

Implication of ICR for the 2010 Population and Housing Census: Thai Experience^{1/}

Sue Lo-Utai
Secretary General
National Statistical Office
Thailand.
loutai@nso.go.th

A. Introduction

1. Thailand's first population census was conducted in 1909 by the Ministry of Interior. Four subsequent censuses followed in 1919, 1929, 1937 and 1947. Since 1960, the National Statistical Office (NSO) has been responsible for undertaking population censuses every ten years under the 1952 Statistical Act (revised in 1965). In accordance with the United Nations' recommendation that countries should undertake national censuses in the year ending with 0 (zero) for the purpose of international comparison, Thailand has conducted its census in 1970, 1980, 1990, and 2000. In 1970, the first housing census was conducted simultaneously with the population census. The tenth population and fourth housing census was carried out in April, 2000. The NSO is currently planning to conduct its eleventh census in the year 2010.

2. For the 1960, 1970, 1980 and 1990 Censuses, the main method of data collection was field interview. After the data collection in all provinces was completed, questionnaires were sent to the central office in Bangkok. Manual editing, keyboard data entry, and other steps of data processing, including tabulation were then carried out.

3. The quality and timeliness of the data provided by the census can always be improved in order to meet the needs of various users. The timeliness of the census information is also important for a public relations campaign. The general public will acknowledge and use data if it is current and will then become more aware of the importance of statistics. Consequently, the statistical efforts of the country will improve, and in turn, the quality of statistics produced will be strengthening.

B. Image Scanning and ICR

4. In general, the population and housing census is a large-scale data collection project that encompasses an entire country. It is expensive and time consuming and requires a large number of field staff and a systematic and effective data processing system. Although results should be published as soon as possible after the field work, there is usually a significant gap because the data entry and cleaning process is very lengthy. To shorten this process, an alternative solution between to increase the number of staff members or to deploy new technology for data capturing such as the image scanning has to be chosen. The image scanning technology has the advantage that it can also be used for other subsequent surveys or censuses.

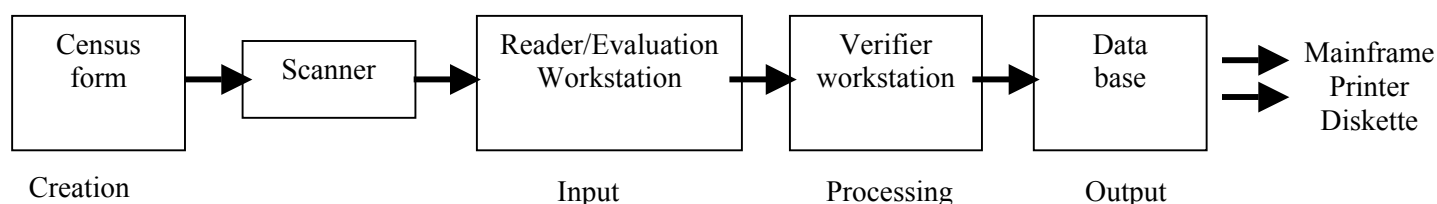
1/ Paper prepared for the Population Census Conference in Seattle, Washington, March 7-9,2005

5. Image Scanning technology is a system used to capture data from a questionnaire (form), fax, or Internet quickly and with a minimum amount of human intervention. It utilizes a scanner that quickly reads questionnaires or data forms, and a software application that automatically reads or evaluates the data or figures recorded in the forms and transforms them into an ASCII data file which can be used for further data processing. The Image Scanning can be OCR (Optical Character Recognition), OMR (Optical Mark Reader), BCR (Bar Code Reader) or ICR (Intelligence Character Recognition).

6. At this time the Image Scanning technology is considered a better option for data capture of large-scale survey and census data as it can reduce the time for data entry, requires less personnel and is cost-effective in the long run. Of the four types of Image Scanning, ICR and OMR are the methods of choice for statistical surveys and censuses, because survey or census forms are designed to record numbers, figures and /or blocks.

7. The ICR process involves the use of a powerful software application that addresses each of the four steps illustrated in Figure 1.

Figure 1. The steps involved in ICR processing of survey data



C. Assessing the experience of TNSO in using ICR

The 2000 Population and Housing Census

8. In Thailand, the ICR was firstly used by the National Statistics Office (NSO) for the 2000 Population Census. Its use allowed the NSO to release the census results within 1.5 years as compared with over 3 years for the 1990 Census and to spend only 9 months instead of 30 months for data entry. The cost of the data entry in 2000 was reduced to less than a fourth of the cost in 1990 (from 108.5 million baht to 23.2 million baht).

