STATISTICS DENMARK

Statistisk sentralbyrå
Statistics Norway

Statistiska centralbyrån
Statistics Sweden

MZ:2012:07e

CSPro and data processing of the Continuous Multi-purpose Survey INCAF

Report from a short term mission to the National Statistical Institute of Mozambique, Maputo Mozambique

27 August - 5 October 2012

within the frame work of the

AGREEMENT ON CONSULTING ON INSTITUTIONAL CAPACITY BUILDING, ECONOMIC STATISTICS AND RELATED AREAS

between

INE and Scanstat

Julio Ortúzar

INSTITUTO NACIONAL DE ESTATÍSTICA

Ref: Contract DARH/2008 /004

Oct, 2012

SERPRO S.A.

Julio Ortüzar

joc@serpro.com

Contents

Summary

During the six week period at INE-Mozambique, the following tasks were performed:
(1) A complete re-structuring of the old INCAF system, providing more freedom to the personal interviews in the sense household members could be selected to be interviewed according to the interviewers' needs.
(2) Implementation of the Operation Control System to add a systematic administration of the interviews and facilitating the flow of data from the interviewers to the supervisors and from them to the central server.
3) Using CAPI to improve the quality of data (implementation of the assisted coding and consistency checking to the INFOR module (INFORmal economic sector of the population).
(4) Development and implementation of the INFOR module.

Tasks remaining for future activities are:
(1) Improvements to the previous system implementation.
(2) Periodic maintenance of the current system.
(3) Preparation of the common format and coding system for the different waves of the survey.
(4) The production of tables for analysis and control of the fieldwork.
(5) Data dissemination using the web as the most convenient way to spread out the tabulations and other documents.

Resumo em Português

Durante o período de seis semanas no INE-Moçambique, foram realizadas as seguintes tarefas:
(1) A completa reestruturação do antigo sistema INCAF, proporcionando mais liberdade para as entrevistas pessoais no sentido que membros do agregado familiar poderiam ser seleccionados para ser entrevistado de acordo com as necessidades dos entrevistadores.
(2) A execução do sistema de controle de operação para adicionar uma administração sistemática de entrevistas e assim facilitar o fluxo de dados dos entrevistadores para os supervisores e para o servidor central.
(3) Uso de CAPI para melhorar a qualidade dos dados (implementação da codificação assistida e verificação para o módulo INFOR (INFORmal sector económico da população) de consistência).
(4) Desenvolvimento e implementação do módulo INFOR.

Tarefas restantes para actividades futuras são:
(1) melhorias para a implementação do sistema anterior.
(2) Periódica manutenção do sistema actual.
(3) Elaboração do formato comum e sistema de codificação para as diferentes ondas da pesquisa.
(4) A produção de tabelas para análise e controle do trabalho de campo.
(5) Divulgação de dados usando a WEB como a maneira mais conveniente para difundir as tabulações e outros documentos

1. TASKS CARRIED OUT DURING THE 6 WEEK PERIOD

The main objective of this mission was to help INE in the implementation of a continuous multi-purpose system, more specifically, a data capture system. A secondary objective but not less important was/is to help INE in designing the basis for a complete processing system for this important survey, including data consolidation, data conversion to a unique format for all the different waves, data tabulation and data dissemination.

During the first days of the mission, a quick assessment of the Data Processing situation was carried out. The results were not promising since it was apparent that: (1) the data capture implemented for the first wave was simply a collection of programs without any link among them; (2) the programs were electronic questionnaires without any consistency checks, capturing the data of each questionnaire: the dwelling (Agregado Familiar), the questionnaire to capture data of all household members 5 years old and older, and the questionnaire to capture the data of all household members 5 to 17 years of age working; (3) the last two questionnaires, being of the same household members enumerated and completely identified by the first questionnaires should defined who should and who shouldn't be part of each of them; however, not having any link among them, this important task was not checked.

The first evaluation gave around 2.5% of the persons of five years and older (second questionnaire) and a 17% of the third questionnaires (5-17 years of age working) missing.

Another important issue was also evident: there was lack of organization in the fieldwork operation.

After the first assessment, a meeting with the Director, Arao Balate, was hold to explain the current situation and to propose my work plan to overcome at least the most important deficiencies of the INCAF data processing operation.

During the meeting, the main problems of the data processing operation were explained and the following plan was presented:

1.1 Main Task: Organization and Order to the Fieldwork (Data Capture)

The most important and urgent task to be carried out was to organize and bring order to the fieldwork and data capture which at that point was rather chaotic. To achieve this goal, I offered the Operation Control System designed and developed by SERPRO. This system has been implemented in several countries –many of them in Africa- and there would be a very important saving in time and resources if we could implement it for the INCAF project. The Control system would require an important adaptation phase since the survey was already in progress, but the results should compensate the

effort. There was however some basic requirements that we would need to fulfill in order to succeed with the adaptation:

- A sample data file would be needed with all the information required by the very demanding goals of the INCAF panel survey. Since it was a continuous survey, not only the geographic locations of the households to be interviewed were needed but also the addresses of each household along with the names and basic demographic characteristics of their household members would be provided. At the end, I decided to use the household data file produced during the first round to generate the Control data file. During this effort, several deficiencies of the data were unveiled but finally, we succeeded creating a partially completed file. Partially because the first wave was still underway and therefore, the final Control file would need to wait for the completion of this round.
- Detailed information regarding the composition of each group (names of the supervisors with their respective interviewers' identification) as well as the clusters assigned to each of the supervisors. This information is crucial to create the correct environment for the Control system to operate successfully.
- The Control system had been designed to operate using the FTP (File Transfer Protocol) to transfer data and software from the head, the IT Director in charge of the INCAF project, down to the interviewers and vice versa. Thus, the supervisors and interviewers needed to establish a simple network connection to be able to use our connection utilities (server and client). The simplest and easiest way was the connection using a twisted pair cable through the RJ-45 interface. Since the tablets in use by the interviewers did not have the RJ-45 to connect to the netbooks used by the supervisors, INE agreed to get one adaptor per supervisor. There would be plenty of time to get the adaptors and, at that point, there was no problem envisaged.
- A complete explanation and documentation of the Control System was given to INE. The documentation provided the guide to manage the whole system and it is simple enough to be understood by any IT person but particularly by the IT project director. This document is attached to this report as Annex 1.

1.2 Using CAPI to improve quality of data

As stated before, the current implementation was just an electronic questionnaire and therefore, there was no gain in the CAPI type interview. To take full advantage of the opportunities that CAPI offer to improve the quality of the data gathered, there were two tasks that could be incorporated to the current programs: the assisted coding to open-ended type variables that in this project were way too many, and the consistency checking during the interview. From these two tasks, only the first one could be implemented during my first visit, due to time restrictions. The second would be left pending for the next period. SERPRO is the owner of the CSProX software, an extended version of CSPro, which is capable of executing the assisted coding (among many other extensions that the public system doesn't

have). Thus, it was also agreed that CSProX would be used to take full advantage of the CAPI system (at no cost for INE).

The assisted coding algorithm created, designed and implemented by SERPRO is a powerful methodology that allows getting to the proper code of open-ended variables in a simple and quick way as follows:

- The user, in this particular case the interviewer, enters one or more keywords that describe best the description or gloss of the variable being coded. The system applies to each keyword the auto-complete algorithm displaying from the dictionary of terms stored for the variable, all those that have the same prefix; as the operator adds more letters to the keyword, the list of terms decreases in number until it arrives to only one term/keyword. Normally two or three letters produces a short list from which, the proper keyword can be selected. If the keyword entered is not in the dictionary of terms (created for each variable to be coded based on the code description), the term entered does not typify the description of the variable and it should not be used. Many times, a synonym can be used instead (i.e. teacher or schoolteacher, etc.). Once the first keyword has been entered, there are two options, search for all the descriptions than contain the term just entered, or add one more keyword to better typify the description. In many cases, one keyword will produce a list of descriptions short enough to select the desired one; if the list is too large, the system will ask for a more suitable or better qualification of the description. In our specific case, some of the lists were very complicated since included chemical products that usually require of two to three keywords to produce a short list of alternatives.
- The method as explained will analyze the keywords entered by the interviewer and if they are vague as to not producing a selectable list, will require a better description. This is important since the alternative, to simply enter the description to be coded later in the central office, will inevitably produce a code that is not exact. Thus, this approach is one step forward in the task to improve the quality of the data gathered.
- Besides the benefits already pointed out, the autocomplete algorithm provides a quicker alternative to typing the complete keywords, particularly using the tablets available for conducting the interviews.
- Lastly, there is also a very important saving in time at the central office since the data coming from the field are ready for analysis.

1.3 Restructuring of the existing data capturing programs

As mentioned before, the three main programs to conduct the interviews were used independently one from the others. This implied that in many occasions, one person, that had to answer for instance the three questionnaires, had to wait until the proper turn to answer the first, then the second and finally

the third questionnaire. This fact was cause of discomfort, fatigue and a good reason to reject the long interview.

One important task was the restructuring of the programs into two applications: the household data capture module and the Personal interviews, which merged into one application the three questionnaires already mentioned: 5 years and older household members (one interview for each member), the 5-17 years of age currently working, and the newly INFOR or Module for the INFORmal sector of the population (members who were working for their own –with or without employees). The following sequence of screen shots will help to understand the importance of the restructuring task.

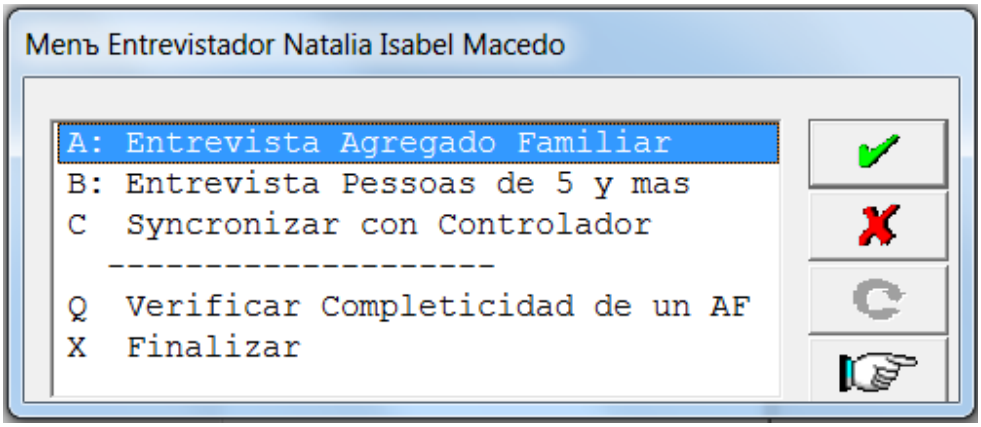

Figure 1: Interviewer Menu Screen

The first two options offer the interviewer to carry out the household interview and the Personal interview (for 5 years and older). The third option permits to synchronize with the supervisor machine to transfer the data from one machine to the other and/or to update the software in the event an updated version was produced. The fourth option, a very important one, permits the interviewer to check all the interviews conducted in one particular household and see if there are one or more incomplete interviews.

