



2010 World Population and Housing Census Programme

## 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

**Kodak Digital Science 3520 Scanner on SCSI 1**

**Mode:**  
Noir et blanc

**Mode de numérisation :**  
Recto-verso

**Points par pouce:**  
200

**Format de page :**  
A3 - 297 x 420 mm

**Numérisation(Z):**  
Seuil adaptatif (ATP)

**Mise en page**  
 Portrait  Paysage

**Tramage:**  
Aucun

**Contraste %**  
 Manuel  Automatique  
50

**Seuil**  
120  
Plus clair Normal Plus sombre

**Côté**  
 Les deux  
 Recto  
 Verso

OK  
Annuler  
Plus d'infos...  
Zone...  
Valeur par défaut  
A propos...



## Scanners operators

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

Vibrating paper jogger:  
Align edges of A3 questionnaires

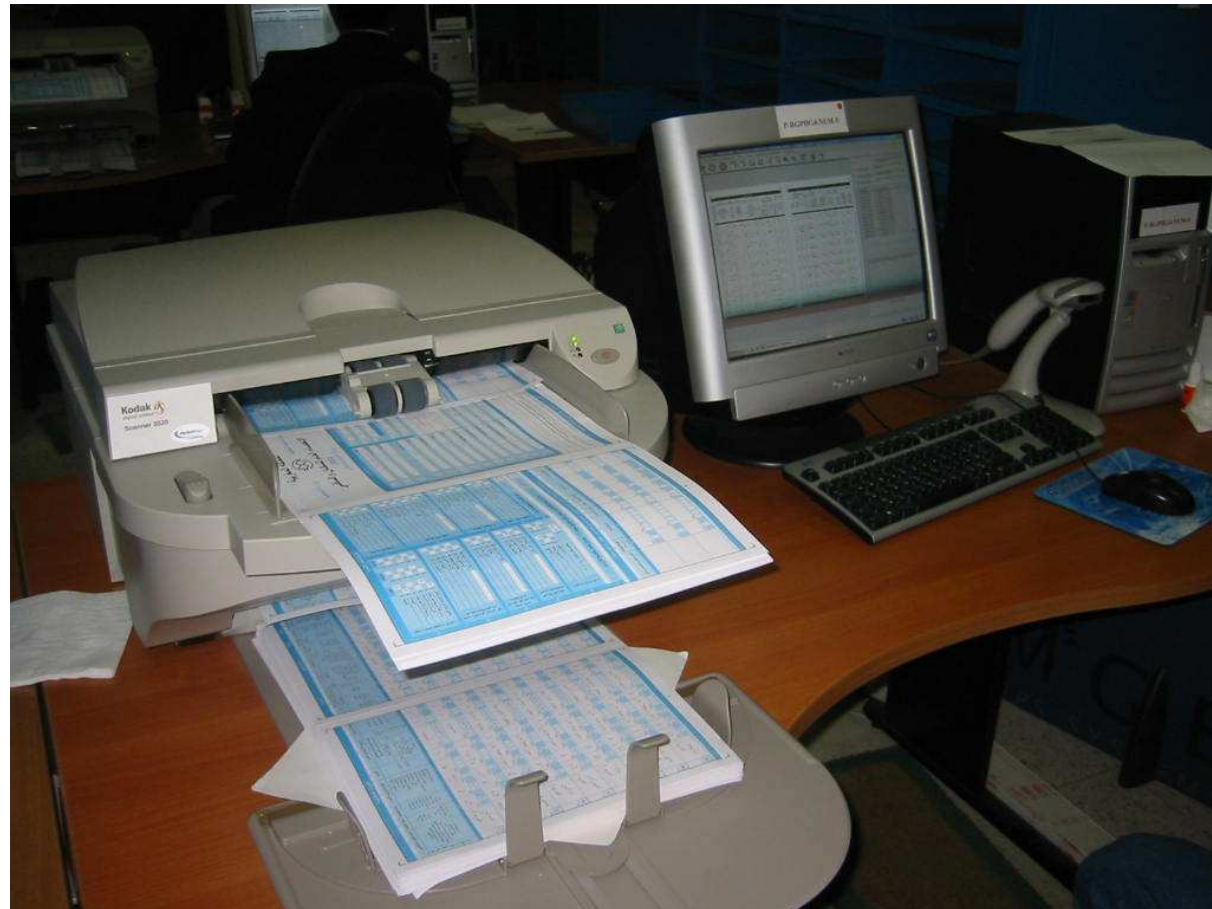




## Scanner

**Kodak Ds Digital Science Scanner 3520 handles 52 questionnaires A3 par minute.**

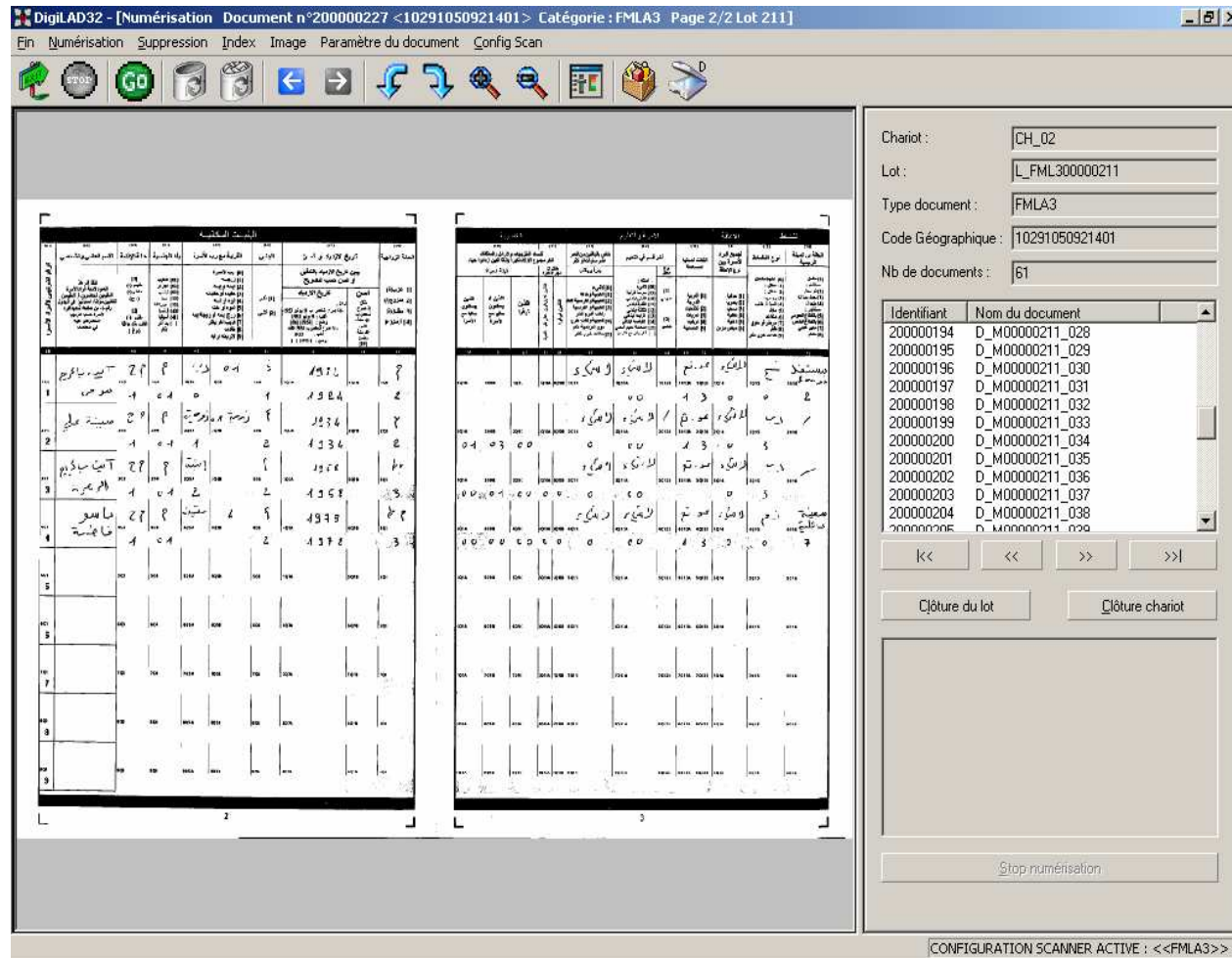
**Scanner operators verify on real time quality of scanned images. If quality deteriorate, scanners are cleaned and scanning is redone.**





# Scanning GUI

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.

