Expert Group Meeting on New Techniques in Population and Housing Censuses

Abu Dhabi, 22-23 May 2006

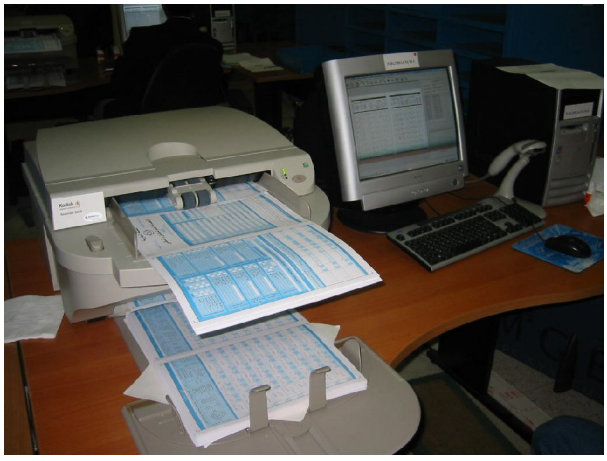
High Commissioner Office for Planning

Department of Statistics - Center of Automatic Reading of Documents (CLAD)

USING OPTICAL CHARACTER RECOGNITION IN CENSUS OF POPULATION AND HOUSING 2004 OF MOROCCO

By Oussama Marseli[1]

Abstract: This paper documents the use of Optical Character Recognition technique to capture data from all questionnaires of Census of Population and Housing 2004 of Morocco. Data extraction was separated into three periods to allow for timely diffusion of readily available data. Population count was available within 2 months of the beginning of data collection. A snapshot of socio-economic and demographic characteristics, and housing conditions were available for every district in the kingdom within 6 months period. Profession, economic activity, diploma and migration codes corresponding to open question answers written in Arabic letters were keyed in using both questionnaires images and integrated electronic dictionaries. The later phase, took 12 months of work to be completed.



[1] Manager of the Center of Automatic Reading of Document, smarseli@hotmail.com, http://www.clad.hcp.ma

CENSUS 2004 OF MOROCCO:

On September first 2004 , 53 601 persons, including 41 130 enumerators, 11 516 controllers, 883 supervisors, 72 provincial supervisors, accompanied with 17 698 auxiliary authority officers spread throughout the country to fill out Census of Population and Housing 2004 questionnaires from all households. Within 20 days, this task was achieved. On the 22nd of December 2004, the nation's population count was published, which was 29 891 708, an increase of 14.6% from 1994 census. State-of-the-art software (i.e. Artificial Intelligence and Image Analysis) and hardware (i.e., scanners, servers and computers) as well as trained personal were used in well organized manner to capture data, accurately and in time.

CHOOSING AUTOMATIC DOCUMENT READING TECHNOLOGY

With only nine months ahead of the planned schedule of the beginning of Census of Population and Housing 2004, the administration was studying two strategic choices to capture data, exhaustively and within a short period of time. The first choice was to deploy large human and material resources to key in data, while the second was to use automatic document reading. The former, used in prior censuses, although on samples, offered a tested and reliable method. The latter technology, used in developed countries (e.g., USA, France, South Africa…) promised a high level of data quality within a record time frame. The High Commissioner Office of Planning of Morocco decided to use automatic reading of documents in order to provide accurate radioscopy of demographic, socio cultural and condition of housing of Moroccan households, much needed on the process of development of Morocco. Secondary Objectives were to acquire and adapt new technologies, which could substantially enrich the statistical system as a whole in the country.

OUTSOURCING TO A PRIVATE CONTRACTOR

The first task was to identify an external partner from the private sector. A committee representing different government structures was appointed to this mean. The objective was to choose a well experienced firm to work closely with, in order not only to achieve Census of Population and Housing 2004 objectives but also to master a technology that could be used to capture data from other surveys.

The technical offer of Digitech - a French company- stood out despites higher financial offer. The international firm agreed to provide a complete solution as well as to insure efficient knowledge capacity transfer to the administration engineers. The project was conducted in three phases.

