

Applications of Information Technology in the Hong Kong 2006 Population By-census

A paper prepared
for the 23rd Population Census Conference
16-18 April 2007
Christchurch, New Zealand

by
Dominic K T LEUNG
Deputy Commissioner for Census and Statistics
Hong Kong, China

INTRODUCTION

1. It is an established practice from 1961 for Hong Kong to conduct a population census once every ten years and a by-census in the middle of the intercensal period. Following this practice, a population by-census was conducted in Hong Kong in the eighteen-day period from 15 July to 1 August 2006.

2. A by-census differs from a census in not having a complete headcount of the population but simply enquiring on the detailed characteristics of a large sample of the population. The size and characteristics of the entire population are inferred from the sample results. The sizeable scale of a by-census, as compared to other household sample surveys, can facilitate the provision of statistics of high precision even for population sub-groups and small geographical areas. Such information is vital to the Government for planning and policy formulation, as well as to the private sector for business and research purposes.

3. Hong Kong has been adopting Information Technology (IT) to support the conduct of population censuses/by-censuses since 1970's. In the earlier years, IT was mainly applied in processing the census/by-census data. In recent rounds of censuses/by-censuses, IT has been applied to support various aspects of the project from sampling, data collection to data dissemination. For this by-census, we have developed/enhanced a suite of computer systems to facilitate its planning and execution. This paper discusses the applications of IT in different stages of the 2006 Population By-census (06BC) in four broad areas: (i) supporting field operation with Geographic Information System (GIS)¹ technology; (ii) allowing sampled households to make appointment for enumeration at preferred time slots through an Internet application "Appointment Request Service"; (iii) providing electronic questionnaires to respondents for electronic reporting; and (iv) using Intelligent Character Recognition (ICR) and Optical Mark Recognition (OMR) technologies to capture data on completed questionnaires.

¹ A Geographic Information System can be seen as a system of hardware, software and procedures designed to support the capture, management, manipulation, analysis, modeling and display of spatially-referenced data for solving complex, planning and management problems. (Goodchild & Kemp, 1990)

GEOGRAPHIC INFORMATION SYSTEM TECHNOLOGY

4. In the 2001 Population Census (01C), GIS technology was applied to support various aspects of work through a computer system named “Digital Mapping System (DMS)”. This is the first time that GIS techniques were used to support a population census in Hong Kong. The DMS enables more efficient updating, production and maintenance of maps and effective monitoring of fieldwork progress. This provided the capabilities to produce large volume, multi-format and multi-scale maps efficiently. On field operation, the tailor-made and up-to-date maps of good quality produced by the DMS allowed field operation to be conducted more efficiently. On data dissemination, data analysis on small geographic areas could be performed more conveniently through the use of thematic maps.

5. In this By-census, though only a 10% sample (around 280 000 quarters in the sample) was taken, some 5 000 temporary field workers were still needed to interview 0.7 million people in 230 000 households living in around 50 000 buildings spread around 4 000 street blocks throughout the territory in a short span of time. This diversity poses particular challenge for the fieldwork operation.

6. Adoption of GIS to provide a computer-based design of enumeration areas and automation of map production tasks led to considerable cost saving. Given the successful experience in 01C, the DMS was enhanced to enable more extensive applications of GIS technology to support fieldwork operation of the 06BC in addition to production of maps. In particular, the applications of GIS techniques in three stages of work: allocation of assignments, itinerary planning and monitoring and controlling fieldwork progress, are highlighted in the following paragraphs.

Allocation of assignments & Itinerary planning

7. Similar to past censuses/by-censuses, the 06BC fieldwork was under a 3-tier management, from the lowest level of “division” to the middle level of “working district” and to the highest level of “region”. Each division comprised 3 supervisors and 14 enumerators. Each enumerator had to visit a number of households and interview all members therein.

8. Ensuring that the enumerators could easily locate and move between quarters to be visited was central to efficient fieldwork operation in the 06BC. Nevertheless, determining the number of households and also the specific households to be visited by each enumerator was a very difficult, tedious and time-consuming task. The assignments were first sorted into an “assignment sequence” based on their geographical locations in order to ensure that the assignments of each enumerator would not be widely scattered. Afterwards, an enumeration effort index, which reflected the amount of efforts required to perform enumeration, was determined for each unit of sampled quarters based on a number of parameters before allocation work started. Once the enumeration effort index for each assignment was calculated, divisions were formed systematically such that each division has about the same total enumeration effort value. The assignments for each division were then divided into 14 parts such that each enumerator’s set of assignments was comparable in terms of enumeration effort.

