Sub-regional Workshop on Census Management and Data Processing – Doha, Qatar 18-22 May 2008

Data capture processes of large scale survey questionnaires: Case study of Census of Population and Housing 2004 of Morocco

Oussama Marseli
Manager of the Center of Automatic Reading of Documents
Plan

- **Data processing steps**
  1. Reception of questionnaires
  2. Questionnaires preparation
  3. Scanning
  4. Image processing and OCR
  5. Normal video coding
  6. Inter-questionnaires control and correction
  7. Quality control
  8. Logical tests video coding
  9. Data export

- **Processing of Census of Population and Housing 2004 of Morocco**
  1. Testing
  2. Implementation
  3. Production
1- Reception of questionnaires

- The first step is to receive lots of questionnaires with an electronic file indicating the identification number of each lot.

- Each lot contains questionnaires of a statistical area, which is about 180 questionnaires.

- The number of received lots as well as their content is verified against the delivery sheet.

- Identification number is keyed in and an identification code bar is generated for each lot.
1- Reception of questionnaires

Computer operators register received boxes (lots) of questionnaires and print an identification code bar on a separate sheet and put it inside the box.
2- Questionnaires preparation

- Questionnaires composed of many pages are cut into individual pages (for easy scanning).

- Questionnaires of A3 or A4 paper size can be scanned, without further processing.
Drying storage zone

Questionnaires are stored in controlled temperature area to reduce their humidity content.
Questionnaires handling

3 storage zones: Each zone is large enough to hold one-day load of received questionnaires.

Lots of questionnaires are placed in wooden boxes.
Carts carry questionnaires

Large carts are used to organize questionnaire transport to scanning area.

Each cart carry 30 lots, each of about 180 questionnaires.
3- Scanning

- Lots of questionnaires are identified by their code bars.

- Questionnaires are scanned with industry type scanners, Kodak ds Digital Science Scanner 3520:
  - 40 to 85 page per minutes depending on resolution, feeding orientation and documents size.
  - Resolution 200 or 300dpi.
  - Documents input: min: check, max: A3.
  - Feeding capacity: 250
Scanners configuration GUI
Scanners operators

Scanner operators identify lots of questionnaires by checking their code bar.

Vibrating paper jogger:
Align edges of A3 questionnaires
Scanner

Kodak Ds Digital Science Scanner 3520 handles 52 questionnaires A3 par minute. Scanner operators verify on real time quality of scanned images. If quality deteriorate, scanners are cleaned and scanning is redone.
Scanning GUI allows scanner operators to verify image scanning quality.
4 - Image processing

- Automatic processing of the images (4 images are produced from each A3 questionnaire)

- Recognition of image delimitations

- Localization of cells

- Optical Character Recognition.

- Some images are rejected, in which case, computer operators identify images delimitations and submit images for OCR/OMR. If an image can't be fixed, the questionnaire is rescanned.
Questionnaire of households and housing A3

<table>
<thead>
<tr>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
</tbody>
</table>

Questionnaire type number

Image delimitations marks
Questionnaire of Households and housing

Each cell has unique coordinates with regards to questionnaire delimitation marks.
Automatic image processing

Application recognize cells locations with regards to image delimitations.

If application fails, an operator execute this task through GUI.

If image can’t be fixed (e.g. only part of questionnaire is scanned), questionnaire is rescanned.
OCR Engine

- A2iA FieldReader combines OCR, ICR, IWR to capture written/printed data within structured forms.

- Input: Supported image formats: tiff G4, bmp, Jpeg or Jpeg 2000 with a minimum resolution of 200 DPI

- Output: Data associated to a confidence score
No cell borders are allowed (to avoid recognition noise).

Light colors help enumerators write responses within white boxes.

Scanning contrast is adjusted so light colors don’t show on images.
Scanned images are black and white.

Areas to be recognized are completely white except for the black handwriting.
Optical character recognition

OCR engine recognizes characters with an associated score. Then, apply logical error tests.

Cells that are recognized with low score or trigger logical errors are presented to computer operators.
5. Normal video coding

- Computer operators, validate/correct OCR suggestions with lower score.

- Two thresholds are used
  - 95% for cells not associated to logical errors tests
  - 85% for cells associated to logical errors tests.

- Requiring a 95% confidence rate on all cells increases considerably video coding load and cost subsequently.
Normal video coding GUI

OCR suggests 8 where as true value is 2.

Image shows 2

OCR Reads 8
Normal video coding GUI
Normal video coding

- Coding open question answers written in Arabic.

- Questions on profession, economic activity, diploma, migration.

- Code entered using questionnaire images and integrated dictionaries.

- Retrieval of record variables to improve quality of coding.
Normal video coding GUI
Normal video coding GUI

Operator search dictionaries (activity, diploma..) using key words, and validate answer.
Video coding lab
6. Inter-questionnaires control and correction

- are done within each lot to verify that all questionnaires within a statistical area have been processed.
7. Quality control

- Quality control is about producing data with a minimum accepted error rate.

