DATA CAPTURE AND PROCESSING 2006 POPULATION AND HOUSING CENSUS, NIGERIA,

9-13th June, 2008, Dar es salaam, Tanzania

By

Ms Adesola Fatilewa,
National Population Commission, Nigeria

Summary of Paper
This paper describes the data processing methodology adopted for Nigeria's Census 2006. It particularly focuses on data capture and first level editing carried out at the Data Processing Centres (DPCs). It briefly touches on second level editing at the Data Validation Unit (DVU) based at National Population Commission (NPopC) headquarters, Abuja.

Acknowledgements
I acknowledge and appreciate the CTA's (J K Banthia, UNFPA) insistence that I present this paper to this audience. I also acknowledge the assistance of Mr. Henry Mbene of headquarters, NPopC DV., He provided information used in this paper on second level editing. I also acknowledge the assistance of Mr. Francis Idume of Beta Systems for providing details of the equipment and softwares supplied by Beta Systems.

Background

The road that led Nigeria taking the bold step of deploying scanning technology for data capture of 2006 Census, dates back to late nineties. It was since then that NPopC had been inundated with proposals on various document scanning systems. As close to the commencement of pre-censal activities as 2005, unconfirmed statements were making the round, suggesting that the idea of using scanning technology was utopia. Word was around that two African countries that attempted using the technology during the 2000 round of Census, had to abandon it mid-way into processing.

Hence when the decision was finally taken by NPopC to use scanning technology for data capture, we in data processing felt like guinea pigs and were quite sceptical. Thank God today for doggedness of NPopC decision makers persistently nudged by indefatigable Chief Technical Adviser, Dr. J K Banthia; we can heave a sigh of relief because we were proved wrong.

Processing Pre-test and Trial Census

Two experimental tests were carried out with scanning technology before the conduct of the 2006 Census: the first test for processing of the census forms of the second pre-test in April 2005 and the second test was carried out for the census forms of the Trial Census conducted in August, 2005.. The first step was taken by acquiring and utilizing scanning system on demonstration basis to process the second pre-test of April 2005. The number of documents processed was in the order of 100,000 forms as survey covered one local government area (LGA) in each of the 36 States of the country and the Federal Capital Territory. Processing was in one location using 2 scanners with a speed of 2000 to 4000 forms per hour. The forms were only optical mark readable and editing was mainly to correct alignment errors.

By now we had the feel that it was possible to use scanning technology. Apparently the Commission was not confident that the system could handle the volume of documents expected during the main census. Hence after final tendering and bidding, another solution provider was sought who provided five scanners along with two servers which were deployed for the processing of the Trial Census. Trial Census which took place in August 2005 covered selected LGAs and the volume of forms translated to about 10 million population.

The processing was distributed between two data processing centres (DPCs); Lagos and Kano; located at two extremes of the country.

Lessons learnt

1. Staff were identified for suitable roles in data processing of the main census
2. Staff gained experience on the various aspects of the new technology
3. Alignment and recognition problems detected and rectified
4. Decision taken on appropriate archiving system for storage and retrieval of documents
5. Need to have various reports to enable management follow progress of processing
6. Decision to completely eliminate manual coding and editing

Data Capture 2006 census

Scanning technology was fully deployed in processing for the 2006 Population and Housing Census of Nigeria. This was achieved with 21 scanners distributed in 7DPCs located strategically across the country. Immediately after the census, OMR/ICR forms (questionnaires) used to collect data started arriving at the DPCs. These were checked into the archive using designed tracking forms which had a list of the EAs within LGAs and within States to guide as to the genuineness of EAs.

The steps leading to data capture with this system are as below and a schematic diagram of the activity is in Figure 1.