Selecting the first option, the interviewer will get a list of households to be interviewed as follows:

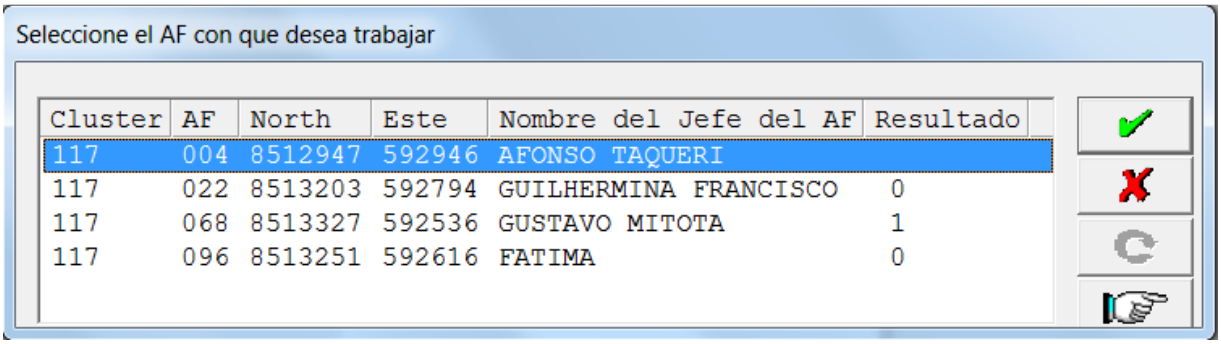

Figure 2: List of households to interview

This interviewer had four households to interview identified by the cluster number, the household number followed by the GPS coordinates and the name of the head of the household. At the end, the result of the interview was also displayed so the interviewer could see how much work was left (the third member shows that the interview has been completed and the other three had not started yet).

By selecting any of the three households with no interview yet, the system will start a new interview providing all the basic information for that household obtained from the control file. On the other hand, if the interviewer selects the one with the completed interview, the system will issue a warning message and will retrieve the finished interview to inspect or to make any changes needed.

If the interviewer selects from the initial menu the option to interview 5 years and older (personal interview), the system will display first the same screen shown on figure 2 so the interviewer has the chance to define the household s/he wants to work with, and then, the list of households member(s) that are 5 years or older.

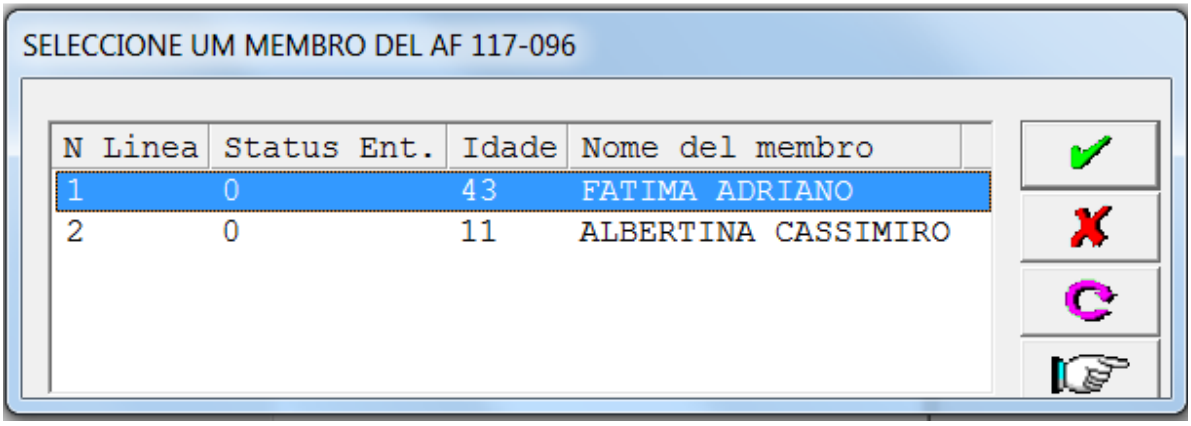

Figure 3: List of household members to interview

The screen shown as figure 3 is the corresponding to select the Fatima household (last shown on figure 2) and corresponds to all the household members older than 4 years of age.

It is important to point out that here the interviewer is free to select any of the list members disregarding the order in which they are listed. This is indeed a great advantage since the member that is currently available for interview can be started first. The selected member will pass through the different questionnaires in a continuous stream and, if s/he doesn't qualify for a particular questionnaire, it will be simply skipped.

Again, the result of the interview is also displayed (Status Ent.) so the interviewer can be aware of the work status.

Finally, the interviewer has the option to check the completeness of interviews for a given household. The household to check is selected from the same screen shown on figure 2 and the system, after checking the Status or results of the various personal interviews, will display the following screen:

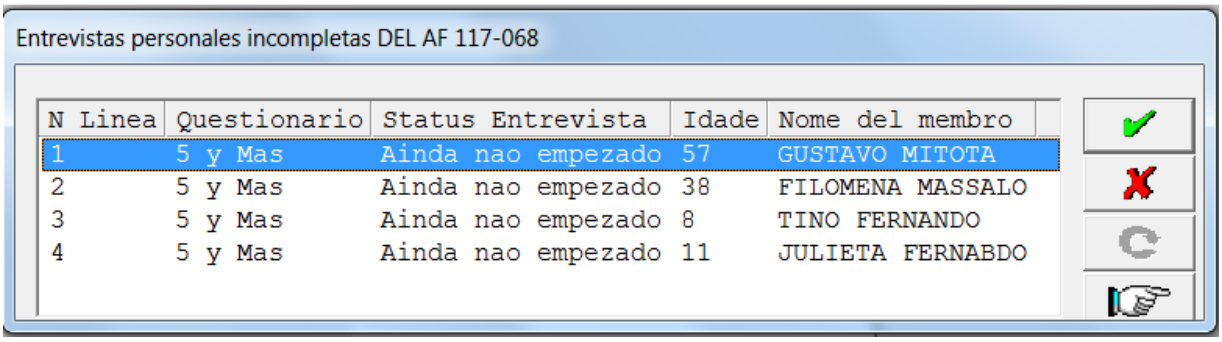

Figure 4: Completeness of the individual interviews

This long explanation is given because it clearly shows the importance of the restructuring performed and focuses the effort in providing the tools to conduct the interviews in a simple and efficient way. This schema assumes that the information provided by INE regarding the household composition and addresses are complete and correct, and that the teams' information with the clusters assignment is trustful. Those two files are the foundation over which the system has been designed and implemented.

1.4 Developing of the INFOR module

This module has great importance both for INE and for the government since it focuses the attention on the study of socio-economic factors of the people who belong to the economic informal sector (most unstable) needing special attention from the government.

The module is rather large and complex; it has 5 data entry screens for the Agricultural module and 19 for the non-agricultural module, meaning approximately 30% of the whole survey.

One of the mayor problems found in the development of the CAPI software for this module was that at the time, the questionnaire was still in that stage of making refinements and changes. This implied that in some occasions, the approach to implement the application had to be drastically changed. It is important to point out for what it implies that the last changes to the questionnaire were done on October 10th and the new questionnaire was sent to me on October 24th. I only heard about the existence of this new questionnaire by talking with the project IT chief regarding some problems with the INFOR module. He was referring to this changes but he never sent the new questionnaire to me.

Well, the development of this module was painful because of lack of help of my counterpart. It should've been easier if I would've counted with the necessary contribution, particularly testing the system.

2. FUTURE IT ACTIVITIES

2.1 Improvements to the previous system implementation

As implied by the preceding paragraphs, the work already done is pretty large considering the time frame in which it was performed. However, improvement of the previous system still needs to be done in terms of adding the necessary consistency checking logic. At least one week of the estimated three weeks to add the necessary checking has been used in the final tuning of the current system. In terms of total time allocated to this task, at least 4 weeks should be budgeted.

2.2 Maintenance of the current system

INCAF is a large an ambitious project and moreover, it changes from wave to wave adding new modules to the existing ones and/or suppressing past modules. Therefore, the system is constantly changing, requiring a constant maintenance activity.

Although the sample over which the survey is performed should not change, some new households can take over the place of old households that have moved out. Thus, there is a future problem in terms of analysis and comparability of different waves unless some actions are designed and implemented.

I would allocate another 4 weeks of time to (1) strictly maintenance; (2) future module developments for the third wave although to make fair estimation, it would be necessary to see the questionnaire.

2.3 Preparation of the unique format or layout for the different waves

In order to prepare the reach information gathered, it's necessary to design the common denominator of the data files gathered on each phase. This implies that, physically speaking, the same data dictionary will work for any of the various wave's data file.

This will help to have only one export application to the three main statistical packages (SPSS, Stata and SAS) to any wave's data file. Furthermore, the resulting data file on any wave should be merged into one final data file having as identifier the wave number. It will also facilitate the time comparative analysis between the different waves.

I would allocate a 1 week time for this task.

2.4 The production of tables for analysis

Time should be allocated to this task but not having the specification of them, is difficult to make any estimation. What is clear is that judging by the amount of information gathered, the number of tables will be immense.

My idea would be to create an integrated system that could be easily handled by a non IT specialist, menu driven, that could take care of the data integration, data export and data tabulation. Some of the first tables should be oriented to provide indicators about productivity of the different teams, completeness of the interviews making sure that there are no results of the interview "in progress", etc. These tables in fact should be designed and implemented as soon as possible so the current wave can be under constant vigilance soon. Since there are two types of questionnaires, the productivity table should contemplate different units by team and by interviewer: number of households interviewed and number of household members 5 and older interviewed.

2.5 Data dissemination

If I understand correct, the intention of INE is to create a web design system to diseminate the reach information gathered. Obviously this effort could be shared with all the other projects or studies carried out by the National Statistical Institute.

Once again, here we have a very big difference in cost depending on what we want to do. Probably, the most conservative idea would be to simply publish the tables or most important tables classified by subject. Maybe, we could have a small engine to search by different keeys (i.e. subject -education, activity, year, etc.-) and the search engine would provide a list of the information available for each set of parameters.

This project could be a three or 4 year project that could start providing the minimum required, and adding more complexity or power each year.

3. FINAL REMARKS

I think it's important to emphasize that the task undertaken was just too large to be carried out in the estimated time. There were just too many loose ends and perhaps, it was a misjudgment of my part to think that the task undertaken was going to be less time consuming. The fact is that it demanded an incredible effort of my part to have the system working as it is now.

Perhaps I was expecting a greater help from my counterpart than what I ended up getting. Throughout my long and vast experience, I have learned that is not good or not constructive to express a negative opinion about one's counterpart but in this particular case, I feel obliged to do it since I think he is not the proper person to direct the IT activities for this important project. My opinion is that he either

didn't want to cooperate or he was unable to do it. The simple fact is that the few tasks that I specifically ask him to do, he either didn't do them or simply were not done properly.

The main problems found and made more difficult my work there can be synthetize as follows:

3.1 Testing the system

A crucial task that needed to be done was a thorough test of the whole system. In a CAPI survey, different issues need to be carefully tested: the logic and particularly the skips pattern, the error messages (still there are error messages in Spanish since I'm unable to write Portuguese, and the questions that need to be displayed for the interviewers (same as before), etc. The fact is that only once I was here at my office I had the feedback of the different tests conducted by him or his team.

Issues like the assisted coding required special attention in the testing, given the great responsibility it implied. The fact is that the descriptive text of the different variables to code was in Portuguese and the special characters like accents and inflections do not have an ASCII representation. Thus, I ask him to get them reviewed so those especial characters were suppressed. Only when I was back here at my office, one file out of three was sent as required.

