

22nd Population Census Conference,
7 - 9 March, 2005. Seattle, U.S.A

**The new types of data input system in korea :
The internet survey & web based data entry system.**

Nam Hoon Kim

Deputy Director, Population Census Division

Korea National Statistical Office

The new types of data input system in Korea : The internet survey & web based data entry system.

Kim, Nam-Hoon

Korea National Statistical Office
920 Dunsan-Dong Seo-Gu
302-701, Daejeon, Republic of Korea, e-mail : nhkim@nso.go.kr

1. Introduction

Population and housing censuses are a major source 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 census is still conducted in Korea, even if the relative cost of a traditional census has risen to a level more and more 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 interval since 1925 in order to obtain reliable data about the structure of the population, household and housing.

With every new round of census, remarkable developments have been achieved not only in content but also in technique. The OMR(Optical Mark Recognitoin) 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 intened 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.

From 2002, the KNSO has been preparing for 2005 a population and housing census of Korea. 2005 Population and housing census will be taken as of 0:00 a.m November 1, 2005. It will be conducted for 15 days. Census items of 2005 population and housing census will be mainly focused on the low fertility rate, the quality of living conditions and related issues. Also, to cope with hard-to-enumerate environments and solve the deteriorating enumeration conditions, The internet survey method and web based data processing method will be introduced. In this presentation, I would like to briefly introduce on the internet survey and Web-based data entry system of Korea.

2. Internet survey

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 from the households, and that was 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. The increase in one or two person households with busy lifestyles also makes it difficult to contact some households.

To overcome these hard-to-enumerate circumstances, and 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 provide more effective methods of data processing. The KNSO has 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 will remain for the majority of households, but the KNSO will provide an internet census form for those who are prepared to complete the census in this way. So under the slogan “e-census”, the KNSO has developed an internet survey system for the 2005 census which will enable a small proportion of the population(about 2%) to submit their census questionnaires electronically, via the internet, instead of using the conventional written method.

2.1 Pilot survey

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 pilot survey in Nov. 2004.

2.1.1 Operational flow

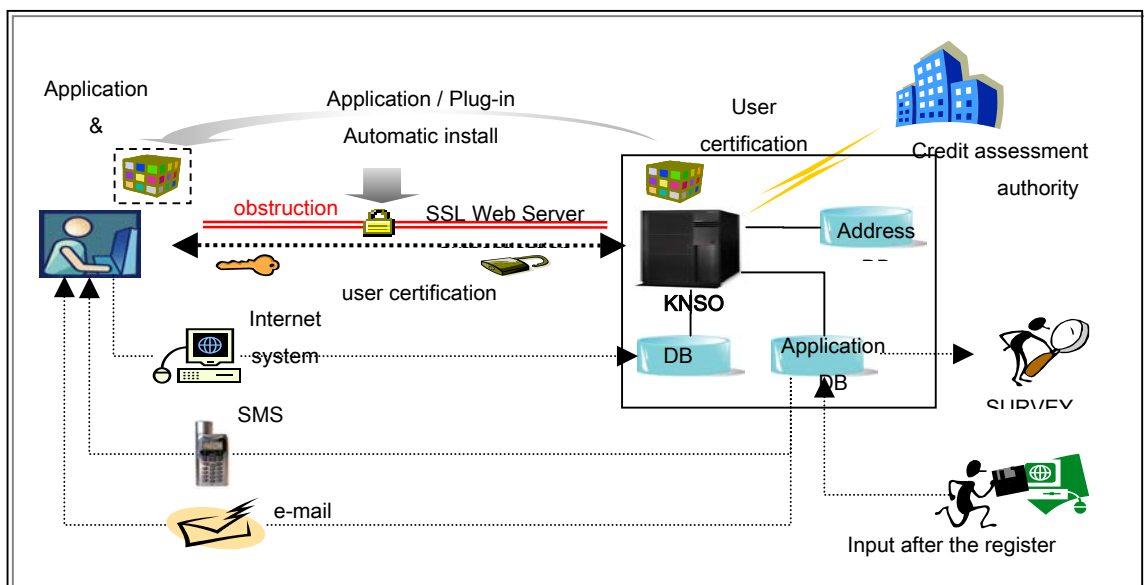
The operational processing of the internet survey has several steps which the respondent must follow :

First, respondents who want to complete the census questionnaires through the internet must make an application.

