phone operators to SURS's proposal for cooperation appeared to be much related to the profile of the persons met. Namely, people form business intelligence, market research and analytics were more prone to cooperate, whereas the legal matters service people were rather reserved.

One of the operators eventually agreed to provide the data. Several rounds of meetings to discuss the details were conducted, some also in the presence of the Information Commissioner. The issues discussed were legal basis, technical solutions, data content, and commercial interests and risks. Participation of the Commissioner was substantial since the operator explicitly conditioned the transmission of the data with Commissioner's favourable opinion. All parties recognised that the National Statistics Act enables SURS to obtain and process practically any data controlled by operators about their users or created by their users. It was also established that the Electronic Communications Act precisely specifies the application of the mobile data and it does not mention the application for statistical purposes. The opinion of the Commissioner was therefore that there is no obligation for the operator to report its data but that it can still do it on a voluntary basis. Data finally obtained covered only variables easily provided by the operator and they were used to make an initial analysis mainly regarding the population mobility.

SURS is currently striving to achieve a sustainable agreement for obtaining the mobile phone data. Although the general attitude of operators is in favour of further cooperation, this is not their priority and, moreover, it is very difficult to preserve the continuation of the cooperation due to the dynamic nature of the mobile phone industry.

Online job vacancies data

In the project a job vacancies scraping process was implemented with the final goal to analyse whether the data on online job vacancies could be used to produce statistics on job vacancies instead of direct data collection.

There are two job portals in Slovenia that advertise the largest share of published job vacancies. The contract was made with one portal which gave the permission to scrape its site for job vacancies once a week. The pages of the other portal are also scraped although without explicit agreement. In doing so, SURS adheres to the 'netiquette': the Robots.txt exclusion protocol of the pages is followed and the scraping is executed during the night/early morning, so as not to interfere with server traffic. Until now the scraping process has been developed and implemented but the practical usability of collected data will be in more detail assessed in the future.

3. Lessons learned

The case studies have pointed to a number of lessons; some of them are detailed below. They cover only the strategic, management type lessons and not the methodological or technical ones.

Access to data is a complex issue and not only a legal one

In Slovenia, official statistics has a strong legal basis for obtaining data from public and private sources. The National Statistics Act provides a clear and broad legal mandate to collect information by stating that in order to implement the programme of statistical surveys SURS shall have the right to collect data from all existing sources. Nevertheless, there are difficulties with obtaining these new privately held data. Mainly there are two types of issues:

- The first one is about data from private enterprises that relate to these same enterprises, for example data on prices and quantities of products sold by retailers. The current legal basis is enough for obtaining the data, but there were big concerns expressed by the enterprises mainly about the possible revealing of business secrets. It took SURS about three years of constant consultations and discussions explaining its data protection regime, how their data are going to be protected, that SURS is not their competitor, what is its role in a society, why they can trust it – SURS was and still is building partnerships with the enterprises.

- The second one is about conflict of laws. For example, the National Statistics Act allows SURS to obtain data on mobile phone movements, events in mobile phone usage, on mobile phone users, etc., but the Electronic Communications Act prevents the use of these data for statistical purposes, also due to underlying EU legislation. Closely connected is also very strong and very limiting personal data protection legislation. As a result, SURS was able to get data only once from one mobile operator and in a highly anonymized form.

There is also the third issue. Namely, the question is how far one can go in the interpretation of “data from all existing sources”. Does it cover only the mentioned cases or also cases where enterprises’ main activity is collecting and selling the data? For example, an enterprise collects data from many real estate agencies on accommodations offered for sale or lease and their prices, and publishes them on the internet. In this case, the enterprise is doing “the data collection phase” of statistical production. It is of the same nature as interviewing individuals for household surveys. If one is willing to pay for data collection in household surveys (field interviewing), then it should not expect other data collections to be available for free. In dealing with such cases, SURS recognised this fact. Instead of an authoritarian approach and broad interpretation of the legal mandate, it opted for a partnership approach which, in some cases, even led to data being offered for free.

Organizational considerations

When facing novelties of any kind, an institution needs to decide how to adapt to them with internal organization being an important element in this regard. For statistical offices, privately held data are certainly a novelty that requires careful consideration of whether the existing organizational structure is still suitable.

For long, SURS has been considering development work to be an integral part of each statistician's work. It therefore does not have special development units that would deal only with innovation, progress, improvements, etc. It believes that separating development and regular work leads to segregation between them, to creation of new stove-pipes, to alienation of "development statisticians" from practical issues and "regular statisticians" from progress, and to additional problems in introducing changes. There might be benefits, of course, but the inclusive approach has served extremely well so far.

It is therefore no surprise that no changes were introduced in the organization of SURS's work. After initial, somewhat confusing years, it was decided to establish a formal project for each task dealing with privately held data. The project structure complements the regular organizational structure; it is used to accomplish defined tasks of a development nature, within the defined time limit, with defined resources and with participation of employees from different organizational units. Work on privately held data thus makes a perfect case for the project structure. Currently (as of August 2018), there are three new projects already approved (on satellite images, job vacancy statistics and HICP) and one pending initial proposal (on financial transactions). It is appreciated, however, that there are still a lot of unknowns around privately held data (be it in terms of organization, knowledge, process or access). In time, by knowing more and by moving from development to regular production the current approach will need to be evaluated and modified, if needed.

