

# The evaluation of the capturing system using scanning technology: Challenges in measuring completeness and quality of Census data processing in South Africa

Hakizimana, Jean-Marie, Statistics South Africa, [E-mail: Jean-MarieH@StatsSA.gov.za](mailto:Jean-MarieH@StatsSA.gov.za)

## Abstract:

Statistics South Africa has used scanning technology in the last 2001 Census as well as the subsequent surveys as means of capturing the large-volume data. The implementation of the scanning technology has allowed Stats SA to draw some important lesson in the implementation of an innovative approach in capturing system. Thereafter, corrective measures were implemented with success in the large survey called the community survey. Stats SA has acknowledged that the lack of proper assessment, planning, testing and development of a full complete system at pilot phase has resulted in some weaknesses of 2001 Census data processing. The common challenges in the developing countries are the absence of good local knowledge-base in information technology, the lack of thoroughly tested systems, lack of good management skills of the systems and the unavailability of enough financial resources supporting the information technology locally.

The aim of this paper is to give some innovative approach in the evaluation of the capturing system using the scanning technology. It is trying to give an acceptable level of the completeness of the capturing system as well as the acceptable level of the quality of the information captured using OMR, OCR and ICR. There is a dependency on the how well the scanning system has been able to adhere to the defined requirements taking into account the type of the form, the size of the paper, the graphical layout of the form, the paper background colour-dropout and the identification of the form (i.e. barcode). First, the indication of the completeness is achieved through proper document tracking and performance evaluation across the different processes. Second, the quality is measured by comparing a sample of the scanned / recognised information with similar information recaptured using any another acceptable capturing approach. The acceptable data quality should give an agreement rate above the minimum threshold (i.e. 95%). The quality of information is improved by the corrective measure such as key correction of false-positive reading cases and the extent of imputation applied during data editing. The examples used in this paper are based on the experience of Statistics South Africa from 2001 Census and 2007 Community Survey.

**Key words:** scanning, capturing, completeness, quality control, Census, agreement rate, capturing accuracy, performance management, system, management, data evaluation, data editing, data imputation

## 1. Background

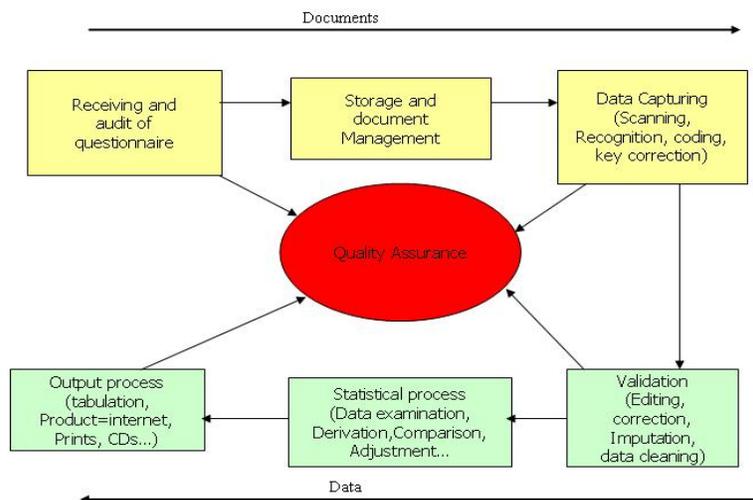
Statistics South Africa has been conducting Censuses every five years. 1996 Census became the first post-apartheid period that targeted everybody living in the country. The Second Census took place in 2001 followed by a large sample survey called “Community Survey” in 2007. Although the objective of having an accurate count of people in the country has been maintained, there has been tremendous change in the methodology taking advantage of the newly available technologies. In 1996 Census, the Census data processing was distributed in the 9 provincial offices using the old technologies of keying from paper. The organisation took advantage of setting a first Census relational database model using the SYBASE platform. In 2001 Census, the focus shifted toward the scanning technology.