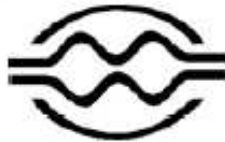




## Questionnaire of Households and housing

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

المملكة المغربية



المنذوبية السامية للتخطيط

الإحصاء العام للسكان

2004

ورقة الأسرة والمسكن

الرمز الجغرافي		
جهة :	(15) E7	مشيخة، دوار :
إقليم أو عمالة	(19) E8	رقم المسكن في منطقة الإحصاء :
دائرة :	(22) E9	عدد الأسر بالمسكن :
مقاطعة، بلدية	(24)	رقم الأسرة في المسكن :





## 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.

المسكن		السكان	
نوع المسكن	عدد السكان	عدد السكان المؤقتين	عدد السكان الدائمين
عدد الغرف	عدد السكان المؤقتين	عدد السكان الدائمين	عدد السكان المؤقتين المؤقتين
عدد الغرف المأهولة	عدد السكان الدائمين	عدد السكان المؤقتين المؤقتين	عدد السكان الدائمين المؤقتين
عدد الغرف غير المأهولة	عدد السكان المؤقتين المؤقتين	عدد السكان الدائمين المؤقتين	
عدد الغرف المأهولة جزئياً	عدد السكان الدائمين المؤقتين		
عدد الغرف غير المأهولة جزئياً			
عدد الغرف المأهولة مؤقتاً			
عدد الغرف غير المأهولة مؤقتاً			



## 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



## Questionnaire of Households and housing



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.

الرمز الجغرافي		
(01) E1	1 6	جهة : طنجة تطوان
(02) E2	5 1 1	إقليم أو عمالة : طنجة أصيلة
(04) E3	0 3	دائرة : أصيلة
(06) E4	0 9	بلدية أو جماعة قروية : الخلوة
(10) E5	2	مركز :
(11) E6	0 5 0 8	رقم منطقة الإحصاء :
(15) E7	0 4	مشيخة : المركز
(17) E8	0 8	نوار : أولاد حباس



## Scanned image

Scanned images are black and white.

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

المملكة المغربية  
المنذوبية السامية للتخطيط  
وزارة الداخلية

**الإحصاء العام للسكان والسكنى 2004**  
ورقة الأسرة والمسن

الرمز الجغرافي

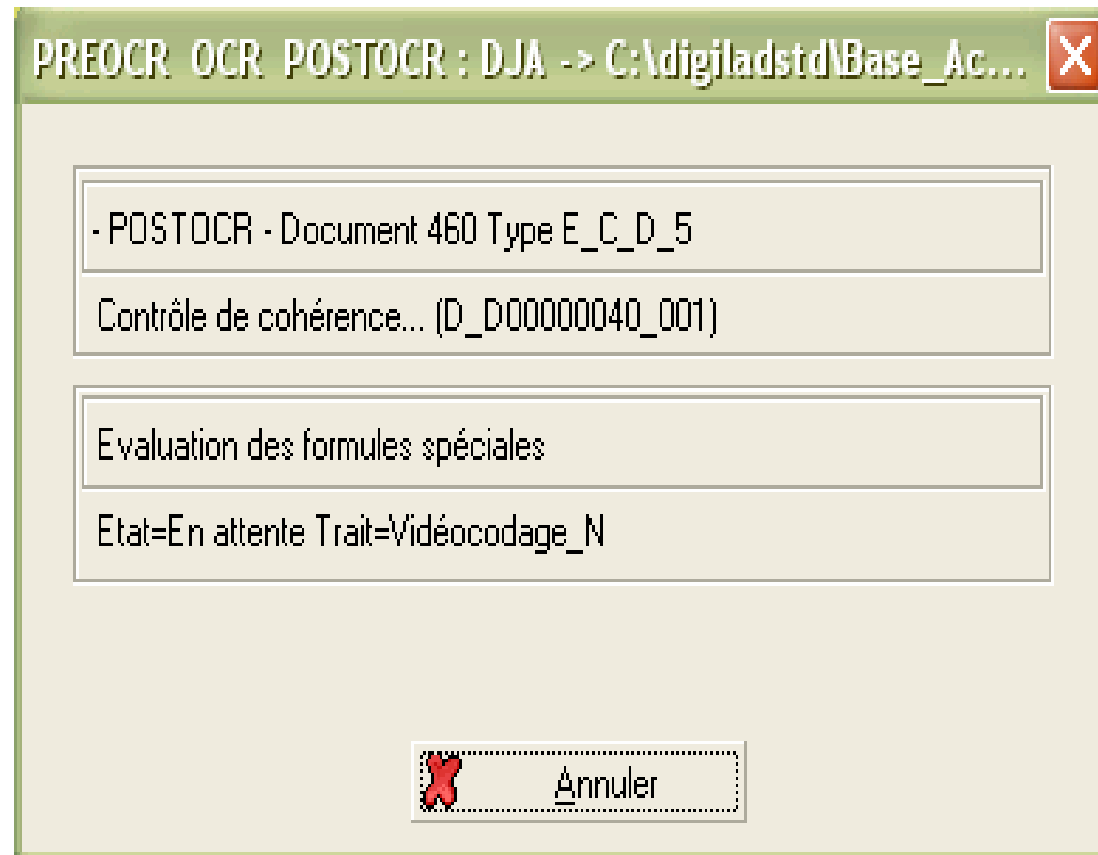
13	جهة: مكناس - تافيلالت	مشيخة:	
061	إقليم أو صفاة: مكناس	لوار:	
01	دائرة: 01	رقم المسكن في منطقة الإحصاء:	001 001
01	بلدية أو جماعة قروية: 01	عدد الأسر بالمسكن:	01 01
1	مركز:	رقم الأسرة في المسكن:	01 01
0004	رقم منطقة الإحصاء: 0004	عدد الاستمارات التي ملئت لهذه الأسرة:	1 1
		رقم الاستمارة:	1 1



## 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.



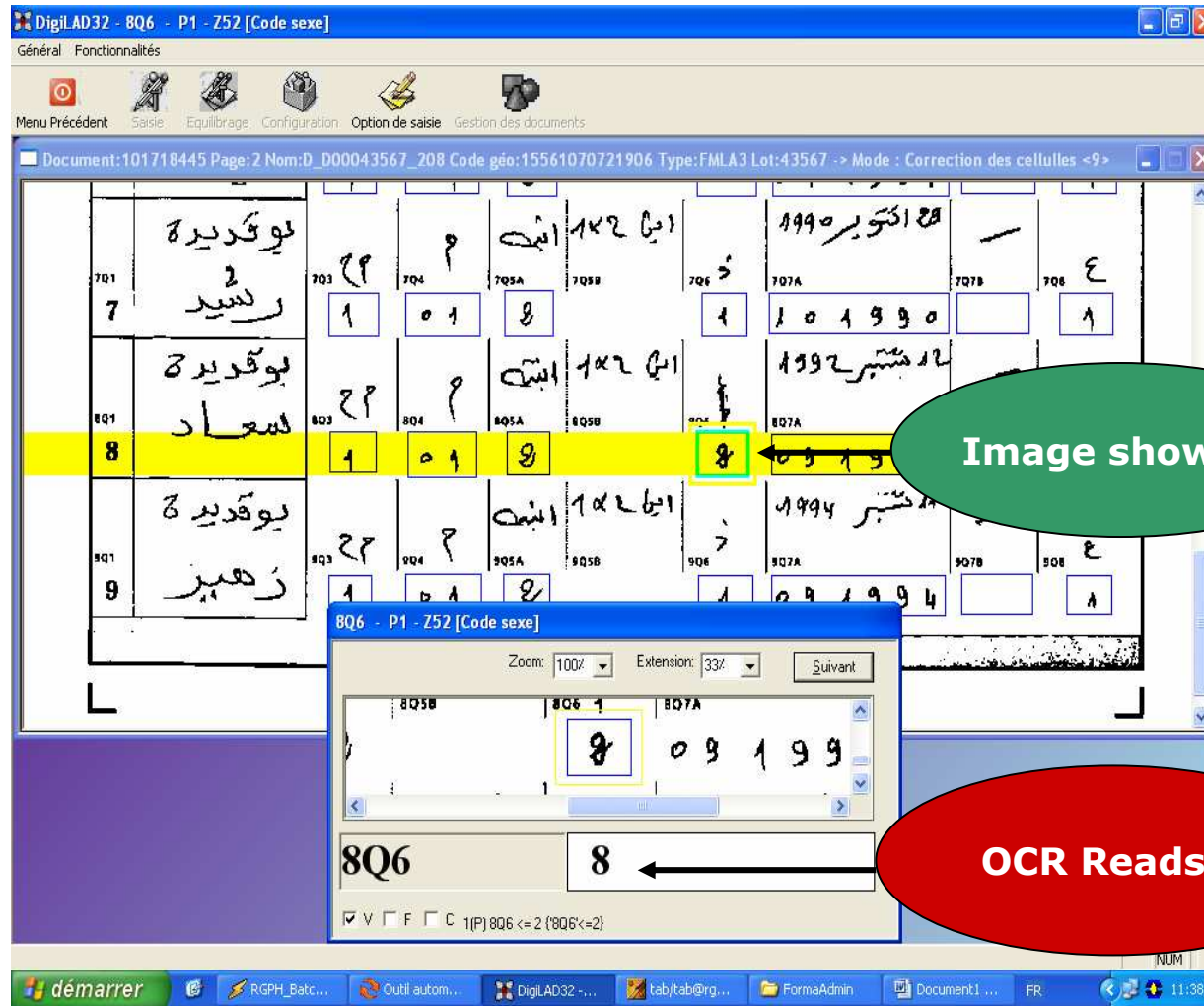


## 5. Normal video coding

- ❑ 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.

