United Nations ECE/CES/2019/30

Economic and Social Council

Distr.: General
2 May 2019

English only

Economic Commission for Europe

Conference of European Statisticians

67th plenary session
Paris, 26-28 June 2019
Item 2 (a) of the provisional agenda
New data sources – accessibility and use
Session 1: Accessing new data sources

Experiences and challenges on the use of new data sources in Statistics Korea

Note by Statistics Korea

Summary

This document presents the strategy and projects on new data sources in Statistics Korea (KOSTAT) since launching a new division on big data in October 2015. This strategy focuses on data linkage between public sector data (i.e. KOSTAT's data such as administrative data and census) and private sector big data (e.g. mobile phone data, social media data) as well as establishing institutional framework and cooperation. KOSTAT has implemented projects successfully such as linking personal credit evaluation data as well as mobile phone data with KOSTAT's data, organizing big data forums, and establishing international cooperation. However, KOSTAT still faces challenges to overcome: limited access to personal information in private sector data due to strong privacy protection law, lack of cooperation from private sector data providers, shortage of experts such as data scientists, and low quality of private sector big data. Against all challenges, KOSTAT should continue to communicate more with stakeholders from politics, policy makers, business, academia and NGOs in order to make them understand the importance of new data sources for official statistics, and enhance the internal capacities on big data infrastructure. In addition, it also is important to communicate with international organizations for solving big data issues.

This document is presented to the 2019 Conference of European Statisticians seminar on "New data sources – accessibility and use", session 1 "Accessing new data sources" for discussion.

GE.19-07177(E)
1907177

I. Introduction

1. In order to change statistical production paradigm from traditional field survey to new methods for data collection, Statistics Korea (KOSTAT) has continued to make efforts on using administrative data for compiling official statistics. As a result, register-based population census was conducted for the first time in 2015. Basic twelve variables of total population such as name, age, gender and household characteristics were collected using twenty-four administrative data sources from thirteen government agencies. Fifty-two variables not obtained from administrative data were collected using a field survey from 20% sample population. In addition, KOSTAT has implemented a comprehensive statistical register database project to establish four sectoral databases using administrative data: population/household, housing/building, business/enterprise and economic activity.

2. Recently big data has received high attention as a new data source in statistics as well as in business. In terms of statistical aspects, big data can provide more relevant and timely data for decision making through linking various data and reducing the statistical production cost without field survey for data collection. In terms of business aspect, big data create new growth engines as a core of 4th Industrial revolution such as big data analytics for Internet of Things (IoT) and Artificial Intelligence (AI) technologies. In this context, KOSTAT established new division on big data in October 2015, and implemented many projects for developing official statistics under the new big data strategy. However, there are still many restrictions to use big data for official statistics. Therefore, this paper examines KOSTAT experiences in facilitating use of big data and the related challenges.

3. The structure of the paper is as follows: section 2 presents KOSTAT strategy on big data and projects. Section 3 shows the challenges faced in using big data for official statistics. The final section presents a summary and conclusion.

II. KOSTAT experience on new data source

A. Strategy on big data

4. Although there is no affirmative definition on big data, it usually refers to data sources described as high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making (UNECE, 2013).

5. KOSTAT launched a new division on big data in October 2015 to facilitate use of new data sources, i.e. big data, for official statistics. After defining big data as “statistical information” which refers to useful data through data linkage and analysis, KOSTAT has established big data strategy consisting of two approaches: producing various statistical information and establishing institutional and cooperation framework. The strategy has four tasks: 1) linking the public and private sector big data, 2) providing new statistics and supplementing existing statistics, 3) establishing legal and institutional framework, and 4) enhancing external cooperation. Under this strategy, many projects have been implemented.

