

**Use of Information Technology in the Census Operation of Japan
– Recent Improvements and Challenges for the Future –**

Masao Takahashi
Statistics Bureau of Japan

I. INTRODUCTION

1. The Population Census has been conducted by the Statistics Bureau of Japan (SBJ) every five years since 1920. The last Census was conducted in 2005, which was the 18th undertaking, and the next is scheduled for 2010. With a population of over 120 million, the Census is the largest statistical undertaking in Japan. To collect and process the vast amount of data within a limited time, the use of information technology (IT) is essential. This paper describes how IT was used in the 2005 Census of Japan, and explains the lessons learned and the challenges for the future.

II. OUTLINE OF THE 2005 POPULATION CENSUS

2. The 2005 Population Census was conducted as of 0:00 a.m. on 1 October 2005, in order to assess the number and basic characteristics of the population usually living in Japan. One of the characteristics of the 2005 Population Census was, as in the past, data collection by enumerators, whereby questionnaires were delivered by enumerators, filled in by households and also collected by the enumerators.

3. While the planning and overall management were conducted by the SBJ, fieldwork for the Census was legally entrusted to local governments, such as prefectures and municipalities (cities, towns, and villages). The whole of Japan is divided into 47 prefectures, which took charge of Census operations, including distributing Census documents such as questionnaires to municipal governments, supervising the fieldwork of municipalities, and collecting Census documents from municipalities. Municipalities, totaling about 2,000, were in charge of field operations that included establishing Enumeration Districts, selecting and training supervisors and enumerators, and collecting Census documents such as questionnaires.

4. During the enumeration period, enumerators visited all the households in the assigned Enumeration District and asked them to fill in the questionnaires. After the reference day of the Census (1 October 2005), the enumerators revisited the households and collected the completed

questionnaires.

5. The questionnaires collected by the enumerators were checked at local governments (prefectures and municipalities) and sent to the SBJ. The tabulation, including data capture, coding, data editing etc. was planned and managed by the SBJ, although its operation was performed by the National Statistics Center, which is an incorporated administrative agency specialized in data processing. The results of the tabulation were then sent to the SBJ for analysis and publication.

6. The results of the 2005 Population Census have been released by the SBJ in series since December 2005, when preliminary counts based on summary sheets on household and population by sex were published. In June 2006 the results of a 1% sample tabulation were released, aiming to clarify the situation of Japan according to basic characteristics as a prompt tabulation. The results of primary tabulation, in which all the questionnaires collected are used, have been released in three stages from October 2006 and will be finished in December 2007, followed by tabulations on small area statistics, on detailed statistics using a sample of less than 10% of the questionnaires, and on special topics, such as places of work.

7. The dissemination of the Census results has been performed in several manners, with most of the results posted on the SBJ's web site (<http://www.stat.go.jp/english/index.htm>). Besides, the results have also been disseminated in electronic form, such as CD-ROM, through an affiliated organization. The SBJ has also been publishing statistical reports of the Census in the traditional way. In addition, the SBJ operates the "GIS Plaza of Statistics" (<http://gisplaza.stat.go.jp/GISPlaza/>) to disseminate small area statistics using a geographical information system.

III. RECENT IMPROVEMENTS BY THE USE OF INFORMATION TECHNOLOGY IN THE CENSUS

8. There are many possibilities for the use of IT in the Census operations. In the recent Censuses of Japan, major improvements have been brought about by the use of IT in the following areas:

- A. Mapping of enumeration districts
- B. Data capture from questionnaires to computer
- C. Dissemination of small area statistics

A. Mapping of Enumeration Districts

9. As mentioned above, the municipal governments are in charge of field operation for the Census. They employ and supervise approximately 900 thousand enumerators nationwide during the enumeration period. Because the enumeration is the crucial element of the Census work that affects

the quality of the resultant statistics, the SBJ supports the local governments by providing tools that help improve the quality and efficiency of their work.

10. One of the most important tools is the geographic information system (GIS) that is used for demarcating Enumeration Districts (EDs) and producing ED maps. The work of demarcation and mapping of EDs was computerized for the first time in the 1995 Census on a limited scale, and the system coverage has since been extended more widely with every subsequent Census.

