Using OCR for Large Volume Census Data Capture

-------The Experience of China

Li Xiru Xia Yuchun
National Bureau of Statistics of China

National Bureau of Statistics of China (NBSC) has used OCR technology in two cases of large-volume census data capture; one is the fifth national population census, the other is the second national agricultural census.

The fifth population census was conducted in 2000. During this census, four kinds of forms are used, including "short form", "long form", "death population form" and "temporary resident population form". The "short form" has 49 census items, and there're totally 360 million A4 forms. The "long form" has 95 items and there're 40 million A3 forms totally. For the other two types, there're totally 10 million A4 forms.

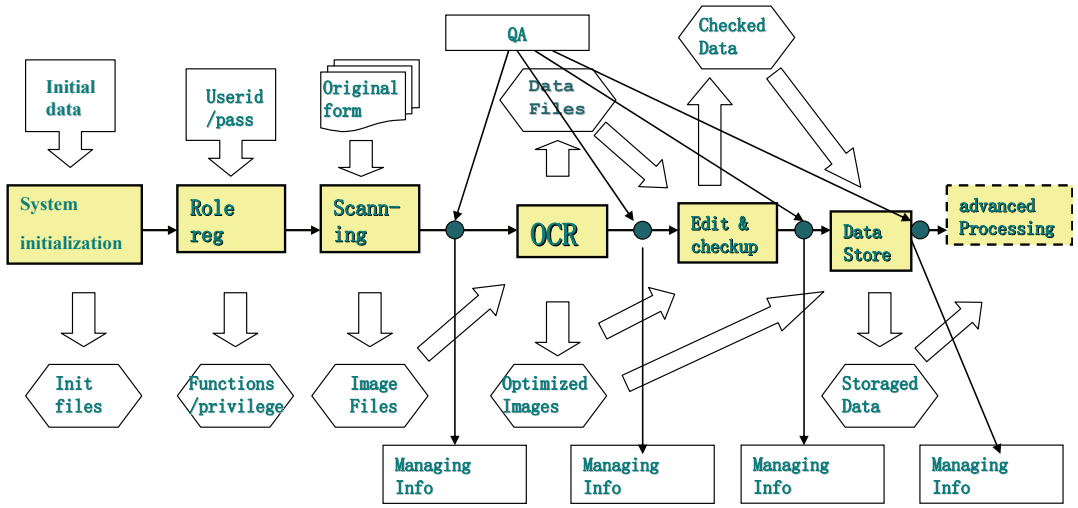
The data capture work for all 4 kinds of population census forms took 10 months. From January 2001 to October 2001, in more than 300 cities, data capture was undertaken simultaneously by OCR method, the original census data is about 64GB. As OCR was utilized, the work efficiency for the fifth national census has been greatly enhanced. The number of the hired staff to capture the data has greatly decreased. Sampling quality inspections during and after data capture indicates that OCR capture has effectively reduced errors and has improved quality of data. Meanwhile, cost accounting indicates that employing OCR method has reduced census cost.

The second national agricultural census was conducted in the end of 2006, which involving about 250 million households in the whole country. There're totally more than 500 million census forms, all the work of data capture was accomplished within 100 days. There're 8 census forms in this project, the number of census items amounts to 541 items. Compared with the fifth national census, the data volume is even larger with more form types and census items. As the fifth national census has been successful in using OCR method, we still employed OCR data capture method in the second national agricultural census. From the end of March to July in 2007, OCR is employed simultaneously in more than 300 cities and regions all over the country for data capture, finally, the scanned image files for the original census forms amounts to 40Tb, data volume through image recognition based on census forms amounts to 300GB, which is nearly 5 times that of the data volume for the fifth national census. Thanks for the advancement of equipments and technologies, we obtained the higher efficiency of data capture.

Chart 1 is the OCR data capture system's processing framework adopted in the second national agricultural census. The target is to achieve timely, safe, stable and effective data capture. The framework includes data flow control, process flow control, quality control; system management and task management of OCR capture system.