- This step comes right after recognition engine and normal video coding correction by operators.

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

- The acceptable quality level is 0.52% errors for filled cells.
According to Afnor norm (general level control II) for lots of 10001 up to 35000 observations, 315 observations are to be sampled on normal and reinforced modes.

### Table 1 — Lettre-code en fonction de l’effectif des lots et du niveau de contrôle

<table>
<thead>
<tr>
<th>Effectif des lots</th>
<th>Niveaux de contrôle spéciaux</th>
<th>Niveaux de contrôle pour usages généraux</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S-1</td>
<td>S-2</td>
</tr>
<tr>
<td>2 à 8</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>9 à 15</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>16 à 25</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>25 à 50</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>51 à 90</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>91 à 150</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>151 à 280</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>281 à 500</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>501 à 1200</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>1201 à 3200</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>3201 à 10000</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>10 001 à 35 000</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>35 001 à 150 000</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>150 001 à 500 000</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>500 001 et au-dessus</td>
<td>D</td>
<td>E</td>
</tr>
</tbody>
</table>

### Correspondance entre lettre-code et effectif d’échantillon du plan simple, contrôle normal et renforcé

<table>
<thead>
<tr>
<th>Lettre-code</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>P</th>
<th>Q</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effectif d’échantillon n du plan simple (1)</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>20</td>
<td>32</td>
<td>50</td>
<td>80</td>
<td>125</td>
<td>200</td>
<td>315</td>
<td>500</td>
<td>800</td>
<td>1250</td>
<td>2000</td>
</tr>
</tbody>
</table>

18/05/2008     www.hcp.ma
Quality control using Afnor norm

- **Normal mode:**
  - Controls 315 observations/lot
  - Rejects lot if 5 errors are found

- **Reinforced mode:**
  - Controls 315 observations/lot
  - Rejects lot if 3 errors are found

- **Reduced mode:**
  - Controls 125 observations/lot
  - Rejects lot if 4 errors are found

- **Production is halted.** Problem investigated.

- If 5 lots in a row are rejected:
  - If 2 out of 5 consecutive lots are rejected
  - If 10 lots in a row are accepted

- If one lot is rejected:
  - If 10 lots in a row are accepted
  - If one lot is rejected

18/05/2008 www.hcp.ma
Quality control GUI

Operator verifies that values in data file is identical to value in image. In order to easy comparison, mark presence is converted to one.
Storage area

Questionnaires are kept in processing center until they reach quality control step.
8- Logical errors video coding

Allows to computer operators, with special training, to correct errors triggered by the test.

Logical test script GUI. Programmers write logical tests using application language.
Logical error example: Although this house is declared as empty, type of ownership is declared as owner. Computer operator corrects inconsistency.
9- Data export

- The last step on OCR/OMR data processing results in exporting data on plain text files and corresponding images of questionnaires on DVDs.

- Steps to follow involve control of exhaustiveness and imputations techniques to produce tabulation-ready data file.
Data is exported in plain text file with dictionary, for further processing using CSPro/IMPS.

<table>
<thead>
<tr>
<th>Record type</th>
<th>Data</th>
<th>Year</th>
<th>Code</th>
<th>Value1</th>
<th>Value2</th>
<th>Value3</th>
<th>Value4</th>
</tr>
</thead>
<tbody>
<tr>
<td>02321010110013</td>
<td>1230101111101101</td>
<td>1968</td>
<td>2</td>
<td>21411</td>
<td>06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02321010110013</td>
<td>123010111210112</td>
<td>1969</td>
<td>201000000000</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02321010110013</td>
<td>123010111310122</td>
<td>1995</td>
<td>1</td>
<td>1111</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02321010110013</td>
<td>123010111213301313331</td>
<td></td>
<td>1</td>
<td>1</td>
<td>12200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02321010110013</td>
<td>12401011140202000000</td>
<td></td>
<td>271</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02321010110013</td>
<td>1240101111111010</td>
<td></td>
<td>321113</td>
<td>06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02321010110013</td>
<td>124010111210191</td>
<td></td>
<td>322113</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02321010110013</td>
<td>124010111213303113331</td>
<td></td>
<td>1</td>
<td>1</td>
<td>11202</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02321010110013</td>
<td>12501011140101000000</td>
<td></td>
<td>3221</td>
<td>06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02321010110013</td>
<td>1250101111120101</td>
<td></td>
<td>1973</td>
<td>1</td>
<td>000</td>
<td>1</td>
<td>06</td>
</tr>
<tr>
<td>02321010110013</td>
<td>1250101112163301333332</td>
<td></td>
<td>1</td>
<td>1</td>
<td>22201</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02321010110013</td>
<td>12601011140505000000</td>
<td></td>
<td>222111</td>
<td>06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02321010110013</td>
<td>1260101111111010</td>
<td></td>
<td>1955</td>
<td>2</td>
<td>22111</td>
<td>06</td>
<td></td>
</tr>
</tbody>
</table>
Disk space monitoring

At this steps, images are deleted from servers, in order to free space for upcoming images.
Statistics GUI

This module offers detailed statistics on production on different steps aggregated, as well as by operator.
Processing Census of Population and Housing 2004 of Morocco

- **Objective:** Capture data from all questionnaires within a short period of time.