TESTING PHASE

The 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. This phase took about three months.

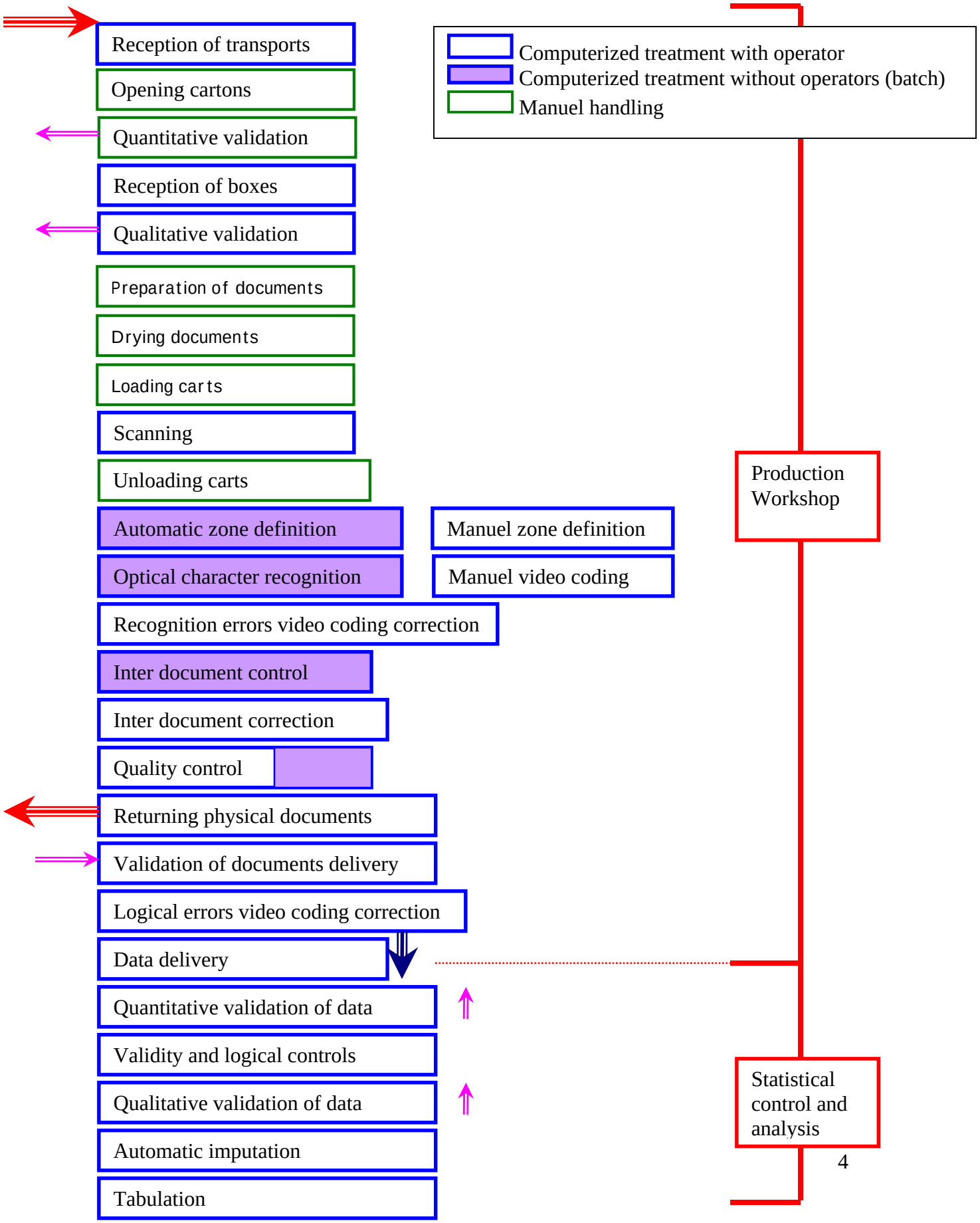
As a result, OCR engine was adapted to account for Arabic hand writing of numbers (i.e. number one). Production procedures and steps were validated (please consult graphic 1). And finally, superior quality of OCR over data keying was confirmed. Indeed, a sample of 400 districts (about 72 000 questionnaires) was filled out. Results showed that with 99.8% confidence level, automatic documents reading error rate was between 3.56 and 3.57 errors per 10 000. Data keying method yielded 5 times higher error rate (19.95 error per 10 000).

