CSES 2003-05

Data Management Manual

Victor Canales

Version 1.0 August 18, 2003
Version 1.1 September 27, 2003
Version 1.2 October 4, 2003
Version 1.3 October 12, 2003
Version 1.4 October 25, 2003
Version 2.0 November 26, 2003
Version 2.1 December 9, 2003

Table of Contents

Chapter 1 The CSES and the data management component

The CSES will visit a very large sample of 15,000 households over 15 months. The data management component must therefore deal with an average of 1,000 households a month. The data management component should be able to meet the following main objectives:

- *Ensure that the workflow will not cumulate any backlogs. In other words, make the datafiles of fieldwork conducted in month-x available by the second week of month x+2.*
- *To minimize data entry errors and allow for the highest quality transcription of the information gathered in the field.*
- *To provide good quality databases to the survey analysis on a timely basis.*

Fieldwork overview

There are 50 teams allocated to the fieldwork. Each month there will be 25 teams working at the field, with a workload covering 2, 3 or 4 sample points or PSUs. The fieldwork plan has been scheduled in order to gather around 60 households monthly per team.

Teams are headed by a **Team Supervisor** and integrated by four **Interviewers**.

For a given month the team moves to the PSU one week before in order to achieve preparatory tasks (enumerate the households and select those to be interviewed). Then, each of selected households is submitted to the first part of the questionnaire, where the household composition is established. For the whole month of operation, the household fills up a diary of expenditures and incomes, and the interviewer visits it repeatedly to complete the remaining portions of the questionnaire.

Once the month ended, the team gets back to the NIS central headquarters in Phnom Penh, where the Supervisor delivers the product of the job to the Data Management team.

Each of PSUs is delivered in a packet including all of the documents used and produced in the fieldwork, including maps, enumeration lists, questionnaires, diaries, etc. In this manual, this is referred to as the PSU-packet, and it is the production for the whole workflow.

Data management overview

The Data Management component must enter the information to data files with the maximum quality possible. To meet this objective, there are two separated teams:

- **Data Entry** room: working with around 20 Data Entry Operators for the household questionnaires, plus other staff to enter the village questionnaires and to capture the information of time-use sheets, this group works with smart data entry programs which control the accuracy and reliability of the entry job. The Data Entry room is headed by a **Data Entry Supervisor**.
- **Quality Control** team: integrated by around 5-6 Editors, this group is in charge of receiving and checking the contents of PSU-packets, coding all the needed fields, and prepares the PSU-packets for the data entry job. Once the PSU-packet entered, they receive the error reports, analyze all the possible errors and mistakes, and determine the corrective actions to be carried out. This team is headed by a **Quality Control Supervisor**.

The activity of both teams above mentioned is coordinated by a **Data Management Chief**, who works continuously with the Supervisors.

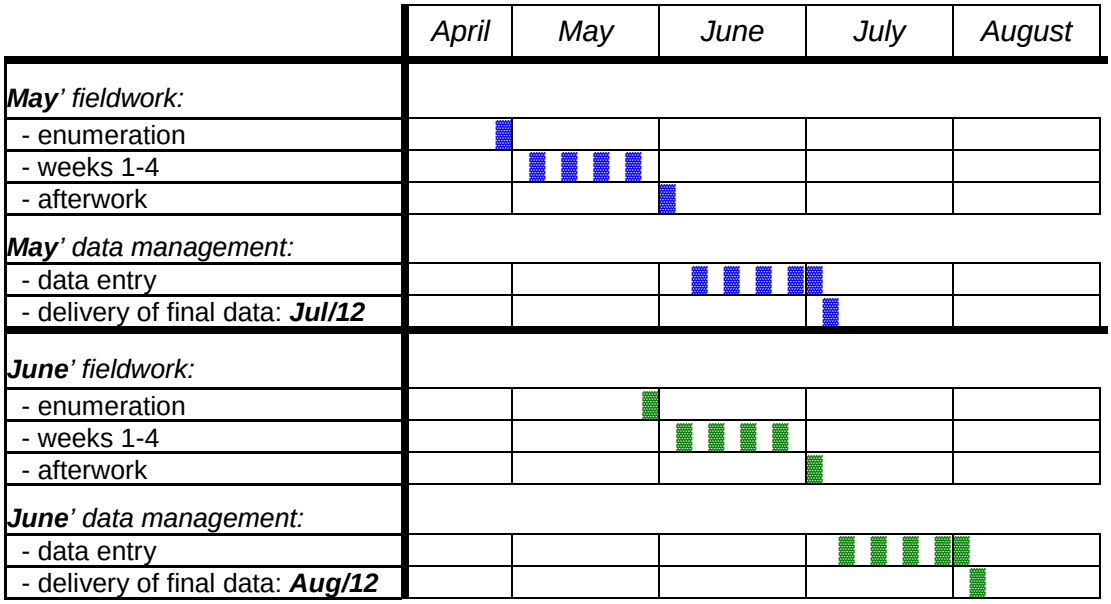
The whole component is in charge of a **Data Management Director**, a senior technician responsible for conducting the relationship with the Direction of the CSES. The Data Management Director seldom involves in the daily operation, and his role is mainly addressed to solve extreme conflicts needing of his intervention.

