UNITED
NATIONS

E

Economic and Social Council

Distr.
General

ECE/CES/2006/SP/9
9 May 2006

ENGLISH ONLY

ECONOMIC COMMISSION FOR EUROPE **STATISTICAL COMMISSION**

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-fourth plenary session
Paris, 13-15 June 2006
Item 6 of the provisional agenda

SEMINAR ON POPULATION AND HOUSING CENSUSES
SESSION III

New types of data input systems in Korea: the Internet survey and web-based data entry system[1]

Submitted by Korea National Statistical Office

I. INTRODUCTION

1. Population and housing censuses are major sources of demographic and socio-economic statistics in Korea. It is also a unique source of geographically detailed data. The demand for this kind of information has increased rapidly in recent years, and this is why a census is still conducted in Korea, even as the relative cost of a traditional census has risen to a level increasingly difficult to justify. The population census in Korea dates as far back as the Samhan Era over two thousand years ago and subsequently to the Goryeo and Joseon dynasties. However, the 1925 census is generally acknowledged as the first population census in Korea from the viewpoint of its coverage and objectives. Sixteen rounds of censuses have been carried out at 5- year intervals since 1925 in order to obtain reliable data about the structure of the population, households and housing.

---

[1] This paper has been contributed by Namhoon Kim, Korea National Statistical Office.

2. With every new round of census, remarkable developments have been achieved not only in content but also in technique. The OMR (Optical Mark Recognition) technique in data capture which was adopted in 1990 has greatly sped up data processing. The OMR is the technique to detect the presence of intended marked responses by using special hardware equipped with light sensors that capture the reflection or absence of reflection on paper. In the 2000 census, a self-enumeration method was partly introduced to cope with hard-to-enumerate areas.

3. Since 2002, the KNSO had prepared for the 2005 population and housing census of Korea. The 2005 Population and Housing Census began on 0:00 A.M. November 1, 2005 and was carried out for 15 days. Census items of the 2005 population and housing census were mainly focused on the low fertility rate, aging population, quality of living conditions, and other related issues. Also, to cope with hard-to-enumerate environments and solve the deteriorating enumeration conditions, an internet survey method and web based data processing method have been introduced. In this paper, I would like to briefly introduce the internet survey and web based data entry system of Korea.

II. INTERNET SURVEY

4. In the early days of census, and almost up until the last census, there was only one possible way to collect the necessary information on persons residing in households. This was achieved with the help of written questionnaires and census takers. This methodology has worked reasonably well but in recent censuses, more households are preferring to complete the census form themselves rather than having it collected by an enumerator. There are some people who have expressed a preference for completing the census through the internet. Furthermore, many households are becoming difficult to contact. One reason is the increase in one or two person households with busy lifestyles.

5. To overcome these hard-to-enumerate circumstances, the ever increasing costs, and pressures to reduce respondent burden, the KNSO has a strong incentive to seek new solutions in census data collection and to provide more effective methods of data processing. The KNSO decided to introduce an experimental internet survey during the 2005 census in order to make a practical use of the high level IT infrastructure and high speed internet which exist in Korea. The traditional written method remains for the majority of households, but the KNSO provided an internet census form for those who prefer to complete the census in this way. Therefore, under the slogan “e-census”, the KNSO developed an internet survey system for the 2005 census which enabled a small proportion of the population to submit their census questionnaires electronically, via the internet, instead of using the conventional written method.

