Census Data Capture with OCR Technology: Ghana's Experience.

1. Background

Population censuses have been conducted in Ghana at approximately ten-year intervals since 1891 except in 1941, when the series was interrupted as a result of World War II but was resumed in 1948. The first post-independence census was conducted in 1960 and the next in 1970, with the expectation that a decennial census programme would be maintained. Due to circumstances beyond the control of the statistical organization, however, the third post-independence census could not be conducted until 1984. It was not until the year 2000 that the fourth post-independence census was conducted in Ghana.

1.1 Why Ghana decided to use scanning technology?

The fact that 16 years had elapsed after the previous census meant that there was anxiety and great demand for results. In that mood of national hunger for data, there was the need to produce preliminary results by December 2000; this was not going to be possible using the traditional method of keyboard and coders. The decision was therefore taken to use scanners to capture the data from the census questionnaires.

A feasibility study was carried out to find out how long the data processing would take and the number of personal computers would be needed, using the keyboard data entry compared with the use of scanners. It was estimated that while scanning would take three months to capture the census data, manual data entry could take more than a year, using 100 personal computers. It was equally noted that scanning would be cheaper than personal computers because the operations would require fewer trained personnel. Another consideration for the use of scanners was that once scanners were acquired, the technology could be used for data capture of future surveys and censuses.

As a result of the decision to use scanning technology for capturing the census data, two companies from the United Kingdom, with branch offices in Ghana, offered their services. Data and Research Services (DRS) Company introduced the Optical Mark Reader (OMR) technology and GAP International introduced the Optical Character Reader (OCR) and imaging technology. Both technologies were tested during the trial census.

Cost and benefit analysis was conducted after the use of the two technologies for the trial census. Consideration was given to the fact that images of the scanned questionnaires could be archived and accessed later without the physical handling of the questionnaires. Kodak scanners were very fast and could be cost effective. Based on these facts, the OCR technology was selected for the capture of the census data.

1.2 Geographical Coding

Administratively, Ghana is divided into 10 regions. Each region is divided into districts, and at the time of the 2000 census, there were 110 administrative districts in Ghana (170 in 2008). Within each district are the localities – towns and villages.

The 2000 Ghana census was conducted at the household level. A 15-digit reference code was used to uniquely identify each household. The hierarchical coding system used was as follows:

Item	Position	No. of Digits	Possible values
Region	1-2	2	01 - 10
District	3-4	2	01 - 18
Locality	5-7	3	001 - 999
EA Num	8-10	3	001 - 999
Structure/Building Number	11-13	3	001 - 999
Household Number	14-15	2	01 - 99

1.3 Results

At the end of it all, the 2000 Population and Housing Census of Ghana used the Optical Character Recognition (OCR) to capture the census forms. About 4.5 million census forms were scanned with three Kodak 9500 document scanners over a period of 12 months. A total count of 18,912,079 persons, consisting of 9,357,382 males (49.5%) and 9,554,697 females (50.5%) were enumerated.

2. Capturing of census data

Capturing of the main census questionnaires involved four processes:

- office editing,
- opening and preparation of forms,
- scanning,
- validation.

2.1 Office Editing

Editing of the census questionnaires involved correcting errors from the field. The volume of work of office editing was underestimated and led to re-scheduling of the scanning programme. Both staff of GSS and temporary staff with a minimum of secondary education were trained for two weeks to assist with office editing which lasted for 14 months, from June 2000 to July 2001.

- In rural scattered EAs, some enumerators gave the same locality (A06) code as that of the base locality. This had to be corrected before scanning.
- Because the scanners could not recognize faintly crossed marks, questionnaire responses had to be crossed again, deep enough for the scanner to recognize. This prolonged the duration of the data capturing process.
- Some enumerators used wrong EA codes for the questionnaire.
- In many instances, after copying the codes, marking them on the questionnaires was done wrongly or not marked at all.
- The front-page of some questionnaires especially supplementary forms was blank.

2.2 Opening and Preparation of Questionnaires for Scanning

After editing, questionnaires were opened, separated and prepared for scanning. A team was trained to be able to check the questionnaires to ensure that the crosses were dark enough and that the 15-digit reference number was on the inner sheet with the household identification on the outer sheet (this was the only link the two forms had) for each household. They were also to check that continuation forms (for households with more than ten members) follow the original and to shelve questionnaires EA by EA within the same district after opening for scanning.

2.3 Scanning

Data capture for the main census began on 21st August 2000, using three Kodak 9500 document scanners with optical resolution of 300 dpi. The scanning software used was Readsoft's Eyes & Hands. Three 8-hour shift groups, each with a scanning assistant and a supervisor initially worked around the clock, 7 days a week for the first 4 months. Later, the duration of work was changed to a 6-day working week for the remaining 8 of the 12 months that the scanning of the census questionnaires took to complete.