The qualifications and responsibilities for each of persons and each of team members are fully described later (please see *Chapter 2 – Organization of the Data Management*).

The life cycle of one month of field workload

A continuous 15-month operation must avoid any delays in producing the results for a given month of job. The data management organization should be able to meet the first of requirements described for this component: "*to ensure that the workflow will not cumulate any backlogs - in other words, make the datafiles of fieldwork conducted in month-x available by the second week of month x+2".*

This mandatory objective may be best explained by looking at the desirable calendar of activities of both fieldwork and data management components:



The **Data Management Chief** must be in situation to solve all of the difficulties that may oppose an obstacle for this cyclic schedule. For example, if the primary data entry I delayed, he should have the resources to quickly implement one of these solutions:

- to open a second shift to enter the data
- to add more data entry stations to the data entry room

If, on the other side, the editing job may delay the production of the final database in time, the **Data Management Chief** should have the authority to allocate more staff to this task.

All this requires a tight control and careful daily evaluation on the workflow situation to adopt the best measures leading to ensure the final delivery date (the 12th day of the month x+2) will always be respected.

Documentation for the data management component

There are three main manuals on the data management system. All are in standard word' document files, and have names starting with the prefix CSES, followed by two letters referring to the specific manual, ending with two digits referring to the version of the document. These manuals and their residence files are as explained below.

➔ ***dm*: the Data Management Manual**

This document, the **Data Management Manual**, is stored in files named ***CSESdm****xx.doc*.

➔ ***de:* Data Entry Operator's Manual**

For the **Data Entry Operator's Manual** , look at the file ***CSESde****xx.doc*.

➔ ***ed:* Coding and Editing Manual**

The **Coding and Editing Manual** is placed into files ***CSESed****xx.doc*.

To print his manual, it is also required to print some of coding tables non embedded in the document. All of the related coding tables are also included in the accompanying excel sheets described later in the Chapter 2 of this document.

Each version is date-stamped in the cover and footers of the documents, and also in the *properties* of the file.

Although only the last version of each manual should be used at any moment, it is useful to keep track of former versions.

Newer versions are to be periodically issued in response to changes, enhancements and improvements on the elements of the system.

Chapter 2 Organization of the data management

To get an efficient data management workflow, the data management needs to work following a model of organization, covering from the smaller detail up to the most global tasks to be performed.

Framework for the data management support

Main features of the data management support are as follows:

- Availability of coding tables in datafiles to be managed in CSPro
- Suitable datafile naming conventions based on the sample design
- Control of the workflow based in the sample design
- Fully-compliant data entry program
- Extended consistency program
- Effective control of database contents
- Repeated training of the data entry operators
- New roles: editors of the information in files, instead of traditional coding staff
- Automatic selective capability for producing intermediate joined databases
- Automatic reporting on the progress and critical delays
- Reporting fieldwork-quality tables aimed to global supervision and feedback of fieldwork staff
- Random checking of data entry quality by means of secondary entry of 5 to 10 percent of the households
- Global conception of a contemporary data management organization for complex surveys

