

**ESCAP Expert Group Meeting on Effective Use of IT
in Population Censuses, Bangkok
10-12 December 2007**

The Use of OCR in 2000 Population Census of China

Li Xiru

National Bureau of Statistical Bureau, China

Population census is one of the most complex operations conducted by any national statistical offices, especially in China with population more than one billion. The People's Republic of China has conducted 5 population censuses since its inception in 1949. The last one was conducted in 2000. With the advancement of science and technology, the use of IT in population census has been promoted. Compared with the 1990 census, more advanced information technology has been employed in different aspects of the 2000 census operations.

1. OCR: A Proper Choice

Manual keyboard entry was used for data capture during the 1982 and 1990 censuses of China. With the development of information technology, more techniques such as OMR and OCR have emerged. These new techniques provide a faster and cost effective way for census data entry. After a series of extensive tests and evaluation, OCR was chosen as the data capture method for the 2000 census. OMR has its advantage of faster scanning and more accurate recognition, but these advantages are based on higher quality paper and stricter accuracy for printing and cutting the census forms. It was difficult to meet these high requirements since the amount of forms were huge, and the provincial census offices wanted to print their forms within their own provinces. In addition, Forms for OCR method is more suitable to be designed use friendly for enumerators.

OCR can also provide the opportunity to store the census forms in electrical format since it is based on image scanning.

2. Distributed Data Capture Centers

The data capture of 2000 census was in a decentralized way since the huge amount of census forms need to be treated. The national census office developed the data capture software system and established 360 data capture centers which located in 360 prefecture level cities, each data capture center was equipped with a set of data capture local network system which consists of one Fujitsu M4097D scanner and 3 computers, one computer was used for scanning and recognizing and the other two were for manual verification and correction.

3. Special Trained Coders for Coding

During the enumeration, the enumerators visited all the households and gathered information for each person and filled in the forms, but the Boxes in the forms designed for coding were just left blank. After enumeration, the forms were collected to the county census offices for coding. The special trained coders in county census offices filled in the blank boxes in the forms according to information written by the enumerators to ensure that all the information needs to be captured are transformed into 10 digits from 0 to 9, so the recognition engine just needs to capture these 10 digits, then the county census offices delivered the census forms to each data capture center at prefecture level there the forms were scanned, the image of forms were stored and the information were captured.

4. Well Designed Software and Proper Workflow

The software system and the workflow for data capture were tested extensively and designed properly. An address code file was prepared beforehand and loaded in the system as a control for scanning control;

each EA has a unique ID code in the address code file. The forms for each EA were packed with a cover page which containing information such as the name and the ID code of the EA, number of household and persons by sex enumerated in the EA. When the forms for each EA were scanned, the ID code in the cover page was captured firstly and matched with the ID code in the address code file. The number of household and persons scanned were also used to check with the information in the cover page for scanning control.

During the capture, the system recognized all images as certain digits among 0 and 9, and then gathered and sorted the images need to be confirmed into 10 sets from 0 to 9 for manual verification. The operators can examine all images recognized as each certain digit and find which one need to be corrected or just press <enter> to accept wholly if all of image had been recognized correctly, which made the confirmation and correction easier and faster.

After data capture, the records of all persons in each EA was transformed to the database, and the edit program was loaded for logical check among items and persons with the same household.

5. The Results: Successful Data Capture

The data capture process of 2000 census of China was finished within 6 months, about 4000 persons were engaged in this task, compared with 11 months and more than 10 thousand persons involved in manual keyboard entry during the 1990 census. According to the results of quality check, the data capture was very successful. The rejection rate of the recognition was less than 1 percent and the total error rate was less than 0.5 per thousand.