9. In the past censuses/by-censuses, the above-mentioned assignment sequence and some of the parameters for estimating the enumeration effort index like the average traveling distances from field centre² to buildings with sampled quarters and that between buildings with sampled quarters were determined by the field officers manually based on their fieldwork experiences. It was unavoidable that the estimates varied from officer to officer and this process should take a very long time. In this by-census, GIS techniques were applied to tackle these tasks scientifically and efficiently, and to minimize manual intervention in these processes. These tasks were accomplished more efficiently and consistently using the Network Analysis tools in the sophisticated GIS module “Network Analyst”³ as developed in the DMS. **Figure 1** below illustrates how the shortest route (in terms of walking distance) can be generated making use of the Network Analysis tools in DMS.

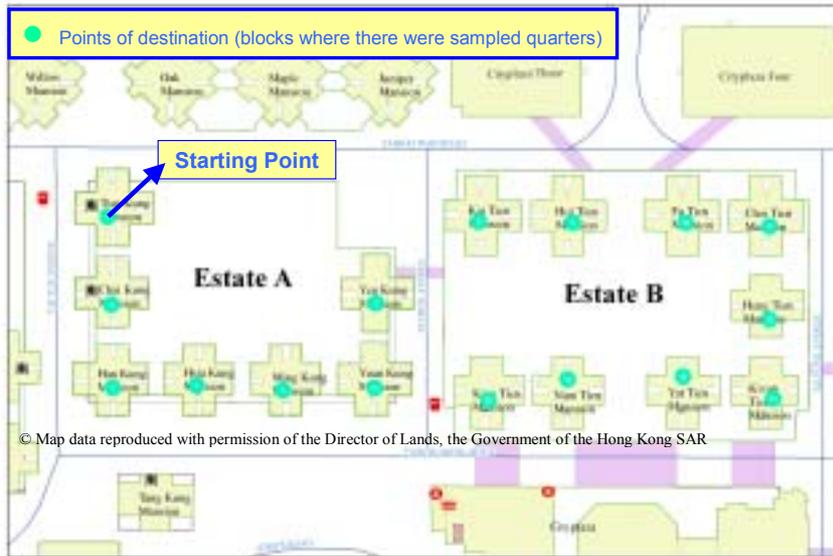
10. After allocating the sampled quarters to the enumerators, an assignment list showing the details of the quarters to be visited was provided to each enumerator. On average, around 70 sampled quarters were visited by each enumerator in the 06BC. Based on the above-mentioned assignment sequence, the assignments as shown on the assignment list were arranged in such a way that the shortest route (in terms of

² One field centre was set up in each working district to facilitate data collection work during the fieldwork operation period.

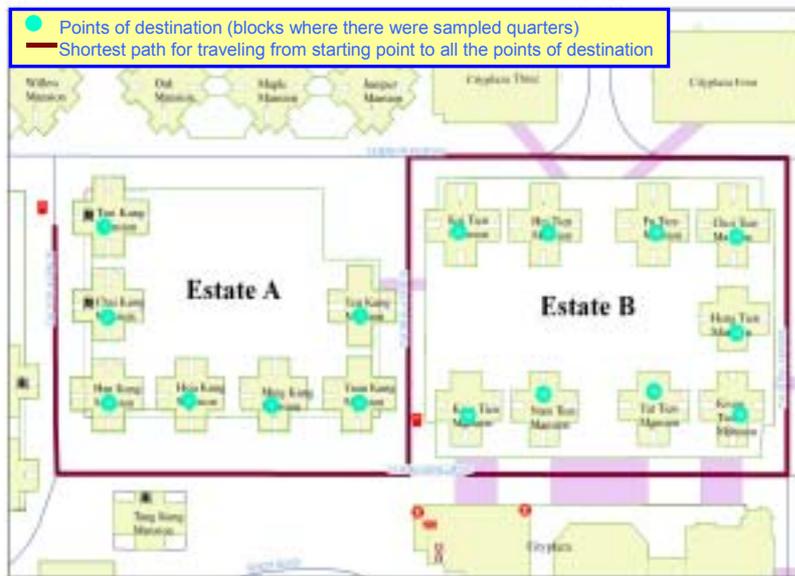
³ It provides tools to solve common network problems, such as finding the best route across a city, finding the closest emergency vehicle or facility, or identifying a service area around a site.

walking distance) to visit the assignments would be incurred. Making available a suggested itinerary to enumeration enabled them to better plan their household visits, and hence, led to higher productivity and efficiency. This information was particularly useful for the enumerators without local geography knowledge.