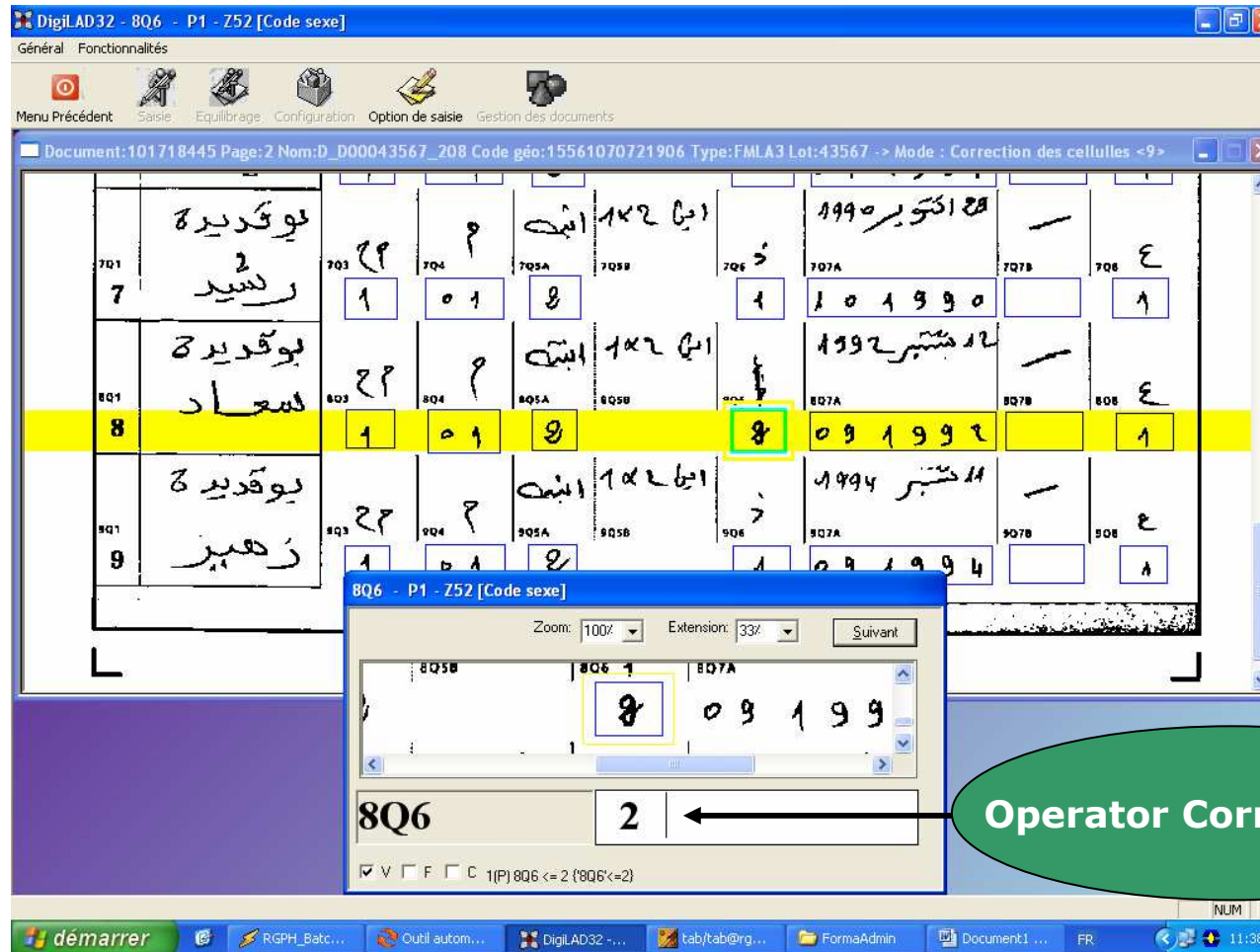


The screenshot shows a software interface for video coding. The main window displays a table with handwritten entries in Arabic. The table has columns for various fields, including names and dates. The entry for '8Q6' is highlighted in yellow. A green oval points to the handwritten '2' in the '8Q6' field, with the text 'Image shows 2'. A red oval points to the OCR result '8' in the '8Q6' field, with the text 'OCR Reads 8'. A smaller window in the foreground shows a zoomed-in view of the '8Q6' field, with the OCR result '8' and the true value '2' visible. The OCR result '8' is highlighted in yellow, and a red oval points to it with the text 'OCR Reads 8'. The zoomed-in window also shows the date '09 1999' and the field label '8Q6'.

Image shows 2

OCR Reads 8

# Normal video coding GUI



The screenshot shows the DigilAD32 software interface. The main window displays a grid of handwritten data in Arabic. A specific cell containing the number '8' is highlighted in yellow. A smaller window is open over this cell, showing a zoomed-in view of the data and a field where the value '2' is being entered. A green callout bubble with an arrow points to this field, containing the text "Operator Corrects value".

Row	Cell	Content
701	701	بوكدية 2 السيد
701	702	22
701	703	1
701	704	01
701	705A	8
701	705B	1
701	706	ع
701	707A	101990
701	707B	
701	708	1
801	801	بوكدية سعاد
801	802	22
801	803	1
801	804	01
801	805A	8
801	805B	8
801	806	091992
801	807A	
801	807B	
801	808	ع
901	901	بوكدية زهير
901	902	22
901	903	1
901	904	01
901	905A	8
901	905B	1
901	906	091994
901	907A	
901	907B	
901	908	ع