The model of the data management organization

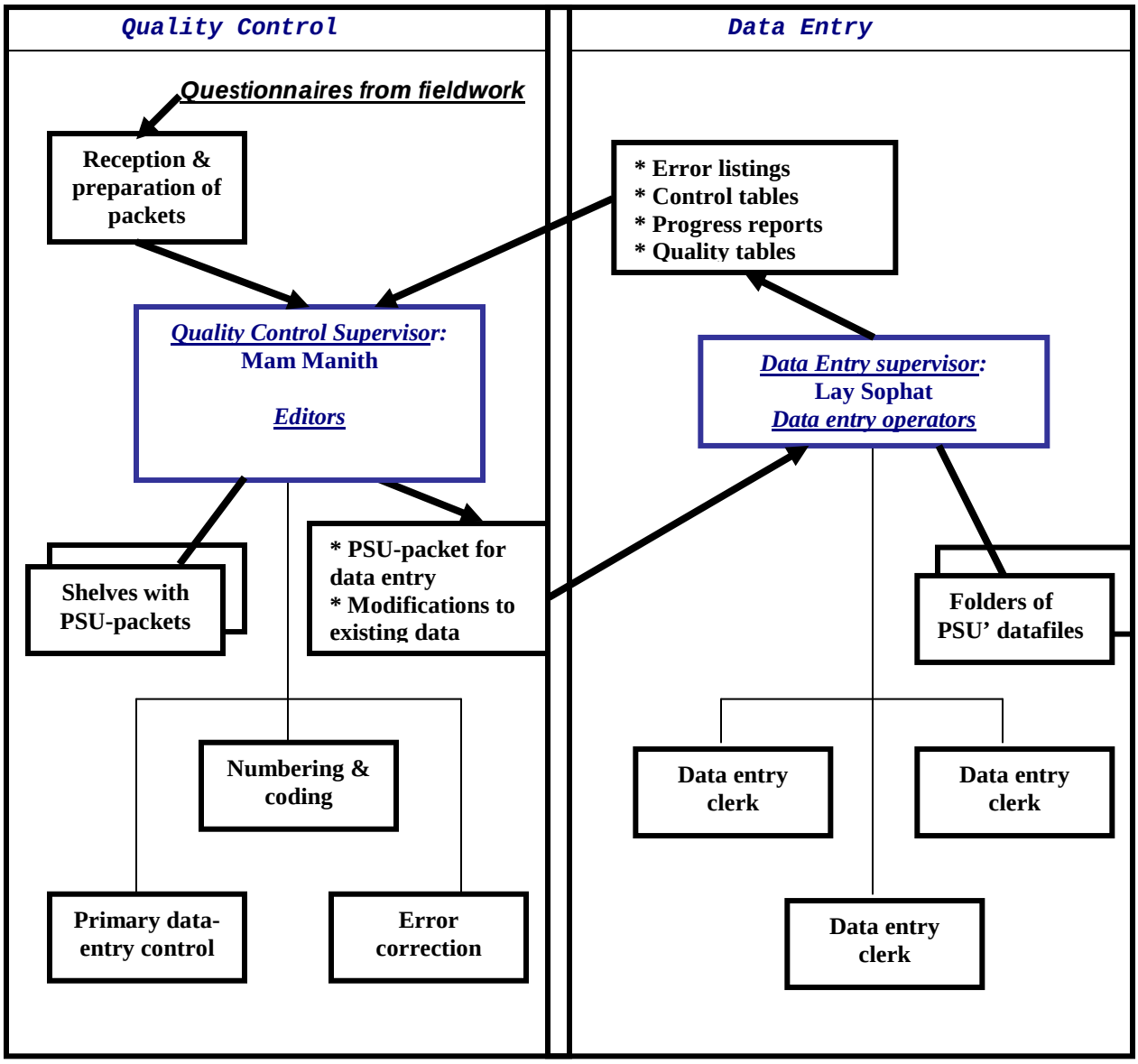
The responsibility for the success of this component is in charge of the **Data Management Director**, *Mr. Saint Lundy*, senior IT specialist of the NIS. He involves on the strategic technical issues and conducts all the relationship with the Director General of the NIS, *Mr. San Sy Tanh*; the Deputy Director General in charge of the operation, *Mr. Seng Soeurn*; the Survey Director, *Mr. Tith Vongh*; and the Sampling Specialist, *Mr. Mich Kantoul*. He also supervises, at a general level, the job made by the staff under his command.