The adaptation of the scanning technology as new approach resulted into some unexpected problem due to the lack of fully tested system. However, the revelation of these problems has assisted in in-depth assessment of the weakness of the process design trying to overhaul the system as part of the improvement. Thus in 2007 Community Survey, the data processing system and operation were almost well prepared that the focus shifted toward data quality and completeness measure as proof of accuracy of data capturing system.

Hence, the current adopted data processing is based on lesson learned mainly from 2001 Census. The driving force is to achieve data accuracy, production speed, stability of the system and effectiveness of the overall data capture. The framework of the Census data processing is structured into process flow control, quality control of each process and performance management of the overall capturing system.

## 2. The process flow control of the data capturing system

The data capturing system was subdivided into many processes determining the different activities or tasks to be performed. The graph below gives a summarised view of the high level processes adopted.



At the beginning of the production flow, there is the receiving and accounting of the forms. The handover from the field operation to the data processing centre is managed using a Census and Survey Administration System (CSAS) where each form (i.e. questionnaire) is referenced in the database using a unique number (barcode). The system makes sure that there is a relationship between the form and the container (i.e. box). This database allows to establish a tracking mechanism of each document item (box, form...) across the different processes. Stats SA uses the document management system to manage the movement of the document, image or data across different production stages. This reporting module of the system gives different performance reports necessary for the management of the data capturing production. Moreover, the system gives warning in case a particular document moves into the stage without the necessary preceding requirement from prior processes.

Although the majority of form goes through the scanning stage, a provision of exceptional cases is always prepared to allow those forms that damaged or bad to be captured manually using the traditional approach. Only the good forms are allowed through scanning creating both the image and data. The preferred image file is a multiple-page file with a link to barcode number. The file size is also reduced due to the background colour dropout for better recognition. The storage of the image allows any post-capture verification easy access to the raw information on the form.

Each field area on the form is submitted through the recognition process depending on whether it is optical mark recognition (OMR) or optical character recognition (OCR) or intelligent character recognition (ICR). Since the Census forms are designed with mixed type of fields, the South Africa has adopted eFlow system because it allows the creation of voting rule between different recognition engines. Those cases not recognised are prompted for correction by the operators. After recognition and correction, the quality control is implemented to determine those cases requiring improvement or more attention. The most important part is to account of each form that went through the process. Stats SA has a set of sign-off rule that permit to balance the form /data as complete from capturing system. Any editing and imputation is applied on data that has passed the completeness checks.

### 3. The Quality control measure in data processing

The adopted approach is to put each process through a reversal process that can yield similar results. However, some reversal processes are done on sample of cases available to each stage. For instance, a sample of images goes through determination rules of how good images are or not.

The quality of character recognition is done using different recognition engines. Table 1 gives the proportion of cases recognised by the four engines giving the true positive reading by digit value. Most of the false-positive readings are first submitted through voting rules to determine the most plausible reading and later shown to an operator to key in the value read from the image. A sample is subjected to a second key correction to confirm the correctness of the value and efficiency of the operator. The acceptable threshold of double keyed is set also above 95% overall.

**Table1: Proportion of sampled cases by true or false reading recognition**

Digit value	% False-positive reading	% True positive reading
0	36.4%	63.6%
1	4.1%	95.9%
2	4.7%	95.3%
3	3.4%	96.6%
4	1.1%	98.9%
5	13.6%	86.4%
6	4.9%	95.1%
7	2.9%	97.1%
8	6.5%	93.5%
9	0.0%	100.0%

The overall quality of the capturing system was measured by comparing the keyed value from the image with the recognised value exported. In order to fit within the timeliness, a sample of form and field was selected within grouped form into enumeration area or group of enumeration areas. The quality measure is calculated using the agreement rate as indicating the proportion of cases where value recognised by scanning is the same as the value keyed from image. The group is of good quality if the agreement rate is equal or above 95%.