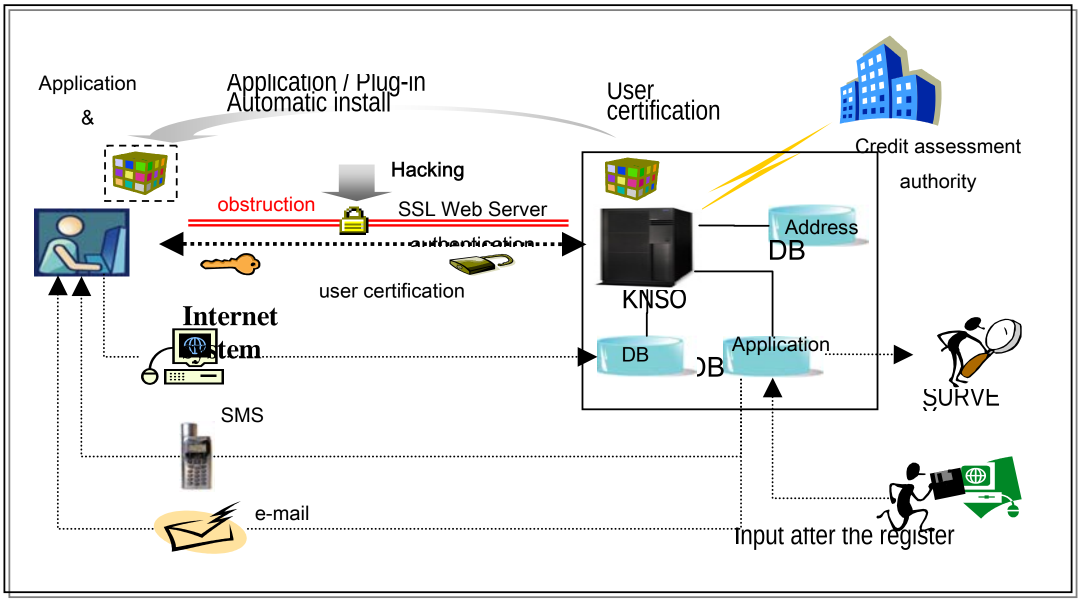
A. Application

6. The KNSO studied the feasibility of the internet survey on the basis of the following criteria: the efficiency of data entry, response rate, accuracy, and network capacity through a pilot survey in November 2004. After testing and improving the system, the KNSO applied the internet survey as one of the data collecting method in the 2005 Census.

(i) Operational flow

7. The operational processing of the internet survey has several steps which the respondent should follow.:

8. First, respondents who want to complete the census questionnaires through the internet must submit an application. Second, the KNSO certifies his/her real name using their Resident Registration Number through a Credit Assessment Authority. Third, the applicant inputs other necessary information such as an ID, password, telephone number, e-mail, address, etc. Last, a short or long form questionnaire pops up on the screen after matching an applicant's residence and the address D/B. Then, the applicant can complete and send the questionnaire through the Internet.



(ii) Advantages

9. The advantages of the internet survey include:
(a) A faster availability of data through simplification of data entry and editing,
(b) Better coverage for households who have a high probability of non-participation in the survey such as one member households (students, workers, etc.) and hard-to-enumerate households,
(c) An increased level of user-friendliness versus the paper questionnaire,
(d) And an interactive user guide and automatic filter of irrelevant survey items.

(iii) Challenges

10. While the internet survey has several advantages, it also raises new challenges and problems that must be resolved which include:
(a) Rectifying errors such as duplications and omissions of the members within one household

(b) Data security
- Use of ID and password specific to each household
- Encryption required (in Korea, 128 bit)

(c) Calculating the capacity of the server and network itself in order to implement the system
- In the event that a considerable amount of unspecified individuals participate, the system should be designed to accommodate peak user frequencies during the 15 days of the survey period
- Minimum of 10,000 users accessing the system simultaneously

(d) A strategic plan for various obstacles
- If connected to the system and a session is interrupted, the entered data would have to be automatically stored and made available to the user again later

(e) Certifying the residence of the applicants who have completed the census questionnaires on the internet

(iv) Results

11. Regarding the background of introducing the internet survey, limited public relations were applied focusing on the concentrated area of students, workers, and one member households. As a result, approximately 1% of respondents (150,000) completed the questionnaire via the internet. Many of 150,000 respondents, those who preferred not to respond to enumerators through interviews, participated in the internet survey, even though they lived in ordinary households. Therefore, the KNSO is willing to improve and expand the internet survey in the next Census.

III. WEB-BASED DATA ENTRY SYSTEM

A. Historical changes of data input method

12. The data processing cycle of the census generally involves four different independant stages such as initial reading/encoding, data capture, editing, and tabulation/loading into the database. The stage of initial reading and encoding is to pre-edit the questionnaire by physically reading it and to assign classification codes to responses on the census form. The stage of data capture is to digitalize the enumerated data and to create a computer data file. The stage of editing is to check for erroneous data and to ensure consistency of data items. The stage of tabulation and loading into the database is to tabulate the data and to enter it into the publishing Database.

13. Looking into the history of data entry methods in Korea, the Punch Card System (PCS) used punched cards that could be read with a card reader to produce data files. This was adopted from 1960 to 1980. In 1985, the Key Entry System which could input data by key punchers with dummy terminals to produce data files was used. The OMR technique was adopted to speed up data processing in the 1990 and 1995 censuses. As far as the OMR is concerned, it was quite successful in Korea even with some of the operational problems such as transcription errors and high printing costs.

14. However, the OMR system gradually fell into disarray, so the KNSO arrived at a decision to use a PC-Based Key-Entry system in a decentralized manner for the 2000 census.

15. These methods mentioned above are all off-line batch processing in a centralized manner. The files that were entered through data capture were gathered and data editing were executed. Data capture and data editing were separated respectively and a considerable part of the data processing time was consumed by these stages.

B. Characteristics

16. After the enumerators complete and collect the census questionnaires in the survey fields, the questionnaires are sent to the headquarters or the local offices of the KNSO. They are inputted using the Off-line Batch processing method and then are edited through the telephone or by field re-interviewing. However, since the procedure of data processing is complicated and several input/editing stages are needed, it consumes lots of time.