11. For computerized mapping, digital data of the ED boundaries and digital base maps are needed. The digital data of ED boundaries are highly dependent on the Census Mapping System (CMS), which the SBJ developed by applying GIS in the 1990 Census. The CMS initially aimed to enhance the efficiency of the statistical compilation work at the SBJ by maintaining digital data of all the ED boundaries. The CMS was then used for compiling small area statistics, such as grid-square statistics, from statistics of EDs. As for digital base maps, they were initially very expensive when they came onto the market in the early 1990's. As digital base maps have become increasingly popular for PCs and car navigation systems, however, their prices have become more affordable.

12. For the 2005 Census, the SBJ has developed a new system for the computerized mapping of EDs. The system can produce ED maps by placing the digital data of ED boundaries onto digital base maps. The SBJ has provided local governments (prefectures) with the system, along with digital data of the ED boundaries for the previous 2000 Census and the digital base maps, thus allowing them to produce ED maps of their own areas for the 2005 Census.

13. Owing to the computerized mapping, most municipalities can produce ED maps without cutting and pasting paper maps as was previously the case. Consequently, the workload of local governments has been considerably reduced and the coverage of such computer-generated ED maps has increased to cover approximately 80% of all the EDs in Japan.

B. Data Capture from Questionnaires to Computer

14. Data capture from the questionnaires is a crucial part of the census operation because of the vast data volume involved. In Japan, the data capture operation is centrally done by the National Statistics Center.

15. The National Statistics Center used optical mark readers (OMRs) for data capture since 1965 until the 1995 Census. However, Optical Character Readers (OCRs) were used in the 2000 Census for the first time, as well as the 2005 Census, because of the following advantages over OMRs:

- (i) The capability of capturing numerical responses directly as well as marks;

(ii) The capability of capturing image data on a full scale.

16. Owing to the first advantage, it became possible to design a user-friendlier questionnaire in the 2000 Census. In the 1995 and earlier Censuses, the respondents were asked to write numerical responses, such as the month and year of birth, in both numbers and marks. Writing marks is not only an extra burden for the respondents but also a cause of response errors. OCRs have enabled us to eliminate such mark fields from the questionnaire, which has helped make the questionnaire more compact. The second advantage has made the data editing work more efficient because the editing staff can refer to the full image data of the questionnaire on the screen of their own PC. In earlier Censuses, when the staff needed to refer to questionnaires, they had to retrieve them physically, which required time and manpower. With OCRs, the necessary information can be retrieved on the spot electronically, and such extra work has been eliminated; a change which has enabled the staff to make faster and better judgment.

17. OCRs, however, are not free from problems. The main problems are with the accuracy and speed of recognition. As for the accuracy, in some cases, characters cannot be recognized (non-recognition), while in other cases, the recognition is incorrect (false recognition). In OMRs, non-recognitions or false recognitions seldom occur provided the responses are clearly marked in the right position. However, the recognition performance of OCRs depends on not only the ability of the machine but also the quality of the written characters: when characters are not neatly written, non-recognitions and false recognitions will occur. The speed of recognition also varies according factors such as the complexity of the form and the legibility of characters. For the purpose of the Census of Japan, the character set has been limited to numbers, although the OCR models available can recognize both Roman alphabets and Japanese characters as well. By limiting the character set, the recognition rate and speed have increased.

18. To cope with the problem of recognition, great care is taken during every phase of work from planning to implementation. In the planning phase, OCR models are initially selected on the basis of catalog specifications. However, because the performance of OCRs is affected by various factors not specified in the catalogs, the machines are tested with real questionnaires of pilot surveys or a certain sample survey to measure the performance, i.e. accuracy and speed. According to the test result, patterns of handwriting that are unrecognizable or falsely recognized are analyzed, and a recommended form of handwriting is developed for each number. They were included in the instructions to fill the questionnaire.

19. During the implementation phase, unrecognized characters are judged and entered manually, and the recognition performance is constantly monitored on a real-time basis: If non-recognition occurs, the unrecognizable characters are displayed on the operator's screen, whereupon he/she judges it and enters the correct numbers. Moreover, there is another group of operators assigned to monitor the

accuracy of recognition: a sample inspection is applied to batches of questionnaires, and if a sample batch includes more errors than is permissible, the batch is rejected and recaptured. As a result of the control, the non-recognition rate was 0.38%, and the false recognition rate was 0.16% for the 2005 Census. 11 units of OCRs were rented, and it took eight months to process the 60 million questionnaire sheets in the 2005 Census. The average reading speed was approximately 150 sheets per minute.