The outcome of the test was important to scientifically quantify necessary human and material resources to achieve the task at hand within a constrained time frame.

Automatic Documents Reading is not only about scanning and “OCRizing” images. A comprehensive solution is necessary to capture, understand, manage and communicate large amount of data regularly and in time. The whole cycle of production is well organized in many steps.

Reception of transports: The first step is to receive questionnaires while validating the number of received boxes.

Graphic 1. Automatic Documents Reading Line of Production

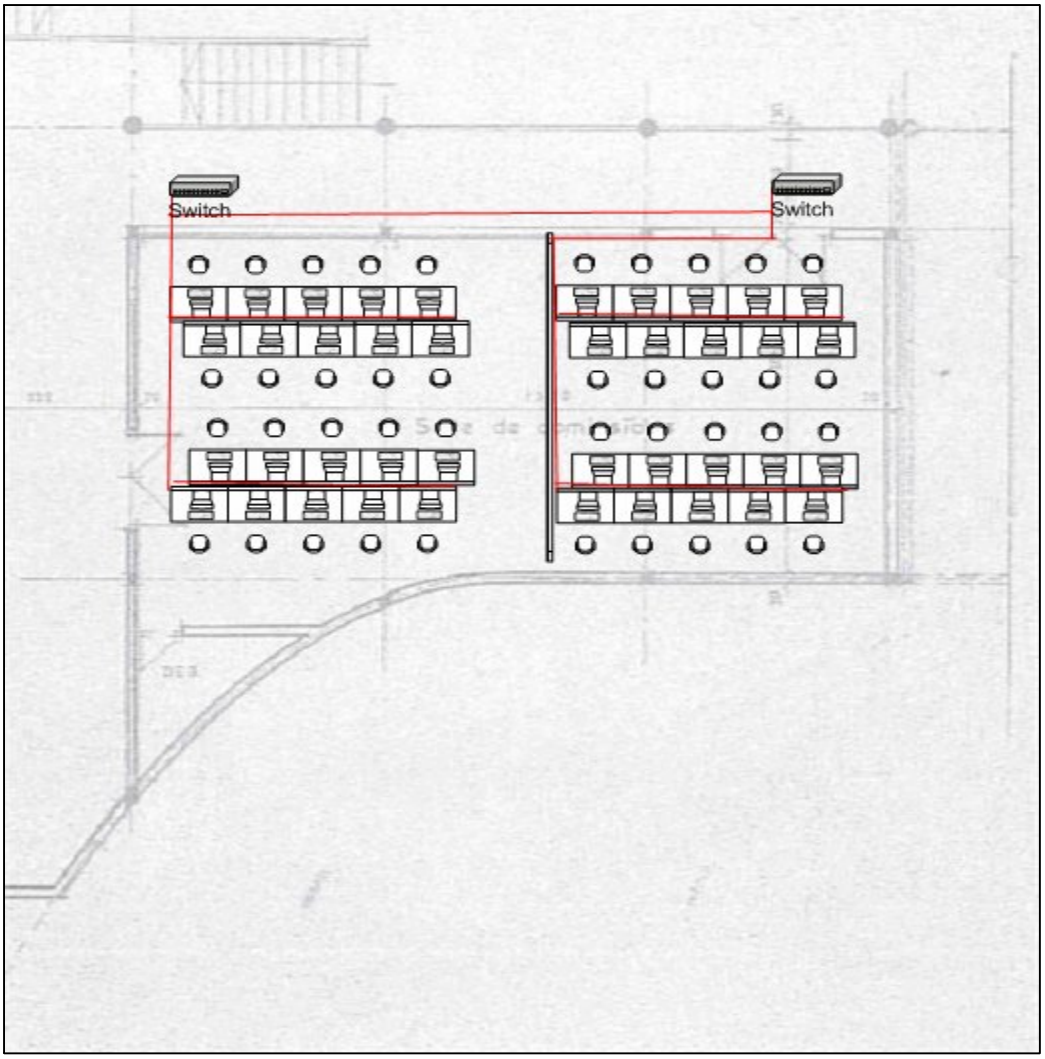


IMPLEMENTATION PHASE

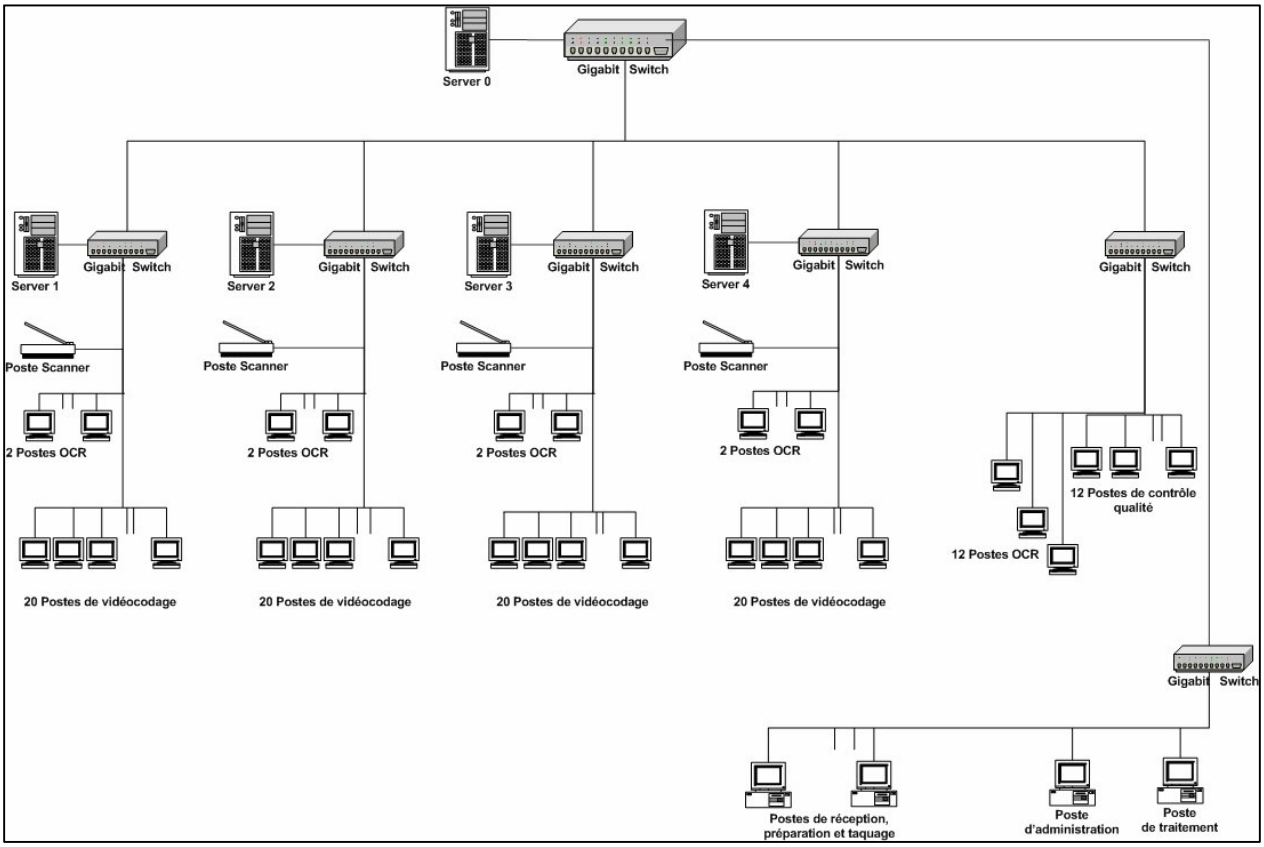
During this phase a new Center of Automatic Document Reading was created. In about two months state-of-the-art software (Digi-LAD) and hardware (110 computers, 5 scanners and 5 servers) were installed in a newly managed area (desktops, shelving, early fire warning system). There after, specialized workshops were organized for potential employees (240 persons). It is important to note that 50% of the work force was temporally hired through a third party.