The data capture involved scanning of the questionnaire, interpretation of the scanned marks, transfer of the data and loading the scanned data into an oracle database. Periodic backups of the data and images were made on compact tapes.

The questionnaires were received and scanned district by district within a region, after which they were stamped, bagged and sent to the documents room.

2.4 Validation

Validation of scanned data was an activity to correct structural and inconsistency problems identified in the dataset. The validation groups ran the same three 8-hour shifts as the scanning groups. For every household that failed the structural and/or consistency checks, the validators recalled the data and image of the questionnaire and made the necessary corrections. This procedure was repeated until all structural and consistency problems of district/region were eliminated. This was a very slow and tedious process since the images, which were on DLT tapes, had to be mounted and there was no direct mechanism to retrieve images of the questionnaires.

3. Control Mechanism

To keep track of the movement of satchels between the documents room and data processing rooms, a number of control mechanisms were introduced. Two control forms were designed for recording the satchel(s) of each EA that were received from, and returned to, the documents room. Centrack, a module program in IMPS, and a logbook kept in the scanning room, were used to track scanned EAs. These were very helpful in rectifying errors during the movement of questionnaires.

4. Consistency Edits

One of the crucial steps in census data processing is the editing process during which changes or correction of invalid and inconsistent data are made by imputing non-responses or inconsistent information with plausible data.

The Census Secretariat carefully developed Editing and Imputation rules with written sets of consistency rules and corrections. These rules were translated into three CONCOR editing applications (Pop-Edit.exe, Hse-Edit.exe and Fertility.exe), which were used to *'clean'* the data. This was done at the Regional level.

5. Difficulties and Challenges

The scanning of the census questionnaires using the OCR technology, turned out to be different from what had been anticipated and programmed.

5.1 Paper weight

One of the major problems encountered was the different grammage of paper used to print the questionnaires. The scanner was programmed to accept a particular weight during scanning, but with the varying weight, the sheets got jammed up in the system. Since three different grammage of paper ($80g/m^2$, $100g/m^2$ and $120g/m^2$) were used in printing the census questionnaires, the scanners had to be fed manually to avoid jamming the muddier. This slowed down the scanning process. Problems with the operation of the scanners prolonged the period for data capture.

5.2 No Barcodes on census forms

An 8-page questionnaire, consisting of two A3 sheets was used to design the OCR readable census questionnaire. Though time was taken to redesign the questionnaires, the company printing them could not print unique barcodes on them. That is there was no form of identifying the two A3 sheets forming one household questionnaire.

5.3 Number of scanners used

A total of six scanners were proposed. These were to be networked to operate in pairs, such that while half (three scanners) would be scanning; the other half would be interpreting and transferring data. Thus, the three stages, namely scanning, interpretation and transfer of data, would go on simultaneously with no break or idleness in the process.

However, only three out of the six scanners planned for the data capture were purchased, hence no advantage was derived from the original setup of "no idleness". All three scanners were used to scan but became idle during interpretation and transfer of data.

5.4 Output from scanned questionnaires

The scanned data was interpreted and transferred into Oracle database. Then an ASCII format was generated to be imported into IMPS for analysis. This process took a very long time. The generated ASCII data file was all numeric and left justified. Fields with 3 or 4-digit had

their leading zeros truncated. This meant that field positions defined for the record changed. The data file had to be restructured.

The census forms were designed to cater for a maximum of ten household members. Continuation forms were used for large household and enumerators were tasked to mark the form as a continuation and insert a 15-digit reference number which identifies a household. In most cases, Enumerators did not mark the form as a continuation; the scanners also could not pick the reference number correctly.
These problems prolonged the validation process, which was slow and time consuming, especially when some of the image files were transferred onto tapes, due to the limited storage capacity of the RAID storage.

Another disadvantage of the OCR systems was that corrections made to the datasets could not be made onto the corresponding images of the questionnaires.

5.5 Power interruptions

Power fluctuations, power cuts and low voltage disturbed the flow of work to the extent that it sometimes became impossible to scan during the day. It also led to the destruction of two motherboards of the scanners and damage to a couple of computers and printers. This problem was however resolved when a 100kVA generator and a stabilizer were installed.

Before the installation of the scanners, recommendations were made to connect the Census Secretariat to a major electricity transmission line to provide enough voltage for the scanners. That could have solved the low voltage and the power fluctuation problems, if provided.

Submitted by:
K.B. Danso-Manu
Director of ICT
Ghana Statistical Service
Accra, Ghana.

June 2008.