2010 World Population and Housing Census Programme

UNITED NATIONS SECRETARIAT
DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS
STATISTICS DIVISION

Census Data Capture Methodology

Technical Report

New York, September 2009

NOTE

The designations employed and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The term “country” as used in this publication also refers, as appropriate, to territories or areas.

The designations “developed regions” and “developing regions” are intended for statistical convenience and do not necessarily express a judgment about the stage reached by a particular country or area in the development process.

Symbols of United Nations documents are composed of capital letters combined with figures.

PREFACE

Data capture in census is the system used to convert the information obtained in the census to a format that can be interpreted by a computer. Whilst it is acknowledged that data capture is only one small part of a national census project it is deemed to be one of the most critical, costly and time consuming activities of a population and housing census. Rapid advances in data-capture technology, especially optical, have greatly increased the speed and reliability of producing census databases in an accurate and timely manner. Nevertheless in the recent past many countries have faced difficulties in mastering these technologies, sometimes by lack of preparation or sufficient knowledge to avoid the numerous pitfalls.

In 2008, the UNSD organized a series of six workshops in different regions of the world. The objectives of these workshops were to:

(a) present and discuss contemporary technologies in census data capture, including the use of Optical Mark Recognition (OMR), Optical Character Recognition/Intelligent Character Recognition (OCR/ICR), Internet data collection, use of handheld devices for data collection;

(b) discuss the process stages for data capture, including best practice, timelines and overall planning;

(c) present an overview of major commercial suppliers for data capture;

(d) offer the possibility to the participants to present and share the experience of their countries in census data capture.

In order to build knowledge on the vast amount of information presented and collected during these workshops, this technical report has been prepared to help countries in their planning of their next population and housing census.

The largest part of the budget for undertaking a national census is used in acquiring the temporary labour necessary to run the census, with the data capture and information technology (IT) systems being a much smaller part. Due consideration needs to be given to both personnel and IT systems to compliment each other and ensure the smooth delivery of the census results. The data capture part of a census should not be viewed in isolation as it and other systems before and after it have interdependencies.

This report is intended to give the reader an insight into the various technical options available for data capture and how they apply to each method. It should be pointed out combinations of these described methods have been used to good effect by some countries; therefore each method is not necessarily mutually exclusive. The circumstances relating to any method/s chosen relates directly to each nation’s specific needs and requirements.

This report has been prepared by Andy Tye, DRS Global Business Leader for Census

Table of contents

1 Technical overview of methods of data capture

There are a variety of methods of capturing data for national census projects. All of the main methods have been used in one form or another and in some cases multiple methods have been implemented to compliment each other. All methods have been used successfully in the commercial environment and all have their own unique technical challenges when being adopted for use in Census. National Statistical Offices (NSO's) that have no experience in these methods should take heart that other NSO's have been in their position before and thus lessons learnt from those experiences may make any technology transition easier. These data capture methods include:

- Manual entry from paper – Operators type in responses they see on the physical census form into the computer system;
- Manual entry from image – Operators type in responses they see on an image of the scanned census form presented to them on a computer screen;
- Optical Mark Reading (OMR) – Using special scanners data is automatically extracted from the census form at the point of being scanned by the recognition of marks (such as tick boxes or multiple choice lozenges) in specific locations on the form;
- Intelligent Character Recognition (ICR) – Software is used that attempts to recognise handwritten text on each census forms' scanned image;
- Personal Digital Assistant (PDA) – A digital handheld device is used to log and record census information by an enumerator (Alternatives are Pocket or Portable PC's);
- Telephone and Internet – Remote capture of data either by automated telephone interviews or entry of data via a dedicated, secure website.

The method/s choices above may be restricted or determined by the logistics of the census project, for example if the NSO is planning to undertake self-enumeration by undertaking a mail-out of census forms for the public to complete, then choice of PDA is not a viable option (PDA's could be used for follow-up of non-respondents, using interviewing staff).

The workflow typically used when paper based data capture is used varies depending upon the method selected. The table below gives an indication of the processes needed to be undertaken and there sequence in the overall workflow

Process	Manual entry (From paper)	Manual entry (Key from Image)	Optical Scanning (OMR)	Optical Scanning (ICR)
Forms arrive at processing centre	Y	Y	Y	Y
Forms are registered	Y	Y	Y	Y
Forms are sorted/sifted	Y	Y	Y	Y
Forms are cut (Optional)	-	Y	Y	Y
Forms are scanned	-	Y	Y	Y
Image archive produced	-	Y	Y	Y
Physical forms moved to storage	-	Y	Y	Y
Background ICR data extraction process	-	-	-	Y
Character inspection	-	-	-	Y
Key correction (Optional QA processes)	-	-	Y	Y
Data entry (Optional QA processes)	Y	Y	-	-
Physical forms moved to storage	Y	-	-	-
Output data produced	Y	Y	Y	Y

1.1 Manual entry

This section gives an overview of the two main methods for NSO's to manually enter the census forms' data into their computer systems. They both require large numbers of staffing and associated computer infrastructure.

NSO's that consider using these methods will require staff with IT skills to set-up a large computer network and all the associated responsibilities that come with supporting large numbers of computer users including and not restricted to, hardware failure, security, backing-up data to offline media periodically, etc. Most NSO's will already employ IT staff that have the competence to scale up systems for this type of project and manage them accordingly, others may require assistance in this matter.

1.1.1 From paper

This is by far the solution that requires the least amount of technical knowledge and implementation and requires an operator to key data directly into the computer from the physical Census form.

A Typical hardware setup for this method may include:

- Networked computers
- Network infrastructure (switches, cabling and domain controllers if required)
- Servers (storage/backup, application, database)
- Backup storage solution (DVD, CD, Microfiche, or other)
- UPS or backup power source

Where a wide area network infrastructure is available installed systems can be easily replicated and connected to create a de-centralized configuration if required by the NSO.

During the data capture phase each completed census form has its response codes manually entered into one of the networked computers by an operator. A software application that relates directly to each response area on the census form is required for this process. Sophisticated versions involve computer assisted key entry where each operator selects a response from options displayed on the screen. For example where a response is either 'Male' or 'Female' the operator may be able to select the required response by using a single key stroke thereby negating the need to type any response.

Typically each operator processes one entire census form at a time and when finished entering responses for that form physically moves it to the completed pile and picks up the next physical form to process. From a quality control perspective double keying may be used so that two operators enter data from the same form and the data they have input is compared to check for accuracy and processes may be adjusted accordingly. This is typically undertaken 'blind' whereby each operator has no knowledge as to who has previously entered any data from that form or what data was entered. It is likely that new or inexperienced operators work is exhaustively controlled/monitored in the first instance during any learning phase. Alternatively or as the work progresses 'expert' operators may have their work checked using sampling where specific data input is checked by a supervisor or administrator for accuracy. In all cases the supervisor is required to make the final decisions on the data input into the system.

This method is also used to process textual responses into classification categories.

Average data input rates vary from between 5,000 to 10,000 keystrokes per hour per operator[i]. Depending on how many responses are on the census forms depends on how many forms can be completed by an operator each hour. Also please note that an operator will use keystrokes to move between forms and between fields. Based on the Papua New Guinea two page census form (Which required approximately 500 keystrokes per form) in the year 2000, equates to perhaps 10 - 20 forms an hour per operator.

Advantages

- Method requires simple software systems and low-end computing hardware.
- Relatively low software costs for example CSPro[ii] is license free and allows for direct data entry.
- Low cost (depending on the costs of local skilled manpower).
- There will be a large number of workstation computers available for other uses after the census is completed.

Disadvantages

- Requires very large numbers of staff, both PC operators and managers.
- Task takes many more man hours to complete as it is labour intensive compared to other automated data entry solutions.
- Potential for errors during data entry so quality control procedures such as double entry or sampling will need to be implemented.
- Standardization of operations is difficult as performance may be individually dependant.
- Staff needs to be kept motivated due to repetitive nature of their work and there is a need for appropriate levels of supervision to be in place.
- Large volumes of paper will be physically moved from input storage to an operators PC to Output storage.
- Physical space to house PC operators and all of the associated requirements large volumes of staff require such as rest areas, toilets, parking, etc.
- There is no backup/copy of the physical paper form. Only data in the computer system will exist after being processed.
- Retrieval of forms to answer queries is time consuming and risks are associated with losing batch integrity.

1.1.2 Key from image (KFI)

This method involves initially scanning the completed census forms using an industry standard document scanner that capture each forms image. These images are then sent in turn to computer screens for operators to select the appropriate corresponding response.

A typical hardware setup for this method may include:

- Networked computers.
- Fast network infrastructure (switches, cabling and domain controllers if required).
- Industry standard document scanners.
- Servers (storage/backup, application, database).
- Storage Area Network (SAN) or Network Attached Storage (NAS).
- Backup storage solution (DVD, CD, Microfiche, or other).
- UPS or backup power source.

Quality control systems used in manual entry from paper (see section 1.1.1) can again be used along with sophisticated techniques such as ‘Seeding’. This is where a sample image is displayed to an operator where the response results are already known (and have been carefully selected) for that image and the operators keyed/selected responses are compared with the expected response. Each operator can be assessed accordingly increasing the potential accuracy of the process.

Scanning of the completed census forms is typically undertaken using a batch control process. For example all the forms that are returned from a specific Enumeration Area (EA) are scanned and all images produced are electronically 'tagged' with the EA code they relate to.

Network and file security relating to the images scanned needs to be considered especially relating to any confidentiality of information on them. Access should be restricted to the images captured as deemed appropriate by any confidentiality/anonymity policy of the NSO.

Advantages

- There will be a large number of workstation computers available for other uses after the census is completed.
- A digital image archive of all completed Census forms is created automatically from the scanning process. The paper forms can be removed for long term storage or disposed of if appropriate.
- IT has been reported that an increased speed from operators can be achieved (compared with manual entry from paper) of between 20%-40% less time/form as they do not need to move physical forms around.

Disadvantages

- The keying process cannot be undertaken until forms have been scanned. Therefore time is needed to scan and then subsequently time is also required to allow data operators to manually enter the data from the scanned images.
- There is a need for a relatively sophisticated workflow to be put in place in order to manage the keying process and smooth flow of images.
- Large numbers of staff required and extra costs associated with the technical infrastructure needed (Hardware and Software as specified above). A significant amount of storage hardware maybe required to appropriately store all images of forms scanned and is likely to be in Tb (Terabytes).
- Finding a suitable use for the document scanners after the processing has been completed.

1.2 Optical Scanning

This section gives an overview of the two main methods (OMR and ICR) for NSO's considering using optical scanning techniques to enter the census forms data into their computer systems. They both require sophisticated software applications and associated computer infrastructure.

- Optical Mark Recognition (OMR) is the term associated with recognising tick box/multiple choice data.

- Optical Character Recognition (OCR) is the term associated with the recognition of machine printed characters, like printed text and as such has a limited application in census projects.

- Intelligent Character Recognition (ICR) is the term associated with recognition of hand written data.

- Barcode recognition can also be achieved with optical scanning methods and may prove very useful in the processing of census forms if they are required to be uniquely identified.

Both of the main methods are well proven with OMR being used for around 40 years and ICR for approximately 20 years. Both technology methods are used successfully today in lots of large volume paper based data capture projects other than census and each have their own set of advantages and disadvantages.

Note - Both OMR and ICR technologies are becoming interwoven with almost all traditional high volume OMR scanner manufacturers being able to capture images for ICR processing. These 'hybrid' solutions offer the best of both technologies for the NSO wishing to gain the benefit of both methods. Even though this may initially sound daunting to any inexperienced NSO they should have the confidence that similar installations and solutions are becoming very commonplace and get the maximum benefit from both methods from a speed and accuracy perspective.

For the successful implementation and use of these methods staff will be required with the necessary IT skills that are familiar with databases, software configuration/support for set-up, management and maintenance of such deployments. For NSO's that do not have staff with experience of such systems a learning curve may be required before adoption of the technology and before the main census activity. Small scale pilot projects are an ideal way to introduce this proven industry technology to all stakeholders concerned and will enable NSO's to optimise and adapt the technology and/or their procedures accordingly.

It should be noted that the resultant quality of data output for any paper based method chosen will be heavily dependant upon how well the enumerators complete the forms and the condition that they arrive at the processing centre. Processing of bad forms will be slow and inaccuracies are more likely. Therefore the most important factors for timely and accurate data capture is to make sure the forms are filled in correctly and are returned in good condition. This means form design and training of the enumerators are both significant factors to consider spending time and effort on to reduce the associated risks as any data capture processes chosen cannot make bad forms good.

1.2.1 Optical scanning – Paper considerations

Paper production is a complex business that involves a large number of variables that produce a huge amount of paper variance in how it performs especially during any scanning process.

There are many properties of paper which can impact the process of scanning. Most scanner suppliers will detail a paper specification relating to some of the properties of paper which their scanner has been designed to optimally scan. Scanner suppliers typically specify a range of paper sizes (minimum and maximum lengths and widths) and thickness, a typical paper grade is 90 GSM (Grams/Square Metre).

Other properties of paper that may wish to be considered to get the best from scanning are properties such as the whiteness and reflectivity that the paper surfaces produces (normally the more reflection during the scanning process the better), the reaction of the paper for the environment it will be used in, such as how it will react to high humidity, heat and light. The amount of paper dust that is generated may hamper the scanning. Scanners that have automatic feed mechanisms will need to separate each sheet before scanning so low friction properties will help in this regard.