Graphic 2. First area: Reception of boxes, shelving, drying, loading carts, scanning, and unloading carts

Graphic 3. Area 2: Video coding and quality control stations



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



PRODUCTION PHASE

Data capture from questionnaires of the Census of Population and Housing 2004 was separated into three periods. During the first period urban and rural population count were established during one month of production. The second period was intended to capture data (only Arabic numbers) from questionnaires of households and housing A3 during a 6 months period. The last period, of 12 months, was left to capture data from questionnaires of households and housing A4 containing both Arabic numbers and letters. This separation, allowed for a timely diffusion of available data. Indeed, open questioned to be answered in Arabic letters were on a separate sheet to be exploited last.

Table 1. Census of Population and Housing 2004 questionnaires types, volumes and allocated time to extract information

Documents	***Type***	***Volumes***	***Fields /document***	***Deadline***
Urban and rural population count	A4 : Arabic numbers	38 000 documents (1 document = 21 pages A4 two sides)	3051	t0 + 1
Questionnaires of households and housing	A3 : Arabic numbers	6 800 000 documents (1 document = 1 questionnaire A3 two sides)	248	t0 + 7 months
	A4 : Arabic numbers & letters	5 800 000 documents (1 document = 1 questionnaire A3 two sides)	12 54	t0 + 18 months
Separately counted population	A3 : Arabic numbers	12 500 documents (1 document = 1 questionnaire A3 two sides)	260	t0 + 7 months
Nomad population	A3 : Arabic numbers	40 000 documents (1 document = 1 questionnaire A3 two sides)	245	t0 + 18 months
	A4 : Arabic numbers & letters	40 000 documents (1 document = 1 questionnaire A3 two sides)	12 54	t0 + 18 months
All documents		39 888 000 pages (A4)		t0+ 18 months

The following table shows that a dynamic planning of production was in places to adjust for gained or lost time with regards to the final deadline. During the first month, only 40% of the planned production was achieved due to organizational issues. Therefore, objectives were raised in the following months to catch up. Since January, a steady production insured an exploitation of about 330 districts, the equivalent of 60 000 documents (2 pages of A3) per day.

Table 2. Dynamic planning of data capture from questionnaires of households and housing A3 by district (180 households on average)

	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%

* District: 180 households on average

EMPLOYEES OF THE MONTH

In order to improve production a rigorous procedures of quality control as well as productivity measurements were in place. Computerized modules were used in daily basis to identify employees of the month to whom incentives of as much as 20% salary increase were provided. A monthly news letter was published to keep employees informed of work progress in its different steps.