For data flow control: mainly to control the flow of census data and image data in various phases in the system and also control the generation of management information. Generalized management information includes three aspects of information: the information of task and state management, system management, quality management.

Chart 1 Processing framework of OCR data capture system



Note: Data means Numeric Data

For process control: The system sets current processing status in the phases of scanning, recognizing, editing, data checking, Chinese name editing, data warehousing, importing/exporting; after finishing processing each phase, set actual status of accomplishment in batch based on the results and execute task distribution according to the workflow predefined by the users. Meanwhile, adjust the workload for recognizing, editing and checking on client side, achieving load balance; Make the best utilization of system resources and the whole efficiency of the system to ensure the work to be executed normally and orderly.

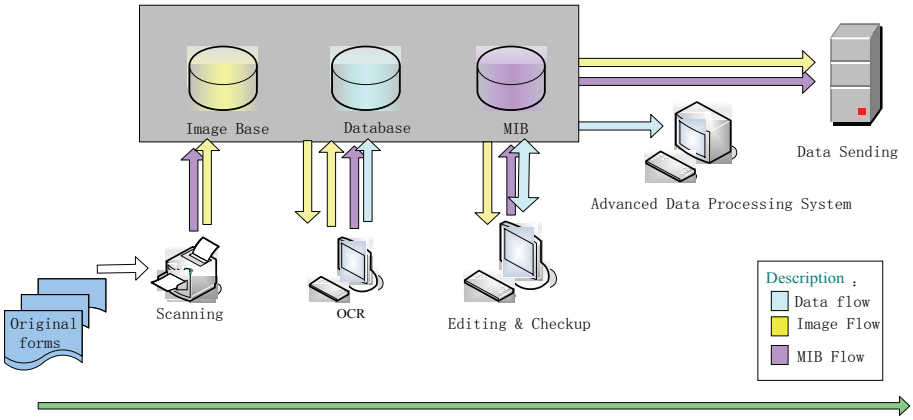
For quality controlling: checking the scanner periodically. Perform selective examination in batch over the census forms. After each batch has passed the phases of scanning, recognizing, editing and data checking, quality examination should be executed. Quality examination and process flow had been combined to ensure that the low quality processing results not enter the next processing phase.

For system management: Include user management, role management, authority management and work log management to ensure the security of systems and operations.

For task management: including system management, template management, workload monitoring and progress management, etc. For workload Monitoring and progress

management, three levels of management are provided: national level, provincial level and on-site level.

Chart 2 Framework of data flow



OCR data capture process

Chart 2 is the implementation of data flow. The management of census data, image data and management information is achieved based on databases, image repositories and management information warehouse.