A few countries in the past have suffered with obtaining paper from more than one source and variances seen in quality have been significant which have had a negative impact during the scanning process. Even variances in paper thickness may cause problems as the majority of scanners try to detect if a double sheet has been fed. If a significant thickness variance is detected from one sheet after another most scanners will stop and ask for operator intervention slowing the scanning process down.

Also other considerations in the design of the form will help the scanning process. If folds are to be introduced then the form can be designed in such a way to assist the scanning process. By the introduction of pre-folded forms or perforated forms, anyone folding the form is likely to fold it along the pre-fold or perforation. Folds may result in skewed (twisting of the form) or stretched images and/or missed data, so care must be taken if folds are to be added. Open discussions with your scanner supplier should help you decide what is optimal for your specific requirement. If a multi-page booklet is to be used then consider the location of staples or binding for either removal in part or complete using spine trimmers, guillotines, perforations, etc.

If double-sided printing of the census forms is to be undertaken then consideration in the forms' layout specific to response areas may improve data accuracy. It is recommended that response areas on one side of the census form are not positioned directly over response areas on the rear of the form as 'show through' could result in inaccurate data.

The background colour of the form may be printed in a colour that is not seen/invisible to the scanner. These are called 'drop-out' colours and if you plan to use them consistency in the ink used will be required so that all of the background remains unseen by the scanner.

More complex approaches relating to uniquely printing either a barcode, number or address onto every form produced can be achieved by specialist printing suppliers. If one requires uniquely identified census forms then one will need to identify printers who can provide 'over-printing' services either directly on the printing presses or post-printing via laser printers or other such means. Other such approaches are 'Crash box' printing where a sequential number is stamped onto each form and litho-code which is similar. Uniquely identifying each form in some way can help during the data capture process especially if the form is mistakenly scanned more than once. If recognition of this unique data is undertaken then the application software should be able to identify if forms have been scanned more than once and take corrective action so that the data is only counted once.

Unique numbering of forms also provides the advantage of an audit trail for quality assurance purposes. Forms with unique number ranges may be allocated to a specific EA area thus allowing identification if the forms are returned with missing or incorrect Geo-coded data.

There are international standards (ISO) that printing companies can conform to, giving a greater consistency of output of their product by having detailed and defined internal procedures. Selection of such suppliers of large volumes of printed forms may be worthwhile considering to this end.

It is advantageous to try and limit the questionnaires to a single page form that can be scanned without being split into two or more parts. As soon as a questionnaire 'spills over' onto multi-part forms the scanning process for that questionnaire is increased by a factor of two or more depending on how many physical pieces require scanning.

1.2.2 Optical Mark Recognition (OMR)

OMR is a form-scanning method whereby "tick box" style responses are interpreted by a specially configured OMR scanner (using predefined rules and tolerances to gauge the significance of the marks made) and are automatically and immediately passed into a computer systems file or database without the use of a keyboard.

This is achieved very fast and very accurately. The scanning speeds can be up to 15,000 forms per hour with accuracy running at a range of between 99.99% & 99.5%. Although care should be taken with figures quoted for form throughput from suppliers as real world speeds will be a lot lower than the maximum speeds quoted due to loading and unloading of forms, paper jams, routing of forms, etc. OMR is the fastest method of automated data capture.

OMR technology reads marked responses to questions on specially designed and printed paper. The design, print and cutting of the forms is particularly important to make sure that the scanner has the best possible chance of capturing the intended mark on the form. 'Timing marks' or 'Skunk marks' are required to be printed down the edge of the form that relate to the position of the data area.

Only the presence or absence of a mark is detected by the machine. Each mark that is detected may be 'scored' (internally within the configured system of the OMR scanner) based on a number of criteria such as darkness, area, sharpness, etc. Therefore rub-outs or dust/dirt can be automatically excluded from the resultant data produced using tolerance levels. i.e. if a mark is detected for both Male and Female against the record for an individual the scores of each mark can be compared and a decision made by the scanners' software as to which mark was intended.

For example:

> *If an individual on the form has a mark for Male which scores 2 out of 15 and there is also a mark for Female which scores 11 out of 15 it is likely that the intended mark is the much larger scoring mark and that the intention of the enumerator was that this individual is a female.*

Such automatic decisions are only made where specific predefined tolerances and rules are applied. These decisions can be made extremely fast within the scanner allowing for the re-routing of forms to a dedicated output hopper if required, thus separating forms based on the rules applied.

The scanned marked responses are transformed into the resultant output data that a computer can interpret depending on what rules and definitions have been applied.

For example:

If a mark is made against 'MALE' for an individual, the resultant output data may be required to be, but not limited to, one of the following:

- *1*
- *M*
- *Male*

In fact most OMR scanner software will allow for any type/format of output data based on a mark made.

Traditional OMR scanners are limited to the capture of tick box style responses although many will also happily scan barcodes (with possibly some restrictions on barcode type, orientation, placement and bar-thickness). Therefore any handwritten responses on the form must be manually entered or coded using computer-assisted methods. This part of the process should not be underestimated and could require significant planning and resource.

Where OMR scanners have the ability to capture the image of the form during scanning the ability to onscreen key correct any missing, multi-mark or un-validated data should be available. Quality control using techniques such as sampling, double keying and seeding (as described in section 1.1.1 & 1.1.2) can also be effectively implemented to increase accuracy.

Advantages

- Very accurate and very high speed processing can be achieved.
- Equipment is relatively inexpensive.
- Relatively simple to install and run.
- A well-established technology that's been used in many countries.
- If OMR scanner is used that has image capture ability then the digital filing of questionnaires resulting can be achieved allowing for the storage and retrieval of questionnaires images for future use.

Disadvantages

- Requires specially printed and cut forms and scanners.
- There may be restrictions on the background colour choice.
- Tick box responses are not suited to all types of questions.
- The forms are not easy to fill in for the public and usually need a small amount of training for enumerators to complete the form.

1.2.3 Intelligent Character recognition (ICR)

ICR systems interpret hand written number and letter character responses from electronic images of forms scanned. ICR technology interprets responses in pre-defined specific locations on the form and transforms any responses into output data for a computer system to use. For census applications their use should only currently be considered for interpreting characters that are not connected or joined together (Cursive).

Providers of ICR systems can offer solutions that will be able to interpret most commonly used scripts (Roman, Arabic, Cyrillic, etc.). This is based on the ICR engine that they use with their application software.

All ICR applications will require the use of one or more ICR engines. This is the core software that will try to recognize each hand-written response. To improve the recognition process many systems will enhance the image prior to passing it to the ICR engine. Typical enhancement processes include:

- De-Skew – This means the image may be rotated so that the responses are more vertically positioned.

- Shift and Stretch – If the image is offset in one direction it may be re-positioned or adjusted due to a stretched image.

- Contrast and Brightness – These may be automatically adjusted to give a clearer and more defined edge to each character.

- De-speckle – To remove background noise such as dust or dirt and clean-up the area for recognition.

Typically ICR engines will expect to receive a bi-tonal image *i.e.* pure black and white - like a fax image, for processing. Enhancements will improve the engines' ability to recognize characters correctly.

An ICR engine may use a number of methods to interpret each characters image, these include; Neural networks and Feature recognition.

An ICR engine will produce a confidence level for each image of a hand-written character passed to it. For example if a hand-written image of an 8 is passed to it the confidence levels produced may be something like the following:

- 8 = 93.4%
- 6 = 44.2%
- 3 = 34.9%

It is then a matter for the ICR application software to apply its rules to the engines output in deciding if the image of the 8 is actually an 8 or a 6 or a 3. Clearly each ICR system will have its own set of internal rules to apply to these confidence levels, which may or may not be defined or adjustable.

False positives or substitution errors where a character is incorrectly recognized are expensive to identify and to correct, therefore for quality control purposes most ICR

application software will still ask for human confirmation of those characters it thinks the ICR engine has high confidence in. This is called mass verification or character inspection. It typically means that images that the engine thinks are the same characters (for example all images of 8's from many census forms) are grouped together and displayed on screen for a human to confirm before any data is written. This means that a single operator could potentially verify thousands or even tens of thousands of characters each hour.

Those characters for which the ICR engine returned low confidence levels (as defined in the ICR application software) are typically passed to a key correction process. This is where an operator will decide if the character is as expected by being presented with the image of the response with the ability of the operator to correct the data to be output. Quality control using techniques such as sampling, double keying and seeding (as described in sections 1.1.1 & 1.1.2) can also be effectively implemented to increase accuracy.

The drive for increasing automated accuracy within the ICR system can mean that some industry proven methods are implemented such as:

- Pre-defining a restricted character set for a response area – If a response area is only ever going to have a response of 1 to 5 in it then the ICR engine/application is pre-programmed with this knowledge giving more accurate recognition. As a guide here are some indicative figures of ICR engine accuracy for western script hand-written characters:

    - Numeric only in isolated (a single response box for each character) fields 98%
    - Numeric only in semi constrained (a single response box for a number of characters) fields 95-96%
    - Alpha upper case only 90%
    - Alpha lower case only 85-87%
    - Alpha mixed case 75-80%
    - Alpha/Numeric mixed case 50% or less (reduce by 5% if there are special characters not a-z and 0-9)

- Undertaking Multi-level comparisons – The data at the response level can be compared to the data at form level and/or batch level. i.e. Does the response fit the data expected for this form or batch of forms?

- Lookup tables and dictionaries – By having a set of reference data, responses can be compared which may improve accuracy by assisting operators in decision making processes. Implementation of fuzzy logic can be used for this purpose especially when looking at names, places or words.

- ICR engine training – Most current ICR engines have been well trained with various language sets, and training of ICR engines is not commonly undertaken. Some specific character sets may benefit from ICR engines being trained and this requires a large sample of hand-written characters to perform that are representative of those to be recognised

- Voting systems - Multiple ICR engines are used within an ICR application all providing the confidence levels for the same specific characters image. A

voting system may be used internally within the ICR application to make the most appropriate decision based on the engines' outputs.

It is likely that the gains in automated accuracy may be small compared to the work and investment required to implement some of these more complex methods. It will be a trade-off between what level of accuracy is desired and the cost/time associated with achieving it. There will be a limit to the capability of the chosen system due to the quality of handwriting from the enumerators. ICR solutions cannot make bad forms good. The number of staff required to verify and correct the data and to fill the gaps, should not be under-estimated.

To give the ICR the best chance at recognizing the handwritten characters correctly, the NSO may consider training all enumerators on how to best write characters into the response areas on the form. This training can be reinforced by the addition of an example being printed onto every census form, assuming there is room to facilitate it.

Scanner maintenance should be planned and exercised regularly due to build-up of dirt and paper dust from the forms being scanned. If the scanners' paper transport systems are inefficient, recognition could suffer. Also the condition of the census forms may not be conducive to scanning if they have suffered physical damage or have been exposed to bad environmental conditions. It may be that these forms' data are better being manually entered into the system or should be transcribed onto spare blank forms. This is true for all scanning methods not just ICR.

To implement an ICR solution then the following hardware and software is likely to be required:

Hardware

- Image Scanners (TWAIN or ISIS interface)
- Database Server (Full redundancy)
- Storage Server – Terabytes (Raid 5, Mirrored, etc.)
- Network (Gigabit preferred)
- Administrator PC's
- Analysis and reporting PC's
- Key correction PC's (Verification)
- Character Inspection PC's (Mass verification - optional)
- Scanner PC's
- Automatic data capture PC's

Software

- MS-SQL, Oracle or other database
- Data Storage, Archive and Retrieval
- Backup Software
- Software for Administrator PC's
- Analysis and reporting software
- Software for Key correction PCs
- Software for Character inspection PCs
- Software for Scanner PCs
- Software for automatic data capture

As ICR solutions work from the image of a form, de-centralized scanning can be undertaken if appropriate. If this is to be considered, thought should be given to the extra audit trail and communication needs to preserve data integrity, quality and accuracy of the entire process. Also the logistical and security issues associated with moving large amounts of data and images files around (Maybe gigabytes)

As discussed all ICR applications use ICR engines and there are many available in the market for suppliers to choose. Some engines are built to recognize specific languages and others will recognize multiple language/character sets, so your chosen supplier should be using the one (or more) that best fits your local requirements. As time progresses the differences in accuracy between ICR engines is diminishing and so the key differences between suppliers systems are now the methods and workflow within their applications. All census applications tend to be tailored to the local country requirements and as such 'off the shelf' ICR applications are difficult to integrate. The ICR system will require careful planning and design with significant tests using representative questionnaires.

Advantages

- Form design is not as stringent as traditional OMR forms.
- Processing time can be reduced due to automated nature of the process compared to manual entry method.
- Allows for digital filing of questionnaires resulting in efficiency of storage and retrieval of questionnaires for future use.
- No specialist hardware required.
- Very high speed scanning can be achieved.
- Forms designed for ICR are relatively intuitive to complete. Locally printed forms can be used.

Disadvantages

- Comparatively higher costs of equipment (sophisticated hardware/software required).
- Significant hardware/software and trained IT staff will be required to support the system.
- Handwriting on census forms needs to be concise to avoid recognition error.
- Possibility for error with character substitutions which would effect data quality.
- Tuning of recognition engine and process to accurately recognize characters is critical with trade-off between quality and cost.
- Accuracy dependant on manual intervention.