Graphic 5. Employees of the month in different production line steps

Technical supervisor
Mr. Majid MRANI

Fonctionnel supervisor
Mlle. Zohra KARIM

Reception
M. Rachid BOUDERSA

Inter-document control
Mr. Mohamed AYAT

Scanning
Mme Saida MEKTOUM
Mr. Ali AGOUZOUL

Quality control
Mlle. Hanane ELHAIRECH

Recognition errors video coding
Mlle. Naima TAOUFIK

Logical errors Video coding
Mlle. Hanane EL Kasbioui

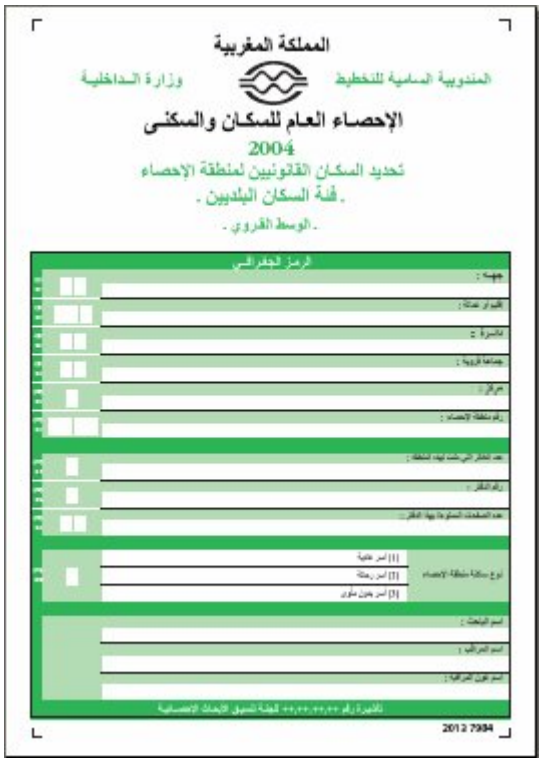
Data export
Mr. Adil Messoudi

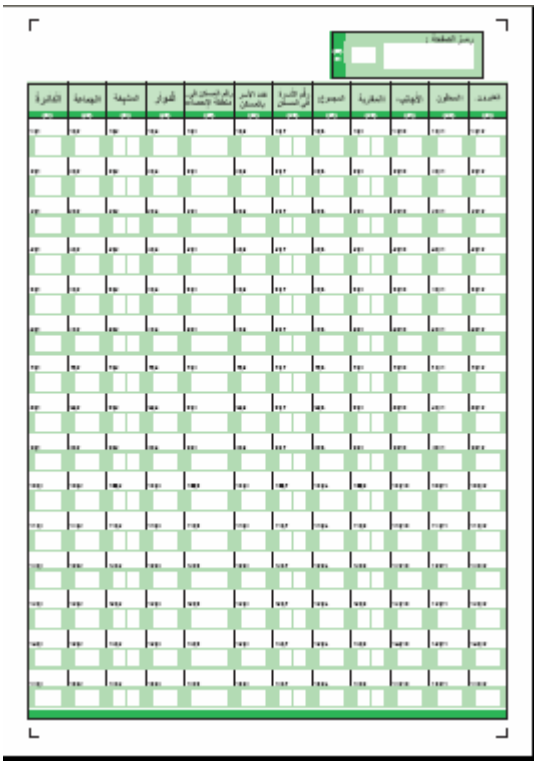
CONCLUSION

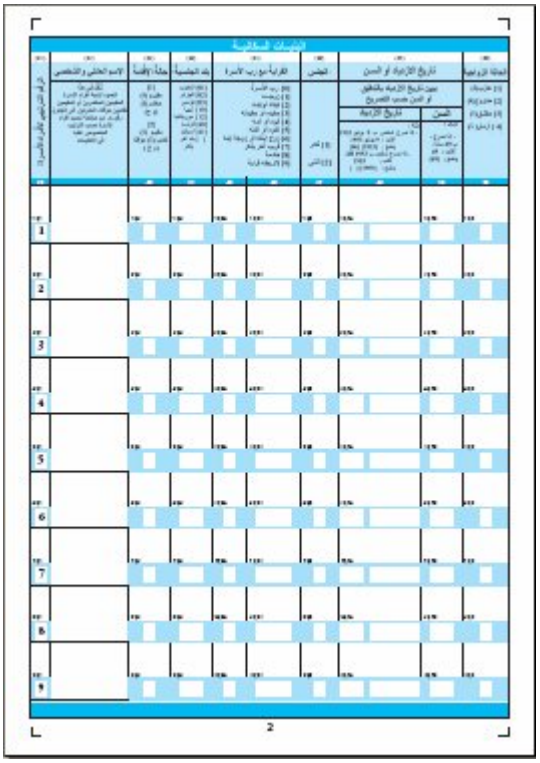
Moroccan experience in capturing data from all questionnaires of Census of Population and Housing 2004 was deemed as a successful one, with objectives achieved ahead of schedule. Furthermore, data from various surveys (e.g., survey on older people) are being now captured using the newly acquired technology.