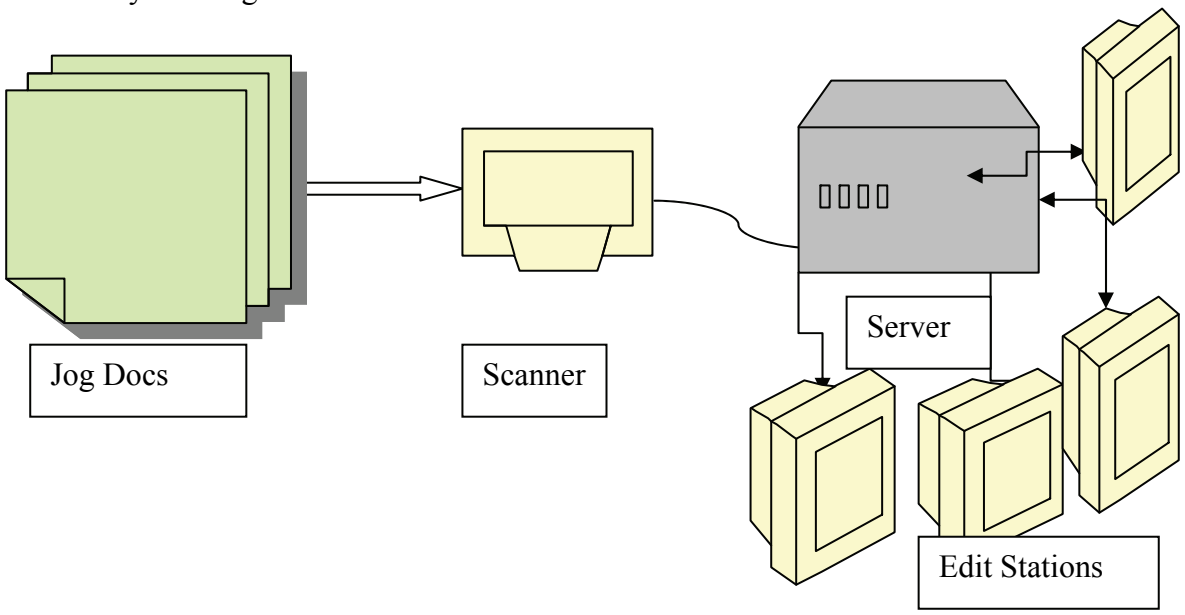

Fig.1. Stages in Data Capture

1. Document Preparation
   - Separation of forms
   - Print batch header
   - Jog sheets

2. Feed sheets through scanner- Scanners deployed were **SC80H with speed of 8000 doc/hour** barring jams and other loading difficulties. Sheets loaded on the feeder in batches separated by batch headers were passed through the transport system and

collected at the output tray. The sheets were returned into their envelopes and sent back to archive.

3. The scanned forms were stored in XML format in servers networked to the scanners- output of the scanner are gray-scale image of each form (black and white image) and an XML information covering each of the fields on the form that was captured. This was imported into the server where software programs with OCR & ICR capability read through and compares with the rules in the database and then highlight any violation for editing.
4. Forms in XML files were imported into server and loaded onto edit stations – (See Fig.2) Scanned forms are loaded in batches of EAs, onto edit stations which were networked terminals.