Figure 1:



To specify the starting point and points of destination by clicking on the map in DMS

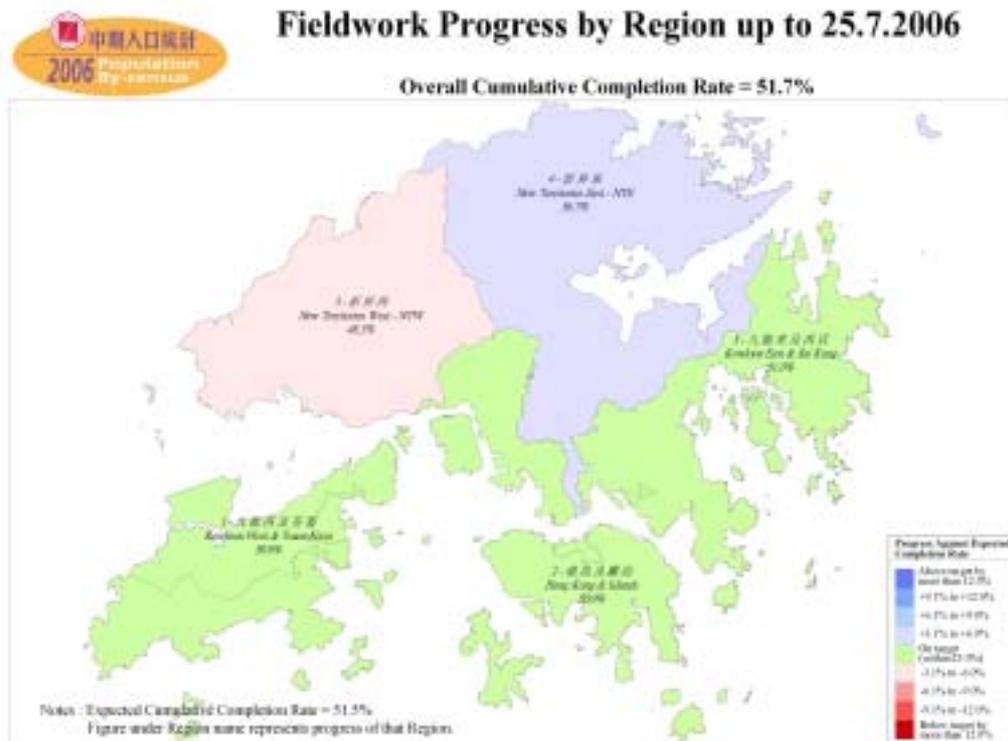


To display the shortest path for traveling from the starting point to all the points of destination using the Network Analysis tool in DMS

Monitoring and control of fieldwork progress

11. Thematic maps⁴ on fieldwork progress were produced based on the records of completed enumerations everyday during the operation period to facilitate the monitoring and controlling of the progress of enumerations. **Figure 2** is a thematic map showing the overall progress of fieldwork in the 06BC for each region which comprised some 3 - 5 working districts. At a glance of this map, the management could quickly grasp an idea of the overall progress of enumerations. For the area behind schedule (i.e. in pink), special attention was taken to identify any possible problems and take appropriate actions to speed up the completion rate. Given the successful experience in 06BC, more extensive use of such kind of presentations (for examples, down to “division” level) for monitoring and control of field progress in the coming round of the population census will be explored.

Figure 2:



© Map data reproduced with permission of the Director of Lands, the Government of the Hong Kong SAR

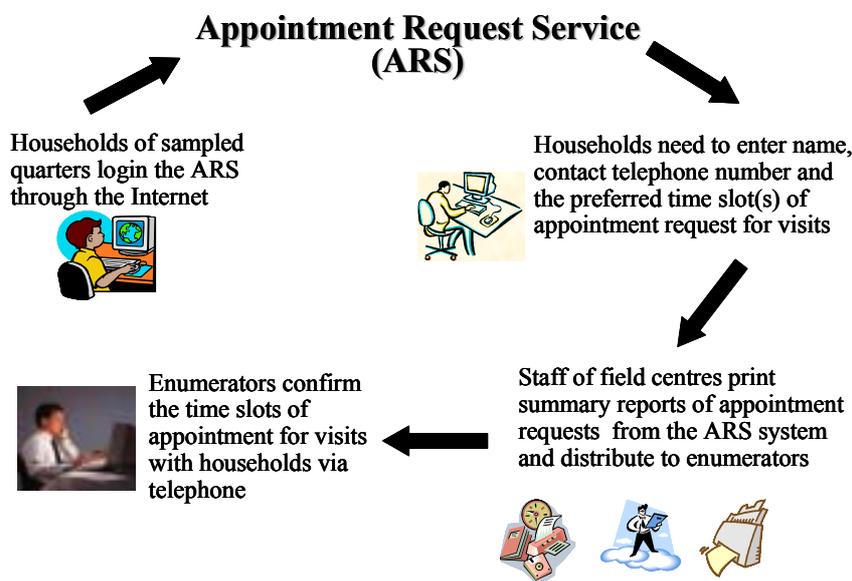
⁴ A thematic map shows the spatial distribution of one or more specific data themes for standard geographical areas. The map may be qualitative or quantitative in nature.