➔ Daily operation management

The daily operations are conducted by the **Data Management Chief**, *Mr. Yip Thavrin*. He works continuously supervising the daily overall workflow, and reports the major incidents to the Data Management Director. He also provides him, on a regular basis, with the reports on progress of the data management. The responsibilities of *Mr. Thavrin* are fully described in the paragraph *Qualifications of the staff* below.

➔ Workflow model

Under the control of the **Data Management Chief**, two separated teams - **Quality Control** and **Data Entry** - share the responsibilities for the data management tasks. The figure below shows the activities and products of each team, along with their inter-relationships.



➔ Qualifications of the staff

The **Editors** are responsible for the reception and preparation of the PSUs coming from the fieldwork. Later, they analyze and fix any error and abnormal information detected by the consistency batch programs. They integrate the **Quality Control** team, headed by the **Quality Control Supervisor**, *Mr. Mam Manith*. Each Editor knows in depth the questionnaire and its codes. They are able to solve any error condition, have a detailed knowledge of fieldwork activities and procedures, and they are the only responsible for "building" the PSU-

packets to be managed in the activity. They must have a good level of knowledge of the activities at the Data Entry team and, under the command of the **Data Entry Chief**, must record the information regarding the reception of fieldwork documents and the preparation date of the PSU-packets into the Control Catalog. Since the Editors are the only responsible of editing and controlling the quality of the data, they ***must*** know, best than any other person, the possible variations of an interview. They ***must*** have a full mastery in analyzing all complex relationships of different pieces of data. They ***must*** be able to fully understand the error listings, deal with them either in paper or in their computers, and ***always*** take the best possible decisions to solve the errors or warnings issued by the checking programs. They work under close supervision of the Information Manager.

The **Data Entry supervisor**, *Mr. Lay Sophat*, is in charge of the **Data Entry clerks** and manages the **Data Entry** team. The DE supervisor is a specialist in data processing, having skills enough to well manage and take care of the PCs used in the activities. He is able to ***organize his staff*** and distribute the tasks in order to reach fast and accurate results and efficient levels of control on the data management. He ***must*** be able to provide a complete training to each one and all of the data entry clerks. He must know in depth the questionnaire (and the dictionaries and related applications). At any given moment, he must solve even the smallest doubt of the operators, mainly regarding the interpretation of error messages issued by the data entry program. He must also have the best possible communication with the Quality Control team and each of Editors, in order to solve and fix the most complex errors reported by the system.

The **Data Management Chief**, *Mr. Yip Thavrin*, is in charge of all of the activities of both preceding teams. He has the responsibility to prepare the ***weekly reports*** (the progress report and the fieldwork quality tables) to be delivered to the Survey manager, and to provide fresh copies of them on demand. He must have a ***good knowledge of the software*** tool or tools used in the data processing system, in order to install the modifications suggested by the project advisor (either in person or by e-mail). Because of his higher skills, he should know and master the use of the data entry programs, far best than any of data entry clerks. He is the ***only*** responsible of the integrity of the information stored in the whole set of data files and, at the same time, he has ***exclusive*** rights to access and manage the Control Catalog. Specifically, he ***must*** insure the Catalog is always up to date, and the information there placed is completely exact, as well as for the activity dates as for the actors involved in each step. To best achieve this mission, he will keep a high level of integration and control on both the Quality Control and the Data Entry supervisors, mainly when recording the reception of the documents arriving from fieldwork, and when insuring the contents of each PSU. Finally, the Data Management Chief must plan and implement a support to get the whole data file collection properly backed-up by the most suitable means (CD's, network backups to a secondary concentrator, etc).

Control Catalog

The Catalog records the following key activities performed with each of the PSUs included in the sample:

1. Starting date of the fieldwork
2. End of fieldwork
3. Reception of PSU' documents at the office
4. Preparation of the PSU-packet
5. First data entry of the PSU-packet
6. Final approval f the PSU-packet

The maintenance of the Control Catalog is the only mean to produce reliable progress reports and, while it is the only source to get information on how to verify the integrity of the database. The Catalog must be able to insure the generation of fully reliable reports on the status of each PSU, and to control that the number of questionnaires in the data files is the same as in the PSU-packets stored in physical shelves.