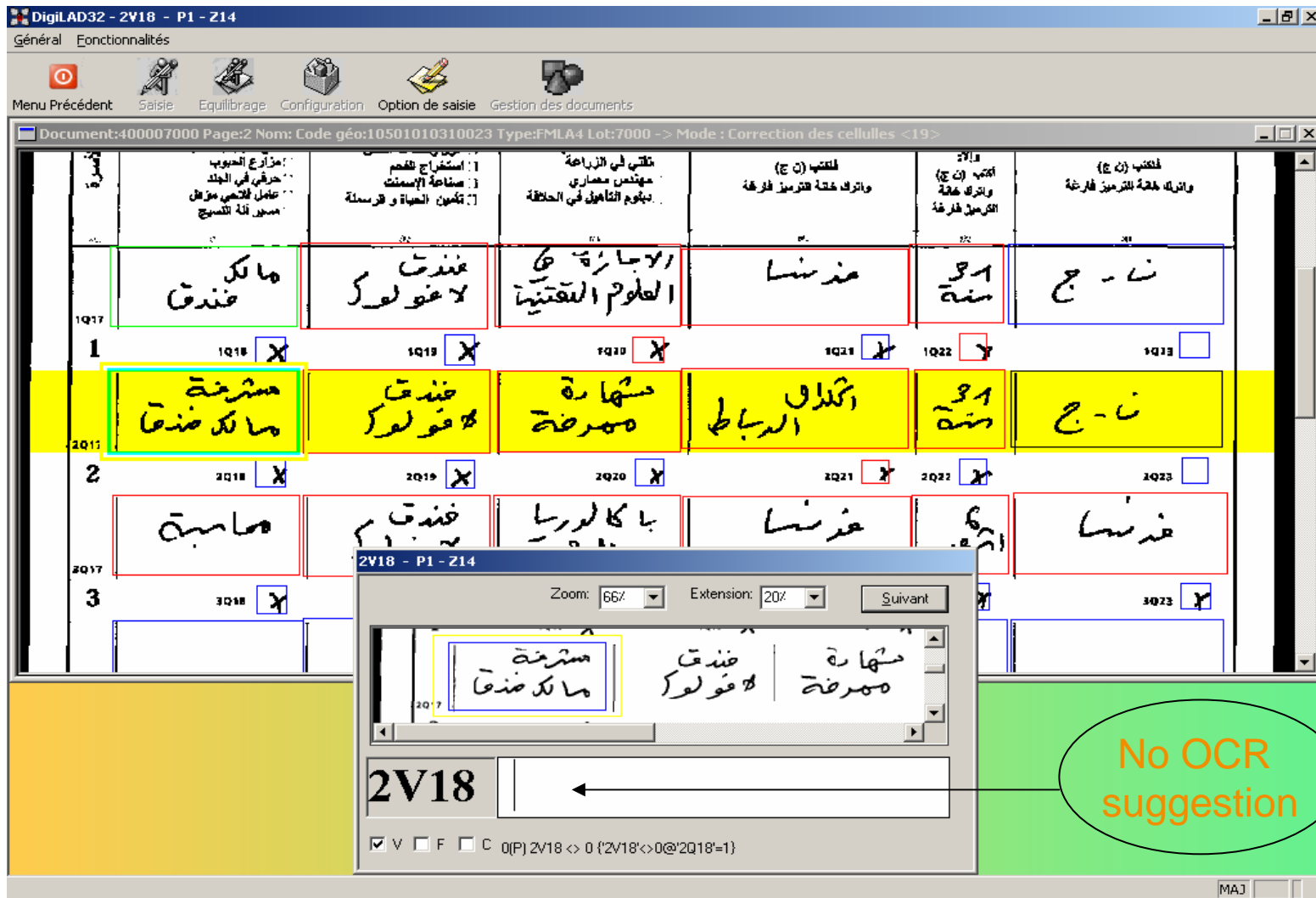




## Normal video coding

- ❑ 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



The screenshot shows a software interface for video coding. The main window displays a grid of handwritten text cells. The grid is organized into rows and columns, with each cell containing a different piece of handwritten text. The cells are color-coded: some are highlighted in yellow, some in red, and some in blue. A zoomed-in view of one cell is shown in the foreground, displaying the text 'مشرفة' and 'ما لك ضنقا'.

The interface includes a menu bar at the top with options like 'Menu Précédent', 'Saisie', 'Equilibrage', 'Configuration', 'Option de saisie', and 'Gestion des documents'. The status bar at the bottom shows '2V18' and 'No OCR suggestion'.

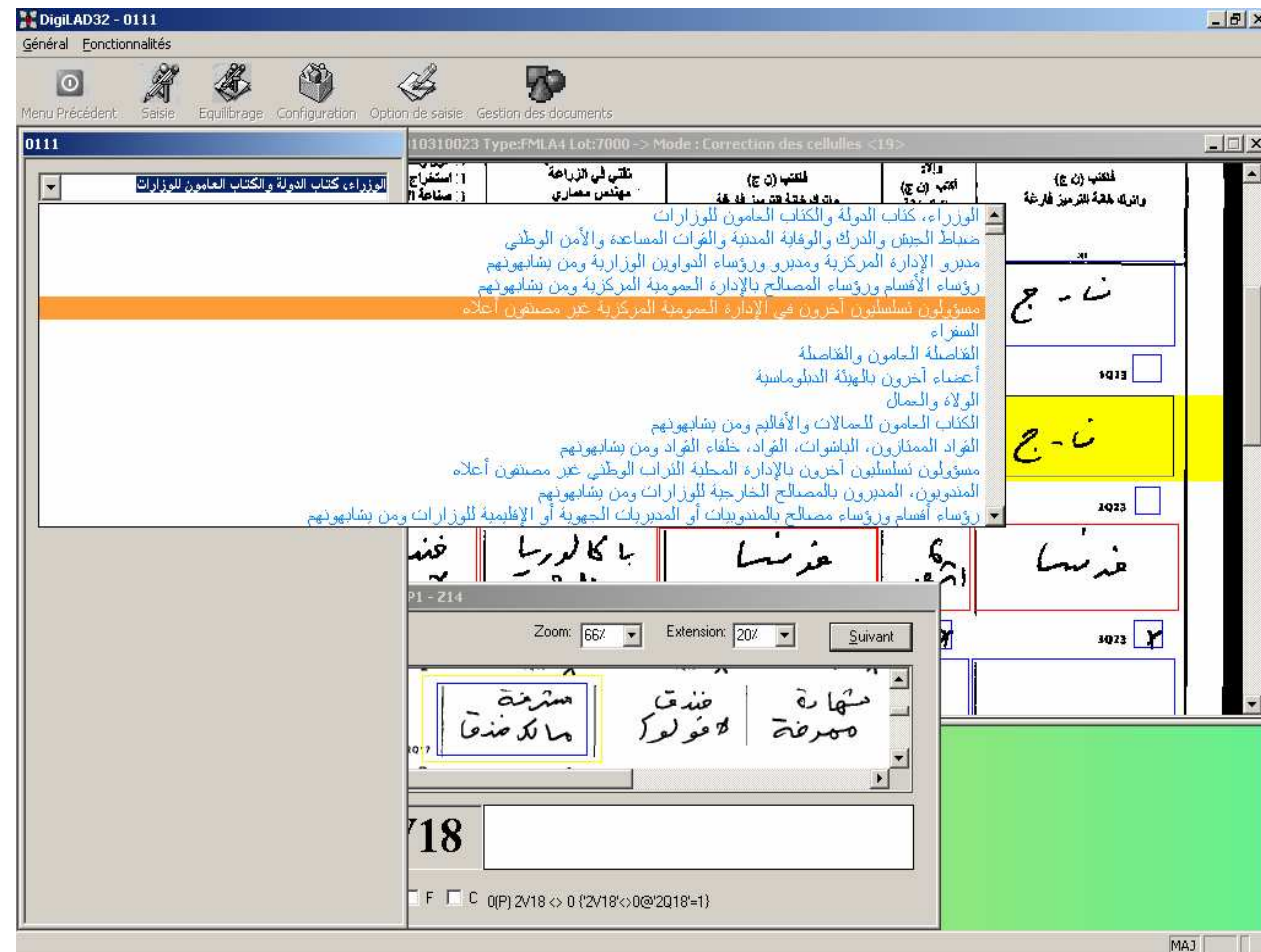
Row	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
1	ما لك ضنقا	خذنقا لا قولوك	الاجازة في العلوم التقنيّة	خذنقا	31 سنة	ن - ج
2	مشرفة ما لك ضنقا	خذنقا لا قولوك	مشرفة مشرفة	انكلاو الرباط	31 سنة	ن - ج
3	مما سبة	خذنقا	باكالوريا	خذنقا	31 سنة	ن - ج

No OCR suggestion



## 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.



## Afnor norm NFX06-022 of October 1991

Table 1 – Lettre-code en fonction de l'effectif des lots et du niveau de contrôle

Effectif des lots	Niveaux de contrôle spéciaux				Niveaux de contrôle pour usages généraux		
	S-1	S-2	S-3	S-4	I	II	III
2 à 8	A	A	A	A	A	A	B
9 à 15	A	A	A	A	A	B	C
16 à 25	A	A	B	B	B	C	D
26 à 50	A	B	B	C	C	D	E
51 à 90	B	B	C	C	C	E	F
91 à 150	B	B	C	D	D	F	G
151 à 280	B	C	D	E	E	G	H
281 à 500	B	C	D	E	F	H	J
501 à 1 200	C	C	E	F	G	J	K
1 201 à 3 200	C	D	E	G	H	K	L
3 201 à 10 000	C	D	F	G	J	L	M
10 001 à 35 000	C	D	F	H	K	M	N
35 001 à 150 000	D	E	G	J	L	N	P
150 001 à 500 000	D	E	G	J	M	P	Q
500 001 et au-dessus	D	E	H	K	N	Q	R