Second, the KNSO certifies his/her real name with a Resident Registration Number through a Credit Assessment Authority.

Third, the applicant has to input other necessary information such as ID, password, telephone no, e-mail, address, etc.

Last, a short or long form questionnaire pops up in the screen after matching an applicant’s dwelling place and the address D/B. Then, The applicant can complete the questionnaire through the internet directly.



2.1.2 Advantages

The advantages of the internet survey include :

- The faster availability of data through the simplification of data entry and editing
- It provides better cover for households who have a high chance of non-participation from the survey such as one member households(students, workers, etc.) and hard-to-enumerate households
- More user-friendly than the paper questionnaire
- Interactive user guidance and automatic filtering of irrelevant survey items

2.1.3 Challenges

While the internet survey has several advantages, it also raises new challenges and problems that must be resolved which include :

- Rectifying errors such as duplications and omissions of the members within one household
- Data security
 - Use of ID and password specific to each household
 - 128 bit encryption
- Calculating the capacity of the server and the network itself and implementing the system
 - Because considerable unspecified individuals had participated, the system therefore should be designed to accommodate peak user frequencies over 15 the days of the survey period
 - Minimum 10,000 users accessing the system simultaneously
- Strategic plan for various obstacles
 - If the session to connect the system was interrupted, the inputted data would have to be automatically stored and made available to the user again later
- Certificate the residence of the applicants who completed the census questionnaires on the internet

3. Web-based data entry system

3.1 Historical changes of data input method

The data processing cycle of the census generally involves four different independent stages such as first reading and encoding, data capture, editing, and tabulation and loading into the database. The stage of first reading and encoding is to pre-edit the questionnaire by physically looking at 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 load it into the publishing Database.

Looking into the history of data entry methods in Korea, the Punch Card System(PCS) used punched cards that could be read with card reader to produce data files was adopted from 1960 to 1980. In 1985, the Key Entry System which could input data by keypunchers with dummy terminals to produce data files was adopted. 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 if it left some operational problems like transcription errors, high printing costs, etc. However the OMR system gradually fell into disarray, so the KNSO arrived at a decision to use PC-Based Key-Entry system in a decentralized manner for the 2000 census.

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

3.2 Characteristics

A main stream of traditional data processing systems has followed. After the enumerators complete the census questionnaires in the survey fields, and collect them, questionnaires are sent to the headquarters or the local offices of KNSO, and they are inputted using the Off-line Batch processing method and then are edited through the telephone or by field reinterviewing. However, since the procedure of data processing is complicated, and several input/editing stages is needed, it consumes lots of time.

However, with the fast development of Information Technology, the Internet is starting to take hold, and the utilization range is becoming wider through a collection of information or share of data day by day.

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.

3.3 Pre-survey

A Web-Based Field Input Method was firstly introduced during the third pre-survey for the 2005 Census. The KNSO successfully carried out this method three times in the last year to test the timeliness and data quality and to find solutions for minimizing the network traffic, data security, optimizing the various kinds of servers, etc.

The Web-Based Field Input system was largely divided into two categories – the data input part and the management part. The data input part that is to input enumerated data consisting of three main menus which are selection of enumeration district, data capture of questionnaire and data editing. The management part has user management, assignment of enumeration district and retrieval of input situation.

3.3.1 Results

From the table 1, the elapsed time for data input between the third and fourth pilot surveys were almost the same but the fifth pilot survey slightly increased due to the change of input method. However, the elapsed time for data input from table 1 shows meaningful results which make it possible to shorten the data processing period considerably compared with that of the 2000 Population and Housing Census . That is, we will be able to reduce the total time of data processing

to about 3 ~ 6 months compared with 12 ~ 18 months of total data processing time in the 2000 census.

It was because of 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 provide rapid publishing census results and contributes to the improvement of the data quality especially when executing the editing process in the enumeration field.

Source : KNSO

Sequence	No. Of Households	No. Of Enumerator	Elapsed Data Capture period
3 rd	12,000	20 person	9.5 days
4 th	5,600	10 person	9.0 days
5 th	40,600	31 person	14 days

Table 1. The results of pre-surveys for the 2005 Census

3.3.2 Challenges

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 will have to be resolved are listed below.