- **Strategic choices:**
  - Data capture using Manuel entry (used in prior census on samples)
  - Optical character recognition (new technology used in developed countries).

  - Partnership with private sector to develop specific solution using OCR.
3 Phases

- Testing phase – 3 months
- Implementation phase – 2 months
- Production phase – 18 months
Testing phase

- Allowed for fine tuning recognition engine, identifying organizational issues, and quantifying resources.

- A secondary objective was to compare OCR accuracy to traditional data keying scenario
Implementation phase

- Center of Automatic Reading of Documents was created

- State-of-the-art software and hardware (110 computers, 5 scanners and 5 servers) were installed in a newly managed area (desktops, shelving, early fire warning system).

- Specialized workshops were organized for potential employees (240 persons: 50% of the work force was temporally hired through a third party).
Centre of Automatic Reading of Documents
Facilities: Scanning area

- Scanners
- OCR/OMR
- Loading carts
- Unloading carts
- Shelving of not yet scanned questionnaires
- Shelving of scanned questionnaires
- Reception
Facilities: video coding computer labs

52 postes de Vidéocodage N
Logical architecture of hardware installation into 4 clusters and one central server

Major data processing steps are conducted in 4 different clusters.

This separation reduces the risks of shutting down all lines of production.
# Human, hardware and software resources used to capture data from census 2004 questionnaires

<table>
<thead>
<tr>
<th>OCR/OMR data processing steps</th>
<th>Human R.</th>
<th>Hard &amp; Soft</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Reception of questionnaires</td>
<td>3</td>
<td>3 PCs</td>
</tr>
<tr>
<td>2. Questionnaires handling</td>
<td>20</td>
<td>Cutter, 16 carts</td>
</tr>
<tr>
<td>3. Scanning</td>
<td>18</td>
<td>5 scanners (1 spare)</td>
</tr>
<tr>
<td>4. Image processing and OCR</td>
<td>4</td>
<td>16 PCs, 12 OCR dongles</td>
</tr>
<tr>
<td>5. Normal video coding</td>
<td>120</td>
<td>60 PCs</td>
</tr>
<tr>
<td>6. Inter-questionnaires control</td>
<td>8</td>
<td>16 PCs</td>
</tr>
<tr>
<td>7. Quality control</td>
<td>24</td>
<td>12 PCs</td>
</tr>
<tr>
<td>8. Logical tests video coding</td>
<td>32</td>
<td>16 PCs</td>
</tr>
<tr>
<td>9. Data export</td>
<td>2</td>
<td>2 PCs</td>
</tr>
<tr>
<td>Shared resources (supervisors)</td>
<td>20</td>
<td>5 servers</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>297</strong></td>
<td><strong>125 PCs</strong></td>
</tr>
</tbody>
</table>

18/05/2008 www.hcp.ma
Production phase

- 3 periods:
  1- Urban and rural population questionnaires processed during 1 month.
  2- Questionnaires of households and housing A3 (only Arabic numbers) processed during 6 months.
  3- Questionnaires of households and housing A4 (Arabic numbers and letters) processed during 12 months.

- This separation, allowed for a timely diffusion of available results.
# Census of Population and Housing 2004 questionnaires types, volumes and allocated time to capture data

<table>
<thead>
<tr>
<th>Questionnaire</th>
<th>Volumes: # of Questionnaires</th>
<th>Type</th>
<th>Fields /quest.</th>
<th>Filed type</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban and rural population</td>
<td>38 000</td>
<td>A4 (21 pages two sides)</td>
<td>3051</td>
<td>Arabic numbers</td>
<td>1 month</td>
</tr>
<tr>
<td>Households and Housing</td>
<td>6 800 000</td>
<td>A3 two sides</td>
<td>248</td>
<td>Arabic numbers</td>
<td>6 months</td>
</tr>
<tr>
<td></td>
<td>5 800 000</td>
<td>A4 two sides</td>
<td>12 54</td>
<td>Arabic numbers</td>
<td>12 months</td>
</tr>
<tr>
<td>Separately counted Population</td>
<td>12 500</td>
<td>A3 two sides</td>
<td>260</td>
<td>Arabic numbers</td>
<td>1 day</td>
</tr>
<tr>
<td>Nomad population</td>
<td>40 000</td>
<td>A3 two sides</td>
<td>245</td>
<td>Arabic numbers</td>
<td>1 day</td>
</tr>
<tr>
<td></td>
<td>40 000</td>
<td>A4 two sides</td>
<td>12 54</td>
<td>Arabic numbers</td>
<td>2 day</td>
</tr>
<tr>
<td>All</td>
<td>39 888 000</td>
<td>pages A4</td>
<td></td>
<td></td>
<td>t0+ 18 months</td>
</tr>
</tbody>
</table>
Monthly processed questionnaires of Households and housing A3 by production steps