- Where OMR is interpreted from the forms image the recognition rate maybe less than that achieved with traditional OMR methods

1.3 Personal Digital Assistant (PDA)

The PDA is a small handheld electronic device which can substitute for the traditional paper based enumeration. It allows for census data to be captured and stored electronically. The traditional census form is replaced by a series of sequential questions appearing on the PDA screen where the enumerator enters the answer by either selecting predefined responses or entering a variable. They are typically used with a stylus (a pen shaped device) to enter data via the devices' touch screen interface. It is acknowledged that technology is constantly evolving and that PC technology is becoming more portable but for the purposes of this document this section will focus on the PDA. Many of the points discussed in this section will also apply to any portable PC technology being considered for implementation.

Data can either be stored on the device locally in its memory and/or be transmitted to a central location if the appropriate communication infrastructure is present. Locally stored data can be transferred using the PDA download function via various ways, including direct attachment to host computer, transfer of data to memory stick/card, device to device transfer, etc.

PDA devices have a number of technical options that can aid the enumerator and census process. They have the ability to make telephone calls (if within the network coverage area) and transmit data, although consideration should be given to data security.

If a device is lost or stolen what procedures need to be in place to make sure that individual's data is not compromised. It could be that the system is setup so that no individual data is stored locally as it is immediately transmitted over the communications network. This would mean that if the device was lost or stolen data would not be found on it. Alternatively data could be encrypted in some way and without the decryption key would be of no use to anyone.

Another security feature on the device could be by restricting who can log-on. Passwords may be the obvious choice but handing out large numbers of unique passwords to thousands of enumerators may be onerous and also problematic from a support perspective. i.e. if an enumerator forgets their password. A possible alternative would be for fingertip log-on but the PDA device would need an integrated hardware fingertip reader and software for this to work.

Enumeration Area (EA) maps and/or address information can also be loaded onto the device and even aerial or satellite photos to help the enumerator find the correct housing units to visit. If the PDA has integrated GPS, tracking could be undertaken to assist the enumerator in understanding their current location and also capture the geographical location of where the census data was captured.

All of the technical features discussed here require the devices to be pre-programmed in some way and for a large number of these devices this exercise should not be underestimated.

Power consumption and the practicalities of charging the device should be tested and checked as failure in the field would also cause support issues.

PDA's are consumer items which make them plentiful and easy to obtain. Consider the risks in giving a sophisticated consumer device to a temporary worker, who could be being paid a lot less than the cost of the device. Also there are the possible security concerns of carrying a device around in the open in some deprived areas of the world.

Advantages

- Instant data capturing at the point of collection, reducing manual input errors.
- Immediate data validation, reducing re-verifications at later stage.
- Time effective with real time logical validation rules, reducing logical errors.
- Faster processing of census information leading to timely availability of results.
- Additional functionalities can be included such as GPS, Camera, Bluetooth, etc.

Disadvantages

- Setting up of process may take a long time as it requires extensive testing.
- Requires that enumerators have ability to use the device which may require administering a test.
- Requires intensive training of enumerators on use of device (training is more complicated).
- Need to recharge the battery which could run out during enumeration.
- Possibility of equipment failure.
- Expensive capital equipment costs with limited high volume use after the census exercise.
- Loss of devices (Not returned)

1.4 Telephone (CATI) and Internet

The use of the Internet and Computer Aided Telephone Interviews (CATI) for census data collection is growing. However, the methods are always complementary to other more established methods.

Similar to PDA's the on-line internet form is not necessarily an exact downloadable version of the paper based form. It is likely that either questions will appear a page at a time or be sequential. The use of user groups and testing of the flow of the on-line questionnaire should be studied and adjusted accordingly. This is because it is a user based experience to be used by a cross section of the general public. For an effective data collection system thought

should be given to the coverage of households that have internet access, and the potential attractiveness of it to hackers. Use of this method will require a level of authentication control in order to validate and grant access to the correct members of the public. Development of the on-line internet system for data collection is generally outsourced for lack of in-house expertise and if complimentary any data collected will need to be integrated with the other collection methods. From the experience of countries that have undertaken this method it is deemed essential that there are adequate levels of literacy for self-enumeration to take place in an Internet census option.

CATI is a method whereby the housing unit is contacted on the telephone by an interviewer who follows an on-screen script on a computer or completes an internet form on behalf of the housing unit. The computer may also be used to call the telephone number of the housing unit if an appropriate list is available and loaded into the system. Confirmation may be required from whoever answers the telephone at the housing unit before data is collected.

Advantages

- Reduced resources necessary for form handling and data capture.
- For CATI there is a better opportunity to enumerate difficult to reach population groups.
- Automatic filtering of irrelevant questions.
- Better quality data due to in-built interactive verification mechanism.
- Faster availability of census results through simplified data entry and editing.
- The running costs are significantly cheaper than paper based methods

Disadvantages

- Requires that respondents have a computer with Internet access or telephone.
- Management of responses can be problematic, e.g., that households have responded once and only once, and that the actual householder has completed it (Security)
- When completing internet census forms there is a requirement for a high security system to ensure safe transfer of data.
- Need to build parallel processing system as not everyone will use the Internet or telephone.
- Requires mechanism to check for omitted and duplicate submissions.
- Is costly and requires a lot of resources for setting up and adequately test the system.

2 Critical elements and key considerations

2.1 Method Selection considerations

The choice of which data capture method to use is likely to be dependant on national circumstances. As such the choice of method should be part of the overall strategic objective of the census in terms of timeliness, accuracy and cost. Maintaining the integrity of the system and confidentiality of the data will also be key in the decision making process. The choice of which processing system and technology to use for the census data capture needs to be established early in the census cycle so that enough time is available to effectively test and implement it.

When any new technology is used by the NSO for the first time for data capture, extensive testing is critical well in advance of the census. The use of pilot exercises, pre-tests and small scale trials will enable the systems to be optimised and adjusted before the main project is undertaken. If the skills for the chosen technology are not held in-house by NSO staff then there is the possibility to outsource all or parts of the process to a third party company.

If choosing a paper based data capture method the design and paper quality of census forms should be linked to the method of data capture. This will aid the data capture process when the forms are returned for processing.

When imaging technology is to be used, adequate training of enumerators on how to properly fill in the forms is crucial. Time and effort invested to ensure that the census forms are completed as accurately as possible and returned in the best condition possible will pay significant dividends during the data capture exercise.

The manual entry data capture method should only be seriously considered where time is not a consideration and skilled labour is cheap. The feasibility to implement more sophisticated technology to improve timeliness, cost and accuracy is a proven and well practiced exercise. The lessons learnt from similar countries experiences that transitioned from manual entry to other technologies can aid and assist any decision making process.

Considering the population to be enumerated may also assist in the decision making process, especially if Internet or telephone systems are to be thought of. Aspects such as Internet or telephone coverage may need to be accurately estimated before any decision is taken. If the country has migrating populations at specific times of the year or predictions of environmental conditions during the census period this may help decide which method is preferred. Also how the data would be collected from special populations such as the hospitalised, prisoners, temporary visitors and nationals travelling abroad.

The skills of the enumertors should be considered as they are a key stakeholder group that will have an overall impact on how well data is finally captured.

2.2 Outsourcing

Outsourcing is the process of using a third party to undertake work on your behalf to your specification and ranges from total outsourcing of the operation (whereby the third party provides all necessary resources and infrastructure) to variants of combining specific outsourcing components of the project and use of in-house developed systems.

2.2.1 Why Outsource?

Outsourcing could be an attractive proposition for NSO's that have a lack of necessary technological expertise, skill set or processing equipment internally. Using expert external resources could lead to improved timeliness and accuracy of the resulting output data. When using outsourcing the NSO can concentrate on their core substantive work and use its own resources more effectively. It is likely that the job to be outsourced is complex and when adopted the NSO gains access to external expertise and knowledge as part of the process.

The purchase of new systems and equipment every 10 or 5 years may not be considered a productive use of funds especially when new skills and technology need to be learnt and likely not used again.

2.2.2 Determining if and what to outsource

A decision on whether or not to outsource should be based on a number of factors including:

- Defining the technical needs of the NSO in terms of expected output.

  The NSO should clearly define the desired output or product for which outsourcing may be needed in a document that should include the objectives of the project, the output or outcome to be achieved and the time frame involved.

- Specifying the requirements for the delivery of the output in terms of timeliness, quality assurance, accuracy, confidentiality, etc.

  The contractor and NSO should have a shared understanding of the requirements of the contract, including objectives, expected outcomes and priorities. Clear specifications are required, including standards to be met to ensure you get what you want and that everyone understands what is expected from the beginning. Specifications should describe in detail the tasks that are the responsibility of the NSO and of the contractor. They should also include detailed milestones with deliverables against which performance should be evaluated. Furthermore specifications for the output should also address requirements for timeliness, data confidentiality and security and quality assurance.

- An assessment of the market. The NSO needs to determine if it would be feasible to undertake the outsourcing.

  It is important that the NSO has a clear understanding of the market before embarking on outsourcing especially for:

  - Assessing technological possibilities for the work to be outsourced;
  - Assess potential competitors for the project;
  - Estimating cost of the outsourcing;
  - Assessing whether the NSO can afford the outsourcing;
  - Helping with preparation for the tender.

2.2.3 Outsourcing procurement process

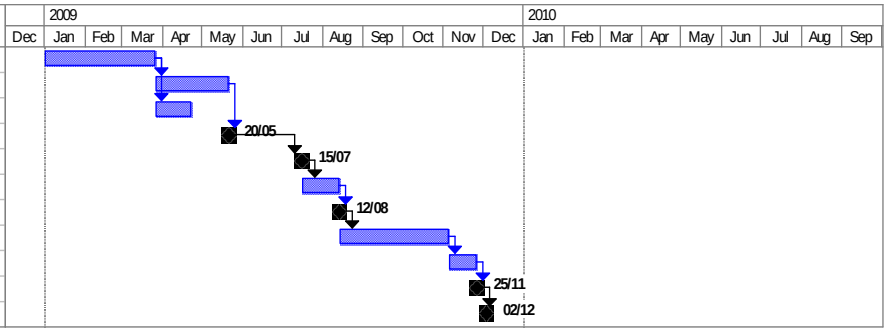
During procurement the bidding process may take a long time as the choice of contractor should be based on a transparent competitive process. The competitive tender process may differ by country, but will need:

- Rules for tender including announcement with deadlines;
- Documents detailing and defining exactly what is being outsourced;
- The NSO may need to collaborate with other government departments in order to learn from their experience in procurement.

When making a choice of contractor it is advisable to choose a contractor with a proven record of data processing (preferably with regional support). The contract should be awarded to those who meet the specified requirements and should not be based only on cost but on other considerations as well, including reliability, proven track record, confidentiality, accuracy and timeliness. It is important to rank the criteria upon which the contractors are being assessed and for transparency convey this process in the tender documents. As part of the selection process a grid may be constructed showing the performance of the contractors against the criteria for the assessment. It may be a useful exercise as part of the process for prospective contractors to undergo a competitive test. For example, having different companies scan a stack of questionnaires and assessing their performance.

Example of Outsourcing schedule and stages (France):

ID	Task Name	Duration
1	Preliminary study, definition	3 mons
2	Writing specifications and requirements	2 mons
3	Administratives clauses	1 mon
4	Official annoucement	0 days
5	Deadline for receipt of bids	0 days
6	Analyse of the bids	4 wks
7	Contract award	0 days
8	Development of the production chain	3 mons
9	Pre-production Test	3 wks
10	Validation of the test	0 days
11	Production starts	0 days



2.2.4 Types of contracts and outsourcing

There are different types of contracting and of outsourcing with advantages and disadvantages such as:

- Full contract/Single source (contractor performs all outsourced work).

- Mixed contract (different aspects of the work are outsourced to different companies).

- Sub-contracting (Prime contractor that contracts out some of the work to another company).

While the mixed contract has the advantage of getting the best for different aspects of contracted work, it presents challenges and risks in terms of management and also assessing responsibility for different contractors. For example if one contractor supplies the census forms and the other the scanning solution, if a problem occurs then each contractor may blame each other.

For administration of the contract it is important to have a detailed legal document (contract) showing obligations and responsibilities of the NSO and the contractor. The contract should:

- Specify clear lines of authority with the NSO retaining management of the overall operation'
- Specify any penalties for breach of contract on the part of the NSO or contractor.
- Specify the possibility for the NSO to break the contract or to delay (if the census is delayed) and the corresponding conditions.
- Be flexible enough to allow amendments as required.

2.2.5 Issues to consider when outsourcing

Adequate planning is needed when considering outsourcing. The plans for data processing should be part of the overall census plan. Any plans for data capture should take into account the census form to be used and vice versa. It may be advisable to choose a data processing contractor who will also print the census forms, to reduce the risk with compatibility issues. Plans for training of enumerators should also take into account the method of data capture to ensure proper filling in and handling of census forms.

When testing the data processing system all aspects should be covered and any problems identified should be corrected accordingly. It is likely that more testing will be required at the beginning of the process to allow fine tuning of the system until the desired result is obtained. The census form design should be finalized early enough to be used in the testing process and it is preferable to start this process early as sometimes long lead times are required.

