Distr.
GENERAL

Working Paper
… February 2014

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

UNITED NATIONS
ECONOMIC AND SOCIAL COMMISSION
FOR ASIA AND THE PACIFIC (ESCAP)

Meeting on the Management of Statistical Information Systems (MSIS 2014)
(Dublin, Ireland and Manila, Philippines 14-16 April 2014)

Big Data and Official Statistics in China

Prepared by Liang Damin, Cheng Jinjing , National Bureau of Statistics of China(NBS)

I. Introduction

1. In modern China, with more than 600 million internet users and the booming internet industries, more and more data are generated on the web than ever before. The huge amount of data produced by the electronic devices is described ‘big data’. So far, there is not a very accurate definition of big data, but it can be generally represented with four v characters: volume, velocity, variety and value. Big data might be stored in the structured, semi-structured and unstructured forms.

2. In our opinion, the outstanding characteristics of big data as follows:

(1) Big data is a digitizing reality world: the various attributes, physical characteristics, activities of people(including family and organization) and matters in the real world are digitized, thus the reality can be recorded permanently and searched, copied, predicted, and shared easily. And that’s why the official statistical department should rely on big data.

(2) Big data is the result of public participation: the data generated from social media and search applications is the main component of big data. That means 80% of big data is produced spontaneously (non-administratively) from the information subjects, and it reflects the real requirement and preference of the public. In statistics, the using of big data is complied with the idea that respondents would be ‘willing to’ instead of ‘asked to’ report data.

(3) The main data sources for statistical analysis are operating data, transaction data from the large-scale enterprises and administrative registers.

(4) The growth rate of the data mining capacity lags behind that of data accumulation in every industry.

II. Challenges and changes

3. Big data brings great challenges and changes to official statistics. In the terms of the statistical business process, a variety of difficulties emerge in every stage. These fall into the following categories:

(1) Program Designing Stage. Under 'big data' circumstance, figuring out where and how the existing data comes from and if it could be reused becomes the primary challenge for the statistical designer. It is quite an opposite flow compared with traditional statistical designing. Besides, 'big data' promotes more statistical methodologies. Sometimes, it makes possible for a survey to change the sampling frame from one to another, and the cost may be controlled as well.

(2) Data Collecting Stage. To a certain content, using big data can help ease burden on respondents and enhance the facticity of statistics. For traditional statistics, the quality of raw data highly depends on the feedback of the respondent as well as high cost. But if part of the raw data is extracted from the records collected by technology, it will weaken the data fraud and lower cost. NBS accordingly must seek pro-per intelligent management system to join the official statistics.

(3) Data Analyzing Stage. The data analyzers face the greatest challenge to untangle the inherent relations when surrounded by thousands of semi-structured and unstructured data. As a result, more professional technologies on data mining and processing will be indispensable for statistical officials.

(4) Data Releasing Stage. Big data require higher transparency and pertinence of the official statistics. On the one hand, official data without detailed explanations and fairly statistical methods lead to worthless will be queried by public and replaced by others. On the other hand, digging out the potential users and providing them with the most concerned data will make official statistics more competitive with authority.

III. Achievements and Experience

4. NBS, as an official statistical office, is confronted with the challenge and changes with big data, and has made many efforts both in the IT infrastructure and the application.

5. From the IT infrastructure side, NBS is devoted to developing computer technologies. The infrastructure projects based on the virtualization, network expansion and remote disaster backup system have been improved. The statistical information security construction based on the classified protection technology has been performed. So far, the cloud computing has been launched already and the virtualization has been widely applied in statistical information computing and storing in NBS.

6. From the application side, NBS has been pacing with the trend of big data in a quick succession. A number of new technologies have been adopted in many statistical businesses.

(1) The on-line reporting system which supports the direct submission of raw data collected from enterprises to NBS, has been broadly implemented in China. It is extremely efficient to avoid administrative intervention and better the data quality.

(2) The PDA(personal digital assistant) system has been applied in several statistical surveys, such as the CPI(consumer price index) survey and the third economic census which is performing now in China.

(3) The corresponding digital administrative registration records from other official ministries have been used to verity the statistical Business Register system and support some surveys like the real estate price survey.

(4) The space information technology, such as RS(remote sensing), GIS(geographical information) and GPS(global position system), and the industry internet system have been adopted in many statistical fields including demography, agriculture, investment and transportation.