3.2 Sample data file

As it has been explained throughout this document, the sample data file, a crucial file to have the whole system working harmonically, was a main preoccupation during my stay in Maputo. This file was generated using as input the household data of the interviews done during the first wave of the survey. However, since the very beginning, there were problems because of duplicated keys or any other reason. Finally, before my departure, we got a working file that apparently met the needs. This file was not complete because the fieldwork was still under way but at least, the remaining part was supposed to be integrated to the main file and sent it to me so I could produce the final Control data file. This was done as planned and finally, not too long ago, we were able to have this important file incorporated to the system.

To my surprise, during the first week of the second wave's fieldwork, one of the supervisors reached me to let me know that a whole cluster had the wrong households' members and characteristics; in simple words, all the households of that cluster had the wrong cluster number. This could be a serious problem or it could just be an isolated problem. Let's hope that the same experience won't happen again. The result of this accident is that once I got the correct information for the cluster, I had to manually fix the interviewer's control file and then the supervisor's control file. This is a very delicate task since the control file is storing all the historic work performed by the different players in this system. If the same

thing happens later, the work will be much more difficult since more information will be added. Let's hope that we won't find any other problem.

3.3 Teams data file

As pointed out earlier, the teams file stores all the team composition providing very important information regarding the cluster assigned to each supervisor, and the machine id and code for every interviewer and supervisor along with their respective names. This file is used every time there is synchronization between two machines and the information provided allows the system to accept or reject the synchronization process. Furthermore, the clusters assignment for each group permits to generate the control file.

The teams file provided to me had several errors like duplicated machine ids and unassigned clusters. The negligence in the generation of this file is simply incredible. Up until now, I have not received the teams' file. It does need to be fixed or otherwise, the missing clusters won't be interviewed.

Annex 1 ToR

June 14, 2012

TERMS OF REFERENCE

for a short-term mission on

CSPro and data processing at the Continuous Multi-purpose Survey

June 2012

within the

AGREEMENT ON CONSULTING IN INSTITUTIONAL CAPACITY BUILDING, ECONOMIC STATISTICS AND RELATED AREAS
between INE and Scanstat.

Consultant: Julio Ortúzar

Counterparts*:* Arão Balate and the Continuous Multi-propose Survey team at INE

Background
The yearly survey program at INE has since 2000 collected information on employment, time use, household budget, health, victimization, sanitation, water supply, basic infrastructure and buildings etc. with a periodicity of 5 years for each module

Since the beginning of 2000 studies have shown the importance of poverty assessments as instruments for monitoring development. The lack of timely information on poverty, and wellbeing, is at the moment discussed among users where the frequency and timeliness of household surveys within the information system for poverty monitoring being an important aspect. Current information is required to monitor and adapt to on-going dynamic changes.

Studies show that the lack of employment is one of the two most important social problems in Mozambique according to the inhabitant's own judgment. This is especially a problem for the households where the head only has short-term employment, paid on a daily/hourly basis. Employment is also a major concern among students 15 years and older.

For these reasons INE has decided to modernize the system by introducing a new Continuous and Multi-Propose Survey (INCAF) having a Core Module based on Labour Force and Expenditure where different modules like Informal sector, Tourism, Education etc. will be introduced quarterly.

Data collection for INCAF will be made using PC-tablets. Initially CSPro will be used as INE already has experience of that software.

Main reasons for the mission
INE has gained a lot of experience carrying out household surveys. However INE still lacks experience from Continuous and Multi-propose Household Surveys and thus requires assistance in order to successfully implement this new chain of surveys taken into consideration the continuous modus operandi needing a streamlined work process avoiding bottlenecks, errors and delays.

Objective
The overall objective of the mission is to assist INE in the implementation of the new Continuous Multi-propose Survey using international experiences regarding data collection and especially CSPro. A second objective is to help INE in creating a layout for a complete processing system.

Activities
The Consultant is responsible for the following activities:

1. Verify, complement and document the data collection system for the survey.
2. Prepare validation specifications for all questionnaires.
3. Develop a validation program from the specifications prepared in step (2)
4. Test and implement the validation system.
5. Assist with training of Central Staff to use the data collection system
6. Propose a layout for a complete processing system with a 3 month repeat time, covering all sub processes and transformations from data collection to dissemination on the Internet.
7. Help INE in setting up a scheme for further assistance regarding the data processing of INCAF.

Expected outputs
A fully working data collection system adapted to INCAF. The consultant will prepare a written documentation of the above activities;

Beneficiaries of the mission
The mission will benefit INE staff, interviewers and the whole National Statistical System of Mozambique including users of statistical information.

Tasks to be done by INE to facilitate the mission

- Elaborate ToR for the training
- Prepare and supply the consultant with necessary documents and information, like the new visual identity of INE
- Supply good communication conditions for the consultant.

Source of Funding
Project: MPD-2008-0006 – Inquérito Sobre Orçamento Familiar – IOF
PAAO12 – 1.4.2 Inquérito Contínuo aos Agregados Familiares

Timing of the mission
This first mission on will start on August 16, 2012 and last for 6 weeks. Further missions will be agreed upon during the first mission.

Place
The premises of INE in Maputo with possible allocations to the provinces.

Language
Portuguese, Spanish and English.

Report
The consultant will prepare a short final report to be discussed with INE before ending assignment. Statistics Denmark as Lead Party will publish the final version on www.dst.dk/mozambique within 3+ weeks of the end of the mission. The structure of the report should be according to Scanstat format.

Annex 2 CAPI Control System

CAPI Control System

July 27, 2011

1. Introduction

A CAPI (Computer Aided Personal Interview) survey is a complex operation which encompasses much more than just the specific application to capture data. Granted the CAPI application to capture data is the heart of the whole system, but as we know, a chain is only as strong as the weakest of its links. Thus, we should worry about all the different tasks that relate directly or indirectly to the central goal, the gathering of good, sound statistical data.

The software developed to capture the data includes different steps that pave the way to the final goal, primarily taking care of aspects that have to do with the following subjects:

- Helping to ascertain that the sample design frame is strictly observed during the fieldwork in the sense that the correct number of dwelling/household units is interviewed per cluster or enumeration area.
- No dwelling/household or sample unit is duplicated or missing.
- If the sample is a dwelling sample, tools are offered to interview all the households belonging to each dwelling.
- Within each household, a methodology is followed to ascertain that all eligible persons are interviewed (if a household sample is used).

Since we are primarily concerned with improving the quality of the data gathered, all activities that have an impact on this goal should be our concern. The tasks that we will describe in this documentation are all instrumental in reaching this goal, although in some occasions the steps described will not show this evidence so clearly.

The different activities involved in the CONTROL system are classified into three different groups based on the staff members who are responsible for them. All the activities carried out by the Survey Administrator (SA) also called Survey Director, are grouped together and they should be completed before the second group, the supervisors, can start with their own activities. The supervisors' activities will pave the way for the third group, the survey interviewers.

The present system was conceived for a dwelling sample and due to late changes in the sampling design plans it had to be transformed to work with a household sample or simply any type of sampling unit. The changes introduced were merely superficial and the same system

could just as well be used in the future for a dwelling sample or any type of survey that is based on a sample.

The CONTROL system is handled by a set of menus that include the different activities referred to above. These menus aim to assist the staff members in the execution of their various tasks and, in many ways, guide them in the timing and sequence in which these tasks should be carried out.

The following figure illustrates the hierarchical structure of the three different groups of staff members involved in a CAPI survey.

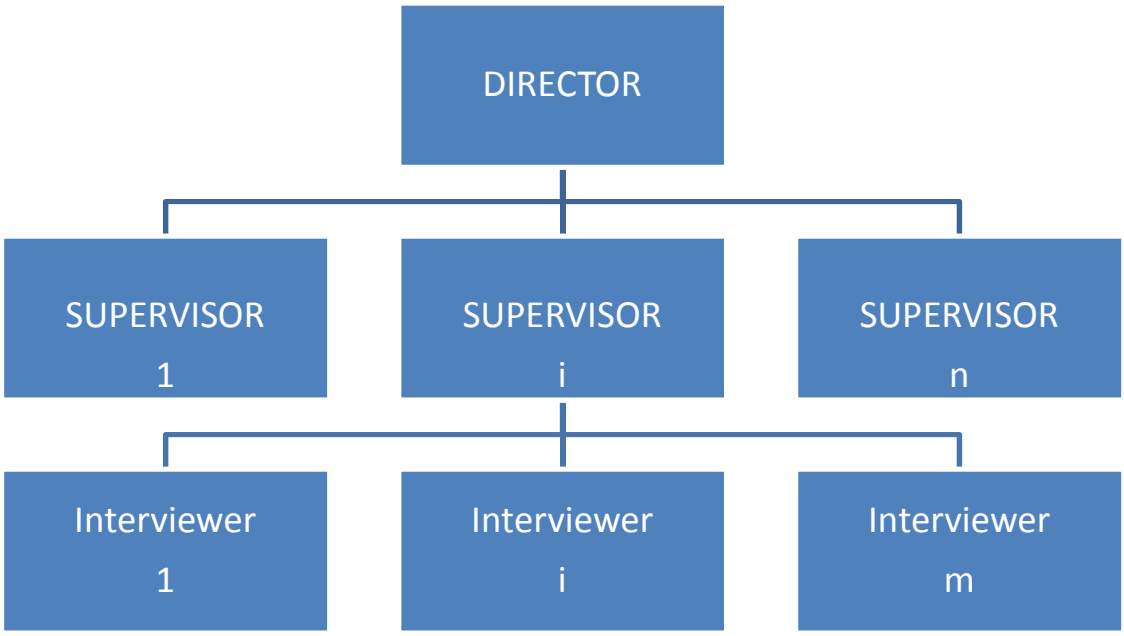

Figure 1: Hierarchical structure assumed by the Control System

The director(s) have one menu that will help them to exercise their director's duties throughout the duration of the project. There might be two types of directors: (i) the development director or also called super-director, who is responsible for development and maintenance of the software applications, the creation of the key data files used by the Control system, the assignation of work to supervisors and the sample data file manipulation. (ii) The director assistant who has only a subset of the super-director's duties.

Normally, the functions of both directors will reside in only one person, the super-director.

2. Machines Setup

Along with this documentation there is some basic software that has to be distributed to each machine belonging to the three different groups identified above on figure 1. The software comes in three ZIP files named: "DISTRIBDIRECTOR.ZIP", "DISTRIBSUPERV.ZIP" and

"DISTRIBINTERV.ZIP". As it is easy to imagine, the first file has to be copied to the Director's computer, the second to all the supervisors' machines and the third to all the interviewers' machines.

2.1 Machine identification

The system requires that each machine has its proper identification and be associated to one staff member. Since we have three types of staff member categories, we suggest that the machine identification (MID) starts with one or more letter related to the group it belongs to. For instance, a supervisor machine might start with the letter "S" or "SUP" showing that it is a supervisor machine. The remaining part of the MID might be up to the director's preference; it might be a code that identifies the machine in the institution's stock or it might be the code of the staff member (we will discuss this concept later in this documentation). The fact is that each machine belonging to the project needs to have a unique MID. In other projects, we have made labels for each machine and we stick it to the computer for easy identification. This will be a physical identification. However, this identification has to be recorded in the hard disk of the netbook/notebook/tablet so it can be accessed by the Control System.