The pilot census is an ideal opportunity to test the data processing infrastructure. It is recommended that the testing is explicitly defined in the contract specification document. If possible the testing process should be based on previous tests and learned experience from other NSO's.

There is likely to be a difference in objectives between NSO and contractor as the contractor is commercially minded to obtain a profit for their company whilst the NSO is interested in paying less for higher quality output produced in minimum time

There are a number of issues to consider relating to data confidentially and security:

- Contractors and their staff will handle completed questionnaires containing individual's data. However, whether census operations are outsourced or not, data confidentiality and security remain of utmost importance.

- NSO's are responsible for data confidentiality in terms of both perception and reality so they should assure the confidentiality of the data even when data processing is outsourced.

- The NSO should ensure that contractors have instituted strict safeguards to prevent unauthorized access to information. The contractors' staff should be subject to the same data confidentiality rules as NSO staff and be required to sign a legal undertaking.

- It is recommended that the contract should include penalties for data disclosure and misuse.

The NSO is responsible for managing the quality of the census output regardless if outsourcing is undertaken, therefore they should ensure that the contractor has captured data accurately and in a timely and cost effective manner. The setting of quality assurance criteria and monitoring systems for data processing are important considerations. Monitoring quality assurance for data capture can be difficult if the NSO and contractor do not have the same set of criteria. It is also possible that the NSO may not have enough experience and expertise to set up the necessary criteria for assessment and monitoring especially with technology that is new.

The ability of the NSO to manage and control outsourced work is critical to the success of the project, so ongoing monitoring is crucial to ensure the continual review of achievements against originally specified milestones. This is particularly critical when the outsourced work is undertaken away from the NSO's site. It is advisable that a system for monitoring progress should be in-built into the contract. Monitoring should also include regularly scheduled meetings with the aim for clear and open communications with the contractor.

With regard to risk management outsourcing does not transfer risks from the NSO to the contractor; actually outsourcing may bring new risks. The NSO is responsible for all problems, so assumes all risks with the entire process and therefore should develop a risk management strategy for any element of outsourced work.

2.2.6 Outsourcing – Conclusions

The final decision to outsource depends on a thorough assessment of the situation including the current resources and needs of the NSO.

There is a need for clear specifications and definitions within the outsourcing contract.

The importance of testing and adjusting the system should be emphasised.

The NSO's role in management and monitoring of outsourcing must be defined and clear from the start.

Experience gained from others should be used to benefit the evaluation, selection, implementation, testing, monitoring and assessment processes.

2.3 Planning

As census dates are known well in advance the NSO has years to plan for its next national census exercise. In the preparation of the census rounds NSO delegates can attend a number of workshops to gain and share experiences of other projects and consultation programmes are established internally to identify what data needs to be captured. The planning process for the data capture aspect of the census will be an integral part of the overall plan. The various stakeholders included in the planning of a national census project may include central and local government, donor agencies, technical advisors, academic researchers and the business community.

Each specific country is likely to attribute varying degrees of importance on the tasks involved in planning for a census. This will relate to their individual circumstances, resources, and political situations. The significant issues that need to be considered when planning for a national population census such as costs, sound project management, legal frameworks and publicity will shape the direction of the decisions made for the data capture process.

2.3.1 Project plan

A detailed timetable of all the activities in the form of a project plan, highlighting the major milestones and the critical path will be required in the first instance and over time refined accordingly. The plan will likely assist the NSO to obtain funding for the project from various donors or government bodies. Their funding decisions are likely to be based upon the cost estimates derived from the prepared plans, so the planning phase is most critical.

The plan will not just focus on the technical aspects such as data loss but will also discuss the need for:

- Quality Processes – The need for an audit trail, where and how this will be implemented. The use of unique identifiers such as barcodes within the system. The reports required to identify progress, exceptions, etc. Also the defined roles of the staff in each part of the process. Other considerations such as sampling and assessing the quality of the data captured need to be factored into the plan.

- Logistic Issues – The initial collation and distribution of enumerator's kits to include items such as; instructions, pens, pencils, bags, uniforms, census forms and/or PDA. The space to undertake such distribution and data processing will need to be identified and this will include:

    - Building space
    - Loading, lifting and moving equipment
    - Manpower – Number and skill set
    - Security provisions
    - Power requirements
    - Fire and emergency procedures

- Risks – Details of the plans required if in the event of fire or flood damage. The overall security requirements and associated risks to the processes

considered. How the system will react to electrical failure and what is needed to allow processing to continue such as portable generators or battery backup via Uninterruptible Power Supplies (UPS).

Within the census project plan the data capture process will ideally be broken down into smaller tasks and may include tasks and timelines for both the pilot and testing phases and the main census project. Tasks within these timelines may include, but are not limited to:

- Funding agreement – Source and release of funding
- Form design and approval
- Hardware purchase
- Software design and testing
- Data processing location and system installation
- Database design (inc. security and backup)
- Staffing and associated training
- Establishment of quality control and assurance systems

It is generally observed as good practice to undertake a pilot census exactly one year ahead of the main census activity as this will give the NSO a good indication of the challenges of running the project at that time of year. Also it will give the NSO an adequate amount of time to make any adjustments to the plans as required. With this in mind the planning of the pilot exercise will need to be undertaken around two years before the actual main census date.

If implementing Geographic Information Systems (GIS) for the first time the conversion and delineation of boundaries will be an activity running in parallel to the data capture plans. However they are linked as the enumerators will need the maps and boundary information to be able to do their job effectively. Planning the various outputs and needs to make the various activities coincide will be required to ensure the smooth operation of the field data capture exercise.

Each enumerator's workload will also be dependant upon the method of data capture chosen and the time frame designated to the data capture process. The number and type of questions, what and how documentation is completed will all dictate the duration of the exercise and the number of staff needed to complete it within the allotted time planned. The enumerators are also stakeholders in the process and getting their input into the data capture process early on in the planning stages may help identify advantages and disadvantages with the systems and structure being considered.

2.3.2 Planning paper based census

In the preparation of any paper based census forms design the physical layout and flow of the questions should be considered, especially relating to the persons completing the forms. This applies to all form types being considered such as long and short forms. The form layout may need to be more user friendly if the general public will be completing them as opposed to trained enumerators. Minimizing complexity and using easy to understand language will help in this regard. If face to face enumeration is planned then misinterpretation of the forms layout, flow and even questions will be reduced assuming adequate training is undertaken. When using a face to face enumeration approach it is much more costly than dropping the forms off for the

general public to complete. The NSO may wish to consider the benefits of each method for their local circumstances. Testing of form competition should be undertaken using a sample of the relevant proposed group as intended to be used in the census to optimise and adjust accordingly. This will effectively improve the response rate and likely increase accuracy and quality of the data collected.

The planned processes involved in paper based data capture at the data processing centre may include:

- Receiving – A large volume of paper will be received at the data processing centre. They will most likely be returned in batches of forms specific to each EA area. Consideration of the method of how these should be returned to limit any transportation damage is needed. Boxes, strapping or envelopes may be considered.

- Registration – The returned census forms will need to be checked, verified and registered into the building so that the computer system is primed and ready to process them. This also enables the supervisor to see and identify any missing or delayed forms and manage accordingly. The use of barcode technology may be considered to register batches being returned as unique identifiers. A batch header may be generated at this part of the process to move around the processing centre with the forms providing a physical audit trail if required.

- Inspection/sifting – Returned forms will need to be initially inspected making sure there are no unexpected forms, notes or other that does not need to be processed. These can be quickly filtered out by staff which are trained specifically to deal with any exceptions including staples and paper clips.

- Separation/guillotining – Required booklets or multi-page census forms may need to be separated into pieces of paper that can be scanned. The use of staff to separate and if necessary orientate forms may be needed. There are industrial machines that can automate this process and if significant volumes of forms are to be processed these could be an attractive cost effective option to manual labour.

- Scanning – The scanners software may need to know which forms are due to be scanned prior to any scanning and this can be achieved in a number of ways. Barcode, scanning of the header sheet (if used), operator selection based on EA listing, etc. It may not be possible to scan every form, especially if damaged so an exception handling procedure may be required. Scanners should also be regularly cleaned and maintained to keep them at their best operational performance level. Scanners capable of double sheet detection and double-sided scanning will increase the quality and performance of the data capture process.

- Storage/archive – Where images of census forms are scanned and archived there will be less of a requirement to retrieve the physical forms, even so correct storage of the forms is required. The physical forms are the ultimate backup and also they may be legally required to be kept in storage for a defined period of time. If images of every form are captured the storage of those images for long term retrieval may be considered. Digital methods have been used in the past, worm drives, DVD, CD, DLT, DAT and reel to reel tape, all of which will have a lifetime dependant upon the media used. An alternative

such as analogue micro-film techniques may offer a preferred long term storage solution of these images.

The quality and timing of training given to enumerators if using face to face enumeration will directly impact the response and quality of data captured. They are the key stakeholders if paper based methods are to be used. No data capture scanning solution can make bad forms good. If forms are not legible, partially completed, physically damaged (torn, folded, been exposed to water, etc.), etc. then the data capture process will be more likely to take longer, be less accurate and consequently cost more. Therefore training the enumerators to the same standard and method to produce consistent quality is paramount. But employing significant numbers of temporary staff and training them effectively is no small undertaking, especially if constrained by the census budget. The enumerator's skills must be taken into consideration. Many countries have used qualified staff such as professionals (Teachers, nurses, etc.) and others have used graduating/graduated students. Also interpreters may be required if various languages are spoken by the general population. Previous censuses undertaken by the NSO will help identify the appropriate group to employ.

The timing of the training will need to be as close to the census date as possible, so what to do it is easily remembered and recalled. An effectively used method of recruitment and training is to initially employ high level mangers who in turn employ enumeration supervisors who then employ the enumerators thus creating a hierarchy where the appropriate training can be undertaken at each level and then passed to each level in turn. Each level is responsible for training their reporting staff to the quality expected to keep consistency in method and approach. At the time of the census some enumerators may not report for work so a contingency will be required to be trained as reserves and used as and when required.

2.3.2.1 Scanner selection

If scanning census documents is to be undertaken a large number of alternative scanners are available in the general market place to choose from. Suppliers may offer a rental or purchase option, depending on time frames and numbers of scanners required. Financial constraints or other planned uses for scanners after the main census exercise may determine if rental or purchase is preferred.

Most initial thoughts of scanners and specifications will relate to the speed at which the scanners can transport forms. Whilst all manufacturers provide statistics on the best speed at which their scanners perform this in real terms maybe never be a reality that can be achieved as stoppages, jams, paper crashes, operator intervention, duty cycles, etc. will all deteriorate the real-world throughput.

If we consider speed there are a number of factors that will have an impact on performance:

- Form condition – If a form is physically damaged, has folds or foreign objects attached (notes, staples, clips) it may jam or crash in the scanners mechanism. Also dirt and dust will have an impact over time if

there is a build up, maybe affecting the image (If collected) or the paper transport rollers, belts, etc.

- Scanner condition – Most manufacturers will specify a maintenance schedule for their products, where a qualified engineer will make sure the scanner is operating as required. Also consumable items will wear on the scanner and will need replacing as the manufacturer suggests. Sometimes this can be achieved by the scanner operator or by technical staff. With some scanners the bulbs or optics may only last a set period of time before degrading. Paper is very abrasive and any surface it rubs against may be impacted and need eventual replacement. In high volume applications operators are typically trained to spot the signs of that slow scanner performance and clean, replace or rectify issues before they get too bad.

- Size of image – If using a scanner to capture images the speed may be impacted by the physical size of image to be captured. The surface area of the form will directly dictate the size of the image to be captured. The resolution of the captured image may impact the speed of the scanner as well. The industry standard for resolution is defined in the number of pixels captured for every inch of paper (Dots Per Inch - DPI). If the DPI is low then the resolution will be more 'grainy' to the human eye but the file size will be smaller. The higher the DPI the better the image will look but the file will be larger. For ICR recognition either 200DPI or 300DPI is currently preferred.

- Type of image – There are a number of options here being mainly; colour, greyscale or bi-tonal.

    - Colour - If a colour image is preferred then this may have an impact on speed and scanner performance. Colour is normally chosen if the images on the forms need to be viewed by an operator and there is a likelihood that data may be written on the form in colour.
    - Greyscale – This is where the scanned image is represented in shades of grey. This includes 256 shades, where lighter colours on the form are represented by light grey and darker colours by darker grey.
    - Bi-tonal – This is where the image is converted to black and white only, such as that seen in print-outs from fax machines.

  If speed is impacted by the type of image captured then bi-tonal will offer the best performance followed by greyscale and then colour probably being the slowest. This is because more work is required by the scanner to capture the extra information for the various types. Also the resultant file sizes for each type will vary based on the amount of information held within them. File compression can be used to decrease the file sizes and should be considered for archiving purposes if required. One method of capturing greyscale or bi-tonal images and displaying colour images to the operator can be achieved using colour template overlays. This is where the colour image on a blank form is merged with the greyscale or bi-tonal image so that it appears on screen to the operator as a full colour image. This removes the need to

capture full colour images and furthermore reduces the file sizes for the forms' images.