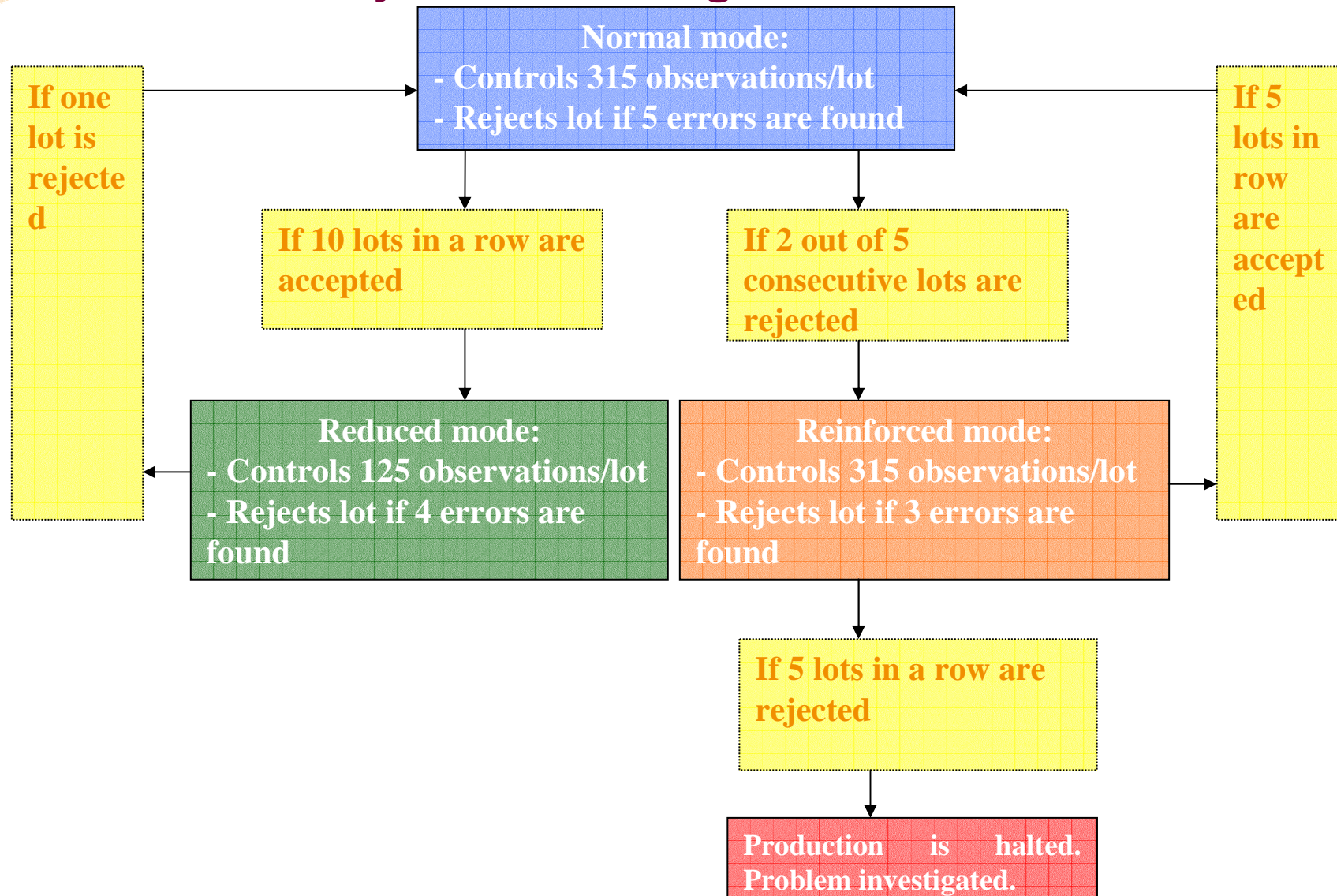
According to Afnor norm (general level control II) for lots of 10001 up to 35000 observations, 315 observations are to be sampled on normal and reinforced modes .

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

Lettre-code	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R
Effectif d'échantillon n du plan simple (1)	2	3	5	8	13	20	32	50	80	125	200	315	500	800	1 250	2 000



## Quality control using Afnor norm

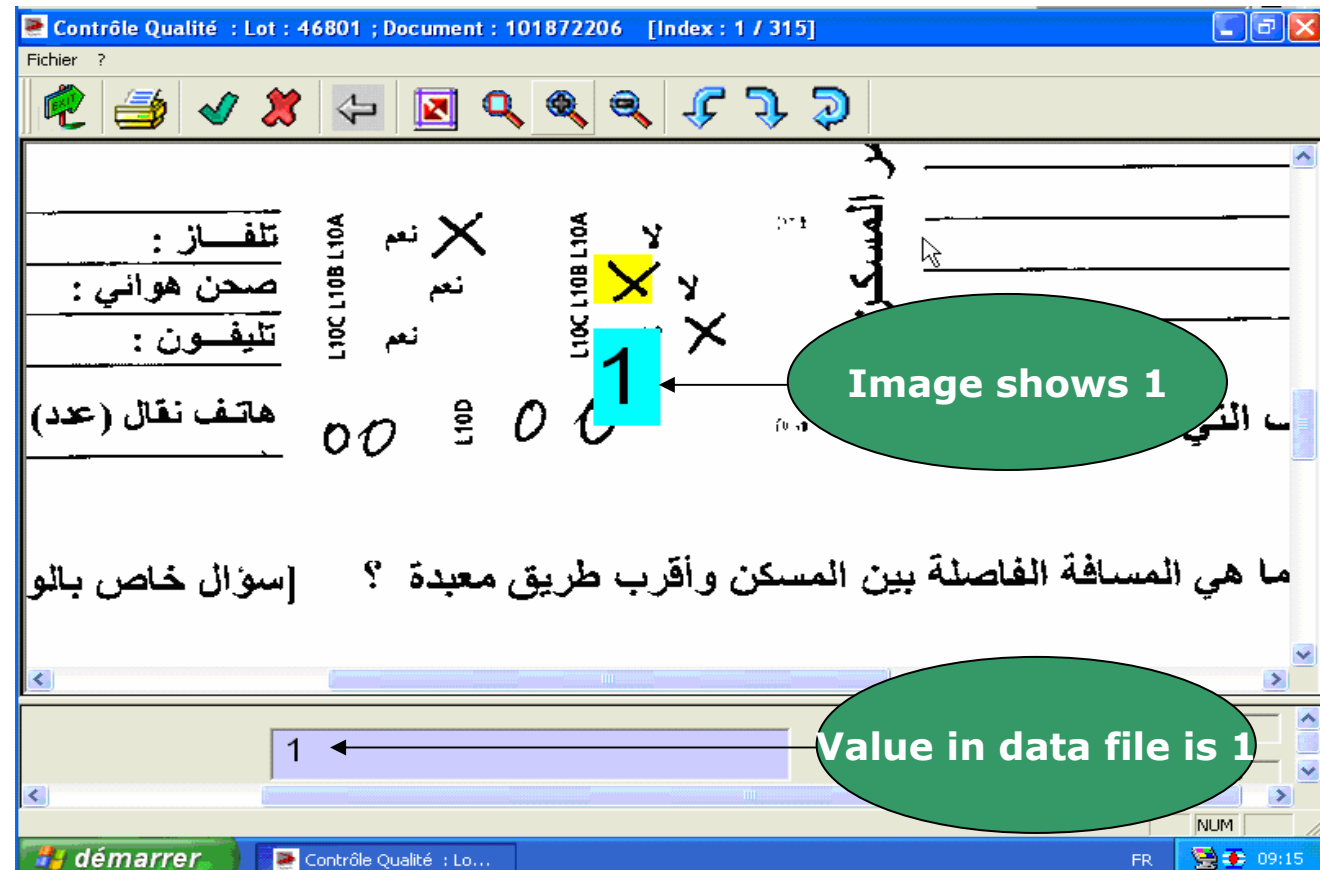






## Quality control GUI

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.

```
Script
## Formules de cohérence FML_A34 Création du : 29/03/2004
##
'E1'=Controle(STR_CLIENT,0,2)[Code région incorrect]
'E2'=Controle(STR_CLIENT,2,3)[Code province incorrect]
'E3'=Controle(STR_CLIENT,5,2)[Code cercle incorrect]
'E4'=Controle(STR_CLIENT,7,2)[Code commune incorrect]
'E5'=Controle(STR_CLIENT,9,1)[Code centre incorrect]
'E6'=Controle(STR_CLIENT,10,4)[Code district incorrect]
'E7'=0@('E5'<>2)
'E8'=0@('E5'<>2)
'E7'>0@('E5'=2)
'E8'>0@('E5'=2)
'E9'>0
'E9'<501
```

Valider Annuler



## Logical error tests video coding GUI

Logical error example: Although this house is declared as empty, type of ownership is declared as owner.

Computer operator corrects inconsistency.

The screenshot shows the DigLAD32 software interface. The main window displays a form for housing declaration with various fields and checkboxes. A red oval highlights a logical error in the form. A dialog box is open, showing the correction of the error. The dialog box contains the following text:

L13 - P3 - Z200 [Nature d'occupation du logement] < MODE FORMULE : 0...

Zoom: 100% Extension: 33% Suivant

2 مسكن سكانه ثانويون بصفة

3 مسكن فارغ

L13 1

1(P)L13=1{(L13=1@((E141?E14)&(E14=0)&(L14=0))}

The dialog box also shows a grid of checkboxes and a red 'X' mark, indicating the correction of the logical error.

Logical error



## 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.



## Exported data file

Data is exported  
in plain text file  
with dictionary,  
for further  
processing  
using  
CSPro/IMPS.