20. In the 2005 Census, the possibility of capturing Japanese characters had been tested in order to introduce the automated coding of hand-written responses for the first time in the Census. The test was targeted for the coding of the “destination of commuting”. This item was selected for testing because the responses fell into one of the approximately 2,000 municipalities, which meant that the words and characters appearing in the response were limited, and that even if one or two characters were unrecognizable, there was the possibility of inferring them from the context by referring to the dictionary of municipality names.

21. The result of the test showed that the accuracy of recognition was not sufficiently high, and it was concluded that the automated coding of hand-written responses should not be adopted in the 2005 Census. The SBJ will continue its study of automated coding from hand-written characters.

C. Dissemination of Small Area Statistics

22. The census data have been made available in the form of various media, such as printed reports, CDs, and the Internet as for other survey data. In February 2004, a new web site named “GIS Plaza of Statistics” was launched to provide small area statistics of the 2000 Census. Users can view the census data of any area in the form of maps. The site has attracted many viewers, with a string of favorable comments fed back to the SBJ.

23. The small area statistics provided in the GIS Plaza of Statistics are compiled on the basis of geographic units called “cho” or “aza” (address blocks). There are approximately 211,000 such address blocks in Japan, each containing an average of around 230 households. For each address block, the statistics of such basic characteristics are made available as population by age group, industry and occupation.

24. The statistics for area blocks are displayed on the base map showing the boundaries of area blocks and geographical objects such as roads, railways, rivers, etc. The site provides various functions to sum up the statistics of two or more address blocks. For example, the population of an area within a certain radius from an arbitrary point can be easily computed. Many such functions are useful for business planning.

25. In designing this site, special attention was paid to privacy concerns. As the area unit is small, there is a risk of disclosing an individual's characteristics inadvertently. Therefore, the statistics provided in the GIS Plaza of Statistics have been limited to basic characteristics without cross-classifications, and no statistics concerning characteristics perceived to be delicate have been included in the site. Balancing the wider use of statistics and the protection of privacy will remain an important issue in disseminating small area statistics.

26. Besides the GIS Plaza of Statistics, the SBJ runs the web site of the Census results, which have more than 1,000,000 accesses every month. People can obtain most of the Census tables through the site free of charge. The SBJ will strive to improve the web site of the Census results to become user-friendlier, by hearing user comments and opinions.

IV. LESSONS LEARNED FROM THE 2005 CENSUS

27. As mentioned, the enumerators collected the questionnaires of the Census by visiting each household of the ED. However, for certain problematic reasons such as absence and/or reluctance to respond to the Census, etc., the enumerators were unable to collect a certain number of questionnaires. The percentage of households of which the questionnaire(s) could not be collected¹⁾ doubled to 4.4 % in the 2005 Census from 1.7 % in the 2000 Census.

28. The factors underlying these problems were identified. These include people's increased awareness of privacy and information security, an increase in certain types of households, such as those with double-income or single, that are difficult to contact, an increased number of households residing in apartment buildings with self-locking systems at the entrance, an increase of the number of younger people who do not understand the meaning of the Census, and increased difficulty in recruiting qualified enumerators.

29. To cope with the difficult circumstances affecting the Census operation, including the collection of questionnaires, the SBJ set up a special committee comprising academics and specialists. Following intensive discussions, the committee issued a report making a number of proposals. These included changing the major collection method of questionnaires from the method using the enumerator's visit to one involving mailing back. The committee also proposed that multi-mode responses, i.e. via the Internet, mail or collection by enumerators be permitted if the household wished to do so. When the questionnaires are not submitted within a certain period of time, the enumerator should visit the households concerned to collect the questionnaires as a follow up enumeration.

¹⁾ When an enumerator was unable to collect the completed questionnaire from a household, he/she contacted the neighbors of the household to obtain basic information, such as the name and sex of the household members within the household concerned.

V. CHALLENGES FOR THE FUTURE

A. Use of the Internet in Data Collection

30. One of the most important issues to be considered for future Censuses in terms of the use of IT is its application in data collection. We recognize that the statistical offices of a number of countries provide options to the respondents to submit their answers by the Internet.