- Real-time validation rules – Some scanners have the ability to apply real-time rules to the data being captured on the forms whilst the form is moving within the scanner. An example of this is in the recognition of the barcode on a form. It is possible to scan booklets in strict sheet sequence and immediately capture the barcode of a form and validate it if required. For instance if a barcode is recognised on a form in a batch it does not belong to, it could immediately be brought to the operators' attention by stopping or out-sorting the form to an alternative hopper, tray or bin. This may slow the scanning throughput but overall increase the system efficiency. Please note that if too much real-time validation is applied scanning will be slow and also the system efficiency could be affected. Testing the system and the shared experiences of others will indicate what level of real-time validation is appropriate.

- Sorting – This relates to the physical extraction of forms to a separate output tray/hopper/bins. This can only be achieved with real-time validation and depending which scanner is selected depends on the impact sorting will have on real-time throughput.

- Doubles detection – All high speed scanners will have a detector mechanism to check for instances of two or more forms being inadvertently fed by the scanners' paper transport system. Scanners may stop to ask for operator intervention to remove the two or more forms or others may automatically attempt to separate the forms and continue without operator intervention. Real-time throughput will be effected with every instance of a detected double so anything that can be achieved to remove this happening in the first place is recommended, such as:

    - Pre-scanning sifting of the forms to remove any foreign objects, folded corners, etc.
    - Jogging the forms to separate them can be achieved with vibration machines; they can also help remove dirt and dust.
    - Using paper that has a low surface friction property.

- Doubles separation – To assist the smooth transport of just one form through the scanner it will have a separation mechanism. This may be a series of wheels, belts or pads to prevent two or more forms passing under the scanners' optics. They may need to be calibrated to the census forms thickness in the first instance, either by mechanical adjustment or electronic. Once setup correctly they will be effective in most instances but may need to be re-adjusted or replaced if wear occurs.

- Operator training – Any unplanned operator intervention with the scanner will slow down the real-world throughput of forms. For that reason reducing the time any operator intervenes with the scanner will increase throughput. This can be achieved by training your scanner operators in the effective use of the scanner so they know what to do under any circumstance that arises.

- Paper crashes and skew – If the scanner is not maintained correctly, dirt, dust or wear can make forms skew (twist) or crash within the scanners' transport path. This may be compounded if there is any physical problem with the form such as folds or tears. Paper crashes will stop the scanner and operator intervention will be required slowing real-world throughput. Skew is unlikely to stop the scanner unless real-time validation is used but may have a larger impact on the processing of data if every image captured is twisted and consequently will need to be digitally 'de-skewed'.

- Environment – All scanners have an effective temperature range and maybe humidity levels that they work most efficient in. It is worth reviewing these specifications to make sure that any chosen scanner will perform in the local environment. Also the amount of space that a scanner and operator is given to prepare the forms and deal with the forms once scanned should be considered.

- Duty cycle - Many scanner manufacturers will define a specific number of forms or hours that the scanner was intended to be used for over a 24 hour period. This duty cycle, if implemented, may mean that scanners stand idle for long periods of time. High speed scanners tend to have a longer duty cycle but this should be considered as it may impact throughput.

2.4 Development & testing

The successful outcome of the census data capture process will be dependant on the development and testing undertaken of the solution. Each NSO will have their own specific criteria and as such there is no readymade solution that will perfectly fit these requirements. Even NSO's repeating a similar exercise undertaken 5 or 10 years ago will have different needs. Therefore any system that is implemented will need to be developed specifically to meet those requirements. Testing of these developments is critical to make sure they meet the required quality standards, timing and system integrity.

Experience shows that it takes considerable time to design, test, evaluate and make any changes to these required developments and as such the appropriate amount of time to undertake these activities is required within the census plan.

Any development should be specified in a requirements specification document produced with the help of suitably qualified technical personnel. This should not only relate to the technical requirements but also its context within the data capture process workflow. The document should focus on what it requires to be achieved and not necessarily how technically the development can be achieved unless this is a limiting factor for the development.

It may be preferable for internal teams within the NSO to develop or alternatively contract a third party to develop, but whichever route is chosen trials, evaluation and testing of any development produced should be undertaken using a well defined test plan, which may include and is not limited to:

- Exception handling;
- Peak load and volume testing;
- System failure;
- Security and confidentially tests;
- Compatibility and data integrity.

System testing should be designed to show any weaknesses in quality, performance and security that will have any impact in the overall census data capture process. Any identified issues should be resolved in a timely manner and re-tested making sure there is no knock-on impact to any of the previously accepted tests.

A perfect opportunity to test the entire process is to undertake a pilot census exercise with a system that is as close to the final proposed system for the main census exercise. Evaluation of as many parts of the system should be completed to identify where improvements can be made.

2.4.1 Paper based

When using paper based systems it should not be forgotten that any implemented data capture process will also depend upon the smooth physical movement of the census forms. Having the appropriate space, storage, staffing, tables, scanners, computer hardware, network, etc. within the data processing building will all contribute to a successful data capture operation. The key here is to make sure that census forms are:

- Kept in good condition;
- Are never inadvertently processed more than once;
- Are not lost;
- Are physically moved the minimum amount of times;
- Can easily have their current physical location identified.

Any data capture workflow process implemented will be similar to that used in any manufacturing or production environment. Therefore the best practice methods used within the manufacturing industry can easily be applied, such as lean manufacturing techniques. To this end an efficient system of workflow will increase productivity and speed in the required output. The physical movement of census forms around the data processing environment in boxed batches (maybe based on EA area) using wheeled trolleys should be considered. Various methods of labelling batches of forms and strict delineation of storage areas will help in this regard.

2.4.2 PDA

If using a data capture method with PDA technology the main developments will relate to the design of the digital questionnaire and its workflow on the PDA device itself. The need to apply quality checks and register data during any downloading process to the central system will also require development. It is crucial that data is correctly identified, verified and validated during the download or transmission process and is expected. Testing the download process or data transmission speeds may identify parts of the system that will require adjusting to accommodate any expected peak load. If the data

is to be held locally on the PDA, tests will need to check the security of that data in the unlikely event that the unit is lost or stolen.

If the PDA is to be used for more than just collecting data then these functions must also be tested. For example if the PDA has integrated GPS and this is going to be used to capture the latitude and longitude of each housing unit, accuracy tests may need to be performed.

2.4.3 Internet

Development and testing of an internet based system is likely to focus on authentication and workflow processes making sure that the right person is responding and they can easily use the system. Peak load tests of the back end hosting systems will be required to account for any eventuality. It would not be to anyone's benefit if the system fails due to underestimated peak load use. As previously discussed it is likely that due to skill sets the NSO will use a third party for this process and discussions regarding scaling and security will be needed. Testing of internet systems can be undertaken using 'robots' which are dedicated computer systems that simulate many users using the same system at the same time. From a security perspective hackers (these may be from anywhere in the world and not necessarily local to the host country) may be attracted to the system as it potentially contains valuable personal information. Testing should be completed to check the security of the site and the use of independent specialised third parties companies or specialists from other government departments would help in this regard.

2.5 System deployment

The pilot census should be a smaller version of the planned main census and therefore plans for hardware, software and logistical systems and installation should be tested. The main census will be on a larger scale but all of the plans will be similar. For those NSO's that are not considering outsourcing their data capture processes deployment of the data capturing centre will be required.

The NSO may already have the physical space to undertake the national census or building/structures may need to be procured with the appropriate space and logistical facilities required.

With the physical structure/building identified, installation of services may be required to support the data capture system, including:

- Tables and chairs;
- Power sockets;
- Lighting and heating (if required);
- Air-conditioning (if required);
- Storage racks and shelving;
- Boxes and storage containers;
- Loading and unloading equipment;
- Security – Access control, staffing, alarms, secure areas, etc.;
- Staff facilities – Toilets, rest area, smoking area, etc.;
- Telephone and internet systems;
- Lifts;

- Computer network cabling;
- Backup and contingency systems – generators, hot or cold swap;
- Fire and emergency systems;
- Staff uniforms and identification badges;
- Room dividers and controlled/quarantine areas.

In an ideal scenario, with the required infrastructure in place, the deployment of the data capture system can be made with the installation of computer hardware and software. One effective way is to run small sample sets with known results through the entire process from start to finish. Testing in this manner will highlight areas which can be adjusted or improved to increase the systems' performance and output. This will also test any systems integration or links between systems, making sure the output of one part of the system is compatible with the input of another.

With initial tests completed the systems should be cleared down ready for operator training. The operators need to be familiar with the systems implemented before any live work is processed. It is an essential part of the process that they know what is expected and when. Training on what to do with exceptions will be one of the critical parts in this exercise to keep data integrity and quality. If escalations systems are to be implemented then these will need to be tested to check their effectiveness along with peak load tests.

With trained operators and the system in place the pilot tests can be undertaken. Running live data through the system in greater volumes will not only test the technology implemented but also the personnel working it. Evaluation of the pilot test will indicate issues that need to be addressed when increasing the load in the system and what to expect when scaling the system up to undertake the main census exercise.

2.6 Management

With a pilot census exercise undertaken and evaluated the main census processing will have a good chance of progressing to plan. The focus on managing each aspect of the data capture process in both terms of staffing, resources and the overall workflow will be required. Accurate and timely reporting will help the supervisors and managers determine the current status of the system. Real-time dashboard and reporting tools offered by many suppliers can assist with this requirement. The key with all the reporting is to identify any bottlenecks, backlogs or quality issues in the system that may impact/delay completion or the integrity of the output data. Managers and supervisors will need to be authorised to take the appropriate action to remedy issues uncovered.

Consideration must be given to the workload staff are asked to complete and the potential risks associated with fatigue or tiredness and their effect on data quality. As the tasks are likely to be repetitive, small breaks and rotation systems can be used to good effect. Other systems to quickly identify operator performance and data quality have previously been discussed such as sampling, double keying (see section 1.1.1) or seeding (see section 1.1.2). Further analysis can be achieved focussing on each individual's performance in the system using sophisticated software tools that look for trends or common errors. Large numbers of temporary staff need motivation and to have accountability for their work, as such supervision in all aspects of their work is highly recommended

Other reports relating to checking the coverage to estimate any undercount will also be required. Therefore the registration and identification of data returned from the field and/or other sources will need to be captured and analysed accordingly. Data may arrive at the

processing centre with information detailing the specific geographical location and summarised totals. Either summarised data or individual housing unit data may be used to help these calculations.

3 Country examples

3.1 Mauritius (Manual entry – from paper)[iii]

Mauritius had an estimated population of approximately 1.3 million at the time for the Housing and Population Census (HPC) in 2000. It was conducted on the Islands of Mauritius, Rodrigues and Agalega. The HPC was wholly funded by the central government.

The census was a large data collection exercise to gather information for future planning and policy making and which cost a significant amount of money. To make optimum use of these resources, it was important to investigate the requirements of potential users in order to collect the most relevant information. This was achieved by requesting inputs from the relevant departments. Once the inputs had been received, various meetings were arranged with the stakeholders in order to discuss the issues and arrive at a consensus on the topics to be included in the census questionnaire.

The Housing census data collection exercise was run separately to the population census data collection exercise. The Housing census was initially undertaken followed later by the Population census.

Enumeration forms, instructions and codebooks were the main source of documentation used during the HPC fieldwork. The printing of census forms was completed by the Government Printing Office.

The first round was the Housing Census where Chief Enumerators enumerated all buildings, housing units, households, commercial establishments, hotels, institutions etc. The enumeration books used each contained 25 Housing forms to be completed.

The Housing Census data processing started during the second week of March 2000 and was completed during the first week of May. 30 editors/coders, three supervisors and about 310,000 Housing Census forms were edited and coded in total. The editors/coders were given a half day training course before they started their activities. Throughout the process, training of these staff was an ongoing process. Short briefing sessions were conducted as and when needed.

All editing and coding was completed manually and included:

- Verification that the geographical identifiers on the cover of booklets which made the EA batch were the same.
- Consistency checks of block numbers, building numbers within blocks as well as the housing units within buildings were also performed.
- Consistency checks were also undertaken and some examples are:
    - If “roof” is concrete slab, then “walls” should also be stone, concrete, concrete blocks or bricks.

- If bathing facilities was marked, then “HU3 – Water Supply” should be marked as “piped water inside housing unit.”

- If tenure is “Owner” or “free” then monthly rent should be zero.

For the data capture of the housing census questionnaires, IMPS 4.1 (Integrated Microcomputer Processing System) was used which was provided by the International Programs Center of the US Bureau of the Census.

The names and addresses of the head of each household collected during the housing census exercise were used as a frame for the Population Census enumeration.

The second round was the Population Census which utilized a multi-page form. Enumerators enumerated all persons present on census night as well as usual residents who were absent on census night. The process used to complete the population census questionnaire was that it together with a census guide (an instruction manual on how to complete the questionnaire) was distributed to heads of households by enumerators during the week preceding the census night. The head of household was expected to fill in the questionnaire. The enumerator re-visited the household just after census night in order to collect the completed questionnaire. He/she verified the questionnaire upon receipt and completed any missing information. In case the householder had not completed the questionnaire the enumerator completed it with the relevant information. However, the majority of households completed the questionnaire themselves.

The population census questionnaire was completed with respect to:

- All persons who spent the census night on the premises whether they were members of the household, visitors, guests, boarders or servants.