7. Last November, NBS signed a series of agreements with 11 large-scale enterprises, in a bid to build a long term cooperation with them on using big data. These enterprises have the same operating model electronic commerce, media technology or information technology but target at different goods and services. They can be divided into 2 groups. One is those who can produce their own data, which formed into big data obviously, like Baidu Inc; the other is those who can collect data, which means they are not only able to produce trade data but also generate some indicators or index, like the Fanya Metal Exchange Co.. No matter which group the enterprises belong to, there is no doubt that all of these enterprises are outstanding in respective field and willing to share data with official statistics to maximize the effect of big data.

8. Take Baidu Inc for example, with tens of billions of queries entered into the research platform every day, it becomes the largest search engine in China. This means that a huge amount of data is genereated by more than 600 million active users visiting more than several hundreds of billion web pages through Baidu everyday At present, the cooperation between NBS and Baidu Inc aims at three aspects: (1) generalizing the official statistical data and programs through the 'baidu' web page; (2) improving the predicting model of macro economy by combining the big data on web with survey data;(3) grasping more meaningful statistical requirement and completing the survey programs by following the netizens' path on the web platform. Other aspects of cooperation will be explored with the development.

9. Another case worthy to address is The FANYA Metal Exchange (FYME), which is the biggest exchange of rare metals worldwide with the largest scale of clients' assets under management on the platform for spot trading. There are eight categories and ten varieties of listed commodities including indium, germanium, gallium, bismuth, APT and etc. The volume of trading and delivery of all the listed commodities ranks the first place in the world except silver. NBS started to cooperate with FYME since September of 2012 on FYME index(FYMEI) program. Last April, the FYMEI was issued for the first time. As the first rare metal exchange index, the FYMEI fills in the gap of the world. Furthermore, the FYMEI and the data related not only supplement the metal database for NBS, but also become an important reference for the government policy making. The agreement signed last year between NBS and FYME illustrates two major issues: (1) the statistical standards(including the indicators' explanation, the statistical classifications, the codes etc.) on which the big data study bases; (2)the specific data that the enterprises should provide as well as the format and procedures the providers should abide by.

IV. Issues and Vision

10. Although NBS adjusted a lot to keep pace with the big data trend, there still remains some challenges yet to be tackled. The challenges are mainly as follows:

(1) Legislation. Bid data has a public participation feature. It is common that most big data are produced or collected by non-government organizations and those situations might be beyond the existing statistical law or others. So enacting corresponding laws, regulations and standards to illustrate the rights, responsibilities and obligations of each participant is really necessary. Typically, the privacy provisions will be plentiful and flexible with more information and more individuals involved in official statistics. And those provisions will empower government to collect big data in a more legal, timely and reliable way.

(2) Management. Bid data impells official statistics to adjust and increase functions to manage more external information. For NBS, a more comprehensive cooperation networking model covered with data production, data integration, data sharing functions should be established under the legal framework. And the networking model should connect each sort of big data sources(including other government ministries, enterprises, research institutions, universities, international, etc.) and remains steady and continuous.

(3) Staff. In order to ensure the networking running efficiently, NBS should start a professional team to work on it. The members should come from different statistical industries as well as the IT department. Firstly, they should develop an implementation plan for NBS to use big data, which means a specific

schedule with detailed big data applications should be considered ahead of time. Secondly, the evaluation system to ensure the quality of big data should be established. Thirdly, few statistical models may be built to aggregate and analyze data. The last but pivotal is specialized computing infrastructures, which required to cope with big data every processing stage.

(4) Technologies and Methodologies. Sometimes, the uses of a certain data is innovative, as various combinations will be created when different sources of information stored in the same database. This leads to several issues: (a). the changes of the database underlying structure may be required for the huge irregular data; (b). the invalidity of the traditional strategies and techniques, such as advising and consenting, fuzzy privacy and anonymity; (c) the question to the combined data between the traditional data and big data beyond preparation. Overall a set of supporting technologies and methodologies ought to be promoted.

V. Conclusion

11. Making full use of big data to promote the scientific development is a world trend for government statistics. NBS will actively perfect the data application legal system, establish technical standards, better the work mechanism, cooperate with related organizations , expand the data sources as well as accelerate the application of big data.