Distr.
GENERAL

Working Paper No. …
… April 2013

ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Meeting on the Management of Statistical Information Systems (MSIS 2013)
(Paris, France, and Bangkok, Thailand, 23-25 April 2013)

Topic (iii): Innovation

Production of Official Statistics by Using Big Data

Supporting Paper

Prepared by Jeong-Im Ahn and Young-Ja Hwang, Statistics Korea, Republic of Korea

I. Introduction

1. Recently a tremendous amount of digital data has been created owing to the boom in smart phones and SNS. All around the world, big data is regarded as a key resource for creating enormous value. In Korea, big data is recognized as a core of national competitiveness in the future to create a new value by both private and public sectors. Accordingly, the Korean government pushes ahead with a “Big Data Master Plan for the Implementation of a Smart Nation”.

2. The Google Price Index and unemployment rates, which are produced by Google, the biggest portal site, are representative examples of using big data for statistical production by the public sector. These trends ask official statistical agencies to find countermeasures against the statistical production by the public sector by using big data and to find ways to use big data by themselves. In the meantime, a problem of checking reliability of statistics that unofficial statistical institutes produce is addressed, too. Currently, statistically advanced countries and international organizations such as OECD and UNECE are discussing statistical policies related to big data, and roles of national statistical offices.

3. Statistics Korea, a central statistical agency, reviews a possibility of combining big data in statistical production to deal with these domestic and overseas trends and to improve efficiency of statistical production. And Statistics Korea designed a Pilot Project for using big data directly in statistical business processes. According to the standard business processes, when producing the Industrial Production Index, every month enumerators visit a sample of establishments. And data on industrial classification, items, sales, etc. are edited, and then the Index is published. In the Pilot Project, the editing process was redesigned to use media data and the big data processing model was inserted. Through this Pilot Project, Statistics Korea aims at establishing a foundation for producing official statistics by using big data.

4. Chapter II describes the background and overview of the Pilot Project, methods to collect and analyze data, and the results of project development. Chapter III presents future plans.

II. Pilot Project for Using Big Data in Official Statistics

A. Project Overview and Direction

5. According to the increase in one-person households and growing awareness of privacy protection, survey environment is getting worse and worse. Under these circumstances, the production of official statistics by using big data has a great advantage in terms of timeliness and cost effectiveness.

6. However, currently there is no rationale for the application of information or results coming from big data to the target population. Accordingly, it is difficult to substitute big data for official statistics. But big data can be used when supplementing existing statistics.

7. As part of this attempt, Statistics Korea plans to develop a pilot project for using big data in official statistics. This system is designed to provide an integrated analysis function by automatically collecting media data, to provide survey data in a visualized way so as to apply a big data analysis technique, and to reduce editing time of the Monthly Survey of Mining and Manufacturing.

8. Project directions are as follows: First, up to now, when producing the Mining and Manufacturing Production Index, much time and effort is needed for data editing and level analysis. Accordingly, to reduce time for editing and analysis, big data will be widely used.

9. Second, a huge amount of media data will be used for data analysis. At first, the project will be applied to major establishments and items in the Monthly Survey of Mining and Manufacturing and then expanded more and more.

B. Analysis Range

10. In the Pilot Project, not only media data but also survey data will be analyzed. Out of the total survey subjects* of the Monthly Survey of Mining and Manufacturing, 4 industry groups (C21, C24, C26, C28), 162 items and 1,438 establishments are selected. Internet news and websites related to those industries, items or establishments will be browsed. And for survey data from 2005 to the current month, indices by industry, item and establishment as well as production tables will be analyzed. [Table 1][Table 2]

* 26 industry groups, 633 items and a sample of 8300 establishment in the Monthly Survey of Mining and Manufacturing

Industry	Source	Contents
C21 (Pharmaceutical products)	Korea Pharmaceutical Manufacturers Association	www.kpma.or.kr/Pharmaceutical news
	C24 (Basic metal products)	Production, goods of pharmaceutical companies
	C26 (Electronic components, computer, radio, television and communication equipment and apparatuses)	Related articles
C24 (Basic metal products)	Korea Metal Journal	new.kmj.co.kr/News per product
	Korea Iron & Steel Association	www.kosa.or.kr/Iron & steel information/Survey report, iron & steel journal
	Internet Newspaper	Related articles
C26 (Electronic components, computer, radio, television and communication equipment and	Ministry of Knowledge Economy	Website/Press releases, notices, policy
	Korea Customs Service	Website/Policy report, press releases
	Korea Semiconductor Industry Association	Real-time news on semiconductor
	Semiconductor Network	Semiconductor Network news
	Korea Display Industry Association	News on display industry

apparatuses)	Internet newspaper	Related articles
C28 (Electrical equipment)	Internet newspaper	Related articles

[Table 1] Analysis Range of Media Data

Table title	Contents
mi_jisu_analysis	Index by industry
mi_jisu_analysis_m	Index by item
mi_dong1_analysis	Table by establishment
mi_dong2_analysis	Table by item