We will take advantage of a commonly used Windows facility called "Environment Variables" to define the MID of each netbook. Then, later, the software will be able to retrieve the MID whenever is needed. To define the MID, you can go through the following steps:

- Click on the Windows "Start" button
- Click on the "Control Panel" on the right side of the Start Menu
- Select the "System and Security" option
- Select the "System" option
- Select the "Advanced System Settings"

At this point, you should have the following dialog displayed on the screen:

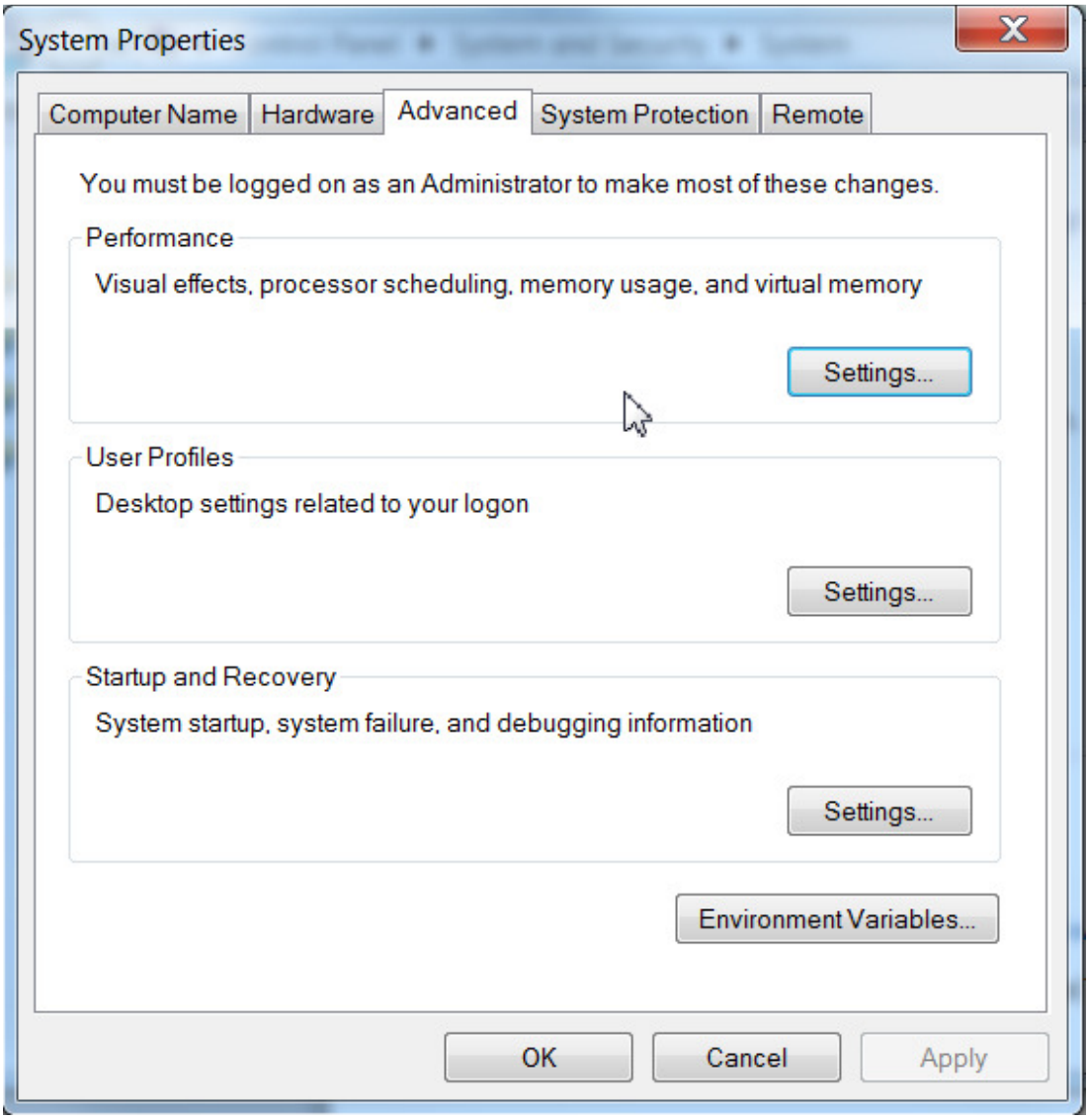

Figure 2: Intermediate step in the definition of the MID environment variable

- Click on the “Environment Variables” tab
- You should have a new dialog like the one illustrated below; note that the dialogs and names might vary slightly depending on the version of Windows your machine has

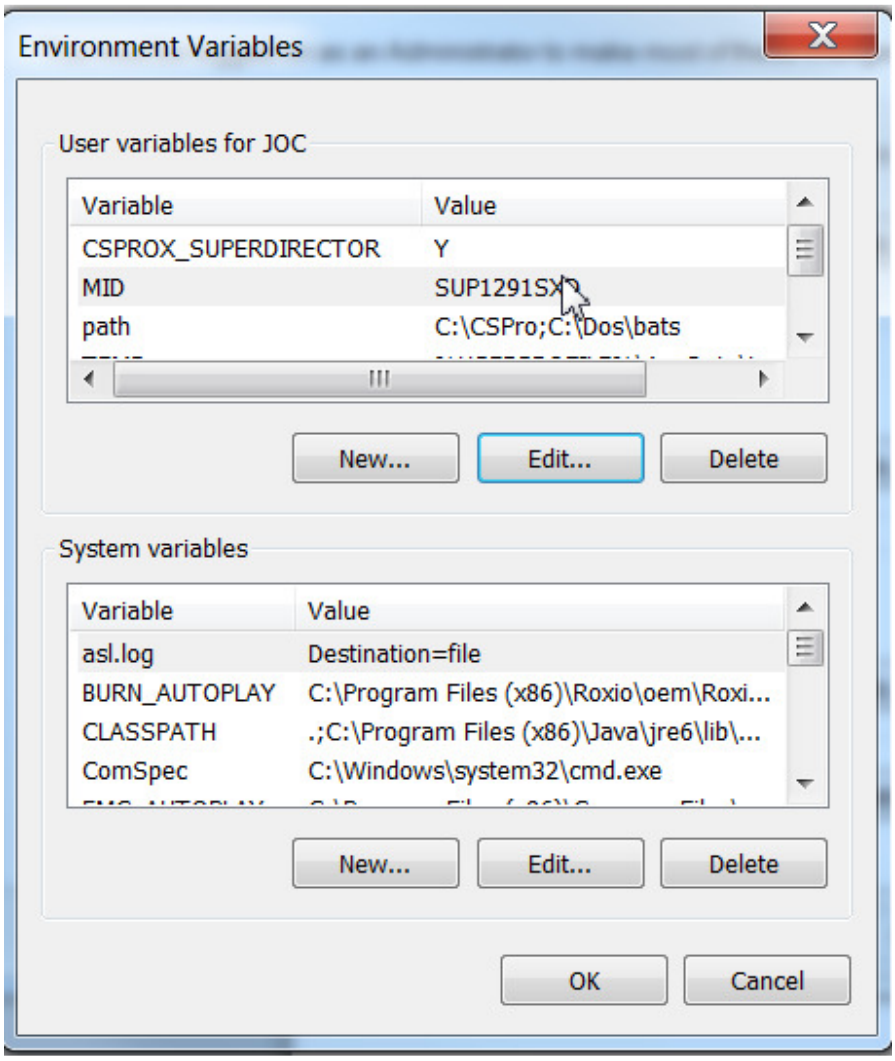

Figure 3: Dialog to define the environment variable MID

The upper window is the one to define the user variables; as you realize, this machine has already the MID variable defined from a previous project (SUP1291SXD); however, the machine you are working with it's unlikely to have an MID variable defied. Thus, click on the "New" tab on the upper window to define the MID variable. Then, you should get a dialog box like the one below:

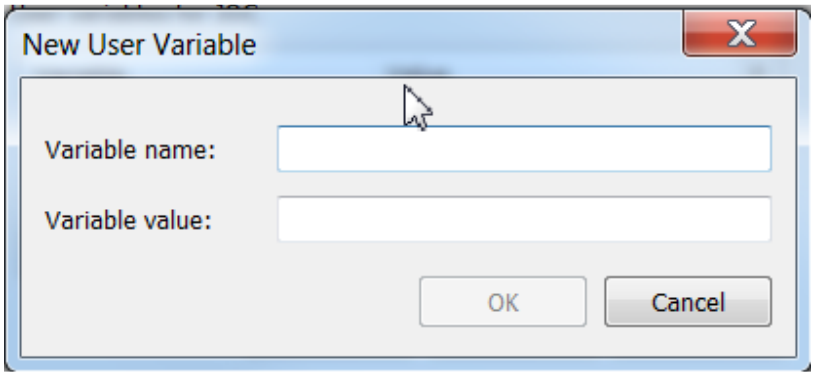

Figure 4: final step in the definition of the MID environment variable

As you might guess, in the "Variable name" box we need to enter "MID" which is the name of our machine id; in the "Variable value" box, we need to define the machine id we have assigned to this machine (i.e. "SUP1291SXD").

After completing these steps, the machine will be identified by the MID name you have selected.

2.2 Definition of the IP number (Windows 7)

The Control system needs to communicate different machines in order to pass information throughout the steps of the pyramid shown on figure 1 above. Interviewers will need to pass data from their machines to the supervisors and supervisors to the interviewer too. Then, if the adequate local communications exist, interchange of information between the project director and the supervisors need to be established.

There are several ways of establishing communication between two computers and the one we are describing here is just one of them; however, we have selected the option that follows because we think is one of the simplest, cheapest and more secure, proving to be a reliable method that has not given any problems in the course of the many experiences in different countries. The method is the connection of two netbooks/machines using of a twisted-pair cable available in most electronic shops through their RJ45 physical interface normally used for Ethernet cables. The total cost of the cable is about US$ 5-10 depending on the country.

The most common network connection will be the one established between each supervisor and his/her interviewers. Since the time needed to transfer the information between them is short, we prefer to do it one interviewer at the time. Furthermore, not all the interviewers come to swap their information at the same time not even the same day.

To get started, we need to setup the network card of the supervisors' machine to a fix IP (Internet Protocol) address: "192.168.1.1" and the interviewers' network cards to "192.168.1.2".