- Data security : The unit record data that is inputted during processing should be subject to strict security rules. Only authorized staff should have access to these record files. Network security will be required to monitor and restrict access by unauthorized staff. It will also need 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.

- 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, or any corruption of data , or other problems.

- Diverse environments of on-site : 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.

4. Planning Model for the 2005 census

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 last year. According to the results of ISP(Information Strategic Plan) on the integrated data processing system includes the internet

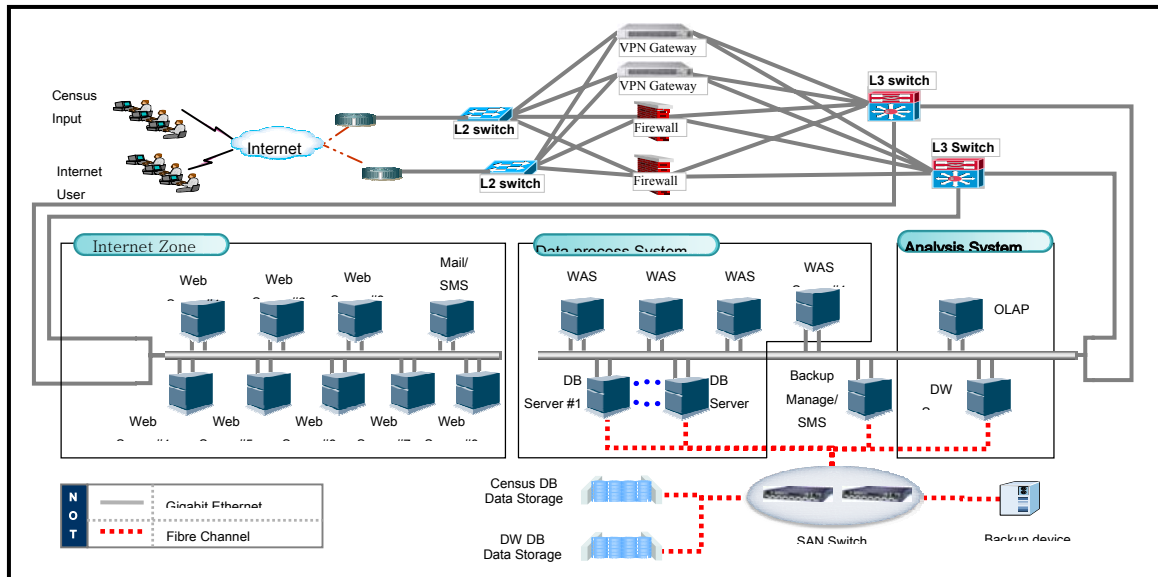
survey, Web-Based Field input system, Data-Ware housing(DW) system for easier tabulation and analysis of the enumerate data, the planning system is basically implemented into HA(High Availability) system with dual system to guarantee the non-interruption over the enumerate and input period even if any kinds of disasters happen to occur.

4.1 Architecture

The Integrated Data Processing system for the 2005 census is made of a 3-tier structure of Web-WAS-DB. The Web-server is composed of multi-nodes to scatter the risk from multi-user's concurrent burden and a balancing of a sudden increased burden. 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. And 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 is as follows.

- Web server : totally 8 servers each with 4 CPUs, 16GB Mem, and 292GB HDD
- WAS server : totally 2 servers each with 8 CPUs, 32GB Mem, and 584GB HDD
- DB server : totally 2 servers each with 16 CPUs, 64 GB Mem, and 584 HDD
- 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.
- 2 DBMS, 1 ETL, 1 OLAP tool, and several kinds of system software

<figure 2> Architecture configuration



4.2 Network and Security

A planning model will be 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 the loss of information. The network between province and KNSO is made of ADSL VPN. VPN is organized by duplex system of ADSL circuit to guarantee the service availability through embodiment fail-over between the circuits.

Details of Network equipments is as follows.

- Network equipments : 2 Routers, 2 L3 Back-bone Switches, and 2 W/G Switches.
- Security equipments : 2 Firewalls, and 2 VPN(Virtual Private Network) Gateways.

<figure 3> Network configuration