ONLINE APPOINTMENT REQUEST SERVICE

Implementation Approach

12. Since the 1991 Population Census, a Telephone Enquiry Centre (TEC) was set up to handle general enquiries and to make appointment for enumeration before and during the census/by-census operation period. Experience from the past population censuses/by-censuses revealed that a large number of incoming telephone calls were received in the first few days after issuing the notification letters to households and near the commencement of the fieldwork operation. In order to alleviate the workload of the TEC of the 06BC, an Appointment Request Service (ARS) application was designed to provide an additional channel to sampled households to make appointment requests for enumeration over the Internet. Through the ARS application, authenticated households could make, change, cancel or enquire appointment requests for enumeration online by themselves. The system flow of the ARS is depicted in **Figure 3**.

Figure 3:



13. Before the commencement of the operation of the 06BC, each sampled quarters would receive a household notification letter which included a 13-digit household reference number and the name of the assigned enumerator beforehand. To make appointment request, households were required to log in the ARS application (via Internet as shown in the screen in **Figure 4** below) by entering the household reference number printed on the household notification letter.

14. After logging in the system, households had to enter the contact details, including the name of contact person and the contact telephone number, and select the preferred time slot(s) for appointment visits (see the screen in **Figure 5** below). A maximum of three time slots could be chosen. On the other hand, if respondents requested for an interview appointment in time slot other than those available in the ARS, they were prompted to call the TEC to make such special arrangement.

Figure 4:

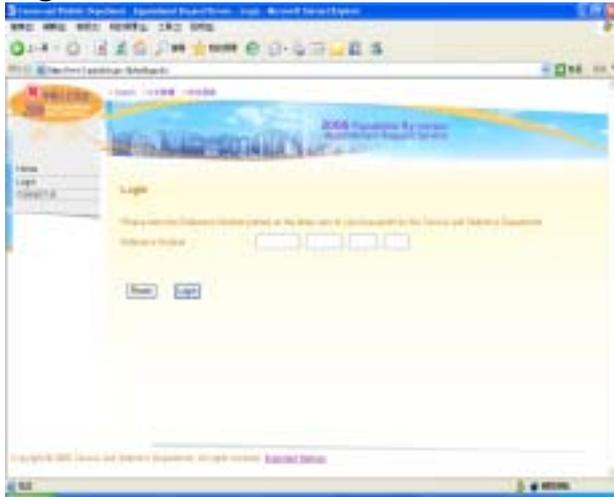
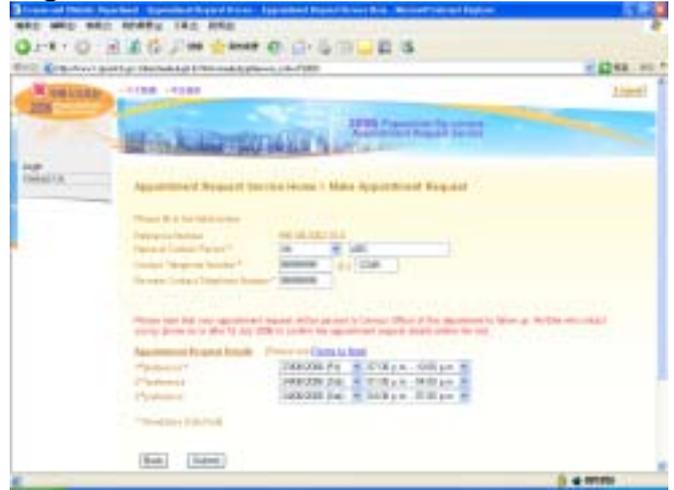


Figure 5:



15. To facilitate the processing of the large number of appointment requests made through the ARS application, a daily cut-off time was set at 8:00 a.m. each day starting from two days before the commencement of the 06BC operation. After the cut-off time of the day, the appointments requested kept in the ARS application would be transferred to the relevant field centres for follow up.