<table>
<thead>
<tr>
<th></th>
<th>Dec05</th>
<th>Jan05</th>
<th>Feb05</th>
<th>Mar05</th>
<th>Apr05</th>
<th>May05</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Working days</td>
<td>23</td>
<td>20</td>
<td>20</td>
<td>23</td>
<td>19</td>
<td>10</td>
<td>115</td>
</tr>
<tr>
<td>Scanning</td>
<td>1 227 321</td>
<td>921 631</td>
<td>1 256 348</td>
<td>1 437 295</td>
<td>1 232 712</td>
<td>762 294</td>
<td>6 837 601</td>
</tr>
<tr>
<td>Normal video coding</td>
<td>1 100 991</td>
<td>1 050 629</td>
<td>1 244 457</td>
<td>1 512 467</td>
<td>1 366 139</td>
<td>909 376</td>
<td>7 184 059</td>
</tr>
<tr>
<td>Inter document</td>
<td>211 360</td>
<td>1 078 449</td>
<td>1 327 056</td>
<td>1 494 357</td>
<td>1 355 789</td>
<td>892 831</td>
<td>6 359 842</td>
</tr>
<tr>
<td>correction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality control</td>
<td>939 540</td>
<td>1 046 538</td>
<td>1 208 747</td>
<td>1 538 407</td>
<td>1 328 688</td>
<td>953 781</td>
<td>7 015 701</td>
</tr>
<tr>
<td>Logical error test</td>
<td>344 807</td>
<td>963 506</td>
<td>801 285</td>
<td>969 561</td>
<td>861 669</td>
<td>629 328</td>
<td>4 570 156</td>
</tr>
<tr>
<td>Data export en DVD</td>
<td>277 739</td>
<td>1 310 466</td>
<td>1 325 687</td>
<td>1 487 121</td>
<td>1 442 570</td>
<td>1 151 285</td>
<td>6 994 868</td>
</tr>
</tbody>
</table>
Dynamic planning of data capture from questionnaires of households and housing A3 by statistical district

<table>
<thead>
<tr>
<th></th>
<th>Dec05</th>
<th>Jan05</th>
<th>Feb05</th>
<th>Mar05</th>
<th>Apr05</th>
<th>May05</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Working days</td>
<td>23</td>
<td>20</td>
<td>20</td>
<td>23</td>
<td>19</td>
<td>17</td>
<td>150</td>
</tr>
<tr>
<td>Objective (# of districts)</td>
<td>3 393</td>
<td>7 191</td>
<td>7 168</td>
<td>7 619</td>
<td>7 115</td>
<td>6 635</td>
<td>37 323</td>
</tr>
<tr>
<td>Achieved (# of districts)</td>
<td>1 370</td>
<td>7 287</td>
<td>7 192</td>
<td>7 626</td>
<td>7 213</td>
<td>6 635</td>
<td>37 323</td>
</tr>
<tr>
<td>Percent (%)</td>
<td>40%</td>
<td>101%</td>
<td>100%</td>
<td>100%</td>
<td>101%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Production in Graph, December 2004

Decembre 2004

Nbr. Of questionnaires per day
Employees of the month

Cluster technical supervisor
Mr. Majid MRANI

Cluster functional supervisor
Mlle. Zohra KARIM

Reception
M. Rachid BOUDERSA

Scanning
Mme Meryem BENMOUSSA
Mr. Abdelaziz EL FAKIR

Scanning
Mme Saida MEKTOUM
Mr. Ali AGOUZOUL

Quality control
Mlle. Hanane ELHAIRECH

Normal video coding
M. Naima TAOUIFIK

Interdocument control
Mr. Mohamed AYAT

Logical tests video coding
M. Driss ELKEDDARI
桂花对于生活品质的提升

為了提升生活品質，桂花被應用於各種產品，包括洗髮精、香水和化妝品。它們的自然香味和舒緩效果備受人們喜愛，特別是對敏感肌膚和頭皮的護理。桂花提取物還被用於食品和飲料中，增添了一種獨特的風味。因此，桂花在現代生活中扮演了重要的角色，不僅是美麗的裝飾品，也是提升生活品質的重要元素。
Thank you for your time