Finally, in an EU perspective and having discussions preceding the ESS Vision 2020 in mind, SURS would like to point out that it does not see a need to change institutional basics of the way we work together in the ESS. SURS will remain responsible for all aspects of statistical production.

From what it is known today, the appearance of privately held data does not justify this to be changed.

The need for a gradual approach

Both, the case of scanner data for the HICP and the case of mobile phone data have illustrated that when dealing with privately held data, their holders, owners and other stakeholders, the gradual approach is the approach bringing results. It is understandable that a statistician would like to obtain data as soon as possible, in a known format and detail, and perform known processing. However, for counterparts in this process it is not that obvious why a statistical office should be allowed to do with data more than anybody else. To build a bridge between the two sides, patience and a gradual approach are necessary. The situation with privately held data today is similar to the situation with administrative data several decades ago. At that time, it was not obvious for SURS to obtain data from administrative registers and databases. It was only possible to get them after several years of persuasion, argumentation, collaboration and adaption of legislation.

There is a trade-off between granularity and level of access to the data

The case of mobile data is a perfect example to show that there is a close trade-off between the desired granularity of data and the ease of access to the data. Aspirations at the beginning were huge: to get all the data, at the most detailed level and with all identifications, and to link them with all other data available in the office. These expectations were very soon significantly lowered. For legal reasons and for other reasons already described above, currently the level of ambition needs to be adjusted as the only way to get any data is to accept less than hoped for.

This might change in the future, as the access to administrative data shows, but it is equally possible that some data will never be available to official statistics in all detail and that some statistical processing of these data will never be allowed. This calls for a contingency plan, such as developing or employing methodologies based on this fact (for example linking data without identifiers, compiling representative statistics with non-representative samples). For statisticians used to work in a world where everything is available with unique (administrative) identifiers and where there are practically no limitations to data processing, it is very demanding to accept the new reality. So it is extremely important to assure that our staff is adapted to the new facts and the requirements of the big data universe.

New data sources complement traditional data sources

While new data sources have the potential to reduce the reporting burden, it already became obvious that they cannot act as a pure substitute for traditional data sources; they could not be an alternative to traditional sources but a complement. For example, for the compilation of the HICP the prices of food, beverages and tobacco are obtained exclusively from the scanner data and web scraped data are used for gathering the prices of computer equipment. For all remaining products prices are collected in a traditional way.

Similarly, only about 40% of all Slovenian job vacancies are published online. Online job vacancies data could thus be used only to improve or complement traditional methods, for example to produce flash estimates, to produce more frequent estimates, or to reduce the frequency of the direct data collection. Yet, the feasibility of these ideas is still to be tested.

Private data holders' opportunities

With new data sources there are not only opportunities for official statistics but also for data holders. The case of mobile phone data showed that there is a clear commercial interest of data holders to use their data also for purposes other than billing and managing the network. Providing the data to a public institution such as SURS is considered by them as a safe opportunity to test the public and institutional reaction to data usage. They also stated that in order to be compensated and in the absence of direct payment, they would like to be trained in secure data handling and acquainted with statistical internal protocols to deal with personal data protection. This would be used as a reference if they decide to produce new data derivatives for the market.

On the other hand, retailers expressed the need for frequent data on the value of sales for detailed product groups and for data on the movement of their prices compared to general price movements. These analyses will be prepared and provided or published, as relevant.

New skills are required

Innovative and automated approaches have to be implemented to successfully integrate new data sources into existing solutions. The advancement of the SOA architectural pattern, assisted by evolving common principles and frameworks designed to promote greater interoperability (CSPA, GSIM, GSBPM, ESS EARF), encourages integration of new data sources by developing reusable and sharable software components for the statistical production process. New skills in the field of mathematics and computer science are required. In order to get new knowledge, experience in the field of machine learning, data mining and artificial intelligence faster, SURS is cooperating with

Slovenian universities and research institutes, mainly with the Jozef Stefan Institute and the Faculty of Computer and Information Science. In addition, knowledge obtained in various ESS activities, and in trainings and courses organized within the ESS is of great importance; these possibilities should continue in the future.

All costs are yet to be assessed

In these early years of privately held data in official statistics, the issue of costs was not given much attention (except for paying for data). There were mainly discussions about possible and expected benefits. However, very soon the issue of all costs (including wages and salaries) relating to the use of privately held data will have to be addressed. SURS is a public institution established to serve the society and paid by taxpayers' money. It is supposed to deliver services expected by the society, mainly the government. It is right to be active and anticipate the needs of users but if these needs prove to be non-existent, the decision to stop producing such products needs to be taken. It is right to experiment but not for the sake of being modern; if experiments bring no results, they should be abandoned. The "temptations" to offer something new, without expressed users' needs, to experiment, to look what's inside the data are very strong with new data sources. As they will become more and more everyday business in official statistics, our accountability towards the society requires that in the following years the issue of all costs in relation to benefits is given much more attention than in the past.

4. Conclusions

New data sources brought with them several challenges (legal, organizational, technical, methodological, ethical, financial, etc.). Many of them, if not all, are still in the process of being addressed and resolved and are thus hindering the possibility for faster progress. The most severe among them is access to data.

However, based on the work done so far it can be concluded that initial expectations regarding benefits (more efficiency, increased capacity to provide new insight, reduce the burden on citizens and businesses, boost new opportunities) remain the same; they were neither rejected nor (very strongly) confirmed. Projects and experiments carried out by SURS have demonstrated that there is a potential of further work on using privately held data for the purpose of official statistics.