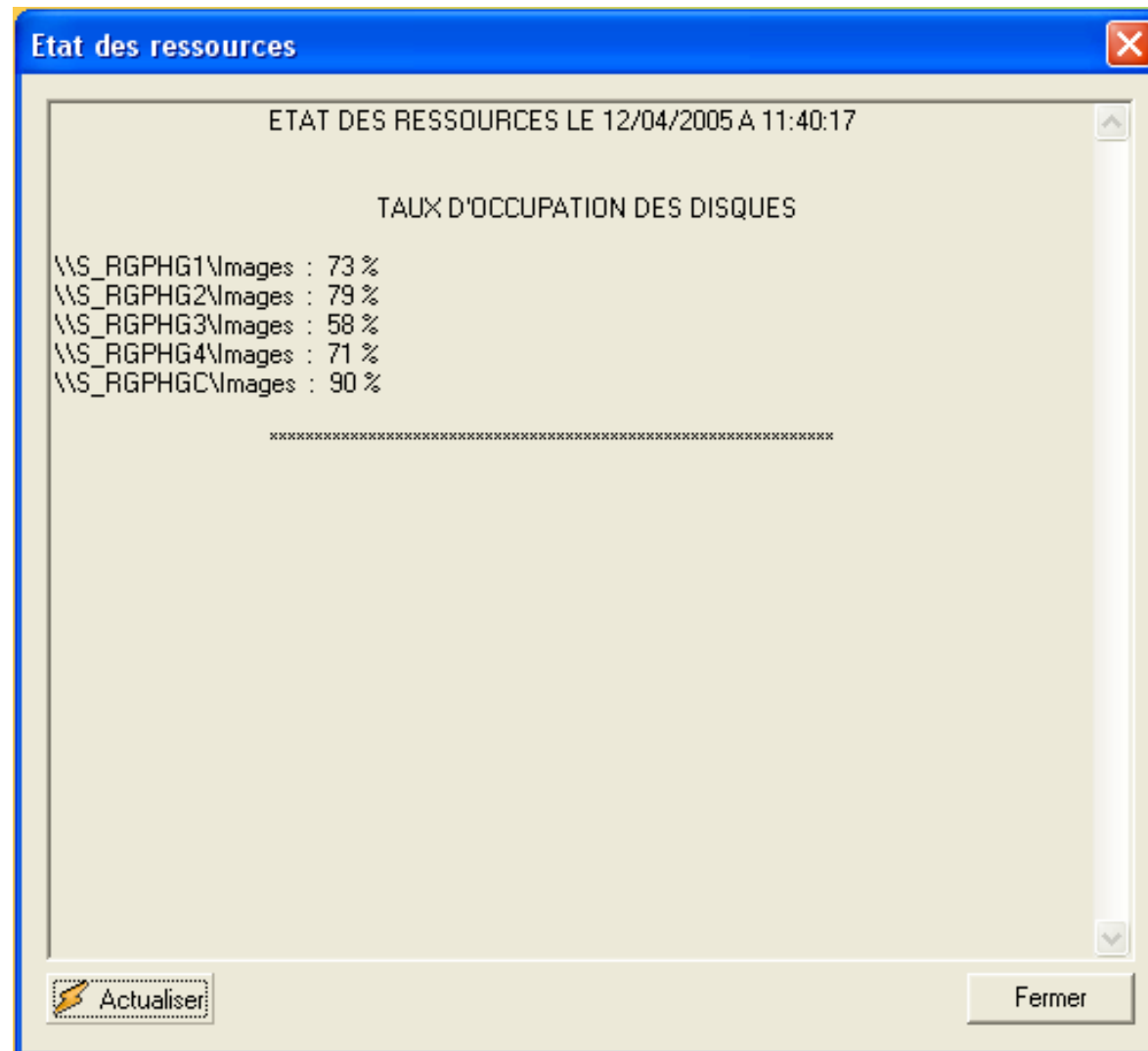
Record type

02321010110013	1230101111110101	1968	2	21411	06
02321010110013	1230101111210112	1969	201000000000	1	3
02321010110013	1230101111310122	1995	1	1111	4
02321010110013	1230101112 133301313331			1 1	12200
02321010110013	1240101114 0202000000				
02321010110013	1240101111110101		271	321113	06
02321010110013	1240101111210191		311	322113	26
02321010110013	1240101112 133303113331			1 1	11202
02321010110013	1250101114 0101000000				
02321010110013	1250101111120101	1973	1	000 1	06
02321010110013	1250101112 163301333332			1 1	22201
02321010110013	1260101114 0505000000				
02321010110013	1260101111110101	1955	2	22111	06
02321010110013	1260101111210112	1961	20301000022211		3
02321010110013	1260101111310121	1981	1	21611	4
02321010110013	1260101111410122	1992	1	21411	4
02321010110013	1260101111510121	1997	1	1011	4
02321010110013	1260101112 133302113331			1 1	11201
02321010110013	1270000004 0000000000				
02321010110013	1270000002 33				
02321010110013	1280101114 0505000001				
02321010110013	1280101111110101081968		2	21411	02



## 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.

[Module de statistiques]

Général

Statistiques journalières

Mass Scan System  
Statistiques

Digittech.

**Statistiques générées le 02/09/2003 à 12:05:56**

**02/09/2003**

Réception (envoi)

Nb utilis	Nb lots	Tps total	Tps/lot	Nb docs	Tps/doc	Nb pages	Tps/page	Nb envois	Tps/envoi	Nb env rej
1	-	07 sec	-	10	-	-	-	1	07 sec	-

Réception (lot)

Nb utilis	Nb lots	Tps total	Tps/lot	Nb docs	Tps/doc	Nb pages	Tps/page	Nb envois	Tps/envoi	Nb env rej
1	2	08 sec	04 sec	-	-	-	-	-	-	-

Préparation des documents

Nb utilis	Nb lots	Tps total	Tps/lot	Nb docs	Tps/doc	Nb pages	Tps/page
1	2	26 sec	13 sec	-	-	-	-

Taillage

Nb utilis	Nb lots	Tps total	Tps/lot	Nb docs	Tps/doc	Nb pages	Tps/page
1	2	04 sec	02 sec	-	-	-	-

Numérisation

Nb utilis	Nb lots	Tps total	Tps/lot	Nb docs	Tps/doc	Nb pages	Tps/page
1	2	14min 05sec	07min 02sec	55	15 sec	402	01 sec