16. In the field centres scattered around the territory, daily summary reports (specimen in **Figure 6** below) on the contact and appointment request details of households by individual field centres and enumerators were generated from the ARS system for enumerators to follow up on the appointments. Enumerators would then confirm the exact date and time of interview with each respondent by phone.

Figure 6:

2004 Population Register Appointment Request Service Application

Daily Summary Report on Appointment Request Details by Field Worker (2004 Last Oct-01)
as on 05/01/2004

Doc: 0007000 Report ID: 004
Time: 09:02:37 Page: 5 / 4

Field Code: 00
Estimate No.: 00100

Quarter No.	Request Date / Time	Req. Type	Name of Contact Person	Contact Phone No./Ext	Field Preference	Second Preference	Third Preference
000700	04/07/2004 20:55:19	M	B. Ross	1254505	06/07/2004	06/07/2004	07/07/2004
000700	04/07/2004 19:10:00	M	B. Ross	1254505+20	06/07/2004	06/07/2004	07/07/2004
000700	04/07/2004 18:44:00	M	B. Ross	1254505	06/07/2004	06/07/2004	06/07/2004
000700	04/07/2004 18:30:47	X	B. Ross	1254505	06/07/2004	06/07/2004	06/07/2004
000700	04/07/2004 18:26:06	M	B. Ross	1254505	06/07/2004	06/07/2004	06/07/2004
000700	04/07/2004 18:25:48	X	B. Ross	1254505	06/07/2004	06/07/2004	06/07/2004
000700	04/07/2004 18:20:26	C	B. Ross	1254505	06/07/2004	06/07/2004	06/07/2004
000700	04/07/2004 18:10:59	M	B. Ross	1254505	06/07/2004	06/07/2004	06/07/2004

Notes:
 1. * Refers to the latest transaction of the quarters.
 2. Type of Request (Req. Type):
 M - New Appointment Request
 C - Change Appointment Request
 X - Cancel Appointment Request

Usage

17. The number of appointment bookings through the TEC and ARS before and during the operation period of the 06BC is summarised in **Table 1**.

Table 1:

Channel	Number of Appointments
TEC	33 152 (83%)
ARS	
(a) within TEC hours	5 599 (14%)
(b) outside TEC hours	1 237 (3%)
Total	39 988 (100%)

18. The functions of ARS were found user friendly to the public and it had achieved the aim of alleviating the workload of the TEC. Among all appointment requests for enumeration made, about 17% of appointments were handled by the ARS. In addition, the ARS operated around the clock and could allow respondents make appointments at any time, especially when the TEC was closed during the period from 11 p.m. to 9 a.m. Some 1 240 appointments, i.e. 3.1% of all appointments done via the TEC and ARS, were made during the non-operation hours of the TEC.

ELECTRONIC REPORTING

19. Data were collected from the households through conducting face-to-face interviews in the past censuses/by-censuses. In the 06BC, respondents were given an option to provide data through electronic means. Sampled households opting for electronic reporting could download their electronic questionnaires (e-Qs) from an Internet application and return them to C&SD after completion through the Internet.

Implementation Approach

20. The e-Q was in Portable Document File (PDF) format, with extension features to enable respondents to use the freely available Adobe Acrobat Reader for self-completion. Key features of the e-Q included:-

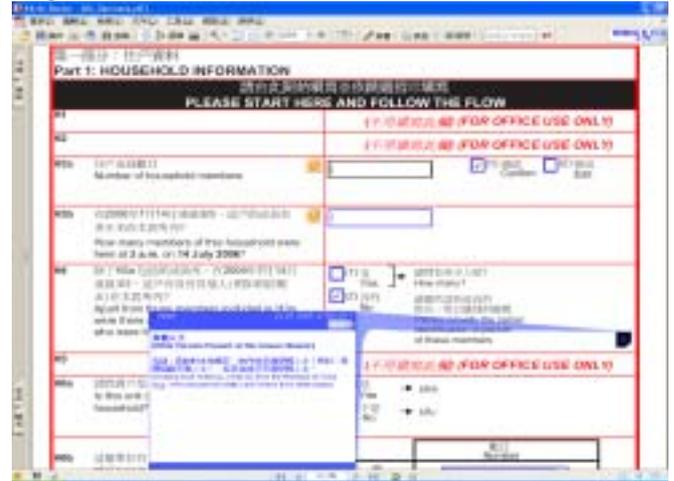
- (i) User-friendliness – the e-Q was an electronic representation of the paper questionnaire with built-in validations/checks and question skipping;
- (ii) Form design flexibility – designing the e-Q did not require complex programming skills; this facilitated making changes to the e-Q during the planning and testing stages;
- (iii) Encryption and password protection – the e-Q was supported by data encryption during transmission and whilst held on the web server as well as unique passwords for individual e-Qs to safeguard data confidentiality;
- (iv) Multi-mode data/file delivery – the delivery of the e-Qs to the respondents and the return of the completed e-Qs were supported by both on-line and off-line modes.