To do this, follow the steps specified below:

- Click on the Windows “Start” button
- Click on the “Control Panel” option
- Select the “Network and Internet” option
- Select the “Network and Sharing Center” option
- Select “Change Adapter settings”
- Double click on the cable network to access the network properties
- Select the “Internet Protocol version 4”
- Define the IP and subnet mask as shown on the next figure

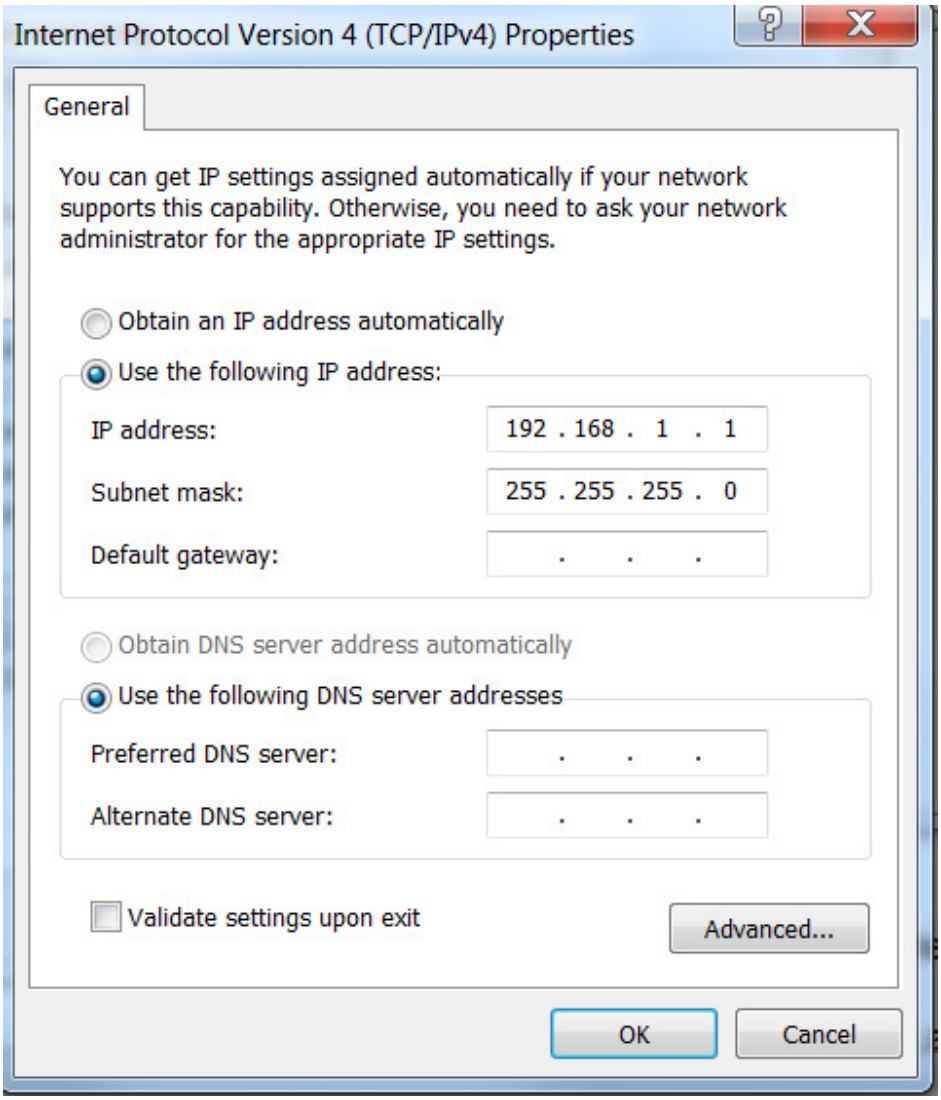

Figure 5: Definition of IP and Subnet mask for the supervisor machine

Follow the same steps to define the IP address and Subnet mask for the interviewers’ machines.

2.3 Folder Structures and files to be transferred

As already mentioned, there are three types of machines in our project: the director's, the supervisors' and the interviewers' machines. Each one of them needs to be set up in similar ways but with some important differences. In addition, we should also be aware that the central server, usually an FTP (File Transfer Protocol) server, plays a very important role in the Control system and will also be analyzed.

Within the machine setup tasks that need to be performed, we need to be aware that each machine type has a predefined set of applications and data files that need to be copied to them. We will also discuss about the files needing to be passed to each machine to guarantee the correct communication between them.

2.3.1 Director's Folder structure

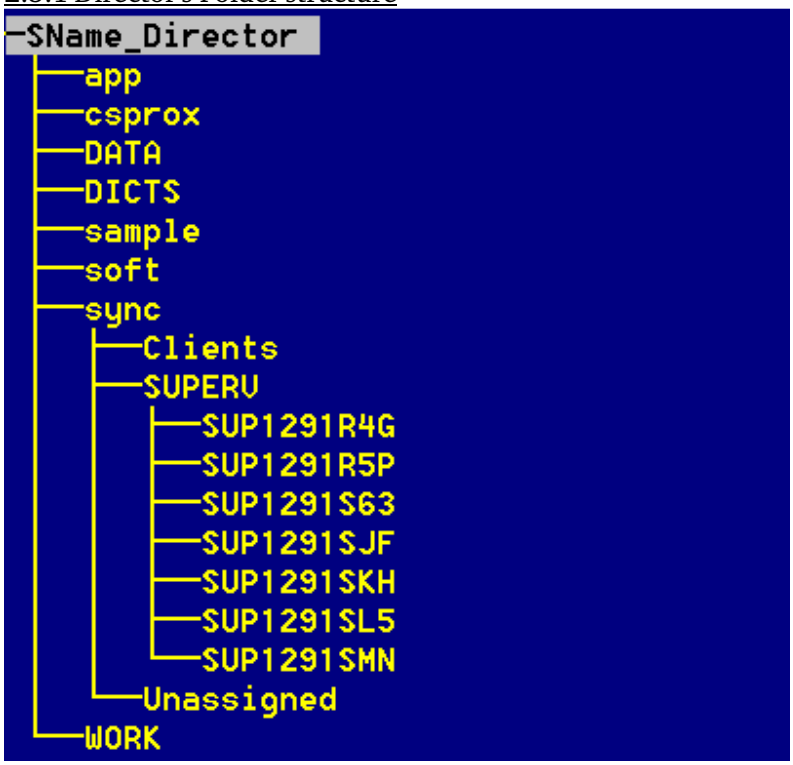

Figure 6: Director's folder structure

App Folder: All the applications that will be used by the Director will be initially copied to this folder; they include application files (*.app) forms or data capture screens (*.fmf), error message files (*.mgf), questions files (*.qsf), and many others. In addition, some specific data files might be copied here if some applications need them.

CSProX Folder: The CSProX folder will be copied here to make sure that the proper version of the system is used. All the applications in App above will execute directly the proper module from this folder.

Data Folder: Generally speaking, this folder will store the two basic data files described later in this documentation (Teams.dat and Control.dat). In some special projects, some other type of data files will also be stored here.

Dicts Folder: All data dictionaries used by the applications in App will be stored here; the exception will be the working dictionaries that are stored together with the application itself (App).

Soft Folder: Some particular CSProX utilities are stored here.

Sync Folder: This folder will store the synchronization utilities along with their respective ".INI" files and their key files to be able to execute them (protected files). If the key is not found, the execution will fail. Two synchronization utilities will be used: (1) the "CSFileSyncClient.exe" which will synchronize the director's machine with the FTP server, and (2) the "CSFILESyncServer.exe" which, at the time, doesn't have any function.

The Sync folder has three subfolders but in the director's functionality, only one is active: the "**Sync\Superv**" folder which is used to store the information from each supervisor. The data stored on each folder comes from the different supervisor machine and has been stored in a zip file and within it, the various data have been concatenated into one file called "AllData.dat". As it is shown on figure 6, under this subfolder, one folder for each supervisor is generated naming them by the respective supervisor's machine id. At the time the director's machine is synchronized with the FTP server, each supervisor's data from the server is copied to the proper folder in the director's machine.

Work Folder: This is a working folder used by the system to perform different operations like copying files for various purposes (i.e. zipping, concatenating, etc.).

2.3.2 Supervisor's Folder structure

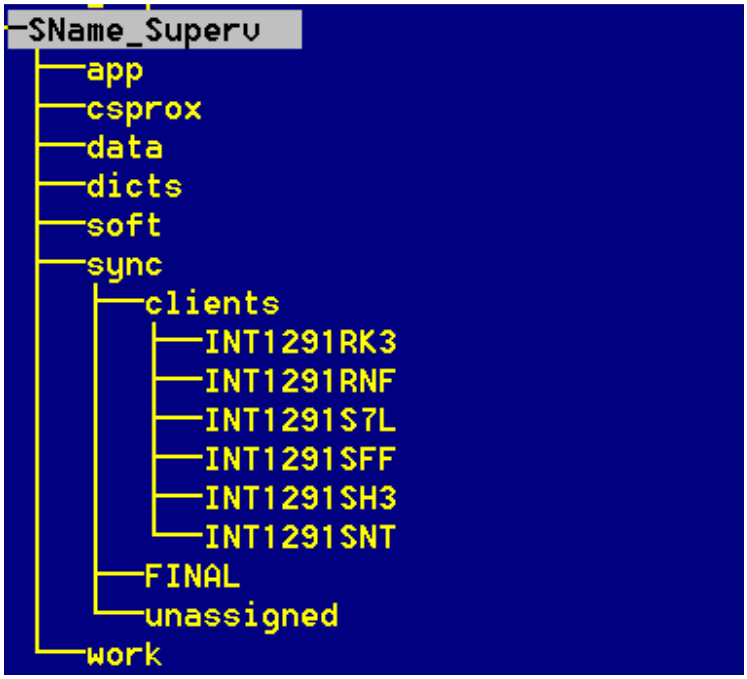

Figure 7: Supervisor's machine folder structure

The supervisors' folder structure is very similar to the director's and thus, we will only refer to the folders that have some difference or peculiarity when compared to the director's folder.

App Folder: the applications stored here have been specifically designed to be used by the supervisor and are just a subset of the ones stored on the director's machine. Most of them are never used by the director although they are also stored on his/her App folder.

Sync Folder: as illustrated on figure 7 above, the Sync folder does not have the "**Superv**" subfolder as in the director's machine; instead, it has the "**Clients**" subfolder and within it, it has one subfolder for each interviewer of the supervisor's team. The interviewers' folders are named using the interviewer's machine id and will stored all the data captured by each interviewer. The interviewers' folders are created the first time the Supervisor_Menu application is run or executed. Note that the communication utilities are stored on the Sync folder as in the director's case.

Final Folder: this is a folder used only by supervisors and it will store all the interviews modified by a supervisor.

Unassigned Folder: It will store those folios or units ids that have been unassigned to an interviewer; these units will add to those units that have never been assigned to conform to a list of potential units to be assigned/reassigned. This folder is only used by the supervisor.

2.3.3 The Interviewers' folder structure

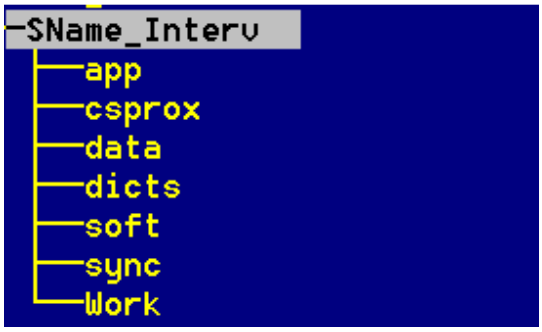

Figure 8: Folder structure of the Interviewers' machines

Data Folder: Besides the "Teams.dat" and "Control.dat" files this folder stores all the interviews captured by each interviewer.

Sync Folder: As oppose to the two folders already mentioned, the interviewers' sync folder doesn't have any subfolders; it just stores the communication utilities.

The remaining part of the folders shown on figure 8 has a similar purpose as the respective ones shown on the director's or supervisors' folders.

2.4 The initial files (Kernel) copied to the different machines

Each set of machines has its own kernel that will provide the basic tools to communicate the various machines and, subsequently, download the complete software. The kernel of the different actors is part of the software that comes with the system and it should be copied to (i) the director/s machine/s; (ii) the supervisors' machines and (iii) the interviewers' machines.