9. Before utilizing the ICR system, the NSO had installed 3 clusters of key stations (1 server and 16 key stations for each cluster). The data entry staff keyed data from the questionnaire at each key station and these data were stored in the server. Another set of staff reentered the data to verify correctness and completeness. Then supervisors retrieved data from the server as a batch file and copied it onto diskettes or sent it to a mainframe for processing. The Manual Data Entry System had been used for capturing data for Population, Marine Fisheries and Agricultural censuses as well as for other NSO surveys before it was replaced in 2000 by the ICR system.

10. Field interview was the major method for data collection but self-enumeration forms were in more extensive use than in the previous censuses. Self-enumeration was more effective in apartment blocks. The period of data collection was 1-30 April 2000 with a census reference date of 1 April 2000. There were about 40,000 enumerators and 5,600 supervisors. For non-municipal areas, school teachers were used for field personnel, while both school teachers and temporary employees were used in Bangkok and municipal areas. Among the temporary employees, supervisors must have at least university education and, for enumerators, at least upper secondary level (Grade 11).

11. A more decentralized arrangement and new technology were adopted in order to improve timeliness of data reporting. Manual editing and coding was carried out at provincial statistical offices (PSOs) in all provinces in Thailand. Then, the questionnaires were sent to the Central Office for data capture using an Intelligent Character Recognition (ICR). Data processing, including tabulation and analysis, was carried out at the central office.

The 2003 Agricultural Census

12. Since first using ICR for the 2000 Population Census, the NSO has used this system for many other subsequent surveys such as the Monthly Labour Force Survey, the Household Manufacturing Industry Survey, and many Quick Surveys (small-scale and quick release of the results).

13. The NSO used ICR for processing the 2003 Agricultural Census. A pilot project was launched in 2000 in Roi-Ed (a province in Northeastern Region) and Nakonsrithamarat (a province in the South Region) to test all steps of the census processing including the ICR system of data capture. The NSO had also reviewed the problems arising during the processing of the 2000 Population Census and of the other surveys and found solutions to overcome them.

14. For the real census, around 35,000 village volunteers were used as the census enumerators. Although a minimum requirement on education background was set for the recruitment of this staff, it was not all in every area that the enumerators would attain such standard. For some with less education, only good hand-writing ones could be employed to be census enumerators. Moreover, in the agenda for training on census field operation, many assignments on writing practice were included. And these helped reduce error in scanning the census questionnaires.

15. For the agricultural census, around 6.7 million forms (47 million pages) were printed. Many printing companies were involved, some as sub-contractors from the main one. A strict quality control was done for every batch of forms to be sure about the quality of the printed questionnaires (size, paper quality and some color). A set of census forms consisted of 4 pages, therefore, it was possible that loose pages might be misplaced. To avoid this mistake, an identification code (page number) for each form was preprinted on top of every page.

16. Before printing the census form, color testing was done in order to ensure that the color selected was able to be processed during the Reader Mode. It was important to control for paper quality and color, ink concentration, and cutting paper. Special A3-size envelopes were also made, to store and transport the unfolded forms. After completing the census forms, the enumerator or supervisor unfolded them and put in envelopes and sent back to the central office to be scanned. Each envelope contained around 10 forms.

17. The completed forms were sent to the central office for data entry and processing. The NSO did not install scanners in the provincial or regional office as it was more costly and inconvenient to maintain the system. The scanners were set to run 10 hours a day under close supervision of the supervisors who had to schedule the form to be scanned and checked for feeding problems during scanning. It was found that the ICR system already installed at that time was rather slow, the speed of work was slower than it should be. It was estimated that with such speed, the data processing might be at least 5-6 months behind the schedule. With some budget left from field operation, NSO had decided to upgrade the ICR system by which the data processing process was kept within the schedule.

D. Lesson learnt from Thai experience in using ICR

The Advantages of ICR System

18. The ICR system of processing survey or census questionnaires has many advantages. The system shortens the time of data capture with fast speed scanning and overnight reading without manual supervision. It is appropriate to use in large-scale surveys and censuses, where the questionnaires are not too complicated but very large in number. Since the demand for more timely data has been increasing, the ICR system can help the statistics agency to release the survey or census results much quicker.

19. The ICR system significantly decreases the risk of manual error at the data entry stage. The scanners will read the questionnaires according to what was recorded in it, while the manual data entry is subject to errors such as typing errors or misreading of data or figure. It was determined that the number of errors was more rapidly when the amount of work was increased. For the large-scale or census projects, less intervention of human or manual operation will reduce the data entry period significantly.

20. With this system, less personnel are required than for manual data entry system. The scanner works very fast and the reader mode can run automatically, only the verification mode needs manual supervision and correction. If it is difficult to recruit qualified staff, it might be better to adopt ICR system.