Chart3 Function framework of OCR data capture

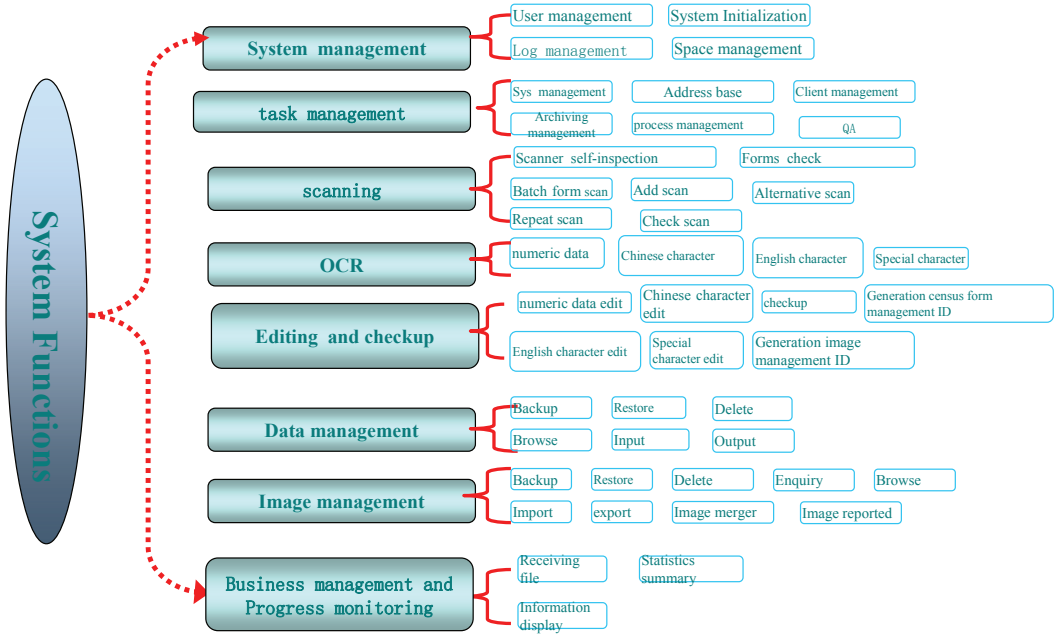


Chart 3 is the function framework of OCR data capture system. It's mainly composed with 8 function modules: system management, task management, scanning, OCR, editing and, checkup, data management, image management, progress monitoring.

The scanning module generates image files and transmits them to image management module with status information transmitted to task management module. The task management module executes task distribution according to the state of vacancy of each OCR clients. The OCR module performs recognition of numerical data and Chinese characters and transmits the data and family name in Chinese characters to data management module with the status information transmitted to task management module. The task management dispatches the data to edit & checkup module for editing and data checking. If original image is needed, corresponding image is fetched by image management module for comparison, the cleansed data after edit checkup are returned back to data management module, the process for family names in Chinese is basically the same with the numerical data, when data capture work is all finished, report upward the data.

To ensure the quality of captured data, quality control is executed in three aspects: before the census form enters OCR system, during the process of data capture, after the data capture finishes.

Before the census forms enter OCR system we check and maintain the equipments and perform quality checking over each batch of original census forms. Concrete operation is to execute one standard form scanning everyday, to check the status of equipments and

perform requisite adjustment and maintenance. Perform sampling scanning and OCR to each set of original census forms to check the paper quality, print quality and filling quality of this batch of census forms. The census forms that don't meet the quality requirements can't enter OCR data capture system.

Before the process of formal census data capture begins, the equipments should pass the checking, the original forms should be eligible and relevant standard documents should be generated. These documents include standard data documents, standard monitoring documents, standard image documents, standard adjustment documents, standard log files and standard temporary files.

Quality control is executed separately in the phases of scan, recognition, edit and checkup in the data capture process.

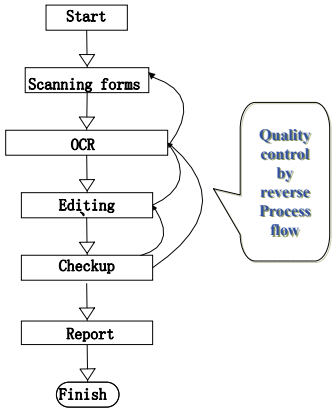
During the scan phase, through recognizing batch cover data and scanner count, as well as table characteristic code, the system checks whether the total page count, subentry page count, total family count for each batch and subentry family count are consistent with the results of scanning. Through comparing the actual address code with address code repository, ensure that the address codes are validity, uniqueness and correctness.

During the recognition phase, collect real time statistics for rejection ratio and suspect ratio. If rejection ratio and suspect ratio is too high, the task administrator checks the reason. If it is not for the census form, execute scanning again.

During the edit phase, check the consistency between recognized record count and the record count in controller document; indicate the rejection items and suspect items; Check the basic logic relationship and value range within the form; indicate the items which have mistakes in logic relationships or value ranges. During this phase, recognition results and corresponding items from original scanned images are displayed comparatively in parallel windows, and convenient modification means are provided for those which need get modified. During this phase, take statistics for false recognition ratio so that evaluation to OCR's recognition capability and the quality of original census forms could be given.

During the checkup phase, checkup the consistency for overall record count; checkup whether there're duplicate codes, whether the addresses are valid and unique, whether illegal characters exist; checkup the consistency between forms; Display the above checkup items and corresponding items in original scanned images comparatively in parallel windows, and provide convenient modification means for those which need get modified.