21. Sampled households/persons wishing to use the e-Q could call the enquiry centre for registration before commencement of the fieldwork. Some basic information of the household and its members was collected during the registration. Separate e-Q, protected by a unique password, was generated for each member of the household. Two screenshots of the e-Q are given in **Figures 7a and 7b** below.

Figure 7a:



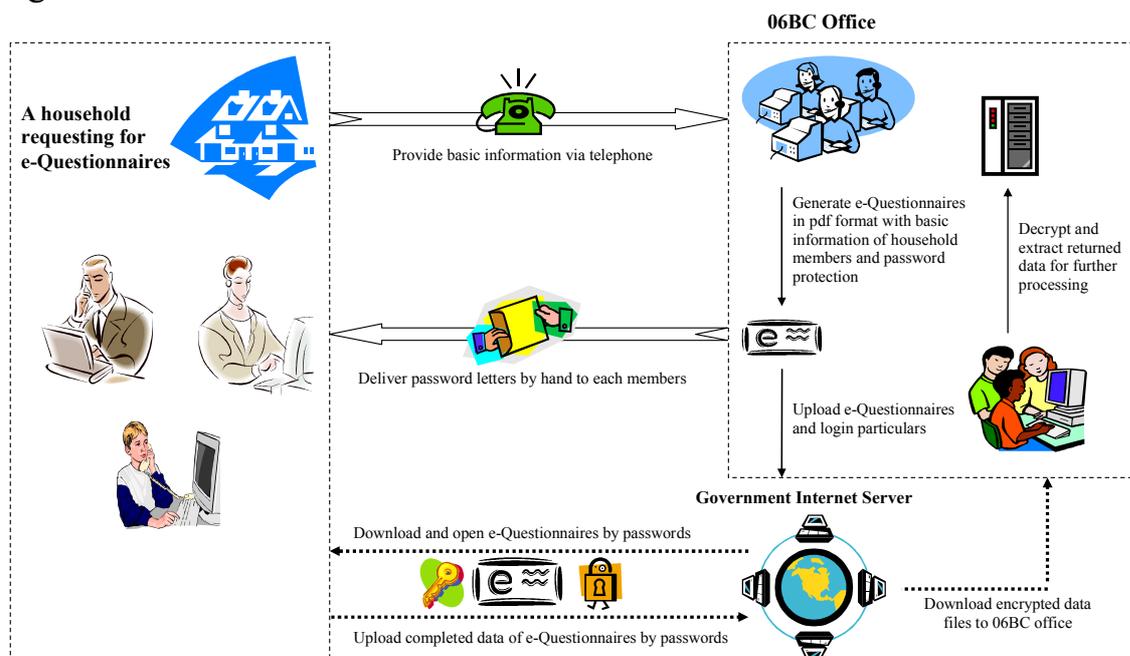
Figure 7b:



22. Census officers would then deliver the password letters to the registered households/persons. Using the password for each household member, respondent was able to login at the designated website to download his/her e-Q and, after completion, submit the data portion via Secure Socket Layer (SSL) of Internet, by simply clicking a button built in the e-Q.

23. The above procedures were specially designed to ensure the confidentiality of data provided on the e-Qs. **Figure 8** summarises the operational flow of the e-Q.

Figure 8:



Usage and way forward

24. The response to e-Q was promising. As shown in **Table 2** below, some 5 400 households registered for using e-Q to report data. Eventually, about 4 500 households had completed the e-Q, accounting for about 2% of all successfully enumerated households.

Table 2:

	Number	Percentage ⁽²⁾
Registered for using e-Q:		
Number of households ⁽¹⁾	5 400	2.8%
Number of persons	15 600	2.6%
Successfully enumerated via e-Q:		
Number of households ⁽¹⁾	4 500	2.3%
Number of persons	13 000	2.1%

Notes

(1) Include households with some but not all members using e-Q.

(2) Refer to the proportion of households/persons among all successfully enumerated households/persons.