➔ Control catalog module: mastering the workflow

This module works on an adaptation of the sample framework. Its visible component is an interactive program to record the actual information on the actual steps performed on a given PSU. An additional batch application must be tailored to produce progress reports on demand.

Progress reports prepared on the information of the control catalog can be obtained with just a couple of keystrokes and produced in a couple of seconds. The information compiled and summarized in control reports is to be delivered to the Survey manager on a weekly basis, or on demand. At the same time, these reports are the only objective tools able to help the Data Management Chief to appreciate and predict any possible bottlenecks in the different activities of the data management component, and to accurately support the reallocation of resources on the organization in order to never miss the delivery date of each month of workload.

➔ How to measure the progress

The workflow organization is based on considering a PSU as a "production unit": each PSU comes from the field to the office in a pack, or PSU-packet. The data entry places all of this info into a separated PSU-file, one for each PSU. Each of editors take one PSU in charge for error correction purposes, and is responsible for fixing the errors either by means of data entry operators or by themselves. Once no more corrections are required, the PSU is declared approved, its PSU-file is stored in a different folder by the Data Management Chief, and the PSU-packet is sealed off and stored outside the circuit of activities.

Using this basic principle, the control catalog has to deal with 900 production units. For each month, the problem reduces to 60 PSUs to be fully processed in just 30 days.

Data files naming convention

A simple data file naming convention, based on the sample identifier for PSUs, has been adopted for the CSES:

- for core household questionnaires, data files are named PSU*nnnnn* (*nnnnn* being the PSU-number according to the sample)
- the related datafiles for the time-use sheets, to be captured by the separated automated module, are named TSU*nnnnnn*

Please be aware that this naming convention is crucial for the integrated management of all the tools and pieces integrating the data management system.

Questionnaires tailored for data management purposes

All the questionnaires bear a common identification based on the five-digit PSU identification (for the village questionnaire), or a combined key based on the PSU-id plus the household sequential number 01-10 or 01-20 (for the household questionnaires, the diary and the time-

use sheets). Time-use sheets bear two more identifiers (the person-id in the household, and the day of the month being recorded).

The household questionnaire has, at the bottom of each significant table containing values, a line for redundant control-totals on monetary values, amounts and quantities.

Both village and household questionnaires have fields for the codes of the interviewer, the team, and the actual month of the survey.

The cover of the household questionnaire includes also a summary of the number of persons in the household roster. This information must be completed, prior to the delivery of the questionnaires to the central office, by the team supervisor in charge of the PSU.

Automated capture of time-use sheets

As the data entry resources are not enough for keying in the time-use sheets, a separated module - aimed to automatically recognize this information by means of operator-driven scanners – has also been integrated to the system. The time-use sheets have been reformatted to permit a suitable rate of automatic capture and to require a minimum intervention of the operator, looking for minimizing the working time devoted to this task.

The final product of this separated module should be just one data file for each PSU (named TUS*nnnnn*, as already explained above). Its contents are to be eventually cross-checked with the corresponding info in the PSU-file for each of households of the PSU.

Evaluating the quality of the fieldwork: quality-control tables

The contents of the quality-control tables have not yet been defined. The main idea is setting up a set of key indicators displaying comparative performance information for the different teams.

This management tool is widely used in similar operations (i.e. DHS) and allow for an accurate evaluation of the quality of the fieldwork. It is normally a set of tables measuring the average size of the household, ratios on critical operative ages (i.e. ratios of ages 5/6 for children may uncover that some of teams are shifting the ages to avoid measuring more children), the length of diaries, etc.

These tables can be easily prepared using the standard resources of CSPro, and should be ready for feeding back the teams before they leave for their next month of fieldwork. Quality tables can be generated on not-fully corrected info, and the design of the tables should avoid relying on too much detailed data.

Folder structure for the CSES

All the computers and workstations involved in the operation must have a main CSES' folder at the root of the system disk. Their minimal structure is as follows:

```
CSES
  |___ Activ
  |___ Base
  |___ Sample
  |___ Docum
```

The role of each of subfolders is the following:

- folder **`Activ`**: it contains the PSU-files in activity.

- folder **`Base`**: it has c copy of the current version of the programs.
- folder **`Sample`**: this folder has a copy of the sample' definition.
- folder **`Docum`**: it should contain copy of the official manuals related to the data entry, preferably in pdf' format. At least the current copy of this manual should be placed here.