[Table 2] Analysis Range of Survey Data

C. Data Collection and Analysis

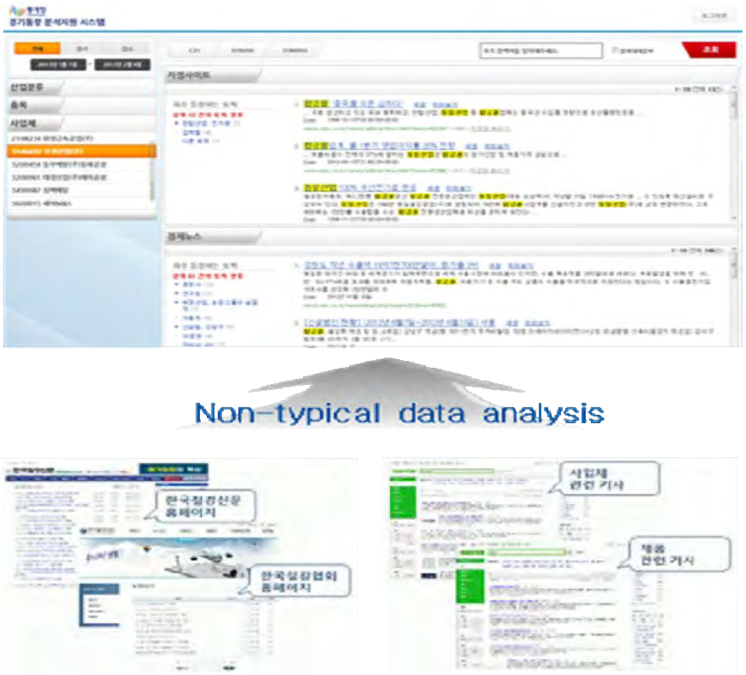
11. Methods to collect and analyze data in the Pilot Project are as follows:

12. First, when collecting media data, related articles on the Internet as well as on relevant websites are scrolled and examined to find words such as increase and decrease in relation to items and establishments within analysis range, for the period of the 1st day of the previous month to the current month. News on the Internet is scrolled on a real-time basis, while attached documents in the PDF or MS word format on the websites are loaded into the analysis server according to the scheduling method.

13. When analyzing media data, attached documents coming from the websites and Internet news scrolled from the Internet in real time are integrated to analyze specific websites, economic news and portal news in the order of retrieval accuracy. To improve retrieval accuracy, similar search words are also registered in advance. [Table 3][Figure 1]

Classification	C21	C24, C26, C28
Common	Increase, grow, rebound, rise or expand Decrease, drop, fall or decline Item name Establishment name	
Additional	Release, operating profit, renewal, futures, export, transfer, public relations, prescription drug, UNESCO	

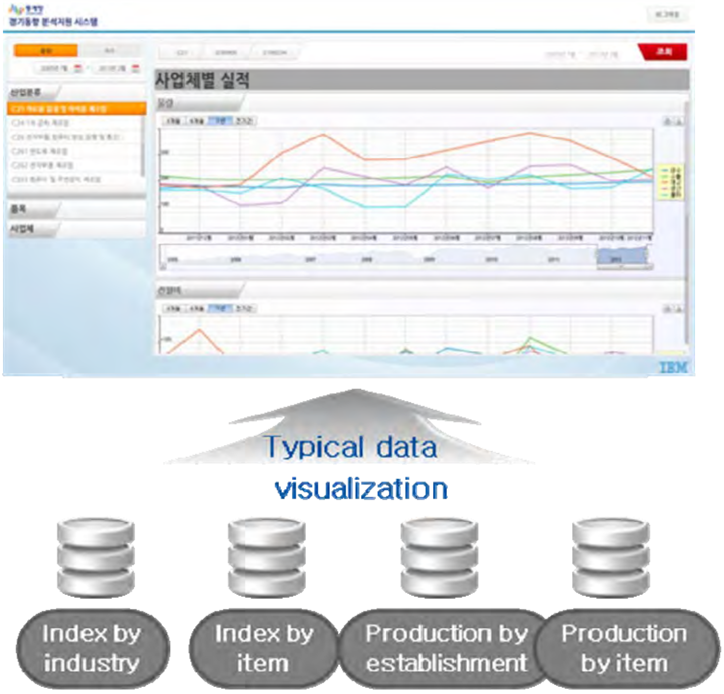
[Table 3] Media Data Search Word



[Figure 1] Collection and Analysis of Media Data

14. Second, when collecting survey data, by linking with the database for the Mining and Manufacturing Survey System, survey data and tabulation data from 2005 to the current period are used on a real-time basis. In addition, responses that enumerators input or establishments input through the CASI (Computer Assisted Survey Input) are used in real time.