17. However, with the development of Information Technology, the internet is starting to take hold and the utilization range is becoming wider through a collection of information and sharing of data.

18. The Web Based Field Input System is a data input method which carries out data capture “on the spot”. This method enables us to input enumerated data on-line through the internet and to input data with the same format in all regions of the nation. An input situation in real time can also be managed with this method. This method can contribute to the improvement of the quality of enumerated data, because some of the enumerators can take part in data capture.

C. Application

19. A Web Based Field Input Method was first introduced in the 2005 Census. The KNSO successfully carried out this method last year.

20. The Web Based Field Input system is largely divided into two categories – data input and management. The data input category consists of entering enumerated data made up of three main menus including selection of enumeration district, data capture of questionnaire, and data editing. The management category contains user management, assignment of enumeration district, and retrieval of input situation.

(i) Results

21. The enumerated data was inputted by the Field organization from November 28 to December 22, 2005. That was, we reduced the total time of data processing to about 3~7 months compared with 12~18 months of total data processing time in the 2000 census.

22. The reason for the increase in efficiency is based on both the decentralization effect of data capture and the reduction of data processing stage such as the preparation period for data editing and field interviewing period. This revolutionary shortening of the data capture period makes it possible to rapidly publish census results and contributes to the improvement of the data quality,

especially when executing the editing process in the enumeration field.

(ii) Challenges

23. Basically, the Web-Based Field Input system is activated in a decentralized manner. Therefore, data management is particularly critical in a distributed processing environment where there may be tens of thousands of PCs in over 250 sites connected on the same network. Some basic challenges which have to be resolved are listed below.

(a) Data security: The unit record data that is entered during processing should be subject to strict security rules. Only authorized staff should have access to these record files. Network security must be required to monitor and restrict access by unauthorized staff. It also needs to provide mechanisms that will prevent unauthorized tampering of the data in the files and provide audit trails of all changes.

Protection against the threat of computer viruses is another important aspect of protecting the data. The introduction, either deliberately or inadvertently, of a virus could have disastrous effects on processing. Therefore, up-to-date virus protection software should be installed on all computers to ensure network security.

(b) Data back-up: In order to recover from the inadvertent loss of data, it is important to prepare a back-up strategy. This strategy may include frequent on-site back-ups of data, and control files, during all stages of processing, and regular off-site back-ups to protect against major disasters.

It is also important to have a recovery strategy in place to be able to reinstate all files back to a consistent state after the failure of web-servers or WAS-servers, any corruption of data, or other problems that might arise.

(c) Diverse on-site environments: To satisfy all of the various network communication bandwidths and PCs that have the different OS, CPU, browsers, etc, it is not only important but also essential to acquire a secure, accurate, and time-friendly system.

IV. MODEL FOR THE 2005 CENSUS

24. With the need to cope with new challenges and problems mentioned above and the priority to find a cost effective system for both the internet survey and the Web-Based Field Input method, the decision to seek external expertise was made at the end of 2004. The ISP (Information Strategic Plan) on the integrated data processing system includes the internet survey, Web-Based Field input system, and Data-Warehousing (DW) system for easier tabulation and analysis of the enumerated data. The results were basically implemented into the HA (High Availability) method with dual independent systems to guarantee the non-interruption during the enumeration and input period, protecting against any disastrous occurrences.