The main computers in charge of the Data Management Chief have a more complex structure. For instance, they must have separate subfolders for the Village questionnaires (**`CSES\Village`**), the management of the time-use sheets (**`CSES\Village`** , (**`CSES\Activ\TimeU`** and **`CSES\Activ\HHold`**), the storage of approved PSUs (**`CSES\Appro`**, with different subfolders by month of the survey), and a couple of other service subfolders.

A later version of this document will describe all other folders participating in this structure, for each type of specific computer.

Coding tables

One key piece of the data management system is the central definition of the coding tables to be used in the CSES. These tables are used by the Quality Control staff to code some textual information of the questionnaires.

This is a summary of the coding tables used in the system:

➔ Occupation's code

This is fully described in the master excel sheet named ***tOccup.xls***. The coding book extracted from this file is used to code the following questions of the household questionnaire:

- Section 13, part B: question 2b
- Section 13, part C: question 3b

➔ Industry code

This coding table is fully contained in the master excel sheet ***tIndus.xls***, whose coding book used by the Quality Control staff to code the following in the household questionnaire:

- Section 04, part H: question 3
- Section 13, part B: question 3b
- Section 13, part C: question 4b

➔ Crops code

This table is stored in the master excel sheet ***tCrops.xls***, and it is used in the household questionnaire for:

- Section 04, part B: question 4

➔ **Cause of death**

This table is fully described in the Appendix 3 of the Coding and Editing Manual. It has no separated master excel sheet. This code is required in the household questionnaire for:

- Section 11: question 7

➔ **Diary/Expenditures**

This table is stored in the master excel sheet ***tExpen.xls***. This code is required in the diary for:

- Expenditures and consumption of own-produced food: question 11

Be aware that this table also contains the definition of the units allowable for each item (as explained for Consumption units below), along with the codes linking the items with COICOP, current CPI and the suggested analytical code for the CSES. (The analysis code is automatically inserted by the system into the expenditure' records of the diary.)

➔ **Consumption units**

This table provides the units to be coded in the diary for Expenditures, and its definition is stored in the master excel sheet ***tEUnit.xls***.

- Expenditures and consumption of own-produced food: question 5

Please be aware that the definitions provided in this table should ideally have been applied to the Expenditures' table above mentioned: each of items of that table should contain a complete definition on the basic unit used (i.e. kilo) and the remaining units allowed for that item (i.e. grams, ounces, pound, etc). *However, these cross links are not available, because of the considerable delays in the definition of coding tables*,

Some of pieces of the original design have not been implemented up to date, mainly because of the considerable delays experienced in the production of coding tables by the NIS staff in charge. Two of these main elements are:

➔ ***<not implemented>*** **Food calories**

The original design considered the specification of a table providing the number of calories per 100 grams of each of caloric foods to be evaluated in a Balance of Calories per Capita (BSPC) automatically built by the consistency program. This information, extracted from standard tables defined by FAO and tailored to the food items used in the country, was expected to be stored in the master excel sheet ***tFoodC.xls***. *However, it will not be available because of the considerable delays in the definition of coding tables.*

➔ **<not implemented> Automated tool for the management of coding tables**

The original design of the data management system considered a separated interactive module devoted to the management of all of the coding tables. *Since the delivery of the tables has been considerably delayed, this piece was suppressed and the overall management of coding tables will have to be made manually on the master excel sheets*. However the NIS should pay close attention to this issue, in order to provide standardized coding tables for any other ongoing or future survey.

Codes for fieldwork and office staff

Staff coding is based on a function-related approach. This means, for instance, that a person working first as interviewer in the field work team 23, and later going back to the office to work as data entry operator, and eventually working as editor from time to time, will have one different code for each function:

- his enumerator's code is ***234***
- once at the data entry room, he is known by the code ***65***
- for his job as editor, he has the code ***19***

The data management system deals with families of pre-authorized staff codes as described below.

➔ Fieldwork teams and fieldwork staff codes

The survey as 50 teams, having codes ***01*** to ***50***.