- All persons who arrived on the premises and joined the household on the day following the census night without having been enumerated elsewhere.

- All temporary absent members of the household, for example, on business trips, in hospital or studying abroad.

The method of data collection used achieved good results given that the literacy rate is quite high among the population. The success of this method was thought to be dependant upon the level of publicity made on the importance of census taking, in order to secure the collaboration of the public.

Detailed instructions were provided to Supervisors, Chief Enumerators and Enumerators. A Census Guide and Instructions was also given to the head of household for completing the Population questionnaire.

Population Census data processing started in August 2000 and was completed in May 2001. 50 editors/coders were involved and about 300,000 population census forms were edited and coded in total. As per the housing census process a half day training for the editors/coders was given with ongoing training being addressed as required throughout the process. Short briefing sessions were also conducted as and when needed.

The editing and coding of the Population Census forms was more complex and time-consuming than that of the Housing Census forms, this included:

- Verification of the EA batches, checking that all census forms in a given batch had their appropriate geographical codes and unaddressed forms had the appropriate code inserted.

- Consistency checking and editing was undertaken on each individual form according to "2000 Population census - Editing & Coding Instructions", which included:

    - Check that there is one and only one head in every household.
    - Check the number of family nuclei.
    - Check for possible inconsistency between name, relationship and sex for every person.
    - Check that age is consistent with date of birth.
    - Check marital status with age.
    - Check if age at first marriage is > 10 years.
    - Check number of children ever born with age and sex.
    - Check education and highest qualification with age.
    - Check the internal consistency of the information on economic activity.

For the coding of both Housing and Population Censuses, international classifications were adapted to the country:

- National Classification of Occupations (Based on ISCO 88);
- National Standard Industrial Classification of Economic Activities (based on ISIC Rev. 3, 1990);
- Codes for Vocational/technical/tertiary qualification (adapted from ISCED);
- Locality codes and country codes.

The data processing of both the housing and population census forms was completed using a similar workflow process, although the various operations involved in the processing were somewhat more complex for the population forms than for the housing forms.

Data capture for the 2000 HPC was outsourced to another government organization as the Central Statistics Office (CSO) did not have its own fully fledged IT department and personnel capable of undertaking data capture activities. Data entry was completed by operators of the Central Information Systems Division (CISD) of the Ministry of Information Technology and Telecommunications. The data entry staff composed of 28 operators and five supervisors. The data entry operators were given a short training session before starting the work and short sessions were conducted as and when needed. All HPC questionnaires were verified using the double entry method.

All coded HPC questionnaires were sent to the CISD in EA Batches for data capture. The data capture for the Population Census started in September 2000 and was completed in July 2001. Data for around 300,000 Population Census forms containing about 1,200,000 records were keyed in during that period. The CISD had recourse to extensive after-office work to complete the data entry exercise within reasonable time limits.

For data validation a programme was run to identify records with errors. Lists of these records were produced with relevant census forms being retrieved and corrections made accordingly.

Once validation of all data files was completed, the data files were concatenated to the country level. Table reports for Housing Census were completed in October 2000 and for Population Census in October 2001.

A post-enumeration survey was not undertaken as past census data which had been evaluated showed that the census data was of good quality. Therefore statistical techniques were used to evaluate as an indirect alternative.

Some issues and lessons learned which occurred during the exercise included:

- Problems of shortage of staff were met at various stages of the census operation, causing delays on the time schedule. (Field staff, editors/coders, data entry operators).

- As regards data capture, this was done by the Central Information System Division of the Ministry of IT and very often the census work was not given priority over other work thus resulting in delays. As both the CSO and CISD formed part of the government service no contracts were put in place which bound the CSO with these departments for the performance of these tasks.

- It is believed that prior to any future outsourcing of any aspect of a census operation to the private sector, the issue of confidentiality needs to be re-assessed in order to maintain trust of users on official statistics.

3.2 Ethiopia (OMR)

The last national population and housing census for Ethiopia was conducted in May 2007 by the Central Statistics Agency (CSA). Ethiopia is the second largest country in Africa in terms of population size.

The previous national census project captured the data manually and took around 18 months to complete for a population of 53 million using approximately 180 staff on 90 PC's, where two shifts were used each day. For the 2007 census there was a need for an alternative solution based on the requirement to have timely census results and limitations such as error rates and management of large staff numbers in the previous exercise.

The data capture system used in 2007 was OMR; also images of every OMR form were collected during scanning to help with their data processing. Study tours were undertaken in Ghana and Tanzania prior to selecting the methodologies to gain the benefit of their experiences. A pilot exercise was completed with both conventional and scanning methods and it was deemed that the scanning approach would be appropriate if properly planned and the necessary cautions could be taken.

Ethiopia had an estimated general population of approximately 77.1 million and 23 million forms were produced to be used in the exercise, being shipped in three separate batches from the UK to Addis Ababa in mid January 2007. The forms were produced in both Aramaic and English and were a combination of A4 and A3 double sided. Using a single dedicated data processing centre, scanning was completed in a total of around 4½ months with 50% of key corrected data being completed by February 2008. All forms had been scanned by the end of April 2008. Preliminary results were disclosed in May 2008.

The data capture system was provided by one supplier as a package that included the provision of scanners, forms, software, support and training.

The types of census forms used were:

- Short questionnaires
- Long questionnaires
- Household Listing Forms
- Summary Forms
- Community Level Forms

- EA Control Forms (Batch Header Forms)
    - EA ID's and number of households were filled in
    - Unique Enumerator number assigned
    - Scanned to create EA Database

The CSA recruited all the necessary temporary staff for the processing centre and made sure they had the appropriate training on both the scanning system and CSPro.

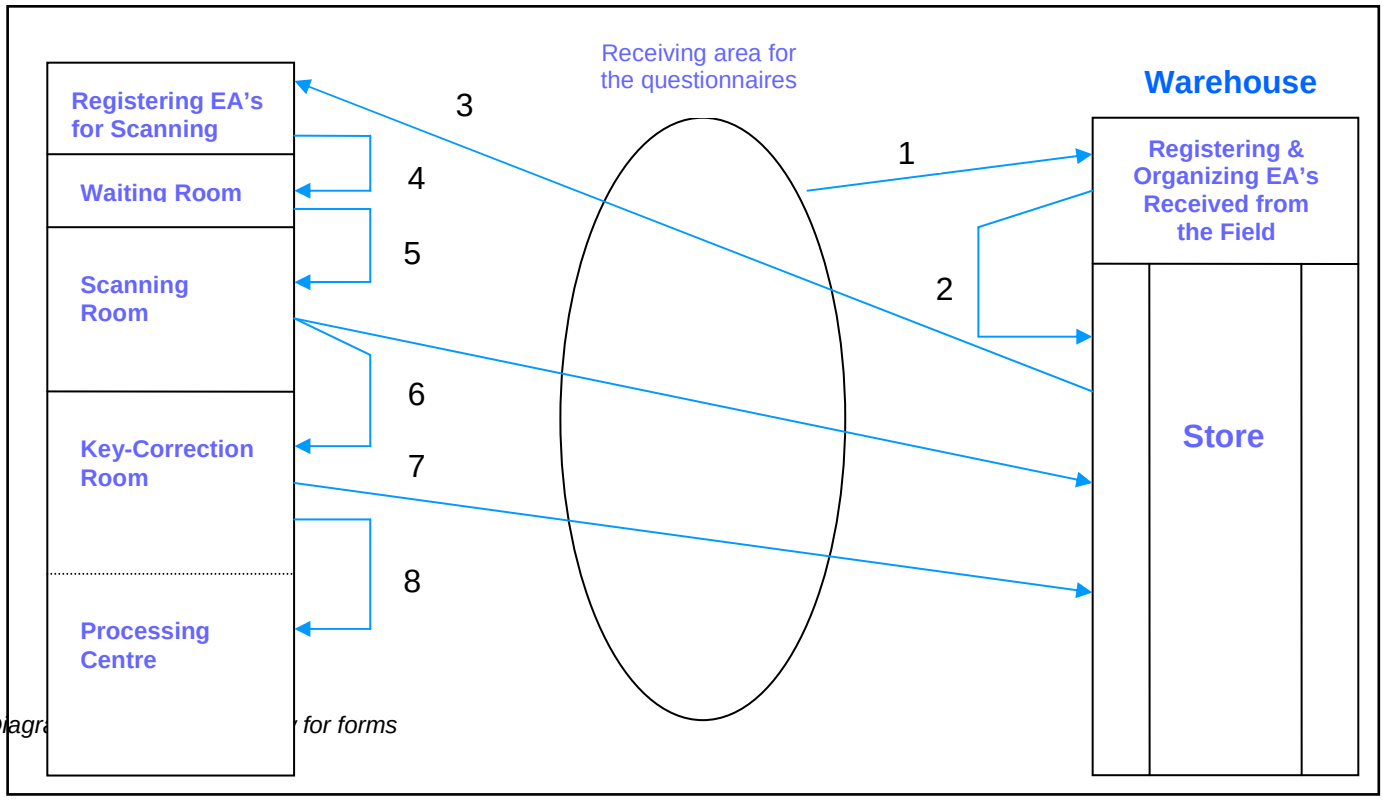
Enumerator training in Ethiopia

There were also significant logistical plans made to the retrieval and re-arranging of the completed questionnaires.

Plans were made and implemented to ensure the smooth running of the data processing centre including:

- The scanners were cleaned on a daily basis;
- An air conditioner for the scanning room was installed;
- A high capacity automatic generator and large capacity Uninterruptible Power Supply (UPS) was installed to ensure power supply in the event of electrical supply problems;
- Regular data back-ups were taken using HP Ultrium 400 Gb data cartridges.

The workflow used in the data processing centre consisted of:



Diagra for forms

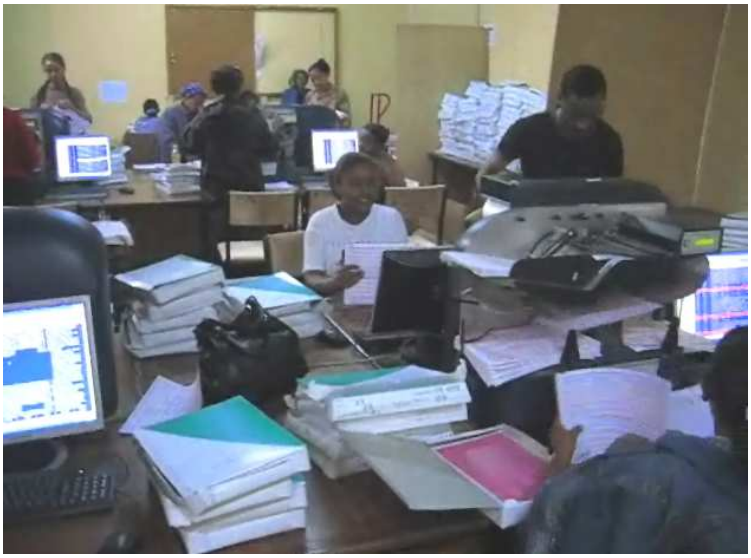
1. Registering and organizing batches received from the field – The forms were received from the field and initially inspected to ensure the contents were as expected (EA forms and a control form). The checking of each batches (EA) ID was undertaken and all forms were prepared and orientated as required. The forms were placed into dedicated boxes to allow for ease of movement around the processing centre. Also, as a quality control check, staff ensured that the EA code on each box matched the one on the questionnaires. About 33 teams for registering and organizing forms were used with three persons assigned per team. Approximately 90,000 batches were received from the field.

2. Storage – The forms were then moved into a dedicated storage area ready for movement into the data processing centre. There was focus on the proper recording of the in-coming and out-going questionnaires from this storage area.

3. Registering batches for scanning - When required the forms were removed from storage and sent to the data processing centre where they were registered onto the computer system. All forms in their respective batches were based on enumeration areas and a batch header barcode was used as part of the system to identify and register the forms. The numbers of forms expected in the batch were also logged at this part of the process. This meant that the system had an indication of how many forms were to be expected during the scanning phase.

4. Waiting Room – The registered forms were stored in a pre-scanning room allowing for a constant feed of forms to the scanner operators.

5. Scanning Room - 11 high speed OMR/Image scanners were used with 44 scanning operators working two shifts per day, seven days a week. Scanning of particular regions was undertaken with the scanning of the 10 sedentary regions from mid August 2007 to mid December 2008 and then the scanning for Affar and Somali Regions from mid January 2008 to mid February 2008. Any forms that were not able to be scanned (as a result of physical problems, tears, parts missing, etc.) had their data transcribed onto blank forms so they could be scanned. The scanners automatically out-sorted specific forms that failed quality control tests during scanning allowing the scanner operator a chance to rectify the problem and re-feed the form.



Scanning room

6. Key Correction Room - The key correction process was based on fixing data for responses that included; missing, multi and partial marks. This was achieved using PC's to display the scanned images captured by the scanner and highlighting individual fields to the operator for an appropriate decision to be made.