Limitations

25. The e-Q technical support line received about 2 000 enquiries during the operation period. The breakdowns by type of problems are given in **Table 3** below:

Table 3:

Type of problems received via e-Q technical support line	Number (%)
Submission of e-Q	448 (23.4%)
System requirements	351 (18.3%)
Completion of e-Q	327 (17.1%)
Extension of submission deadline	140 (7.3%)
Downloading e-Q	99 (5.2%)
Others	549 (28.7%)
Total	1 914 (100.0%)

26. In order to complete and submit the e-Q, respondents had to install specific version of Adobe Reader, which was 6.0 or above. Also, since the e-Q was bilingual (Chinese and English), respondents needed to install the Asian Font Pack when using an English version of Adobe Reader. Most enquiries on using the e-Q from respondents during the operation period were related to this issue. It was anticipated that much more effort would be required to handle these enquiries in the coming round of the population census in view of the much larger number of respondents involved. In sum, it is essential for the e-Q solution to have minimal system requirements on the respondent's computers for it to be successfully implemented.

Salient Characteristics of e-Q users

27. From the results of the 06BC, it was found that domestic households living in housing other than public rental housing were more likely to choose e-Q as the reporting mode. For those living in public rental housing, the proportion of domestic households using e-Q was 1.5%, which is much lower than that for households living in other types of housing, 2.7%. Besides, it was found that more female (8 556) chose to use e-Q than male (7 035). However, it was noted that the return rate for male (82.5%) was slightly higher than that for female (81.2%).

28. Persons with higher educational attainment, especially those with post-secondary education, were more likely to choose e-Q. The percentages of choosing e-Q in the various groups of persons with the same level of educational attainment are shown in **Table 4**.

Table 4:

Educational attainment (highest level attended)	Percentage of population used e-Q
Primary or below	1.6%
Secondary	1.9%
Post-secondary	3.7%
Overall	2.1%

29. As regards occupation, it was found that professionals and associate professionals were more likely to use e-Q whereas craft and related workers as well as plant and machine operators and assemblers were less likely to use e-Q. The percentages of choosing e-Q in the various occupation groups of persons are shown in **Table 5**.

Table 5:

Occupation	Percentage of working population used e-Q
Managers & administrators	2.8%
Professionals	4.9%
Associate professionals	3.4%
Clerks	3.1%
Service workers & shop sales workers	1.6%
Craft & related workers	1.2%
Plant & machine operators & assemblers	1.5%
Others	1.6%
Overall	2.4%

30. The monthly income from main employment of those opted for e-Q (with a median of HK\$13 000) was higher than that of those who did not use e-Q (with a median of HK\$10 000).

INTELLIGENT CHARACTER RECOGNITION AND OPTICAL MARK RECOGNITION TECHNOLOGIES

31. On top of electronic reporting mentioned above, data were mainly collected using a long form (LF) questionnaire via the interviewer method in the 06BC. Besides, a Self-administered Questionnaire (SAQ) was left for completion by those households who could not be contacted during visits made by field staff during field operation period. There were around 284 000 sets of LF (1 set of LF refers to an A4 size booklet with 12 pages) and 12 000 sets of SAQ input form (1 set of SAQ input form refers to a 1-page A4 size form) to be processed within two months after the 06BC field operation. Given the tight schedule and the large volume of data involved, automation of data capturing by the ICR, OMR and bar-code reading technologies was adopted in the 06BC.

32. In the 06BC, questionnaires were specially designed to cater for the operation flow of data capturing as well as for enhancing ICR and OMR data capturing. Special features include:

- (i) *Form Serial Number* – A unique form serial number was printed on the lower left corner of each form page. In the data capturing system, the forms could be identified by this serial number. Besides, the operator could re-arrange the forms in case of any mis-ordering during the operation process.
- (ii) *Cut Angle* – The lower right corner of the form was cut so as to ensure the proper orientation of forms, which further facilitates paper handling and data capturing.
- (iii) *Dropout Colors* – All the ICR and OMR boxes and lines were designed to print with the dropout color. The lines and boxes with the dropout color were removed during image capturing by scanners such that the filled-in information could be recognized more accurately.