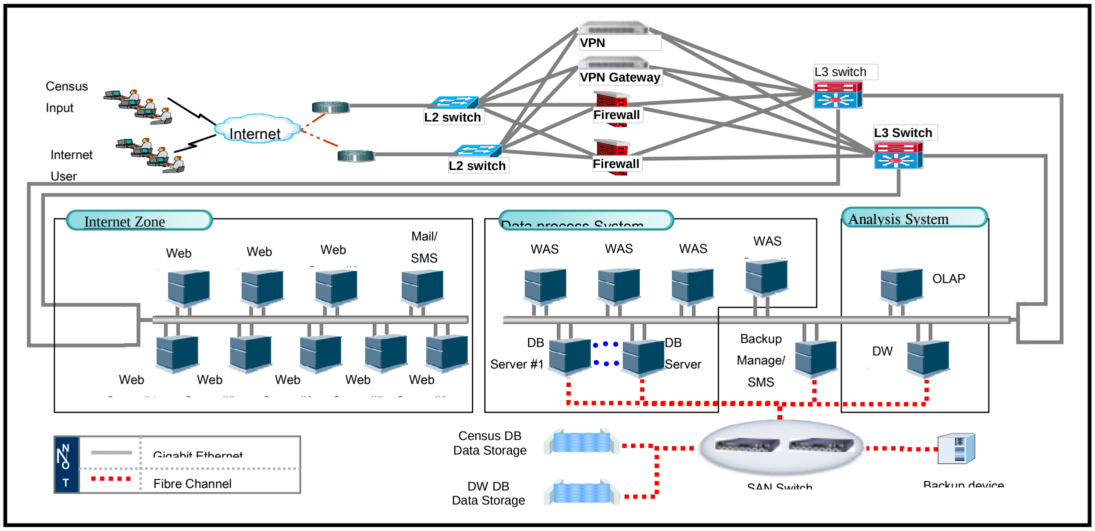
A. Architecture

25. The Integrated Data Processing system for the 2005 census was made of a 3-tier structure of Web-WAS-DB. The Web-server is composed of multi-nodes to scatter the risk of simultaneous access and a balancing of a sudden increase in users. Because the WAS server plays an important role between the web server and the DB server, it is controlled by a technique of logical partition (LPAR). The DB server has influence upon the efficiency of operation. The number of CPUs of the DB servers is calculated by 20,000 tpmc per machine. The DB server is

composed of the method of RAC (Real Application Cluster) to share the physically one DB by multi-server. Storage is composed of mirroring structure with RAID 1+ and RAID 5 to maintain the safety of the original data. Details of hardware and software are as follows:

(a) Web server : in total 8 servers each with 4 CPUs, 16GB Mem, and 292GB HDD
(b) WAS server : in total 2 servers each with 8 CPUs, 32GB Mem, and 584GB HDD
(c) DB server : in total 2 servers each with 16 CPUs, 64 GB Mem, and 584 HDD
(d) 1 DW server with 4 CPUs, 1 SMS server with 4 CPUs, 1 Back-up server with 2 CPUs, 2 Storages each with logically 10 TB, and 2 SAN Switches each with 16 Ports.
(e) 2 DBMS, 1 ETL, 1 OLAP tool, and several kinds of system software

Architecture configuration



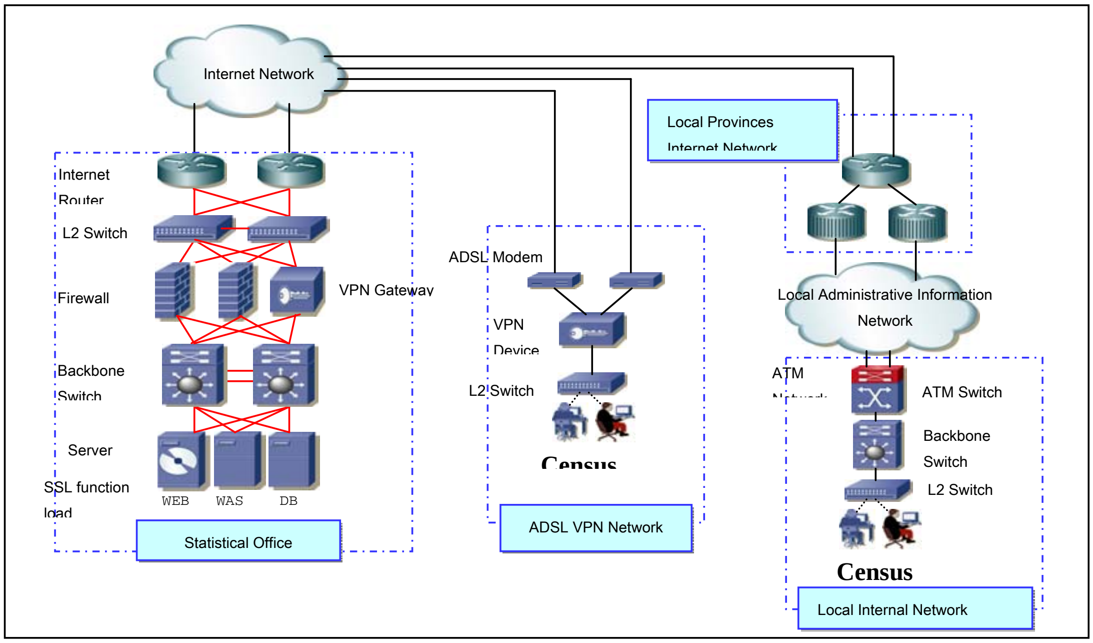
B. Network and Security

26. This model was implemented to achieve the target of 0% interruption. The KNSO considered the effective use of existing IT infrastructure by the provinces and the SSL (Security Socket Layer) encryption technique to eliminate risk on the internet network and to protect against the loss of information. The network between province and the KNSO is made of ADSL VPN (Virtual Private Network). VPN is organized by duplex system of ADSL circuit to guarantee the service availability through embodiment fail-over between the circuits.

27. Details of Network equipment are as follows:

(a) Network equipment: 2 Routers, 2 L3 Back-bone Switches, and 2 W/G Switches.
(b) Security equipment: 2 Firewalls, and 2 VPN Gateways.

Network configuration



* * * * *