B. Projects

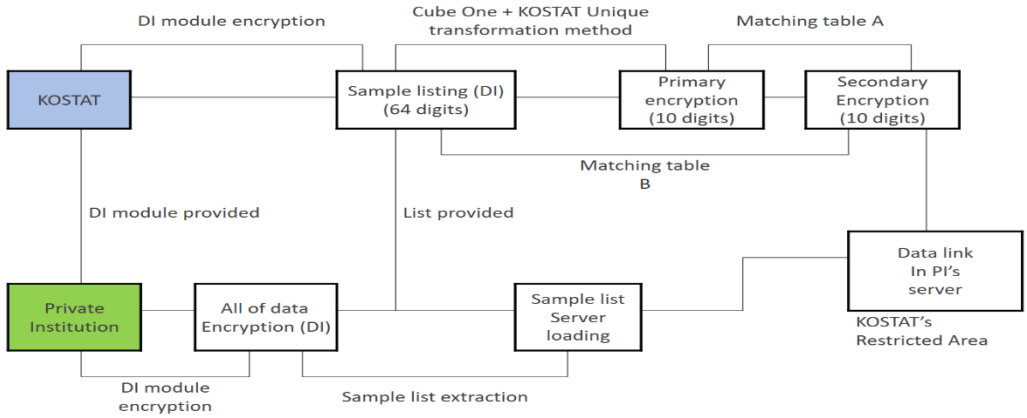
1. Linking public sector and private sector big data

6. KOSTAT has a lot of administrative data (about 89 kinds) collected from other government agencies as well as survey data (about 42 kinds) including population and business census. In order to use and link administrative data to other data, resident registration number (RRN) in administrative data is converted into Statistical Identification Number (SIN). This number is deleted from KOSTAT register database for protecting privacy. As each person has his/her own SIN, thus, data in register database can be used for creating new data or improving official statistics through linking with private sector big data such as mobile phone data, credit card data, personal debt data, etc.

7. However, it is not easy to obtain private sector data because private companies are not obliged to provide data to KOSTAT for other purposes than producing official statistics. In the case of producing official statistics, KOSTAT can get private sector data under Statistics Law. As pilot studies in big data projects are not meant to produce official statistics approved by official process, it is difficult to obtain private sector data.

8. In this regard, KOSTAT established a collaboration framework through signing a memorandum of understanding (MOU) with private sector data providers and implemented data linkage projects together with them. This approach can benefit both sides because private sector can contribute to public good and improve their methodology of compiling big data. To protect privacy in the process of data linkage, a de-identification method of personal information was developed. The same de-identification (DI) module used in KOSTAT for producing SIN is applied to private sector big data to produce the same linkage key (Figure 1).

Figure 1
De-identification process



9. Linkage projects include development of household debt statistics using private credit evaluation data from the Korean Credit Bureau (KCB) and new measurement on leisure and working time through using mobile phone locational data from Korean Telecom (KT).

10. The objective of the project on household debt is to provide accurate debt statistics by household characteristics (e.g. single person household, self-employed, etc.) for policy makers due to increasing household debt in Korea. There are macro and micro household debt statistics in Korea. Macro statistics collected from financial sector reflect the entire volume of household debt but they do not provide information on different types of households. On the other hand, micro statistics from household survey provide debt situation by household characteristics but underestimate the total amount of debt. Therefore, it is useful to combine macro household debt data with KOSTAT data, such as population census to get household information. As a first step, 5,000 newlyweds' (defined as less than 5 years after marriage) debt database from October 2010 to November 2014 was established and analyzed through linking KCB data[1] with KOSTAT data[2].

11. Among other big data sources, mobile phone data have high interest from statistical community because of their high penetration rates and real-timeliness. Their availability for small geographic areas with timeliness provide opportunities of producing disaggregated statistics on population flows, tourism, disaster management, etc. In this context, KOSTAT implemented a mobile phone project to test the possibility and the usefulness of using mobile phone data for producing new statistics that measure the quality of life such as leisure time, commuting time, time poverty through linking KOSTAT data and mobile phone data. There are three mobile network operators (MNOs) in Korea, i.e. SKT, KT and LGU+. Among them,

[1] Variables from KCB data: income, credit rating, loan balance, overdue amount, debt redemption, card usage
[2] Variables from KOSTAT data: date of marriage, age, occupation, education, number of children, housing type, household type, number of owned houses, income, types of job