- **Kernel of the Project director:** The file is a ZIP file "DISTRIBDIRECTOR.ZIP" and it has all the software needed by the Director; since this has to be the first machine to be setup, the software included is not just the kernel but all the applications and data files. If along the life of the project, changes are made to the software, the distribution files (kernel) will need to be regenerated. For this purpose, the system has a batch file called "DistribAll.bat" that is located on the root of the director's folders. This BAT will generate the three kernels automatically. In the particular case of the director's kernel, the list of files that need to be zipped is located in a text file called "DistribDirector.txt". If a new application has to be generated or additional data files are needed, the names of each additional file has to be added to the list in "DistribDirector.txt".

- **Supervisors Kernel:** The file is a ZIP file called “DistribSuperv.ZIP” and it does have the minimum applications to be able to communicate with the FTP server. The director will pass the software and data files to the server and from there to the supervisors. Thus, the supervisors’ kernel has the minimum basic software to synchronize with the server and to download all the software needed, actually zipped into one file called “ApplicationsforSuperv.zip”. The list of all the applications that will be zipped on this file can be found on the text file “ApplicationsforSuperv.txt” and it is used as the guide to generate the zip file. If any new application or data file is needs to be added, it should be added to the list of the text file. When the supervisor machine synchronized with the FTP server, two files will be downloaded: “ApplicationsforSuperv.zip” and “ApplicationsforInterv.zip”. The second file will be downloaded to be transferred to the interviewers when they synchronize with the supervisors. Note that the interviewer’s machine will never communicate/synchronize with the FTP server and thus, it has to be passed through the supervisors.

- **Interviewers Kernel:** The file is a ZIP file called “DistribInterv.zip” having the minimum software to synchronize with the supervisor. The application list identifying the kernel applications can be found in the “DistribInterv.txt” file. The remaining applications and data files needed by the interviewers will be downloaded from the supervisor machine when the synchronization between the two machines takes place. As mentioned above, the whole set of applications and data needed by the interviewers is in the zip file “ApplicationsforInterv.zip”.

3 Some important data files that need to be generated

The Control system will rely on three important data file that will need to be created by the project Director in two cases and by the sampler or the project director too the third one.

3.1 Teams data file

The teams’ data file is one of the three very important sources of information in the frame of the Control system operation. The purpose of this file is to show the teams’ structure; this means, for each supervisor, what are the interviewers that will be part of the same team, showing the code, name, machine id, etc. of each of them. Furthermore, based on the sample file, clusters or units under investigation will be assigned to each supervisor. The list of units that has to be covered by each supervisor is entered and stored on this file.

The teams’ data file is the first important definition required from the project administrator (explained later in this documentation). To be able to provide the required information, it is necessary to have a thorough knowledge of the entire project organization.

A team will be made up of one data manager/supervisor (with expertise in the handling of computers, the Windows operating system, and having passed through the training period in the use of this software) plus the necessary number of interviewers that have gone through training in the use of netbooks and the specific CAPI application. The clusters assigned by the project director to this supervisor are also specified here.

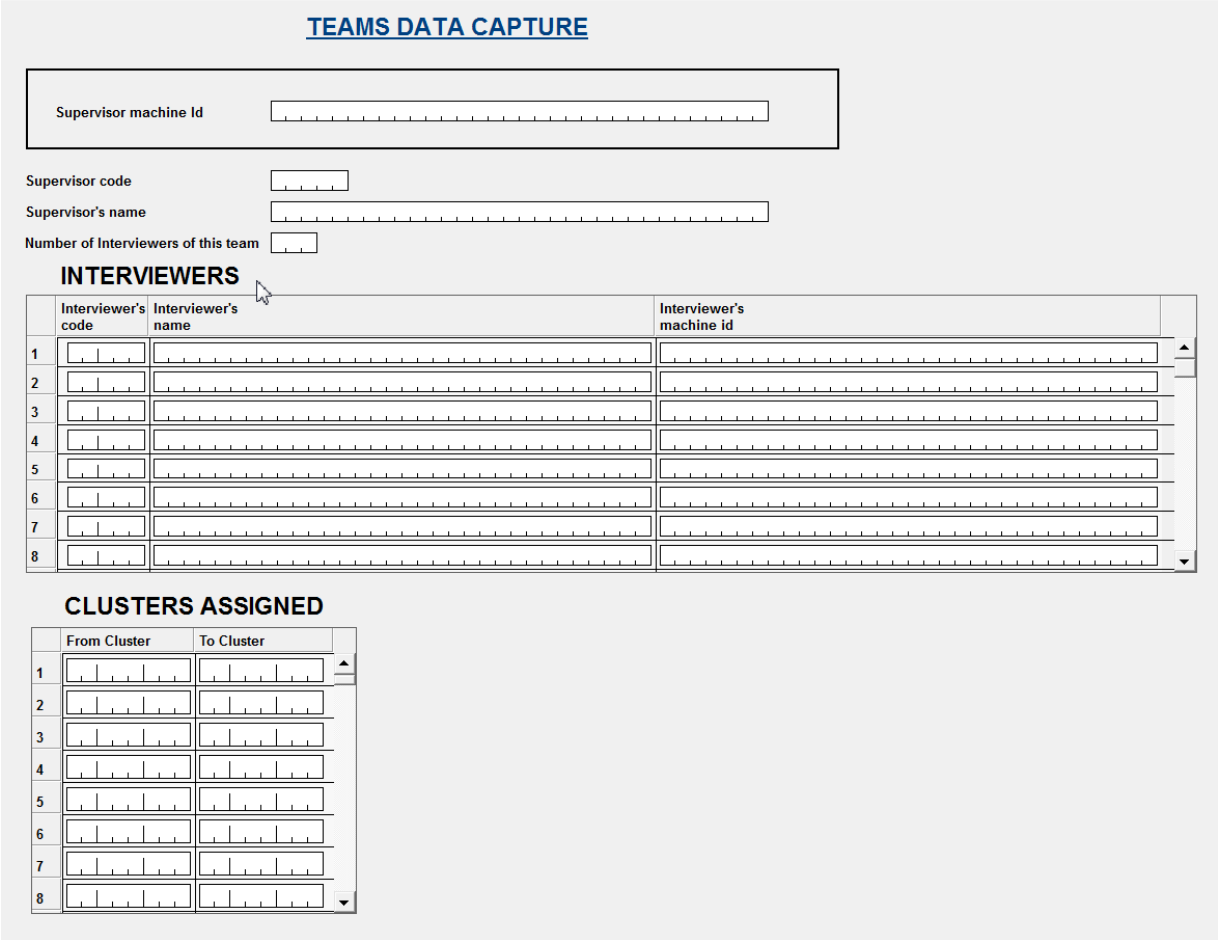

Figure 9: Teams data capture screen

The project director will need to have a clear idea of the geographical area where the supervisor and his/her team will be covering during the fieldwork in order to decide what clusters should be assigned. Although in theory, the clusters can be reassigned later, once the Control data file is generated it's very cumbersome to shuffle clusters around. In fact, every time one cluster is reassigned, the Control file will have to be regenerated.

It's important to point out that the cluster doesn't need to be a sample point; it can be a folio or collection of units that need to be interviewed, hopefully correlated to geographical areas so it's easier to assign to one supervisor/team.

3.2 Sample data file

The sample data file can be the result of an algorithmic/statistical process to determine the units that will need to be interviewed or simply an enumeration of the units that will be visited and the information gathered from. Examples of the type of information that might be part of this file are: dwellings or household units, health facilities like clinics or hospitals, mayor, intermediate and lower geographical/administrative area, urban/rural type location, address, etc.

This file is originated by the sampler or by the survey director and it will be used to generate the second data file that will be our companion for the whole fieldwork operation, the Control file.

3.3 The Control file

This file is the heart of this software and it will be used and updated constantly through the different chain of events.

The generation of this important file is performed using the sample design information together with the Teams data file (already described in 2.3) as inputs, producing the control data file as output. The information stored on this file is relevant for the main purposes of the system and therefore, it is worth to spend some time describing at least the most important one.

There will be a record containing the information shown on figure 7 for each unit of the sample or file listing describing the units that will be interviewed.

The cluster and the identifier have to be interpreted as a unique combination to identify one unit. As it was mentioned before, it can be a folio that will produce a unique identification.

The Geographical levels link the unit with specific geographical area. Since this information is as general as possible, the level 1 might be somewhere the “State” level and level 2 might be the “Province”. The number of geographical levels will be something that is defined by the project/country and if only two areas are needed, the other two should be not applicable.

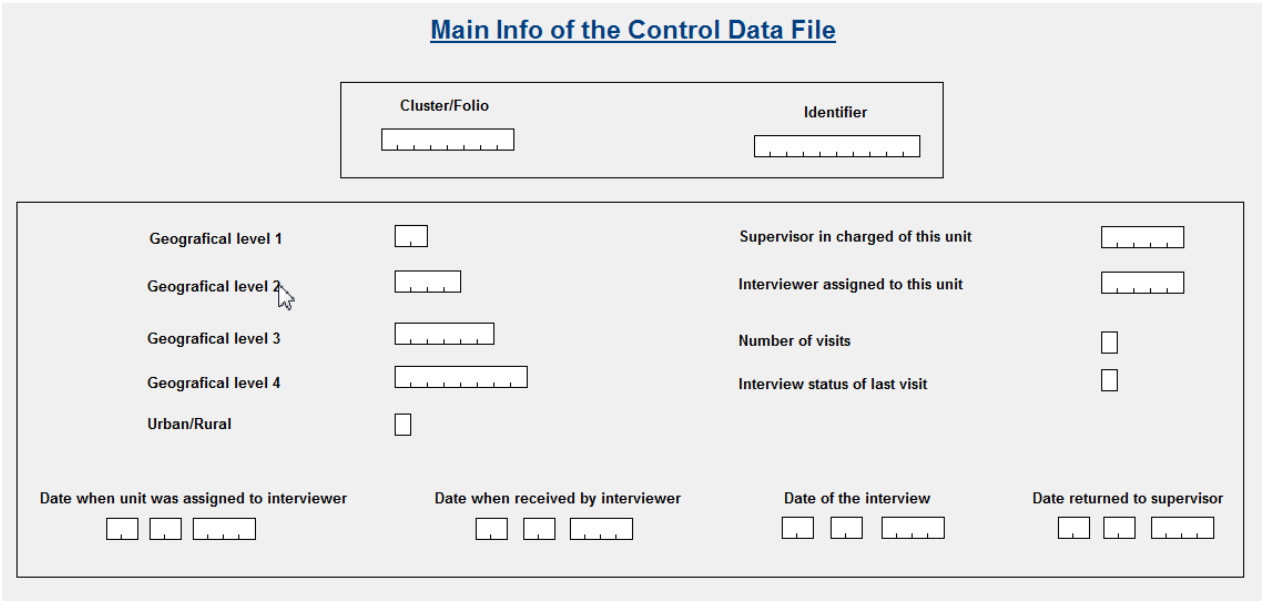

Figure 10: Main info of the Control data file

Again since the variable specifications try to be as general as possible, we define the "Supervisor code" and the "Interviewer code" as a five digit variable. However most of the time the code will be a two or three digit code like we suggest below:

Supervisor Code	Interviewer Code
100	101-107
200	201-206
….	….
900	901-909

The "Interview Status of last visit" is very important since it will be used to show an interviewer and or his/her supervisor the overall status of his/her work for one specific cluster.