33. To date, technologies such as OMR, OCR, ICR and bar-code recognition are commonly deployed to capture data from scanned images automatically:

- (i) OMR, also called Mark-Sense Recognition, is the recognition of marks commonly used on forms, such as check marks, circled choices and filled-in bubbles. The common recognition accuracy rate is 99.5%. This method is still commonly used in the marking of school examination MCQ (Multiple Choice Questions) and in lottery betting slips.
- (ii) OCR, which is how a computer converts words in a scanned image into text. OCR engines can generally recognize only typed or laser-printed text but not handwriting. OCR is usually deployed in applications involving full-text indexing and searches.
- (iii) ICR, also known as Hand-Print Recognition, is similar to OCR but recognizes handwritten characters. Handwritten text is more difficult for computers to recognize and results in higher error rates than printed text. ICR engines usually do best at recognizing constrained printing, which means block printed letters with one letter in each box, or if they are limited to numeric characters.

34. In the 06BC, ICR technology was employed in data capturing. In order to achieve a high accuracy rate, only numeric fields were recognized using ICR, whereas data with multiple choices were captured using OMR technology and the images of textual fields were clipped as images and passed to another computer application for computer-assisted online coding.

Data Capturing Workflow

35. After the 06BC field operation, completed questionnaires were returned from the field centres to the Data Capturing Centre for processing in batches. The LFs returned by each enumerator (around 90 sets) formed a natural batch. A batch header, containing a unique batch serial number, identification of the enumerator, type of questionnaire, and the number of questionnaires in the batch, was placed on top of each batch. On the other hand, information on the SAQs mailed back from households was transcribed to the SAQ input forms.

36. The data capturing cycle in the 06BC comprised the following integral stages:
- (i) Document preparation – The document preparation operators had to check the form condition to ensure a smooth operation for the rest of the tasks. For example, the orientation of each form should be correctly aligned and there should not be any mixed up of other documents. The LF questionnaire booklets were then cut into loose sheets by a cutting machine for subsequent processing.
 - (ii) Document scanning – The forms were scanned by feeding to high-speed scanners Kodak i600 Series (see the photo in **Figure 9** below). The source forms were converted into images. Daily maintenance procedures such as cleaning of all mirrors, glass guide and feeder rollers of the scanner were performed in order to avoid poor scanning quality, such as black lines, overlapping documents, skewed documents, etc.
 - (iii) Data recognition – In this stage, the scanned images first underwent the image enhancement process, including de-skewing, noise reduction as well as line and shading removal. Then, pre-defined form templates automatically authenticated the images to facilitate accurate data recognition.
 - (iv) Completion processing – During the process, all data fields required repairing or additional completion were presented to the operator's computer workstation for action. The image together with its corresponding data files was displayed on the screen so that the operator could rectify any erroneous case manually. Afterwards, the ICR data fields were handled by the Tile Completion processing which allowed the immediate viewing of processing engine results grouped in order by numeric, which made it easier to spot false results for rectification.
 - (v) Quality checking – 2% of the data fields were selected randomly for quality checking. The images of the selected data fields were displayed on the screen and the operators had to key in the data. The data captured were cross-checked with data keyed in by the operators for checking on accuracy.

- (vi) Release of data – Data and images were exported to the server for subsequent data processing.
- (vii) Questionnaire de-preparation – Each set of questionnaires was stapled and delivered to the Questionnaire Storeroom for storage after the whole data capturing process.

Figure 9:



37. Regarding the accuracy rate, the contractor was required to meet the following standard: 98% for ICR and OMR; and 99.5% for barcode recognition. The actual accuracy rates were: 90.4% for ICR before the operators amended the ICR data fields and 100% after the operators amended the ICR data fields, nearly 100% for OMR and exactly 100% for barcode recognition. For those numeric fields recognised with lower than the acceptance rate, they would have to undergo the completion process as mentioned above. Other numeric data recognised would undergo the Tile Completion process.

38. With the advancement of scanning, ICR and OMR technologies, only 3 scanners together with 20 computers were deployed to complete the data capturing work in 34 working days, which was well advance of the schedule. The production hours ran from 8 a.m. to 8 p.m. each working day. On average, 8 339 LF questionnaires (equivalent to 100 068 sheets of A4 papers) were processed everyday.

CONCLUDING REMARKS

39. Results from the 06BC are being released in a series of publications and products starting from February 2007. The planning work for the coming round of population census in 2011 will start shortly. One of the most important issues to be considered is the more extensive use of IT in data collection. It has been recognized that the statistical offices of a number of countries provided options to the respondents to submit their data via the Internet. But there are still many problems to resolve before adopting the Internet e-reporting method in the coming round of population census in Hong Kong.

40. The challenges ahead for us, and probably also other National Statistical Offices, lie in how to meet users' demand for more statistical information, in-depth analysis, with faster turnaround, without unduly increasing respondents' burden. The experiences and lessons learned from the 06BC will definitely be reflected in the planning of the coming round of population census.