KT whose market share is about 31% participated in the KOSTAT project. In this project, only two districts in Seoul (i.e. Gangnam-gu and Dobong-gu) according to Gross Regional Domestic Product were selected to compare a well-being pattern between rich and poor areas. Due to a large volume of mobile phone data, KOSTAT data were stored in KT big data analysis system after being de-identified and linked to mobile phone data. Linked datasets were accessed and analyzed by KOSTAT staff only at a designated place in KT office. The estimation results do not represent whole population in two districts because KT data cover around 30% of total population only. Thus, the aggregated tables were compiled using "Ranking weighting" method considering four variables (region, gender, age, marital status, type of house) through mapping KT data to register based population.

2. Providing new statistics and supplementing existing statistics

12. There is a higher demand from policy makers on timely economic data because most of economic data are released on monthly or quarterly basis. To meet the demand, KOSTAT developed fourteen "timely economic indicators" using various data sources: shopping basket price index, overdue electricity charge, etc. Indicators are released every week.

13. To supplement existing statistics, daily and monthly online price indices based on 284 items of products are calculated using price data from 6 online shopping-mall websites which excludes service prices. However, there are some restrictions: i) data could not be collected when web links were changed by revising websites, or categories changed without notice, ii) cutting-off of collection in the case of seasonal products, iii) no quality adjustment was made like for CPI thus the prices of electronics, clothes, etc. decreased.

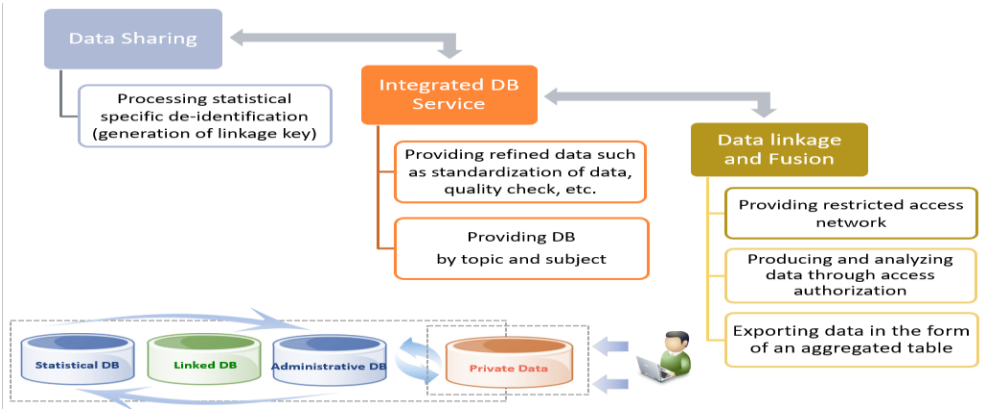
14. In addition, a social media economic index is calculated using social media data (e.g. news, blog, bulletin board, and twitter) related to economic situation in four domains: living conditions, economic situation, household income and consumption expenditure. After collecting documents which contain keywords (138) from blogs, internet café, news and twitter using web-crawling on a daily basis, positive and negative documents are counted, and standardized indices for four domains are calculated. Finally, a total index is derived.

3. Establishing legal and institutional framework

15. KOSTAT continually tries to revise "Statistics Law" to get a legal basis for accessing private sector big data. The current law allows statistical agency to collect private sector data only for producing official statistics. Thus, the revision includes the legal right to collect data from private sector in the case of pilot big data projects that test the possibility of compiling official statistics.

KOSTAT established opening and sharing data infrastructure called "Statistics Big Data Center (SBDC)" with the purpose to support linking public & private sector big data, and to provide de-identification services. Its major function is to implement a quality check on administrative data; to provide register databases by subjects (population, hosing, economic activity, etc.) and statistical survey database; and to provide on-demand linkage services such as de-identification. The customer could link their data with KOSTAT data in a designated place, and export data in the form of an aggregated table. Currently the center is located in three cities: Seoul, Busan and Daejeon. The data processing process is presented in Figure 2.