Chart 4 Quality Management



After the whole set of data has been captured, quality is assured through executing sampling quality check through all phases. This includes sampling check over scanned files, edited files and checked files. The process of sampling check is achieved through reversed flow as shown in chart 4. The result of sampling check is compared to original results to evaluate the quality of the whole batch of work to determine whether the whole batch has been finished or to get back to work from specific phase.

In large-scale census data capture projects, there're three aspects of problems we regard as the most outstanding: 1. How to enhance OCR's recognition capability. 2. Availability and reliability of the system. 3. Project management.

We together with our partners execute following work to improve OCR capabilities:

1) Improve the capability of recognizing numeric characters

Regarding the quality of filling of national agricultural census form varies a lot, two kinds of recognition algorithms and two kinds of recognition engines based on the two algorithms have been developed. A number of onsite tests over actual national agricultural census forms have been executed for these two engines, the recognition engine which better suites the agricultural census projects is chosen. Meanwhile, error rate could be controlled through adjusting the threshold of degree of confidence.

2) Improve the recognition capability for Chinese characters

By collecting large number of actual samples and training the recognizer, recognition capability for Chinese names is improved.

3) Improve orientation capability

Aiming at print deviation and filling deviation, smart locating algorithm has been developed

which has minimized the impact of the print deviation and filling deviation.

4) Enhance efficiency of recognition

Improve the fundamental software of scanner, to achieve the best match between hardware drivers and OCR software and improve the efficiency of recognition.

5) Improve the quality of forms filling from the aspect of management issue

Prescribe the filling standards for form filling so that OCR error rate will be reduced, meanwhile rejection rate could also be reduced.

Before the system is formally used in data capture, we execute test over 2000 actual census forms out of totally 268856 items. Competing double manual data entry results to OCR results, 869 items are difference, in which 793 items from double manual data entry is wrong and 76 items from OCR is wrong. This means that error rate of double manual data entry is much higher than OCR. The improvement of recognition quality can reduce the edit and checkup workload.

In order to ensure the system's availability and reliability, we focused on requirement determination and system test. System requirements are confirmed by experts from NBSC who has got experiences of OCR data capture and the experts from business partners together. During system development, internal tests, third party tests, batch actual data tests and simulated large-volume data tests are executed. Meanwhile, a number of pilot point tests involving representative users are performed. From the perspective of actual use, test the availability, reliability and convenience as well as efficiency of the system.

Regarding the organization and management of the project, first of all is to establish the regulation, working guidance and processes to make the more than 300 data entry site to execute work following uniform regulations, processes and standards.

Secondly, strengthen the training. In order to control the quality of training materials and teaching quality, we carried out the following work:

1) Material writers carefully read technical requirements and software design documents, understand them deeply.

2) Material writers participate in the OCR data capture batch test and manipulate the system actually, get familiar with the actual system.

3) Training materials are firstly used in providers' internal training, then come the training for partners and pilot project training. After each time of training, materials quality and teaching quality will get evaluated for the improvement.

4) Solicit the advices from internal, business partners and users in broader scopes for improvement.

5) Provide network courses on the NBSC website and update the courseware in time.

In order to enhance the quality of actual training, we organized centralized training and on-site training for the users. Lecturing and actual operations are combined during

centralized training, through the combination of these two ways, the familiarity with the system has get deepened.

Onsite training mainly focuses on actual manipulations, including all the sections of system installation, maintenance, operations.

We prepared a batch of agricultural census sample forms ahead of the training; the content involves various situations that may be met during the actual manipulations of the system. We made a number of copies of the sample census forms for the centralized training, onsite training and system tests.

Thirdly, organize multi-target pilot. We organized multiple pilots in many locations aiming at different targets. We organized in the country level:

1) Pilots aiming at integrated management of system installation. Verify the feasibility of integrated installation plan, the cooperation of various parties as well as progress control.

2) Batches of data capture pilot is regarding the availability, reliability and efficiency of the OCR capture system which will be put into actual work.

Meanwhile, the provinces' self-organized pilots for batch of data capture also provide a large amount of experiences and effective advices for the OCR capture system.