21. Since the image of questionnaires or forms can be stored in small files, it is efficient to use this system for large-scale surveys or censuses. Because a large influx of questionnaires comes at once, it is not necessary to provide a large storage area, which later will be left unutilized. It is more convenient to archive and retrieve an image of the form kept in the system than for staff to manually store and search for questionnaires.

22. The overall cost of the system is relatively lower than other methods. The cost for data capture using ICR system is high at initial stage of system installation, but in the long

run only maintenance and development costs are required. Moreover smaller number of staff and less time for data capture does reduce the cost substantially. It is found that, the larger the amount of work the cheaper the cost of data capture for ICR system versus manual data capture. And this emphasizes the advantage of ICR for census and large-scale survey projects.

23. Once the system is installed, it can be used for other surveys with no equipment cost and only processing the system. Therefore, it is a good choice for organizations whose main duties are to produce statistics through surveys or censuses.

Issues to be considered

24. Although, the ICR system has many advantages especially for the processing of large-scale survey and census data, some issues have to be considered when using the system. For example, the distribution and return of the questionnaires must be done with caution, especially from remote areas. For example, it may be difficult for the enumerators to keep the questionnaires dry and unwrinkled during the rainy season. Transferring the forms from remote areas in the country should be done with special care. Otherwise, the forms could get wrinkle and cause a wrong interpretation as well as slow down the scanning time. During the last population census, the census forms were kept in wooden boxes until they were scanned. These boxes helped to protect the edges of the pages and to prevent wrinkles and creases.

25. Another important issue is the control of handwriting of the enumerators. Bad handwriting and improper ways in filling-up the questionnaires cause many errors during the image scanning process. Especially for a census project where many enumerators are involved, it is usually difficult to control for good handwriting. One of the main tasks of the supervisors is to monitor the handwriting of the enumerators.

26. The printing quality of different companies may also be different. It is usually necessary to use many companies in printing the census forms because of the large volume. When printing a large amount of forms or questionnaires using many different printing companies, a systematic arrangement should be made for strict quality control on quality of paper, size of the page and color of the text, figures and answer boxes. When the quality of paper used for the forms was not even throughout the whole batch (some might be too thick some too light), the scanning time was slowed due to feeding problems.

27. The color used to circle or block the recorded data or figure should be constant in its concentration. If the quality is not good and the color level is not even, the reading process could result in a wrong interpretation. Also, the color selected should be tested many times to make sure that the system accepts it correctly. During the reading process, the color will be dropped out before the system reads the data or figure in the circle or block. If it is not dropped out completely, it could again produce a wrong interpretation.

28. It is also important to have future work plans for full utilization of the system, especially during the inter-census period. One might consider on providing the data capture service as a Service Bureau for other Statistical Units either in private or public sectors. It is also possible to use the ICR system for data collected via fax or through the Internet. The decision to install scanners at the provincial or regional offices must be

decided before the project begins. In cases where the branch offices are not normally responsible for processing the survey or census data, it may not be appropriate to decentralize the scanning process. On the other hand, if the branch offices are fully equipped with ICT systems and have processed data for other censuses and surveys, it may be helpful to install scanners there.

29. When Image Scanning is selected for the census project, it is recommended that the system be tested with other survey projects as well as with the pilot census. Any problem can occur, from the form designation stage up to the export of data, and the problems vary with the countries. Moreover, managing and controlling the hand-writing, as well as form care, distribution and return of a large number of field staff is not easy. Problems should be anticipated before the official launch of the Census Project.

E. Conclusions

30. Image scanning system is a technology for efficiently managing the data capture for large-scale surveys and censuses. It provides accurate, timely and reliable data capture from the survey forms with less human intervention than needed for manual data entry. Because the census requires large-scale data collection, the image scanning system to be installed must be sufficiently efficient to process the data within the required period. For such a large investment, the possibilities for using the technology after the census is completed will have to be considered.

31. From the experience that the 2000 Population and Housing Census and the 2003 Agricultural Census had adopted the ICR as the core technology in processing the detailed data. This technology has proved to be reliable and help reduce data processing time. However, there have been some problems. For example, the distribution and return of census questionnaires must be done with caution, especially from the remote areas. Some questionnaires were found to be wet of rain or wrinkled and some were recorded with bad hand-writing, and these had affected the quality of ICR.

32. In Thailand, the ICR technology will be used again in the 2010 Population and Housing Census. The census enumerator may be changed from school teachers to village volunteers following the experience of the latest agricultural census. Since the village volunteers have less education, this change may have some adverse impact on the plan to fully deploy ICR. More training would be to be provided. Better procedures in form handling would also need to be implemented. However, as Thailand is fully committed to using ICT for greater productivity in government operations in both central and local governments together with the lesson learnt from previous work, it is believed that the 2010 Population and Housing Census will be successful as it is expected.