Figure 2
SBDC data processing flowchart



4. Enhancing external cooperation

16. To communicate and discuss with stakeholders from academia, business and government is important to solve the issues related to facilitating use of big data. In this context, KOSTAT organizes "Statistics-Strategy Forums" every quarter since 2015. In addition, KOSTAT co-organized a "Big data Forum" with two big data related ministries: Ministry of Interior and Safety and Ministry of Science and ICT.

17. KOSTAT is also participating in international cooperation related to big data: the UN Global Working Group on big data and with Statistics Netherlands. Statistics Netherlands (CBS) and Statistics Korea (KOSTAT) have built up bilateral cooperation since the agreement on big data was signed during the Dutch-Korean trade Summit in September 2016. In that agreement, joint activities in seven areas are confirmed: i) obtaining big data sources, ii) developing techniques for exploring big data, such as those based on artificial intelligence or data and text mining techniques, iii) methodological/analytical expertise in the selectivity of big data and data processing, iv) e-learning, v) exchanging staff, vi) sharing experiences in public-private big data linkage, vii) big data and privacy.

III. Challenges

18. Regardless of the many success stories in facilitating use of new data sources, KOSTAT still faces many challenges, both internal and external.

19. Firstly, it is still difficult to access personal information in the private sector due to strong privacy protection law in Korea. Personal Information Protection Act (PPIA) is a general data protection law that governs collecting and processing personal data. There are sector-specific laws: Network Act, Credit Information Act and Location Information Act. In PPIA, the definition of personal data is too broad. The use of personal data needs prior consent, i.e. opt-in approach. Personal data[3] is defined as data on a living person that can identify the individual as well as data that can identify by easily combining with other information. Thus, this strong law makes it difficult to use big data for linkage using personal information.

20. Secondly, data providers from private sector have a low perception of cooperation on data. They are reluctant to share data due to strong privacy protection law as well as their passive approach on data sharing.

21. Thirdly, big data might have low quality because they are not collected by traditional survey methods according to official statistical guidelines or quality framework but by ICT-

[3] Any data relating to a living person from which the individual can be identified through the name, resident registration number, visual image, etc. (including information that can be easily combined with other information to identify a specific individual).

based methods such as censors from mobile phones, data vendors, etc. Thus, there is a lack of quality dimensions such as representativeness, consistency and completeness.

22. Lastly, there is a lack of experts such as data scientists and IT infrastructure to handle large data in KOSTAT. To analyze big data demands different skills and IT infrastructures than traditional statistical analysis and data processing. Data scientists should have knowledge on many areas such as Hadoop, NoSQL, data visualization, machine learning and text mining, etc. KOSTAT has limited possibilities to hire new staff with high analytic skills under the current government recruiting system, i.e. restriction on budget and inflexible recruitment process. To train current staff for developing their skills takes long time. Regarding IT infrastructure, huge investments on data warehouse and software for data collection, data storage, data analysis and data visualization are demanded. Thus, KOSTAT could not build its own big data analysis system due to limited budgets.

IV. Conclusion

23. Under big data strategy established in 2016, KOSTAT has implemented projects to test the possibility of facilitating use of big data for official statistics focusing on linking public sector data (i.e. KOSTAT data such as administrative data and census) and private sector big data (e.g. mobile phone data, social media data). In addition, KOSTAT has made efforts in establishing a legal and institutional framework and cooperation with domestic and international stakeholders.

24. KOSTAT has implemented projects successfully such as linking personal credit evaluation data as well as mobile phone data with KOSTAT data, developing cooperative framework such as organizing big data forums for enhancing communication with stakeholders, and establishing international cooperation with UN and the Netherlands.

25. However, KOSTAT still faces the following challenges: limited access to personal information in private sector data due to strong privacy protection law; lack of cooperation from private sector data providers; shortage of experts such as data scientists, and low quality of private sector big data. Against all these challenges, KOSTAT plans to continue to communicate more with stakeholders from politics, policy makers, business, academia and NGOs to explain the importance of new data sources for official statistics and enhance the internal capacities on big data infrastructure. In addition, it also is important to communicate with international organizations for solving big data issues.