Each team is composed of one team chief and has up to four enumerators. For the team number ***nn***, the staff has the following codes:

- Team chief is ***nn1*** (for instance, ***231***)
- Enumerators have the codes ***nn2*** to ***nn5*** (for instance, ***232***, ***233***, ***234*** and ***235***)

The core staffs of the fieldwork operation, *Mr. Tith Vong* and *Mr. Mich Kantoul*, are responsible for the exact allocation of fieldwork staff codes.

➔ Codes for the data entry operators

The definitions embedded in the programs consider a broad range of codes for entry operators in the range ***51*** to ***79***.

The Data Entry Supervisor has assigned the following codes to the data entry operators and other eventually related staff: Please be aware that, although *Mr. Lay Sophat* has not appointed any operator to the job of entering the Village questionnaires, one or two fo them will be exclusively in charge of such an specialized task. *Mr Sophat* will agree later with *Mr. Thavrin* on the person(s) to be assigned to this job, will record the selected name(s) separately, and will report the information to the core survey staff and the advisory personnel.

Code	*Name*	*Duty*
51	Ms Mao Vann Noeun	*Household entry operator*
52	Ms Khon Neary	*Household entry operator*
53	Ms Mey Sokhann Tey	*Household entry operator*
54	Ms Van Camarat	*Household entry operator*
55	Mr. Khiev Khemerin	*Household entry operator*
56	Mr. Nim Sao Mony	*Household entry operator*
57	Mr. Sun Van San	*Household entry operator*
58	Mr. Vy Sovyl	*Household entry operator*

59	Mss Mom Seila	*Household entry operator*
60	Ms Oun Len	*Household entry operator*
61	Ms Kong Srey Ny	*Household entry operator*
62	Mss Khiev Madary	*Household entry operator*
63	Mss Non Thida	*Household entry operator*
64	Mss Rin Sitha	*Household entry operator*
65	Ms Hang Dany	*Household entry operator*
66	Ms Peng Napy	*Household entry operator*
67	Ms Krem Somaly	*Household entry operator*
68	Mss San Sopha	*Household entry operator*
	Other related staff	
75	Mr Mao Chhem	*TimeUse scanner operator*
77	Yip Thavrin	*Data Management Supervisor*
78	Mam Manith	*Quality Control Supervisor*
79	Lay Sophat	*Data Entry Supervisor*

Mr. Lay Sophat, the Data Entry Supervisor, is responsible for keeping an accurate track of the codes given data entry operators.

➔ **Quality Control personnel**

For both coding and editing staffs in charge of *Mr. Mam Manith*, the Quality Control Supervisor, the system has reserved the range of codes going from 11 to 39.

The staff has been distributed in 3 separated groups. Each group is headed by a group' supervisor, and includes 5 Editors. *Mr. Mam Manith* has assigned the following codes to his staff:

Code	***Name***	***Duty***
	Group 1	
11	Ms. Em Samoeun	*Group supervisor*
12	Mr. Sieng Kim Han	*Editor*
13	Mr. Pich Pothy	*Editor*
14	Ms. Orn Davin	*Editor*
15	Ms. Ouch Monisetha	*Editor*
16	Ms. Heng Vicheth	*Editor*
17	Ms. Long Forsevy	*Editor*
	Group 2	
18	Ms. Tong Chhay Rin	*Group supervisor*

19	Mr. Kong Chenda	*Editor*
20	Mr. They Kheam	*Editor*
21	Mr. Yem Sopharom	*Editor*
22	Ms. Tho Samchin	*Editor*
23	Ms. Mak Chantanary	*Editor*
24	Ms. Chhun Chhorvy	*Editor*
	Group 3	
25	Mr. Heang Kanol	*Group supervisor*
26	Mr. Kim Chantharith	*Editor*
27	Mr. Chhuon Sothy	*Editor*
28	Mr. Khin Bunna	*Editor*
29	Mr. Khin Sovorleak	*Editor*
30	Mr. Teav Rongsa	*Editor*
31	Ms. Hout Karolin	*Editor*

Be aware that remaining codes 32-39 are reserved for future use – i.e. replacement of some of involved persons. Quality Control Supervisor must keep an accurate track of the codes given to each person working in his teams.

The programs will always accept the broad range 11-39, and the responsibility for ensuring proper codes will rely on both the Data Entry Supervisor and the Quality Control Supervisor.