- Date when the unit is assigned by the supervisor to the interviewer: is the first action of the supervisor with his/her interviewers and it will show the timing to start following up the unit-interviewer.
- Date when the assignment is received by the interviewer: the data above was just the action/decision of the supervisor to assign the unit to the interviewer; this date is the time when the interviewer get the assignment on his/her machine. From now on, the specific unit will show up on the list to be selected to be interviewed.
- Date of the interview: this is the date when the interview is carried out by the interviewer.

- Date returned to supervisor: this is the time when the interview's data are passed to the supervisor.

With the four dates above plus the unit identification, the supervisor and interviewer's identification we are in the capacity of following up and tell the survey director/supervisors the status of any unit at any time.

4 The Control system driven by menus

As mentioned before, the Control system is entirely handled by menus oriented to three types of staff members, the survey director, the supervisors and the interviewers. Each one of them has a well-defined number of activities that are portrayed on their respective menus.

As we analyze the different menus, we will understand better the concepts and purposes of the Control system. We will start analyzing the Director's menu explaining option by option the different tasks s/he can and should do.

4.1 Super-director's menu

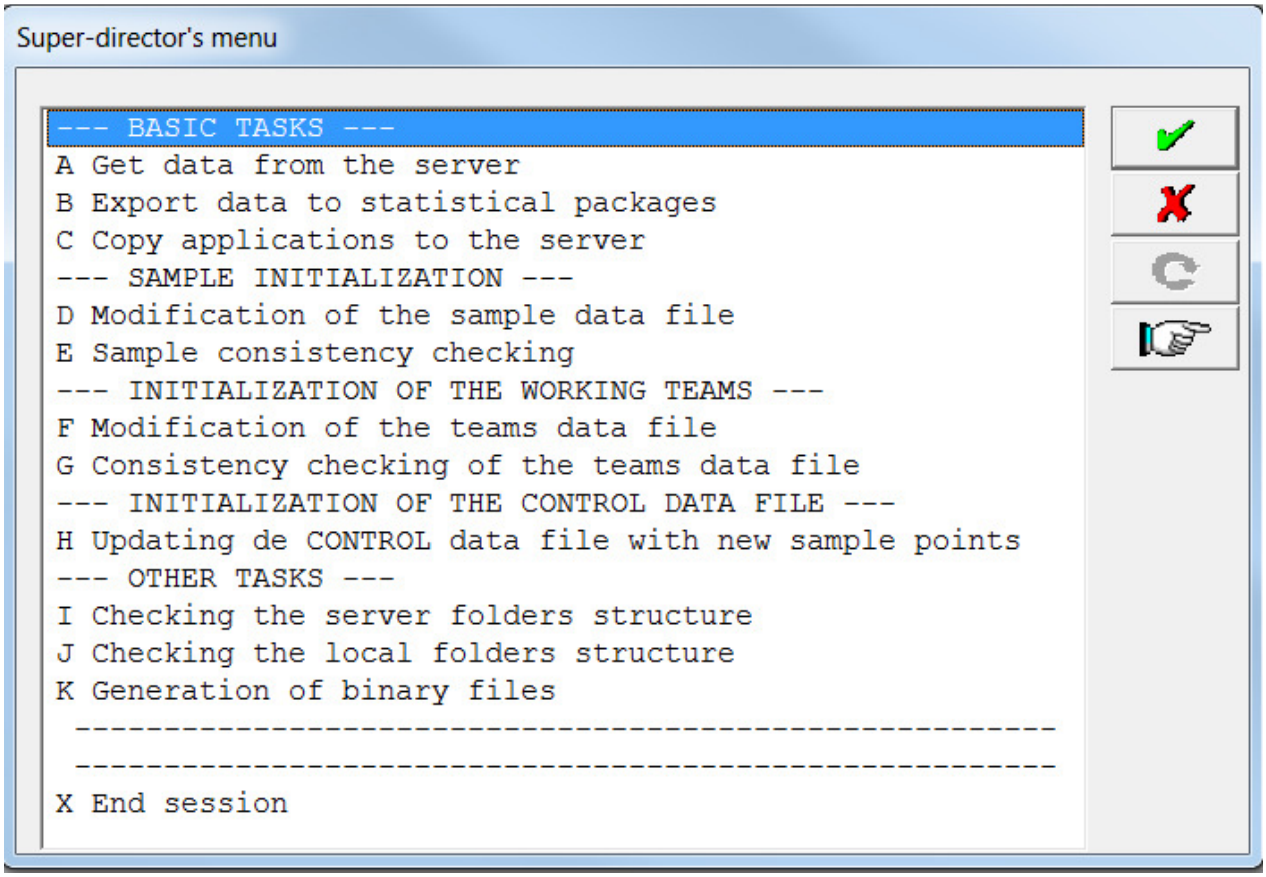

Figure 8: Super-director's menu

To be able to execute this menu, the machine must have the environment variable "CSPROX_SUPERDIRECTOR=Y". Thus, at the same time the director defines the MID variable, s/he should also define this variable. No other project's machine should have this environment variable defined. If the variable is not defined, the subset of the menu for the regular director will be used (options A, B and X).

Option A: Get data from the server

Normally, the data gathered by the interviewers is passed to the supervisors' machines and from there to the central server. We say normally because in many countries the internet communication is not as good in remote areas as it is in major cities. Thus, in these situations, we need to design an alternative to passing the information from the supervisors' machines to the central server.

The central server is therefore the machine that stores together all of the data gathered by the interviewers and using this option, the Director is able to pass all the data to his/her machine.

When the director selects this option, data files from each supervisor will be downloaded to the director's machine and the concatenated into one file called "AllData.dat".

Option B: Export data to statistical packages

The data file "AllData.dat" is exported to SPSS, SAS and Stata.

Option C: Copy Applications and data files to the server

As mentioned before, the director is responsible for the development of applications to be used in the CAPI project and is also responsible for the generation of the TEAMS.DAT and the CONTROL.DAT. An important part of the job will be to passing them to the appropriate machines in the hierarchy later.

The way to distribute the applications and data to the supervisors' machines requires that they are transferred to the central server first. From there, the information needs to be passed to the supervisors (see supervisors' menu later) and from the supervisors to their respective interviewers (see interviewers' menu).

This option will copy all the applications and data files needed to the server so they can be distributed to the supervisors' machines. There are two files that include the whole set of applications needed: "ApplicationsforSuperv.zip" and "ApplicationsforInterv.zip". As the file name indicates it, the first one includes all the applications required by the supervisors' machines and the second by the interviewers.

Note that the server does not distribute files to the interviewers. This can only be achieved through the supervisors.

Option D: Modification of the Sample data file

Normally, the sampler will provide the sample data on a spread sheet or any other format that needs to be adapted to a CSPro data file. In any case, this is an operation that is installation dependent and will need to be prepared and completed by the director.

Option E: Sampling consistency checking

Again, since the sample is installation dependent, this option will need to be completed and refined by the director.

Option F: Modification of the Teams data file

A full explanation of the "**Teams.dat**" was provided before under section 3.1. This option then will permit the project director to create and modify this important data file.

Option G: Consistency checking of the Teams data file

This is a very important option that will check that no cluster is duplicated, that no machine id is duplicated and in general, that the Teams file is sound and ready to be used.

Option H: Updating the Control data file

If new clusters are added to the sample after the control file has already generated, this option will permit to incorporate them to the control data file??????????????????

Option I: Checking the server folder structure

This option will check that all the folders of the FTP server are correctly defined as specified either on the "DirectorsParameters.txt" or the "DirectorsParametersInternet.txt". Which text file will be used will depend on whether the FTP server is remote (accessed through internet) or not. In addition, this option will create all the subfolders under "Sync\Superv", one for each supervisor as defined in the Teams data file.

Option J: Checking the local folder structure

This is very similar to the option above but is performed over the director's machine.

Option K: Generation of binary files

This option will produce the whole set of binary files for the CAPI application. From the security point of view, it's important to distribute only the binary files so the application can't be changed or altered by neither the interviewers nor the supervisors. These files should be regenerated every time that any of the data files of the application is modified.

4.2 The supervisors' menu

Figure 9 illustrates the different options the supervisor has to accomplish the important tasks required from him/her.

Option A: Execute the server

As it was mentioned before, there are two communication utilities distributed with the control system, one is the "CSFileSyncServer.exe" and the second, "CSFileSyncClient.exe". Both should be located/stored in the Sync folder of every machine of this project. The server should be run only when the supervisor wants to start synchronizing with the interviewers; to connect with the central server, option G should be used.

The file "CSFileSyncServer.ini" has important parameters defined, specifying the port that will be used in the communication and the users that will be able to connect. This file is project and installation dependent and thus, it needs to be properly defined according to the local needs.

The users needing to be defined here are the supervisor and all the interviewers working for the supervisor's team.