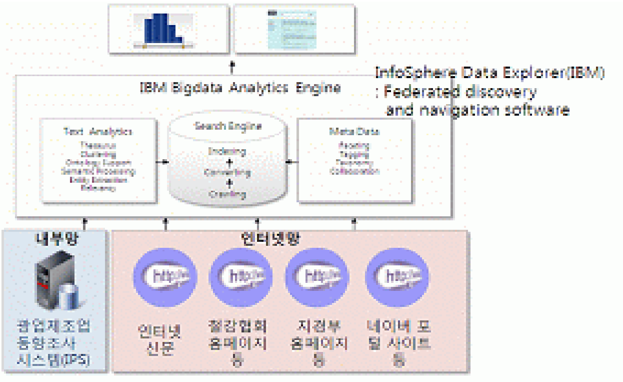
15. When analyzing survey data, for easier understanding of data, the Mining and Manufacturing Production Index and time-series data are presented in a visualized way. For example, as for production by item and establishment, month-on-month or year-on-year changes in production are presented in graphs. And the production indices by industry and item are presented in graphs – as well as month-on-month or year-on-year changes in indices are presented in graphs. [Figure 2]



[Figure 2] Collection and Analysis of Survey Data

D. Analysis System Environment

16. By considering the importance of data security for the Mining and Manufacturing Production Index, the visualized data analysis system was built in the Intranet, while the pubic media data analysis system was built in the Internet. [Figure 3]



[Figure 3] Pilot System Environment

E. As-Is and To-Be

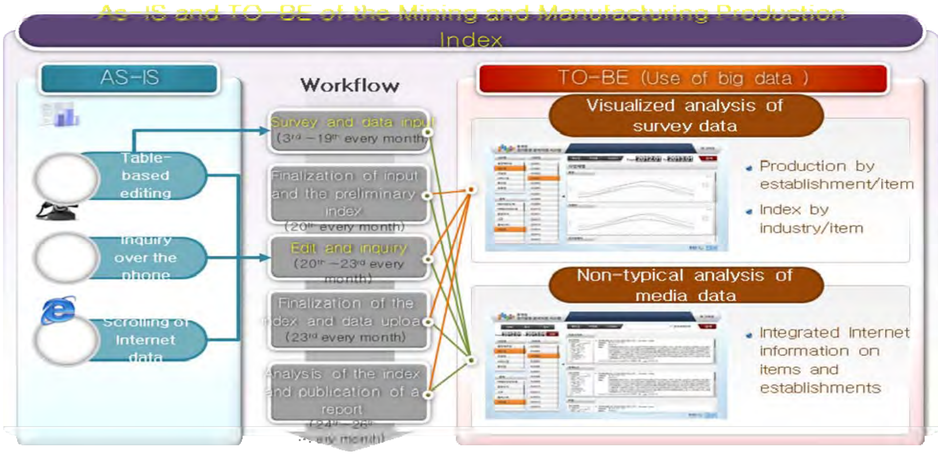
17. For the Monthly Survey of Mining and Manufacturing, data input, editing, inquiry, index analysis and data dissemination are carried out every month.

18. Currently to produce the Mining and Manufacturing Production Index, data are inputted, data are edited by using statistical tables, and phone calls are made to establishments when there is an outlier. For example, if Samsung shows a sharp month-on-month or year-on-year decrease in semiconductor production, calls will be made to get information on the decrease. In addition, to edit outliers, an enumerator visits an establishment, browses Internet news concerning items showing outliers, or browses an association website.

19. Through the Pilot Project where big data are used in statistical production, the following effects are expected:

20. First, from specific websites (e.g. an association website), economic news or portal news, some data related to items and establishments (e.g. data indicating production increase or decrease) are automatically collected and uploaded into the integrated system. Therefore, outliers can be detected at a glance, and then an editing process can be done almost simultaneously. It is expected to reduce editing time.

21. Second, for the data from 2005 to the current period, not only time-series production data by establishment and item but also indices by industry and item are shown in graphs. Outliers can be easily detected. Therefore, it is expected to support editing effectively and reduce editing time. [Figure 4]


As-IS and TO-BE of the Mining and Manufacturing Production Index
AS-IS
Table-based editing
Inquiry over the phone
Scrolling of Internet data
Workflow
Survey and data input (3rd –19th every month)
Finalization of input and the preliminary index (20th every month)
Edit and inquiry (20th –23rd every month)
Finalization of the index and data upload (23rd every month)
Analysis of the index and publication of a report (24th –26th every month)
TO-BE (Use of big data)
Visualized analysis of survey data
Production by establishment/item
Index by industry/item
Non-typical analysis of media data
Integrated Internet information on items and establishments

[Figure 4] Derivation of the Mining and Manufacturing Production Index

III. Future Plans

22. As mentioned above, it might be dangerous to substitute big data for official statistics and it would be desirable to supplement official statistics with big data. In the future big data will be widely used when analyzing current trends, predicting future trends, and suggesting an alternative.

23. Through the Pilot Project, Statistics Korea finds a way to use big data for the production of official statistics and improve the efficiency of statistical business processes, and builds up techniques for utilization of big data. Based on the findings of the Project, for analysis, industry groups will be expanded to all the industry groups for the Monthly Survey of Mining and Manufacturing. And then the Project will be expanded to other economy-related surveys including the Monthly Service Industry Survey.

24. In addition, Statistics Korea will cooperate with international organizations including the High Level Group so as to find ways to use big data for official statistics. And Statistics Korea will carry out a research on the utilization of big data, like the Billion Prices Project, a project started by MIT, which aggregates price information from online retailers and shows daily price fluctuations.