**Table 2: QA outcome on optical character recognition (OCR) for selected fields**

Field name	% of difference between QA keyed value and recognition values	% difference between blank/null value and valid value	% agreement rate (QA keyed value same as recognised value)
Age	0.8%	1.1%	98.1%
Sex	0.1%	0.2%	99.7%
Relationship	0.5%	0.0%	99.5%
Marital Status	0.2%	0.2%	99.6%

The table 2 gives the measure of agreement rate for selected field that are considered having accurately captured (99.7% for sex). The table makes distinction between those cases where the values are different and those cases where the value is different from a null or blank value.

Although the OMR field is assumed to be more accurate, the measure of quality was bias because there were on certain cases dirty marks on the form that are incorrectly recognised. In 2001 Census, there were many false-positive reading from the optical mark reading fields due to poor colour dropout (blue background on white lamp

of the scanner). The problem was resolved in 2007 Community survey by choosing the appropriate background colour at time of form design. There is improvement of quality when the editing rules and imputations were applied. However, the editing strategy has been to avoid any imputation rate above 2%.

Even if there has been improvement in setting up quality control measure in the data process, there quality of the data is still influenced by other aspects from the field operation and the respondents. There is always unexplained reason issue on why the date is undercounted in certain age categories. The quick fix has been to adjust the results to an estimate of true population. The future emphasis should be on total quality management even though the emphasis of getting data with acceptable trend is being achieved.

### 4. Measure of completeness and performance management

The introduction of scanning technology has prompt a review in the design of the Census forms. Most of the forms are currently uniquely identified using a barcode. This gives benefit in setting up tracking mechanism of the forms across Census data processing processes giving the count of completed cases or not at a given time. A document management system has been developed recording a time-stamp of each barcode as it passing through each stage. The measure of completeness is the number of form out of the system over those

counted in the reference database. At the end, there are few cases not accounted for because they have been removed for instance as duplicates.

The measure of performance is based on the time a given form has spent on each process based on the different time-stamp recorded. Every morning, these results were presented as part of daily progress report which allows applying corrective measure. Although these measures are part of best practice in management of production such as Census data processing, they have been not considered as important in many developing countries because the priority is given on the capturing system alone leaving these secondary systems behind.

## **5. Conclusion**

The implementation of scanning technology in Statistics South Africa has resulted in improvement of the system taking into account the lesson learned since 2001 Census. The challenges in the developing countries remain the lack of local knowledge-base in information technology where different systems are properly designed, tested and implemented. The lack of financial resource prompts the management to concentrate only on capturing system with little emphasis on secondary systems dealing with quality control or performance management.'

The experience of Census in South Africa has proven that the measure of quality can be achieved using a sample at selected stage. Since no capturing system can achieved 100% accuracy level, the measure of quality is determined as being within an acceptable threshold. Stats SA has learned that the success in the implementation of scanning technology depend on the form design in terms of its graphical layout and colour. A good form design will reduce the level of false-positive reading that can bring bias in response particularly for the optical mark recognition fields.

The completeness checks can be also achieved taking advantage of the form unique number. The approach used of document tracking at each stage accounts for the form completed in comparison to the reference database. The performance of each process is deduced by looking at time a form or related data has been at a given stage. The lesson learned is that old technology is still required even if new technology has been introduced. The old technology gives the benchmark measure of improvement made using the new technology. The challenges in developing countries are to have locally available capacity to carry out proper the implementation of the new technology.

## **REFERENCES**

Dekker, A: Adapting new technologies to Census operations, Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects, UN Statistics Division, ESA/STAT/AC.84/9, New York (2001)

Stats SA: Census 2001: Post-Enumeration Survey: Results and Methodology, © Statistics South Africa 2003

TIS: Principals of Intelligent Character Recognition, ©Top Image System, 2001.

TIS: Form design tips, ©Top Image System, 2001.

Whitten, J.L.; Bentley, L.D.: Systems Analysis and Design Methods, Irwin/McGraw-Hill, 4<sup>th</sup> edition, Boston Massachusetts (1998).