Reconnaissance de

Terminé

Format XML    Format CSV





## 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).

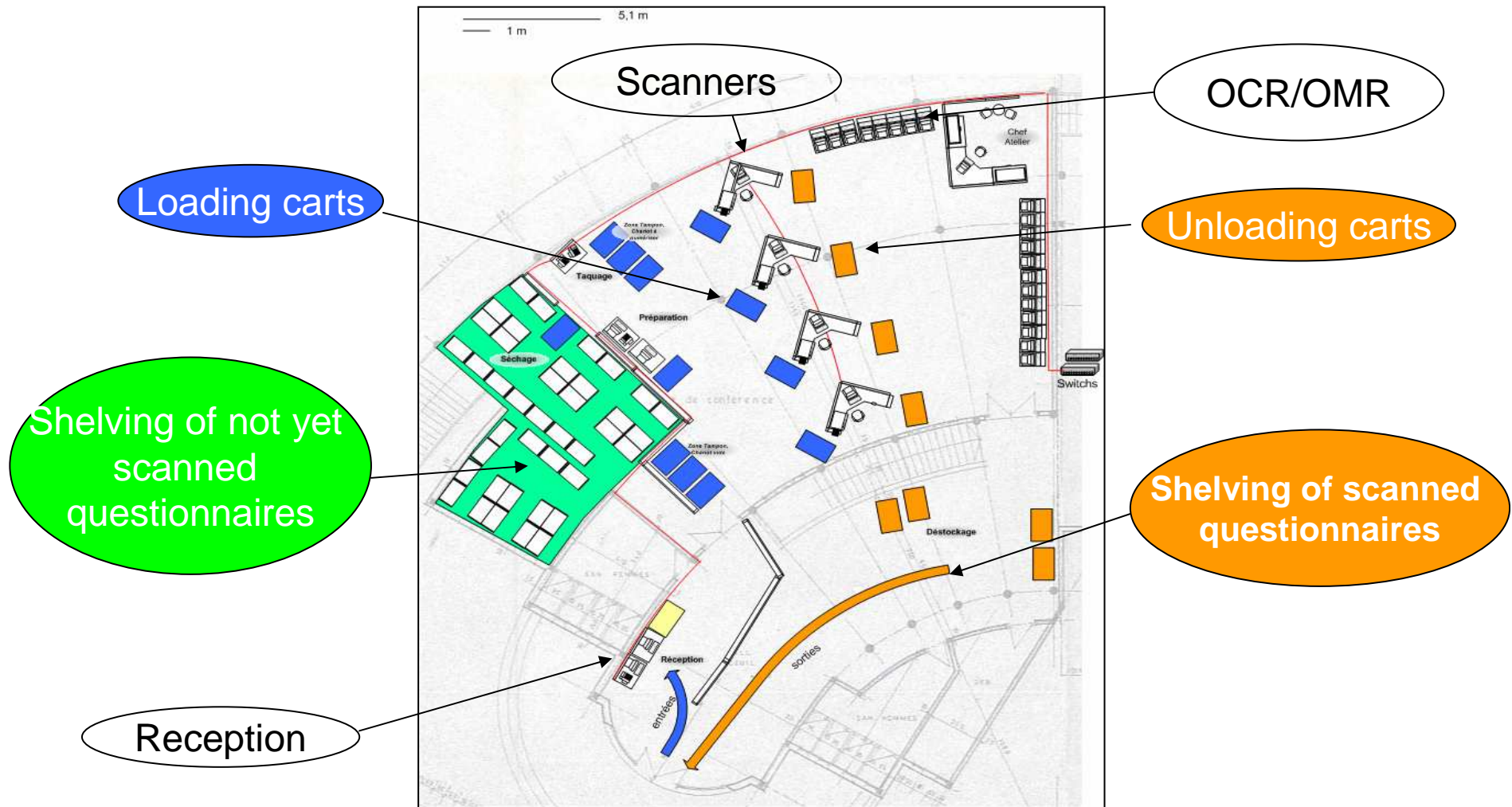


## Centre of Automatic Reading of Documents



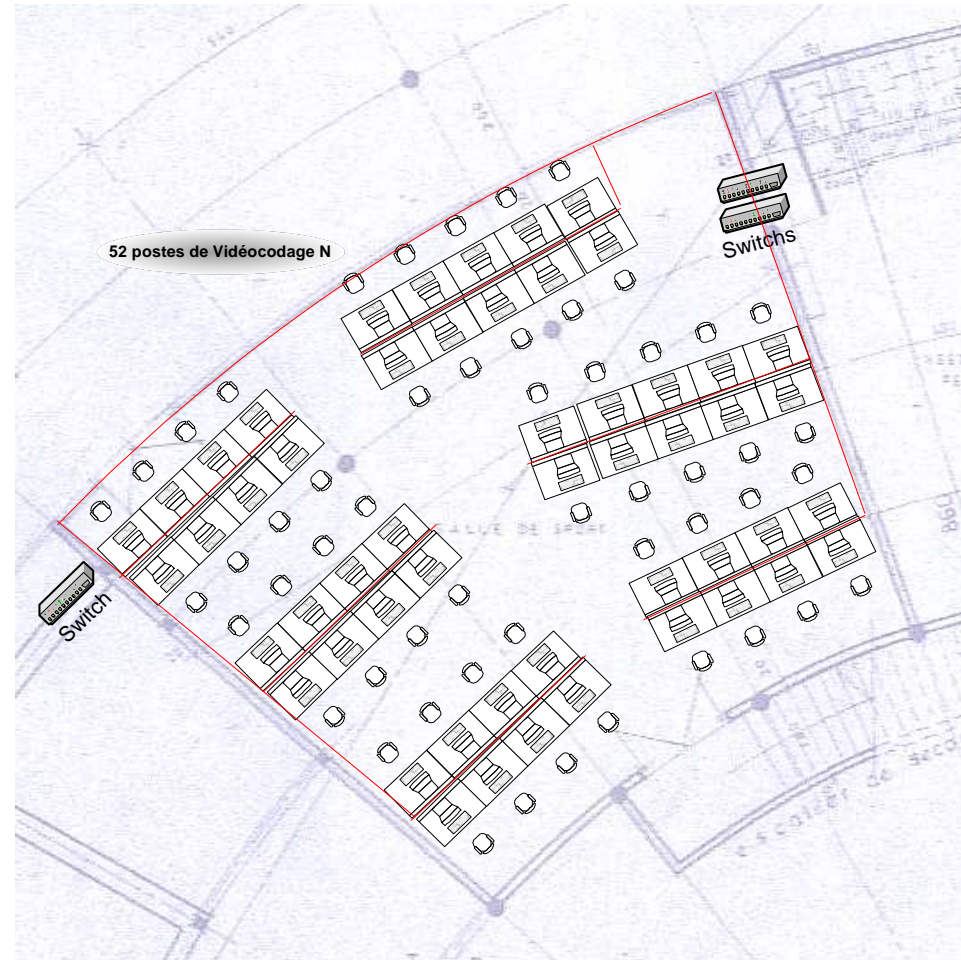
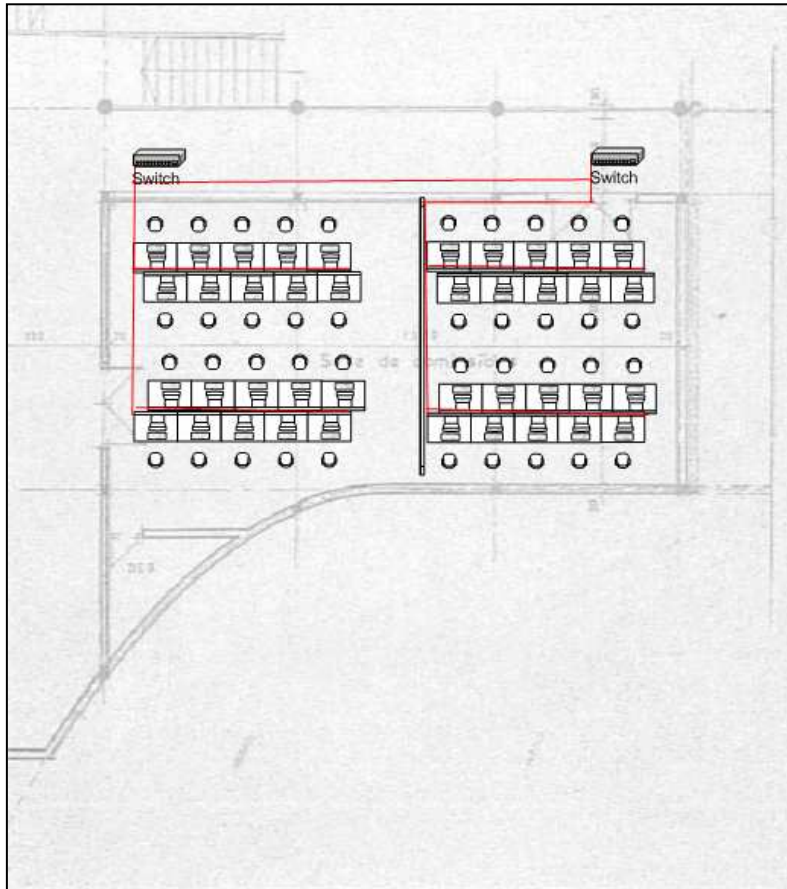