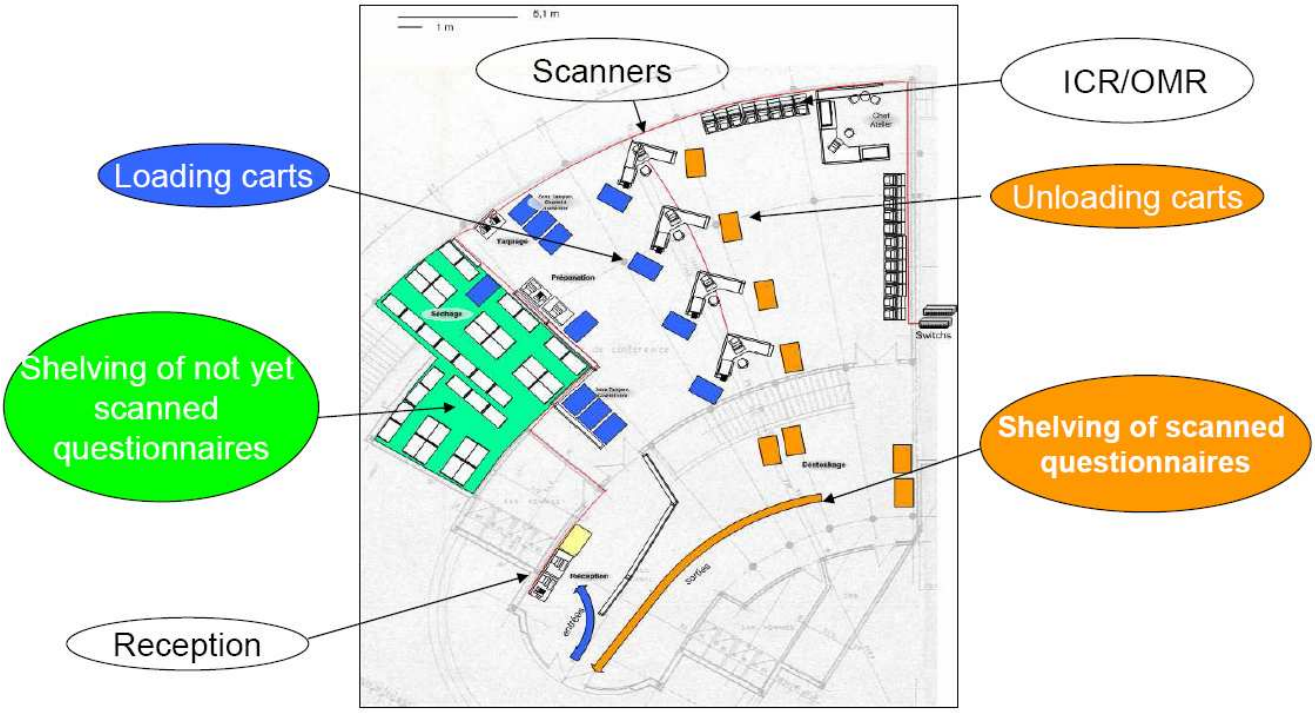
7. Storage – Scanned forms were passed back to the dedicated storage area and recorded as being processed.

8. Processing centre – Corrected and scanned data captured was exported to CSPro for data editing. A batch edit program based on edit specs provided by subject matter specialists was developed and run on the exported data. The batch edit application (.bch) was the component of CSPro which was used to clean the data through editing and imputation processes. Raising factors were attached to the edited long questionnaire data. Tabulation programs (in CSPro) were prepared and tested. Tables in accordance with the tabulation plan were produced with the final data being organized in various formats (ASCII, SPSS) which were sent to the Central Databank for archiving and dissemination purposes.

Various management reports could be created using the software implemented including the ability to see any discrepancies in the number of forms returned for an EA and the number of forms scanned and processed.

3.3 Morocco (ICR)[iv]

For the 2004 Moroccan census an ICR solution was chosen and implemented. Morocco has a general population of approximately 33 million. 41,130 enumerators were used to collect data using six different form types (both booklets and single page forms were used in A3 and A4 sizes). There was the equivalent of 40 million A4 physical forms to process. 18 months was planned for the time to process the forms and extract all of the data. A new dedicated building was erected that was the single data processing centre for the entire country. The data processing centre workflow was designed to keep forms separated at each stage. There were separate storage areas for the forms prior to scanning and after they had been scanned. All forms were scanned in batches that had been registered and all forms were moved around the data processing centre on dedicated carts.



With the building layout decided they implemented a plan that was conducted in three phases:

<u>Testing phase – Three Months</u>

This involved:

- Fine tuning of the ICR recognition engine.
- Identifying organizational issues and quantifying resources.
- Tests were undertaken to compare ICR accuracy against a traditional data keying scenario. This was completed using a sample of 400 districts (about 72,000 questionnaires). Results showed that with a 99.8% confidence level, the automatic documents reading error rate was between 3.56 and 3.57 errors per 10,000. The

data keying method yielded an error rate which was five times higher (19.95 error per 10,000).

Implementation phase – Two Months

The newly built data processing centre was prepared with the installation of:

- Desks
- Shelving
- Carts
- De-humidifiers
- Seating
- Early fire warning system

The number of staff required and computer hardware can be seen in the table below:

Process	**Staff**	**Hardware**
Receiving of questionnaires	3	3 PC
Pre-Scanning Preparation	20	Guillotine + 16 Carts
Scanning	18	5 Scanners (1 Spare)
Image processing and ICR/OMR	4	16 PC's
Key correction and coding	120	60 PC's
Inter-questionnaires control	8	16 PC's
Quality control	24	12 PC's
Logical tests	32	16 PC's
Data export	2	2 PC's
Shared resources (supervisors)	20	5 Servers
Total	**297**	**125 PC's**

A series of specialized workshops were organized for potential employees. 50% of the workforce was temporarily hired through a third party company.

The IT infrastructure was configured so that risks were reduced and would allow for continued operation. The major data processing steps were conducted in four different IT configured clusters. This separation reduced the risks of shutting down all lines of production in the event of hardware failure.

Data Capture phase – 18 Months

The data capture phase was split into three periods:

1. Urban and rural population questionnaires processed during one month.
2. Questionnaires of households and housing (A3 size) processed during six months.
3. Questionnaires of households and housing (A4 size) processed during 12 months.

This separation allowed for a timely dissemination of available results.

A plan was created for processing all of the six form types in the same way. The plan included a number of sequential steps:

1. Receiving of questionnaires - The first step was to receive batches of questionnaires with an electronic file that indicated the identification number of each batch. Each batch of forms contained about 180 questionnaires from a statistical area. The number of received batches as well as any other non-questionnaire content was verified against the delivery sheet. The identification number was entered into the computer system and a sheet was produced with a unique barcode on and placed on top of the batch in a box.

Reception of questionnaires

2. Questionnaires preparation – Questionnaires were initially stored in a controlled temperature area to reduce their humidity content. Questionnaires that composed of many pages were then cut into individual pages for easy scanning.

Controlled temperature storage area

Large carts were used to organize questionnaire transport to and from the scanning area. Each cart could carry 30 batches of forms, each of around 180 questionnaires. Each of the three storage areas was large enough to hold the predicted one-day load of received questionnaires. Batches of questionnaires were placed in wooden boxes for transportation in the carts.

Cart containing batches in wooden boxes

3. Scanning – The scanning was undertaken using five Kodak (Model 3520) Scanners that handled 52 A3 questionnaires per minute. The scanner operators could verify in real time the quality of scanned images as they were presented on the screen at the time of scanning. If the quality of the image deteriorated then the scanners were cleaned and forms were rescanned.

   Prior to scanning the scanner operator could make use of the vibrating paper jogger used to align the forms and remove dust, etc.

   Batches of forms were identified by the barcode sheet produced by the questionnaire receiving process.

   The specification of the scanners used was:

   - 40 to 85 pages per minute depending on resolution, feeding orientation and document size.
   - Resolution 200 or 300dpi.
   - Document input: min: check, max: A3.
   - Feeding capacity: 250

4. Image processing and ICR/OMR – Automatic processing of images was undertaken as a background task (four images were produced from each A3 questionnaire). Recognition of image was achieved using corner stones printed on the form to identify where the response area data needed to be captured. There were instances where the recognition was rejected, in which case, computer operators identified the images corner stones and submitted images for ICR/OMR recognition. If an image could not be fixed, the questionnaire was rescanned.

   During scanning the contrast was automatically adjusted to remove light background colour from the scanned image. Each form was designed without any black borders around the response boxes so they would not interfere during the ICR/OMR recognition process.

   The ICR/OMR recognition engine used was the A2iA FieldReader that combines OMR and ICR to capture written data within structured forms. The engine

supported image formats of tiff G4, bmp, Jpeg or Jpeg 2000 with a minimum resolution of 200 DPI and produced output data with its associated confidence score.

Two thresholds were used for the ICR engine confidence levels:

- 95% for responses not associated with any further logical errors tests.
- 85% for responses associated to further logical errors tests.

By increasing all confidence levels to 95% for all response areas the key correction part of the process saw considerably more load and consequently cost more in time to process. Computer operators validated/key corrected ICR suggestions with scores lower than the levels set.

5. Key correction and coding – As indicated only those responses that fell below the ICR engines confidence levels were key corrected by operators.

Key correction and coding area

Coding of open question answers was undertaken of responses that were written in Arabic. These related specifically to the questions on profession, economic activity, diploma and migration. A code was entered using questionnaire images being shown on the screen and the use of integrated dictionaries. The retrieval of record variables was used to improve the quality of coding. The operators used search dictionaries (activity, diploma, etc.) using keywords to validate responses.

6. Inter-questionnaires control and correction - This process was undertaken for each batch to verify that all questionnaires within a statistical area had been processed.

7. Quality control – The quality control system included sampling. The quality control method used was implemented to produce data with a minimum accepted error rate. This step came directly after the ICR recognition process and key correction/coding.

The Quality control process applied was:

Afnor norm NFX06-022 of October 1991.
(which is in accordance with international norm ISO 2859-1-1989)

The acceptable quality level for response errors was 0.52%. Operators verified that the values in the data file were identical to values in the image. In order to make an easy comparison, the presence of OMR marks was converted to a numeric one. All questionnaires were kept in the post scanning storage area in the processing centre until they passed the quality control step.

8. Logical tests – Tests were conducted to identify logical errors. These errors were corrected by an operator being shown the image and being able to correct the data. A logical error example used was: Although this house is declared as empty, the type of ownership is declared as owner. The operator would correct this inconsistency within the guidelines given.

9. Data export - Data was exported in a text file format with a dictionary for further processing using CSPro/IMPS. This was the last step in the data processing system. The results were also exported in text files and their corresponding images of questionnaires to DVDs for backup and storage.

Other tools were also used to monitor the available free space on the hard disks within the system, to check that they were not becoming full of images of questionnaires. A utility was used to mange the hard disk space on this basis. Also a software tool was used that gave statistics on each part of the process in an aggregated format along with statistics on each operator's performance. Daily reports were run on the number of questionnaires processed each day by form type.

To help keep the staff motivated and improve both quality and production, an employee of the month system was introduced (in each area) and each employee of the month was issued a certificate and incentives worth up to an extra 20% of their salary. A monthly newsletter was published to keep employees informed of work progress in its different steps.

3.4 Brazil (PDA)[v]

The public institution that was responsible for the Census data capture project in Brazil is the Brazilian Institute of Geography and Statistics (IBGE). In 2007 they undertook a combined census exercise capturing data for the agriculture and population census along with national address list data. It was calculated that a cost saving of around 40% would be made if the projects were combined, and this helped meet the budget constraints.

Due to the limited financial resources, only municipalities that had a population of less than 170 thousand would have their data captured in this project. The total number of municipalities in Brazil was 5,564 and it was deemed that those municipalities to be included in the project numbered 5,435 (approximately 97% of the total municipalities), which was estimated at 60% of the total population of Brazil. About 28 million households (approximately 57% of all households in Brazil) would be covered in the data capture exercise.

The resources used in the project included:

- Around 90,000 staff were used to collect, supervise, support and administer in the project
- 82,000 Personal Digital Assistants (PDA) equipped with GPS receivers
- 3,500 networked computers
- 500 ADSL network points – Connected to IBGE
- 700 Satellite points – Connected to IBGE
- 4624 modem dial-up points – Connected to IBGE
- 3,000,000 pages of training material
- 1,200 computerised data collection stations
- Over 5,000 data collection stations

For the population count the data collection was planned to be achieved in the following number of days:

- Urban Areas 30 Days
- Non-Urban Areas 45 Days
- Rural 45 Days

For the agriculture census 60 days were planned for data collection in the rural areas.

There were a number of reasons why IBGE decided to use PDA technology for the data collection. They had previously undertaken smaller survey projects and so had some experience already but the main drivers for using the technology included:

- immediate evaluation at the moment of data collection, allowing the correction of information at the moment of the interview

- the filling out of all the compulsory questions, avoiding the lack of answers due to forgetfulness or mistake by the enumerator

- optimization of the filling out of data through automatic skips in the questionnaire, avoiding covering several items about which, sometimes, there would be no reply; which could optimize time used by the enumerator and the head of household

- The non-necessity of the transportation of a big amount of paper questionnaires and of the handling of these same questionnaires in data capture centres, achieving information precision and better processing time.

Besides all these advantages, handheld computers, equipped with a receiver for GPS signals allowed the geo-referencing of all the units visited in the rural areas, as well as the monitoring of the geographic coverage on a remote basis visually by overlaying digitally the GPS track on orbital images. This was beneficially for quality checking purposes by supervisors and managers to detect collection coverage and detect possible mistakes. Maps and digital orbital photographs were also loaded onto the PDA to assist the enumerator to reach the area for data collection and to assist them when moving around the area.