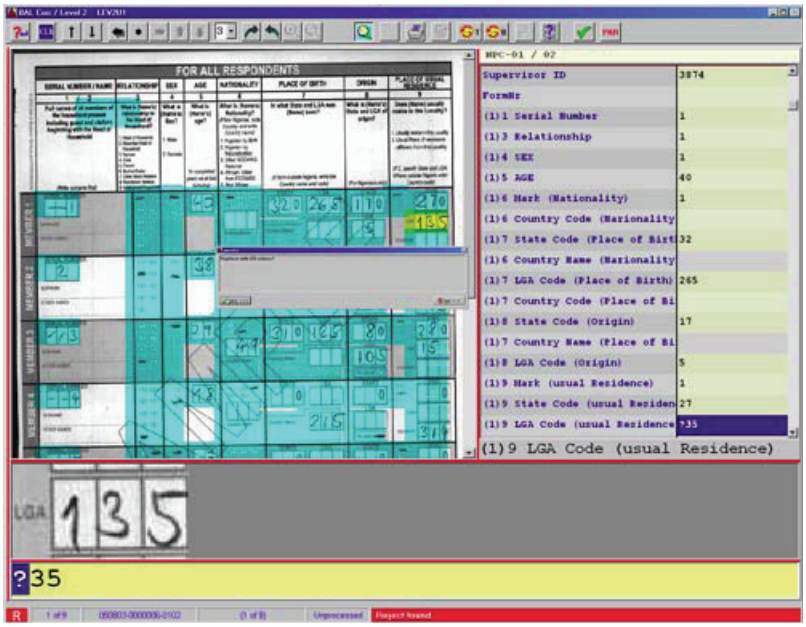

Fig.2. Sample of Edit Station Screen

5. Fields that violated rules as defined by the systems were highlighted on the computer screen and the editor was mandated to key in the correct values. Alsounrecognised data values were highlighted due to recognition problem of scanner and were imputed with correct values as displayed on the screen. The editing system was designed to check geographic identifiers against the batch headers, check 'mandatory fields' of sex, relationship to head and age for every record.
6. The editor was to send a transaction or a whole batch to Supervisor where there is an editing problem or batch did not conform to batch header. This was the 'balancing correction level' which was handled by more experienced staff designated 'Supervisor', so that if the enumerator had filled an incorrect value, the correct value could be 'forced' in.
7. Edited data exported and stored in a Storage Array Network (SAN), in ASCII format
8. Data backed with encryption onto CDs and sent to NPopC hq., Abuja.
9. Daily programmed backup onto LTO tapes was another security measure put in place
10. Completeness Checks at DPC – When all documents in the archive were considered scanned and edited for any state, a list of edited EAs was compared with list of EAs received in the archives. When there was a difference between the two reports, it was necessary to return to the archives to locate or verify any mix up, then re-scan where required.

Occupational Coding

The only field that was not coded during enumeration was response to the question on 'Occupation' for economically active persons. All other fields such as 'geographical identifications', 'sex', 'Country of origin, etc were coded on the field. The 'Occupational Coding' was effected using a computer-assisted coding system, an automatically drop down menu,

Prior to the occupational coding exercise, the system developer was given the coding system to be used and this was an in-house developed one using ILO recommended (International Labour Organization) occupation names. The coding system adopted was loaded in a database that could be called up for viewing.

To effect the coding, previously edited forms were called up and loaded onto edit stations in batches, the editor would view occupation name and enter the code for the occupation as given on the occupation code list. The process is repeated until all the questionnaires in the batch were coded and all available batches completed.

Editing and Imputation

Editing was in two levels. There was the first level editing at the DPCs as previously described. After editing at the DPC, data was converted to ASCII, it was then encrypted, backup on CDs and sent to DVU.The second level editing involved data cleaning, data validation and data consistency checks and was performed centrally at the Data Validation Unit (DVU) of NPopC at Abuja.

At DVU, data was decrypted, validated, collated and further edited. Data is first checked for completeness to ensure that each delineated EA for any local government had data associated with it. Queries were raised, sent to DPCs where a re-check was done to rectify any problems that arose. At this stage, it was possible to collate data and aggregate it by LGA.

CsPro was then applied to carry out various checks on all data items on the questionnaire:

1. Structure checks
2. Range checks
3. Skip pattern checks
4. Inter-record and intra-record consistency checks

The editing team used as guide a previously developed edit specification guidelines produced by Census and Information Technology departments using the UN Population and Housing Editing Book. A combination of 'Hot deck' and 'Cold deck' systems was used when there was a need for imputation. Programs also had to be written to deal with other issues that came up such as linking continuation forms and eliminating data from blank records.

Challenges

1. Ensuring that documents for particular geographic locations were archived in sections of the archive and shelves designated for them
2. That all forms were separated before taking them for scanning
3. Breakdown of jogger
4. Rate of getting documents ready for scanning was usually unable to keep up with scanning
5. Difficulty in maintaining belts and fixing them over pulley
6. That correct batch headers were properly placed on EA batches and that after scanning, EAs were correctly returned to their marked envelopes

7. Poor field work which resulted in 'missing values' of 'mandatory fields', outright wrong values for fields
8. Difficulty in linking forms for households of greater than 8 persons
9. Integration of the two solution providers: form design and equipment and software solutions were provided by two different companies
10. Cleaning of blank records of data associated with them at data capture
11. Dealing with sensitivity of Nigerians to census figures
12. lack of reliable and uninterrupted power supply

Conclusion

In conclusion, the Commission is proud that the decision to deploy a new technology for processing of Nigeria 2006 Population and Housing Census was a success. About 35million forms were scanned and edited using 21 scanners, over 220 edit stations and data in XML format and ASCII stored in about 76TB of SANs. All scanning and first level editing was completed within nine months of enumeration period.

About 1000 Nigerians were trained and gained expertise in various aspects of the scanning technology. However, we are still limited in deploying the acquired technology without external assistance due to limitations in areas of OMR/OCR forms design and development of appropriate scanning softwares. Thus there is a need for intensive trainings in these areas.

Reference

2005 Feb., C. Link, J. Schwaiber and J.Zlesak. 'Training Documentation', Kleindienst Solutions GmbH&Co. KG