## Facilities : Scanning area





## Facilities : video coding computer labs

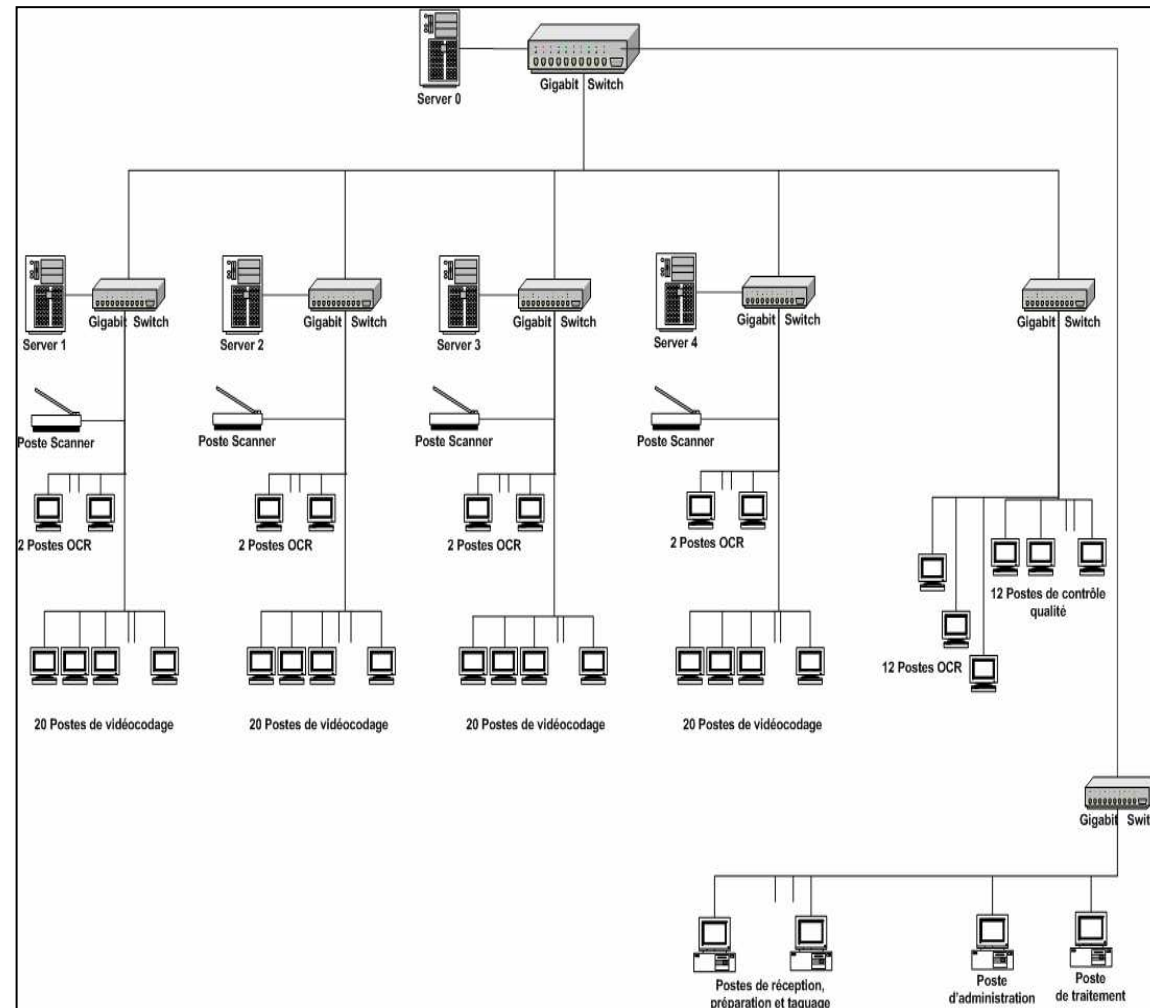




## Logical architecture of hardware installation into 4 clusters and one central server

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

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



## 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

Questionnaire	Volumes: # of Questionnaires	Type	Fields /quest.	Filed type	Time
Urban and rural population	38 000	A4 (21 pages two sides)	3051	Arabic numbers	1 month
Households and Housing	6 800 000	A3 two sides	248	Arabic numbers	6 months
	5 800 000	A4 two sides	12 54	Arabic numbers Arabic letters	12 months
Separately counted Population	12 500	A3 two sides	260	Arabic numbers	1 day
Nomad population	40 000	A3 two sides	245	Arabic numbers	1 day
	40 000	A4 two sides	12 54	Arabic numbers Arabic letters	2 day
All	39 888 000	pages A4			t0+ 18 months



## Monthly processed questionnaires of Households and housing A3 by production steps

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

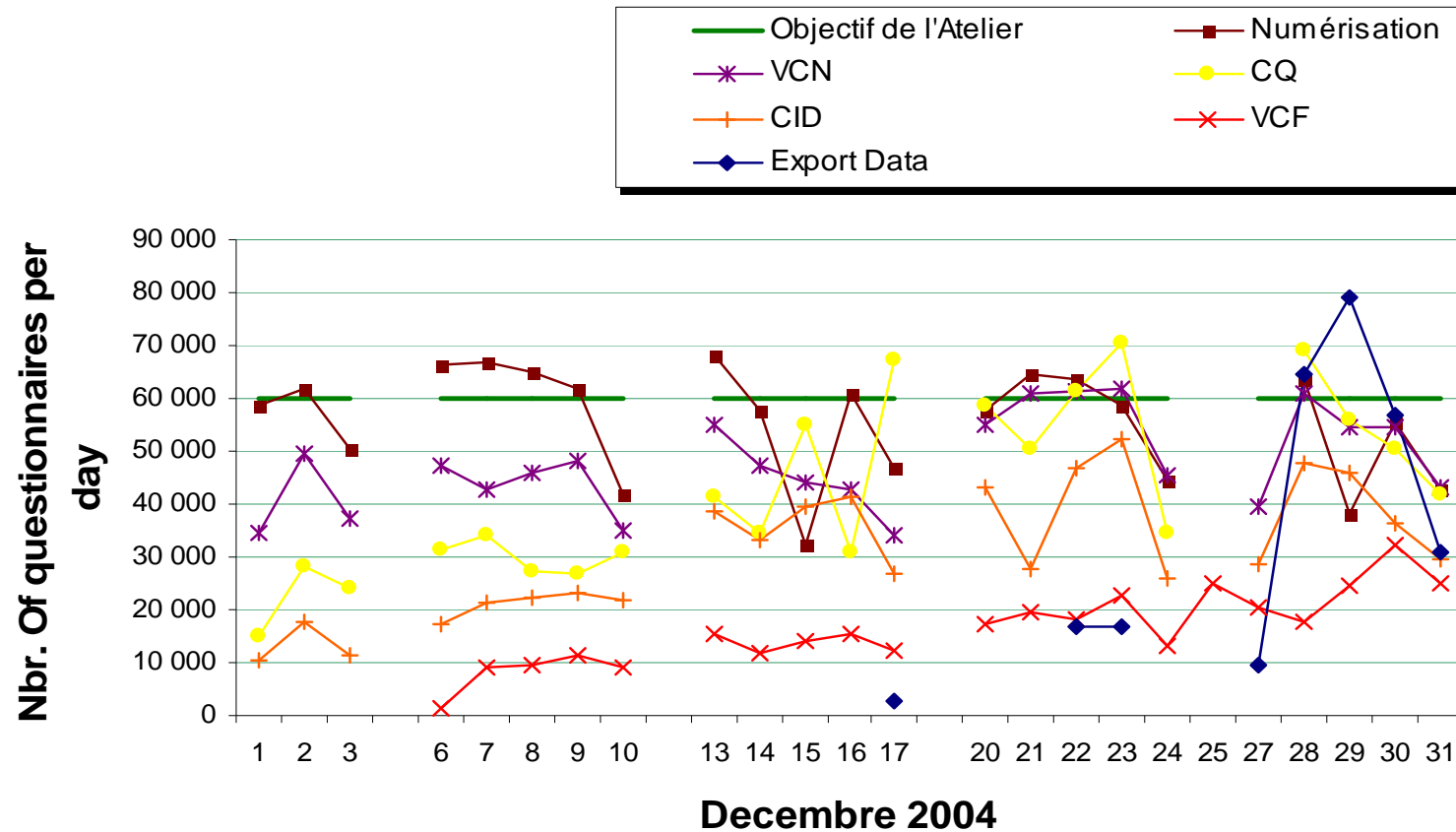


## Dynamic planning of data capture from questionnaires of households and housing A3 by statistical district

	Dec05	Jan05	Feb05	Mar05	Apr05	May05	Total
Working days	23	20	20	23	19	17	150
Objective (# of districts)	3 393	7 191	7168	7619	7115	6635	37 323
Achieved (# of districts)	1 370	7 287	7 192	7626	7213	6635	37 323
Percent (%)	40%	101%	100%	100%	101%	100%	100%

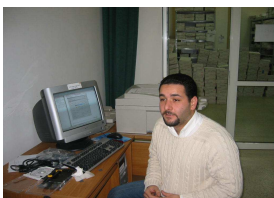


## Production in Graph, December 2004





## Employees of the month



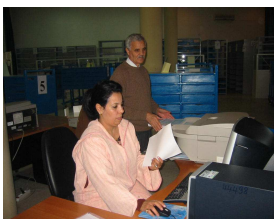
**Cluster technical supervisor**  
Mr. Majid MRANI

**Cluster functional supervisor**  
Mlle. Zohra KARIM



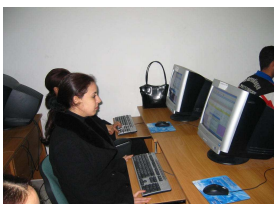
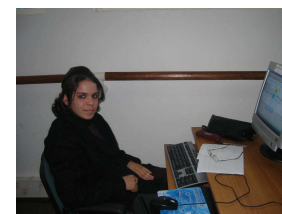
**Reception**  
M. Rachid BOUDERSA

**Scanning**  
Mme Meryem BENMOUSSA  
Mr. Abdelaziz EL FAKIR



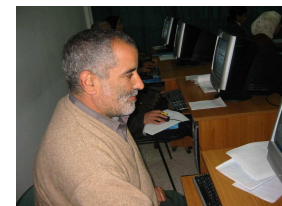
**Scanning**  
Mme Saida MEKTOUM  
Mr. Ali AGOUZOUL

**Quality control**  
Mlle. Hanane ELHAIRECH



**Normal video coding**  
M. Naima TAOUFIK

**Interdocument control**  
Mr. Mohamed AYAT



**Logical tests video coding**  
M. Driss ELKEDDARI



## Employee of the month

**CERTIFICAT D'EXCELLENCE**

*En reconnaissance des efforts soutenus dans l'exercice  
de ses fonctions au sein de l'Atelier de Production*

**Le Centre de Lecture Automatique de Documents**

Mme / M. : *Naima TAOUFIK*

Fonction : *Vidéo codage normal*

se voit décerner le prix :  
"Employé(e) du mois de Décembre 2004"

Fait à Rabat, le 31 Déc. 2004

La Directrice du Centre

Le Chef d'Atelier

Mme. Sabah ELMRINI

M. Oussama MARSELI





## Team of data processing center – Mai 2005





*Thank you for your time*