A combination of PDA and physically printed maps were provided to the enumerators to complete the data collection tasks in the field.

When the publicity campaign began for the census project, it was observed that the PDA had a very strong visual appeal; it soon became a symbol to identify the IBGE enumerators.

PDA used in Brazilian census project 2007

Due to GPS technology being integrated into the PDA device the tracks of the enumerators could be viewed by staff at the IBGE headquarters and managed accordingly. Connection to the IBGE network infrastructure and PDA device needed to be established for this to happen. If no 'real-time' connection was made between the PDA and the IBGE network, the GPS tracks made were stored locally on the PDA and could be seen once connection was established. This gave IBGE a very effective tool to view the activity of their data collection staff as they undertook field data collection exercises.

The Field work started on 16th April 2007 with over 70,000 starting to visit locations to collect data. The exercise was due to end on 31st July but due to some problems relating to communication infrastructure (slow speeds, satellite connections difficulties, etc.) delays occurred. As a consequence most of the data collection took place between May and September of 2007.

IBGE recruited enumeration staff using a number of tests. The first test assessed the candidates general knowledge, Portuguese language skills and mathematics. The second test included a training programme and how to use a PDA, with the eventual test being undertaken using the PDA itself. 18,000 supervisors were also recruited at the same time and they were tested on Portuguese, logic, computer science, general knowledge, administration and management routines. There was a specific effort to try and recruit enumerator and supervisors from the local communities to reduce or prevent evasion.

Training underway with PDA

Training of the large number of enumerators was undertaken using a combination of materials including, Videos, Intranet/Internet, Manuals and other printed materials. A 'chain' training method was used to train all staff involved in the process. 41 specialists trained 332 trainers, who in turn then trained 1,549 instructors who then trained 18,000 supervisors who then finally trained 68,000 enumerators. The training and distance learning materials along with the PDA were all used to support the chain training method to ensure continuity throughout the process.

Distance learning software

During use the PDA held not only the information collected but also the unique identification of the enumerator using it. As a consequence this automatically created an audit trail. Supervisors used a web based system to update the central system of progress and data was uploaded periodically.

The 1,200 computerised collection stations that were located around Brazil were used as a data transmission point for enumerators to send their collected data to the central system. This was achieved using Bluetooth technology to transfer data from the PDA to the local computer which then sent the data to the central system via satellite link or ADSL. Transmission of the data from the local computers to the central system was completed via internet connection and the data was encrypted with the use of a Virtual Private Network (VPN). For backup purposes or in the event of communication issues the data could be written to CDROM and physically sent back to the central office in Rio de Janeiro.

The remaining 4,400 data collection stations used Modems to transmit data from the PDA directly to a host computer in the central office in Rio de Janeiro. A quantity of fixed

telephone numbers were purchased from the national phone company to accommodate this process. The internet was not used for data transfer for these collection stations.

The final results for the population data were released on 21st December 2007.

The use of the PDA's during enumeration raised the profile of digital communication with the general population especially with those who were seeing this type of technology for the first time.

There were a number of lessons learned by IBGE as part of the exercise, the key ones being:

- The necessity of strong investment in the planning stage, during which the equipment to be used is defined, the forms of communication and data transmission, and the software solution to be implanted, are basic factors for the success of the operation.

- The necessity of carrying on dress rehearsal censuses in order to validate all the stages and anticipate alternative solutions. In the case of the 2007 censuses, this work was not done as a consequence of the established turnaround period.

- The necessity to reinforce the training for the use of PDA's and associated systems, which should be including all the involved persons in the operation.

- The urgent necessity of increasing the number of IBGE workers in order to prepare the Institution for the 2010 Demographic Census and the continuous censuses in the coming decade.

The innovations implemented during this exercise were a landmark for the IBGE and it is perceived that benefits will be seen in future exercises including:

- The experience obtained from the large scale process of distance learning of the IBGE personnel and the attendance of a course in which the PDA and other multimedia resources were used.

- The obtaining of co-ordinates of rural area establishments, which will allow a better quality of field operations in the future, including sample surveys.

- The increase of the analytical possibilities of information, provided by geo-referencing and by the combination of information of geographic base associated to address, as well as the unlimited possibilities of the graphic base in construction.

- The rewarded audacity in the digital inclusion of approximately 100,000 Brazilians in a work of this magnitude.

3.5 Canada (Internet)[vi]

For the 2006 census statistics Canada adopted a multi-channel approach of data capture using a combination of paper based questionnaires being mailed out directly to households and hand delivered, along with Computer Assisted Telephoning Interviewing (CATI) and an online questionnaire via the Internet. All methods were complimentary and the system implemented was developed to accept data from the three different methods

for processing. Although three methods were undertaken this section will summarise the internet component.

There was a general expectation from the public that the 2006 census could be taken online due to other government initiatives being made available online and promoted. Consequently the main objective for this part of the data capture system was to provide a secure online based application so that everyone who lived in Canada could complete their 2006 census questionnaire via the internet. The intention was for all private households and agricultural operations to be able to have the option to complete either a long (53 questions) or short (eight questions) questionnaire online. Questionnaires were to be made available in Canada's official languages (English or French).

A previously undertaken study and survey information relating to internet access in Canada to assess coverage that was collected three years prior to the census date was looked at and it was estimated that 54% of all households had internet access that was used regularly. A significant proportion of these used the internet to access online banking services, showing that there was general confidence by the public in security aspects of the internet.

The system was deemed a secure infrastructure. This would be essential to gain the user's trust in the system to provide confidential information. In fact they wanted to show that it was more secure than other online transactions. Different methods of security considered were the ability for the user to download the questionnaire and then allow them to upload it with their response being encrypted. Also SSL128 was considered (typically used by the banking industry). Both of these methods did not meet the objectives for security for the census. Eventually with significant investment (around 50% of the budget) a 'secure channel' approach was used with bi-directional encryption, limited-use digital certificates and isolated network infrastructure. Private sector expertise was used to help in the development of an application and infrastructure after a competence profile was completed internally to determine which skills they had in-house and where they would need to bring in external expertise.

During May 2004 and approximately two years ahead of the main census date, a sizeable dress rehearsal was undertaken. A satisfaction survey was undertaken with a sample from the users to gain feedback and information regarding the system, so adjustments could be made as required. Here are some of the results of that exercise:

- the majority of the respondents (89%) indicated that they completed their census test questionnaire at home;

- 79% of the respondents indicated that they had a high speed. connection;

- when asked why they chose to complete their census test questionnaire online, 52% said because it was easier, 30% said because it was faster, 18% said because of personal preference and 16% said because they did not have to mail it;

- 95% of the respondents rated their overall experience with the census internet application as favourable;

- 88% of respondents felt that the time it took to complete the census test questionnaire was acceptable, of which more were short-form respondents. Respondents who indicated that they did not have high speed were more likely to say that it took too long;

- 98% of respondents indicated they would complete their census questionnaire online in 2006;

- 57% of respondents were not at all concerned that the privacy and the confidentiality of their census test questionnaire data were more at risk on the Internet, while 34% were concerned and 8% were very concerned;

- when asked how secure it is to transmit personal information over the internet, 73% stated that it was secure,

In the 2004 census test, it was observed that the incidence of item non-response was many times lower for internet submissions than for paper. Specifically, for short census of population forms the item non-response rate was 0.01% for internet responses and 2.54% for paper responses.

With the decision to move ahead with the object for the 2006 census, funding was required. Funding for the system was based on investing for the future to reduce costs in the long term. It was anticipated that the response rate via the internet would be 16% so the budget was set at this point as a breakeven; this was purely based on cost savings against other methods that could have been implemented.

The infrastructure was sized to accommodate 15,000 concurrent user sessions with availability 24 hours, seven days a week (for a nine week period). The system faced enormous demand on census night, especially between 6pm and 11pm, and invoked 'graceful deferral' (an automated system that restricted access to only 15,000 concurrent users – asking further users to try later) repeatedly during this peak period. It is estimated that approximately 150,000 users were asked to re-try the application at another time.

Paper based census questionnaires where mailed out to 73% of all households with the remaining 27% hand delivered. All questionnaires had a unique number and an internet access code printed on them to allow self-enumeration to be undertaken via the internet system.

If the questionnaire was self-enumerated using the internet system it left no trace on the host computer. Users who undertook a long form internet questionnaire could save partially completed forms using their own password and unique code for retrieval and further completion if required. Users could also toggle between English and French language versions if required.

One of the major benefits seen was in the response rate/item due to automatic prompts for item non-responses and also the feature of automatically skipping to the next question when responses rendered further questions irrelevant. Radio buttons on the online forms meant no possibility of conflicting responses as the user can only select one at a time and dropdown menus assisted in the appropriate responses being selected by the user.

Responses from paper, internet and CATI were all coordinated to allow for non-response follow-up.

Statistics Canada is looking at the possibility of not delivering paper based questionnaires in some areas for the next census and just send access codes to reduce paper questionnaire production costs. Users will still be given the option to request a paper based version if required.

The overall target for responses via the internet system was set at 20% of all households prior to non-response follow-up. Actual response rate was approximately 22% prior to non-response follow-up. After all non-response follow-ups were completed approximately 18.5% of all responses received were via the internet.

4 Summary and conclusions

It is the NSO's objective to produce census data to a required quality standard, in a timely manner within the limitations of the financial resources made available. The users of any resultant data will likely have expectations of data availability and quality. If the NSO can make any performance or quality improvements in these areas by choosing an appropriate method/technology then the overriding objective will have been deemed to have been successfully met/exceeded.

It is evident there is not just one preferable set of data capture technology for any national census exercise. The challenges faced by a National Statistics Office when planning which data capture method to implement in their national census project varies depending upon a number of factors and external influences including, but not limited to:

- Budget and current funding situation;
- Project timeframe;
- Political requirements;
- Availability of local technical skill set;
- Previous data capture method employed;
- Regional trends;
- Expectations of output data quality.

NSO's should look at a number of factors when choosing data capture method/s to employ as it may not be appropriate just to use the most 'cutting' edge technology. Proven technology also has a big part to play. Commercial suppliers have had the chance to optimise and adapt older technology sometimes making it much less time consuming, onerous and no less accurate than the more sophisticated systems on offer. Anything that previously worked well should not be disregarded without investigation into the impact any alternative may have.

The data capture methods discussed are not new and the NSO is not alone in the decision making and planning processes. Valuable knowledge can be gained by visiting regional or similar countries that have recently implemented methods and technology that is being considered. Information on how and what was achieved elsewhere will guide the decision making process and the various stakeholders in the planning process needed to allow both time and resources for these activities to be effectively completed. Collating any past experience and knowledge of others in country census experience should help indicate the methods and processes that are preferred for the next census. An online knowledge base[vii] is available on the UNSD website where useful documents on methodology for census and other documents relating to countries experiences can be viewed.

Outsourcing has its own set of risks so careful planning (including construction of contracts to minimise misinterpretations and misunderstandings), negotiation, selection, security, quality checking and regular reporting systems are required to ensure system integrity and that performance criteria are firstly established and then met. Nevertheless, it may provide to the NSO the opportunity to be accompanied by an industrial partner and avoid the pitfalls of implementing alone a complex technology that is not fully mastered.

Any chosen data capture method will be directly impacted by the quality, quantity and timing of training given to enumerators. If data collected in the field is inaccurate or incomplete no data capture method will be able to correct this. It is therefore critical that the appropriate resources, funding and time is given by census planners to this part of the overall census plan.

[i] Description of recommended process in the UNSD report *Principles & Recommendations for Vital Statistics Systems - Control of receipt of statistical reports (Series M, No.19/Rev 2, 2001)*

[ii] Provided by the US Bureau of Statistic *CSPro (Census and Survey Processing System) is a public-domain software package for entering, editing, tabulating and mapping census and survey data*

[iii] Based on content of *'Country position paper'* presented at the 2006 ASSD and *'Country presentation paper'* from United Nations Statistics Division (UNSD) workshop on census data processing in conjunction with the National Bureau of statistics (NBS),Dar es Salaam , Tanzania, 9 – 13 June 2008

[iv] Summary of Moroccan census is provided with the content being summarised from the presentation '*Data capture processes of large scale survey questionnaires: Case study of Census of Population and Housing 2004 of Morocco'* delivered at the United Nations Statistics Division regional workshop in Qatar 18th May 2008 by Mr Oussama Marseli - Manager of the Center of Automatic Reading of Documents (Morocco).

[v] Based on information extracted from *"Q10341 - COUNTRY: 2007 Censuses - Innovations and impacts on the Brazilian statistical and geographic information systems"* – UNSD knowledge base

[vi] Based on the paper entitled *'CENSUS TECHNOLOGY: RECENT DEVELOPMENTS AND IMPLICATIONS ON CENSUS METHODOLOGY – Census on the net'* Presented for the ECONOMIC COMMISSION FOR EUROPE, CONFERENCE OF EUROPEAN STATISTICIANS, Group of Experts on Population and Housing Censuses, Tenth session, Astana, 4-6 June 2007

[vii] The UNSD Knowledge base can be accessed using the following website address: http://unstats.un.org/unsd/censuskb/