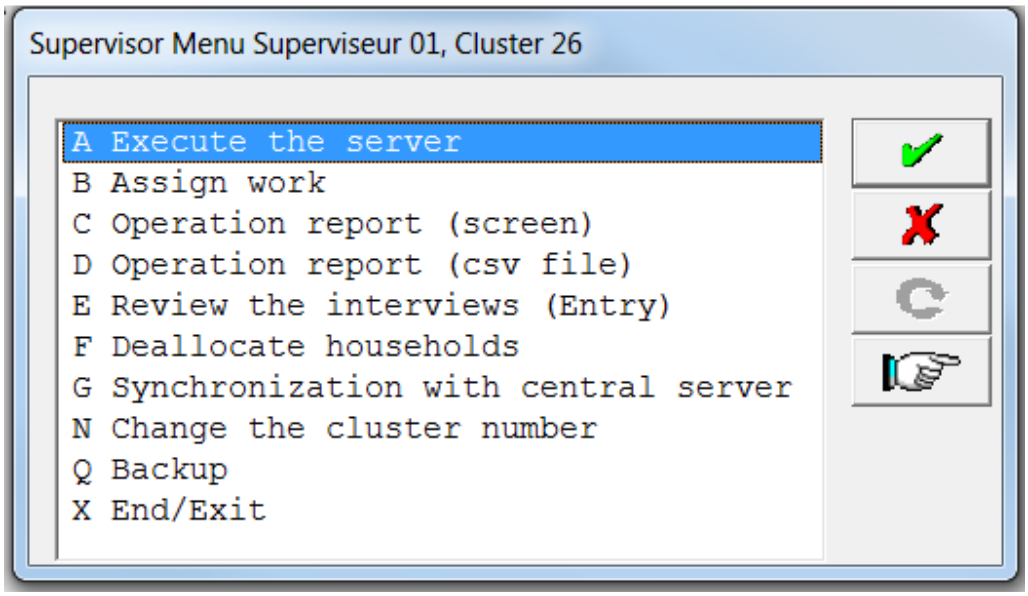

Figure 9: Supervisor's menu

Option B: Assign work

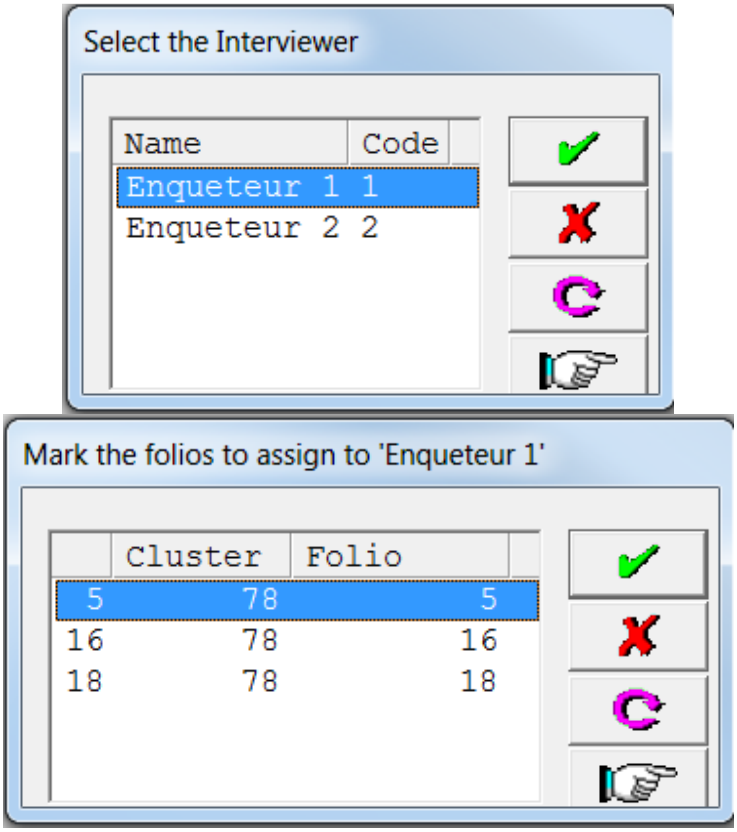

Figure 10: dialog boxes in the assignation of work to interviewers

The Teams data file has, for each team/supervisor, a list of all the clusters the director assigned to him/her. However, those clusters my include one or more folios that are part of the job the team has to interview but they need to be assigned to one or more interviewers. As the fieldwork progress, the list of folios and clusters pending to be interviewed gets shorter; the fieldwork for that supervisor will get to an end only when that list is null.

The supervisor will assign work to the interviewers on a periodic basis based on the speed they perform their duties. To assign work to them, the supervisor will need to perform two steps: to select from all the interviewers belonging to the team the specific one to assign work (first dialog box of figure 10); after the interviewer has been properly selected, then the system will show the list of folios for the cluster selected that have not been assigned yet. In the particular case of the example illustrated on figure 10, the cluster selected is 78. Note that the first step when the "Superv_Menu" is run will be the selection of the cluster s/he wants to work with (see option N, "change the cluster number").

Option C: Operation report

Each supervisor needs to know the behavior of the interviewers regarding the work assignment. The system is able to provide, based on the status of the interviews recorded in the control data file, the amount of work done or pendent by interviewer or by his/her whole team, by cluster or

for all the clusters. To request specific reports, the system provides the following alternatives from which the supervisor can do the selection:

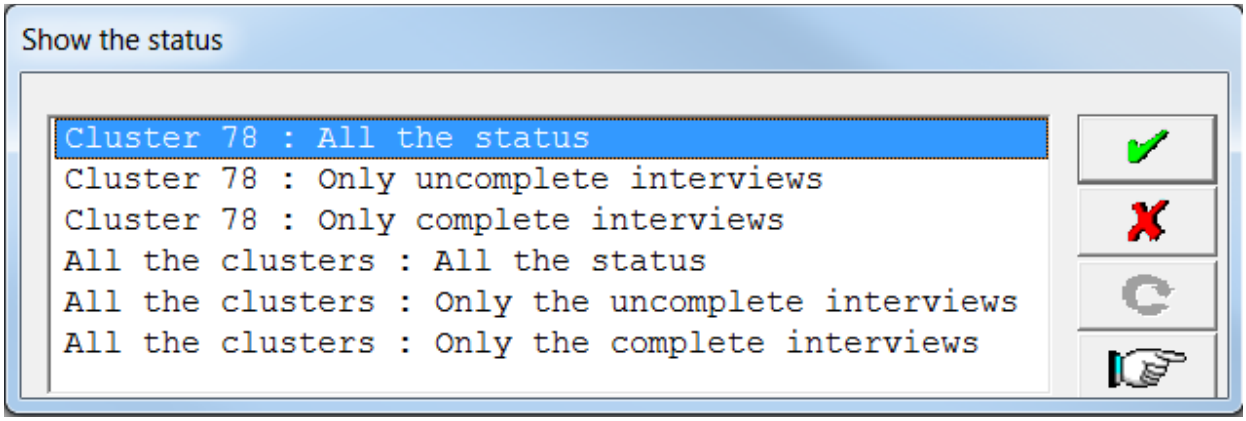

Figure 11: Different work status reports

As in the previous example, the current cluster we are working with is 78. If we need to get the report on a different cluster, there is the option to change the current cluster. The first three options above refer to (i) getting the status of all the folios/interviews for cluster 78; (ii) getting only the incomplete interviews and (iii) getting just the complete interviews. The same reports can be obtained for all the clusters instead of just cluster 78.

Option D: Operation report (csv file)

This option accomplish exactly the same as option 'C' above but instead of being addressed to the screen, it is addressed to a "Comma-Separated Value" known also as a csv data file. This output file can be easily imported into excel for analysis purposes.

Option E: Review the interviews (Entry)

This option will allow the supervisors to check the interviews using the same data capture application the interviewer used in the field.

Option F: Review the interviews in Batch

Here the supervisor can check one or more interviews using the same application to capture the interview data but in batch mode; any errors detected are printed on a report, which is convenient when it comes to analyze the quality of the work done by the interviewers.

Normally, the supervisor will want to check more than one interview at the time; to define the universe which will be checked, the system provides a selection box similar to the report offered to inspect the interviewers' work. Thus, the supervisor might dimension the size of the batch file to check according to the needs.

4.3 The Interviewers menu

The last system menu provides all the tasks we expect the interviewers to perform in a simple and organized manner. The following figure illustrates the duties the interviewer is supposed to perform:

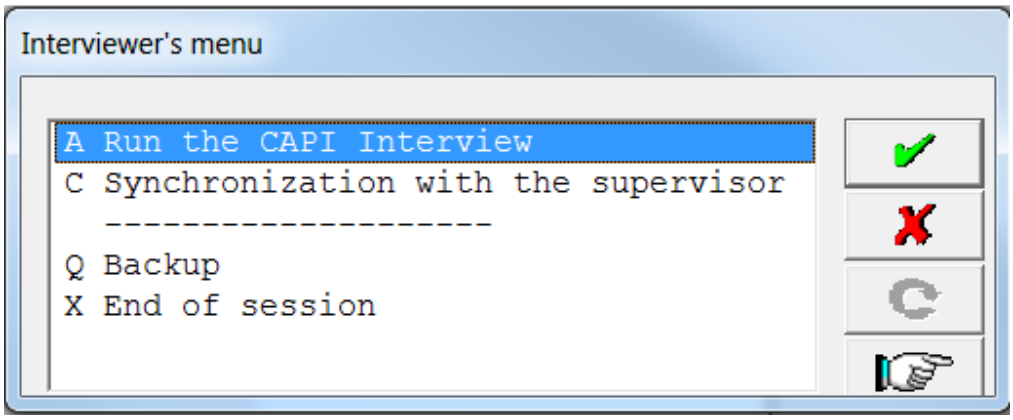

Figure 12: Interviewer's Menu

Option A: Run the CAPI Interview

Selecting this option allows the interviewer to start the CAPI application either with a new interview or with the continuation of an interview that had been previously interrupted. Since an interviewer might have many units/households assigned by the supervisor, s/he will be asked to select which unit from the ones assigned will be interviewed next. The following dialog box will be open to show the units assigned (cluster and folio/unit/household number), the geographical ids and the name of the person/contact at the unit to be interviewed, and the status of the interview.

As seen on the figure below, this interviewer has four units assigned, all of them with an interview status or result zero (meaning that the interview has not been started yet). The interviewer then has the alternative to select the unit that is more convenient and start the interview. By selecting the unit number desired, the system will call the CAPI application and will start the interview for that specific unit.

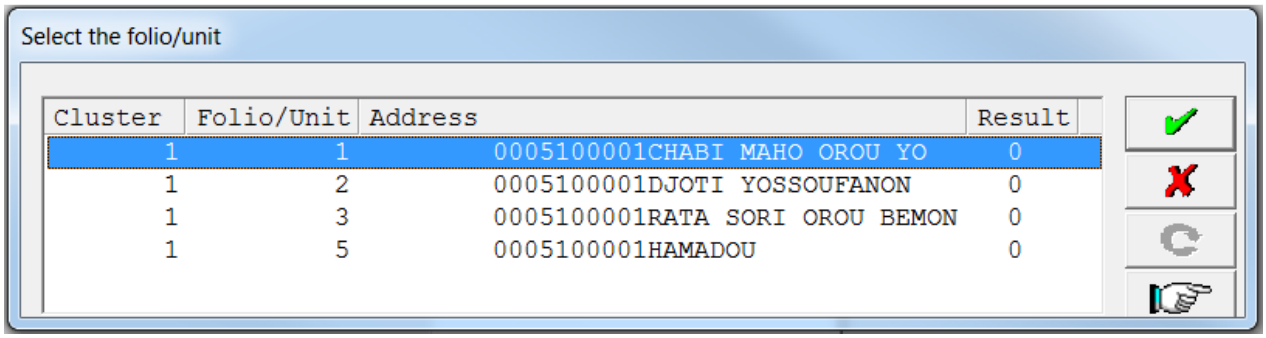

Figure 13: Dialog box to select the unit to be interviewed

The list of clusters and unit numbers are extracted from the Control data file and a few elements stored on this file are copied to the first screen of the CAPI application (i.e. supervisor and interviewer codes, geographic codes, etc.).

Option B: Syncronization with the Supervisor

Once a day or every other day, the interviewer should get in contact with the supervisor to carry out the following tasks:

- Transfer the interview data gathered during the journey to the supervisor with two purposes: (i) to follow the later up so the responsible staff for analyzing the quality of the data can detect any problem related to the CAPI application or to the interviewers that might distort the reality being investigated; (ii) making possible the merging of the different partial data files gathered by each interviewer, facilitating the creation of a master data file.
- Receiving from the supervisor the job assignment for the next journey.
- Updating the CAPI software in the event that any changes had to be made.

To synchronize with the supervisor make sure that (1) the cable to communicate both machines is hooked on both ends; (2) make sure that there is no other interviewer connected to the supervisor machine; (3) make sure that the supervisor's machine is running the FTP server and therefore, is ready to receive your data.

Option Q: Backup

It is advisable they you, once a day, run a complete backup of your data. In this way, in the event that any undesirable thing happens to the interviewer's machine, there will be no loss or at least a minimum of data loss to regret.

The backup device that we recommend to use will be a flash disk or, also called, pen drive. Normally, the device letter assigned to this removable device depends on the machine and the number of peripherals installed; however, commonly the interviewers' machines don't have any devices mounted since they are equipped with the minimum to guarantee that the CAPI data

gathering operation can be achieved successfully. Thus, we can be certain that in most cases, the letter assigned to the pen drive will be “E”. In the event that it was a different letter, the system will give the interviewer the chance to change it.

Option X: End of session

By selecting this option, the system will terminate the interviewer’s session. This option should normally precede the supervisor’s end of session so everything can terminate normally.