31. There are many obvious advantages of the Internet response method. One is the fact that those familiar with the Internet can submit their responses easily and quickly. As increasing numbers of people feel uncomfortable for enumerators/interviewers to view their questionnaires, the feasibility of Internet response is being considered as an option for such people. As the preferences of the people at large are becoming increasingly diverse, providing wider options for response methods other than paper questionnaires is an important consideration in order to obtain good responses. If all or nearly all people would return their responses via the Internet, the workload of enumerators would be reduced.

32. The Internet response method is also advantageous in terms of data processing. If the respondents submit their own data by the Internet, there is no need for data capture. In the Internet response method, functions to check data consistency can be included to ensure accuracy of responses. The Internet responses will reach the processing center almost instantaneously, and will expedite the data processing work, as long as the respondents submit the responses punctually.

33. While the Internet response method is an attractive option as a data collection method, the following issues must be solved before it is adopted:

(a) How to ensure accuracy

With Internet responses, it is quite difficult to ensure punctual responses from all respondents, unless they are highly cooperative. In order to remind the respondents who do not meet the deadline, the most effective way will be the enumerator's visit. However, this will not reduce the workload of the enumerators, and the advantage of the Internet response method will be lost.

(b) How to achieve cost effectiveness

To judge the feasibility of the Internet response method, an important factor is the percentage of the people who choose it. If the Internet response rate is at a low level, the workload and the cost of the enumerators will not be significantly saved, while the development of the Internet response system will simply increase the overall cost.

(c) How to maintain the security and confidentiality of the data

There are also fears concerning the handling of personal information on the Internet. To adopt

the Internet response method, the system should be carefully designed with a view to protect the confidentiality of the respondents' data. It is also necessary to take countermeasures against "phishing" or other kinds of fraudulent attempts.

(d) How to make the method compatible with the enumerators' work in the field

If the Internet response is to be adopted, along with the conventional method, data collection by the enumerators must be well coordinated with the Internet response to ensure that duplicate responses and non-responses can be avoided. How to dispatch enumerators to non-respondents properly and efficiently is particularly important.

(e) How to control IDs / passwords

In connection with the above (d), it is essential to properly control IDs and passwords to access the Internet. If the IDs and passwords are not well controlled and managed, Census operations will fall into confusion due to the occurrence of duplicate responses, non-responses, and fraudulent responses.

34. Technologically, it is possible to develop a system for data collection using the Internet. We have been developing the Internet Survey System under the initiative premised on the "Optimization Plan of Operations and Systems for Statistical Work", which is an application of the concept of Enterprise Architecture. In the Internet Survey System, the necessary measures are taken to deal with the above issues (c) and (d) - maintaining the security and confidentiality of data, and making the method compatible with the enumerators' work in the field.

35. Administratively, however, the feasibility and cost-effectiveness are the crucial factors for adopting the Internet response system in the 2010 Census; and in particular, (a), (b) and (e) of the above issues are closely related to the planning and management of field operation of the Census. We are testing the feasibility in the coming 2nd pilot survey of the 2010 Population Census, which is scheduled in June 2008.

B. A New Way of Using Census Data

36. Another important issue concerns how to meet diverse user needs for the Census results. There are various demands for providing more detailed Census tables to exploit the Census data to the maximum possible extent. However, it is difficult to compile and provide every cross-classification table because the budget and time for compilation are limited.

37. Under the new Statistics Act, which was entirely revised in May 2007 and which is likely to fully come into force in April 2009, a mechanism involving a tailor-made compilation of statistics, via the use of IT, will be established. The SBJ must be ready to start this new means of using Census data.

VI. CONCLUDING REMARKS

38. In the 2005 Population Census of Japan, new methods utilizing information technology were adopted in the mapping of the EDs, data capture, and data dissemination, but not in data collection.

39. The data collection process is the most important process that determines the quality of data in terms of coverage, accuracy, and timeliness. The Internet response method involves many difficulties to be overcome, and was therefore inapplicable in the 2005 Census of Japan. The data collection process depended on the conventional method because it was considered to be still reliable and affordable in Japan.

40. However, the operation of the 2005 Census revealed the limitation of the conventional method and the data collection method had to be modified. When introducing new methods such as the mailing back of questionnaires and Internet responses, it is essential to gain the understanding and cooperation of the citizens. For this purpose, we should make appropriate decisions while hearing the opinions of relevant parties, including local governments and the general public. Accordingly, the SBJ will make utmost efforts to draw up an optimal plan for the 2010 Census to cope with the issues raised at the last Census.