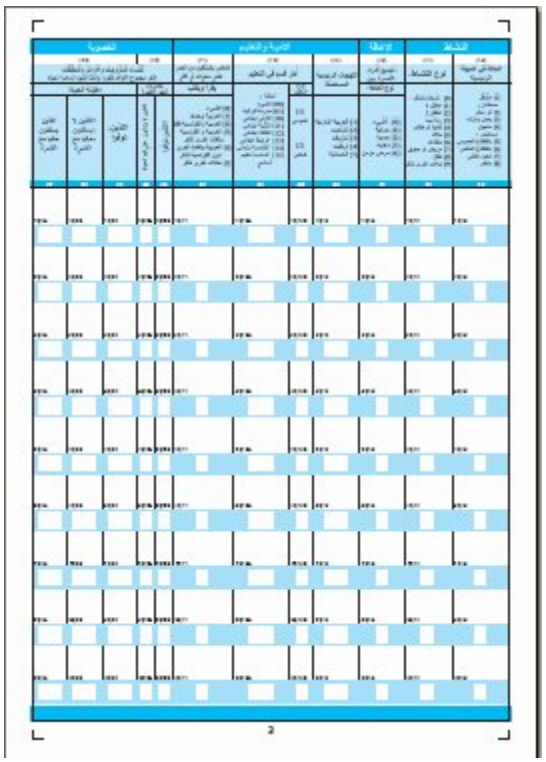
Finally, it is worthwhile to mention that data availability at narrow geographical scale was most useful to the Haut Commissionaire of Planning researchers to elaborate poverty rate and human and social development indicators at the neighborhood level, while combining exhaustively available indicators with income variables that are available only at a higher aggregation level. Furthermore, Geographic Information System was used to insure a wider data diffusion and usability (i.e., Census 2004 in Numbers, Charts and Maps").

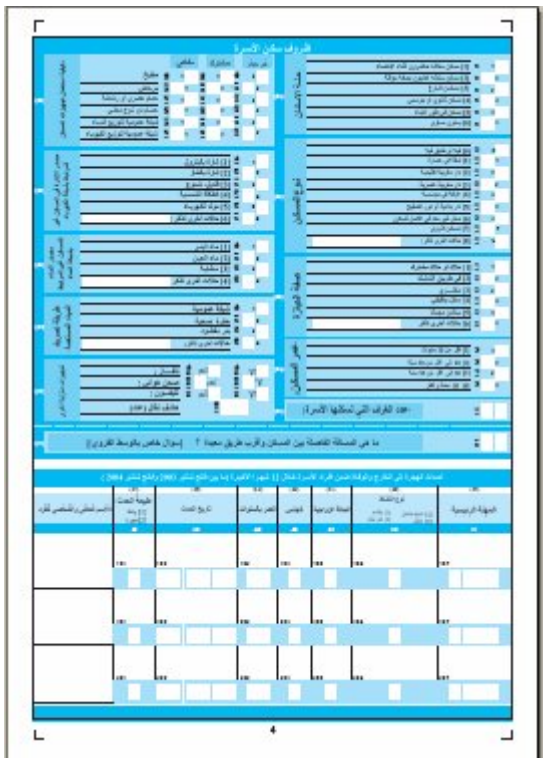
Graphic 6. Questionnaires of Census of Population and Housing 2004 of Morocco












المملكة المغربية
المندوبية السامية للتخطيط
وزارة الداخلية
الإحصاء العام للسكان والسكنى
2004
ورقة الأسرة والمسكن