By using OCR data capture system and effectively solve above technical and management issue, we finished our tasks on time, the data quality is according with the requirements, so the project target has been achieved.

We've gathered following experiences through foresaid two large volume census data capture projects:

1. Utilize advanced technology to enhance efficiency

In past projects we had executed compete tests between manual capture and OCR data capture, specific areas also have once used manual data entry for part of the census forms. Based on the tests and actual use, OCR data capture has the characteristics of high speed, low error rate and it could greatly enhance work efficiency.

2. Combination of technology and administrative methods solve the issues of quality and security.

During large-scale censuses, as there are so many participators, and the involved locations spread in very wide areas, the quality of captured data and system security are two outstanding issues. We adopted the combination of technical measures and administrative methods to solve these problems. Regarding quality issues, besides providing quality controlling mechanism in OCR data capture system, relevant administrative methods is also established. For example, the paper criterion for the census form, print criterion, as well as filling criterion, etc. Regarding security issues, despite providing authentication, access control, data backup mechanisms in the system as well as the technical measures of installing firewalls, antivirus software, intrusion prevention software, security management criterions are also established. Through the combination of technique and

administrative methods, the issues of system security and data security are solved.

3. Choose the partners with the capabilities of development and service

We realized that choosing partners with capabilities of development and service in large-scale projects is very important. If the partner has the knowledge of the census business and has experiences of similar projects, they will be able to provide helpful advices in requirement analysis and design the system convenient, practical, feasible and reliable. The partner who has strong service capability should own call center and customer service system with a large number of branch offices for the service and strong technical expertise. They can give fast responses to service requests to ensure the problem to be solved in time.

4. Execute project preparation as early as possible

Large-scale census project should get prepared as early as possible to ensure the requirements get sufficient analysis and that the system tests sufficient. The administrative system should receive necessary verification to reduce subsequent alteration.

5. Manage projects federally with partners

Managing projects together with partners could reduce the time spent in communication and eliminate communication obstacles in maximum. The agreements on the solution for problems could be achieved in time which will speed up providing response for problems.

6. Training, pilot and management are the key to success

The issues of staff and system, the progress and quality of the project are the key issues for the success of a project. These issues could be solved through enforcing training, pilots and project management. Enforced training could make the staff more familiar with the work processes and the system, reduce the possibility of error occurring, enhance the efficiency of work. Through pilots inadaptable issues within the system as well as the imperfect issues within the management system could be found and adjusted as early as possible. Effective project management could make the resources be effectively utilized, the project progress and quality be monitored in time, the problems could be solved in time.

7. Control the print quality and filling quality of census forms

The speed and quality of data capture is not only depends on OCR's capability, but also the print quality and filling quality of census forms. Besides trying to enhance the capability of OCR, the print quality and filling quality of census forms should also get controlled. Improving the print quality and filling quality of the census forms helps to reduce rejection rate, suspect rate and recognition error rate, reduce the workload of edit and checkup, enhance the efficiency of work.

8. Execute effective change control

For large projects, change is inevitable which includes adjustment of technical system and

management issues. Change control should be implemented to ensure that change is necessary, controllable and manageable with clear objective, so that the success of the project could be achieved.

China will carry out the sixth national population census in 2010, and data capture work will be performed in 2011. As the projects such as the 5th national census and 2nd national agricultural census all adopt OCR data capture and achieves success, all levels of organizations and staff are familiar with the technology, operation as well as the way of management and administration of OCR data capture, according to the draft plan, OCR data capture will still be the main data capture means for the 6th national census. For the hardware, our current assumption is to sufficiently reuse the equipments used in the 2nd national agricultural census, and add some amount of equipments to make them suit the requirements of the 6th national census. For software, our preliminary assumption is to modify and optimize the agricultural census software and make it suit the new requirements of the 6th national census. Having experienced the agricultural census, various levels of staff have been familiar with the OCR capture system used in agricultural census, and have been familiar with its pros and cons. Executing modification and optimization on its basis will make the OCR data capture system for the population census more easy to be grasped, and it will be more easy for the functions and performances to suit the requirements of the 6th national census.