Chapter 3 Components of the data management system

The data entry program for households (*hhEntry*)

Following a close observation of the working practices of the entry operators in the Pilot Survey, the data entry program was built using a native CSPro' behavior known as *SystemControlled*. With this approach, the program follows strictly the logical flow of the questionnaire for each household.

The Data Entry staff should be trained in the utilization of this program according to the guidelines provided below.

➔ Missing values

When the flow of the questionnaire leads the operator to an empty field (i.e. where the interviewer has not recorded the answer to a required question), the field must be filled with 9's. This is the representation of *missing values*, which provide the analysts of the resulting information with an accurate distinction on what was not gathered at the field.

➔ Automatically skipped portions

Mainly taking care of speed factors, the program automatically reacts to the conditions of the information and performs automatic skips when required by the logic of the questionnaire. The operator must carefully check that there is no information recorded by the interviewer in the parts skipped over and, in case of detecting the presence of any information recorded in that segment, immediately ask for the Data Entry Supervisor' advice.

Likely in most of cases, improper skips are due to a mistake of the data entry operator himself. It is highly advisable to check each of the conditions of the data preceding the abnormal skip, and to fix any of the fields incorrectly typed in. When this kind of corrections is applied, the flow of the program will automatically change, meeting the path of the interview again.

In a handful of cases, the improper skips will be due to an improper management of the flow of the interview by the interviewer himself. Whether this type of mistakes is uncovered, the data entry must resume the data entry according to the flow proposed by the program. It is entirely senseless to try recording the skipped information, since it will have no value at all for the analysis. The missing fields that may arise must be filled with *missing values*, as already explained in the preceding paragraph.

➔ Interpretation of messages

All the messages issued by the program have a structure like this:

- a *number*: each message has a number of six digits (although some of them appears in the screen with just 5 digits) with the form ***sspqqx***, where:
  - ***xx*** is the section number. Messages regarding section 11 start with 11, and so. For sections 01-09, the leading zero is not displayed, and the message is actually shown with just 5 digits.
  - ***p*** is the part of the section, if any. When a section has no parts, the digit zero is used. For part A, the digit is 1; 2 for part B, … 8 for part H.
  - ***qq*** is the question number.

- ***x*** is the possible suffix of the question, or one of several messages related to the same question.

For instance, a message regarding the ID code of children in Section 12, question 2, will have numbers like 12002x. Please remark that, as this section has just one part, the digit ***p*** is zero.

A message related to the third mean to avoid the AIDS (section 15, question 4, which may accept up to 5 answers labeled 4a thru 4e) will have the number 150043. Again, please remark that the section 15 has just one part.

A message related to the total of lines 1-10 in question 5 of section 04 part G has a number as 04705x. Please remark that the digit 7 corresponds to the part G. Be also aware that these message will be displayed with just 5 digits instead of standard 6, because of the leading zero corresponding to the number of the section.

- a *qualifier*: messages begin with a letter indicating the severity of the message.
  - letter **W** is used for warnings, meaning that this is just a reminder to the operator, and that he should try to pay attention to the exception explained in the text of the message. Warnings proceed immediately to the next field, but the operator may come back to the affected field and try to fix the problem if any.
  - letter **E** points out to a severe error that must be fixed immediately. The cursor stays placed at the originating field.
- a *text* that explains, as plainly as possible, the nature of the problem found. Explanatory texts normally contain copy of the suspect values involved in the conflict. When the messages arise in a multiple-line table, the message always provide, at the beginning of the text, a description enclosed in square brackets of the involved line (i.e. [ID03], [SN07], [AN2]).

The operator is expected to read and understand the message, and to follow the corrective actions suggested in the *Coding and Editing Manual* (in file ***CSESed****xx.doc*). However, the Data Entry Supervisor, who has participated in the development of the data entry program, knows the general policy to properly react to the messages, and he will insure the training of his staff for these purposes, according to the directives explained in the *Data Entry Operator Manual* (in file ***CSESde****xx.doc*).

➔ Partial save

As this is a long questionnaire, the operator should use the *PartialSave* feature embedded in current versions of the CSPro system.

The Data Entry Supervisor will take in charge the adequate training of the data entry staff to partially save incomplete households in order to prevent any waste of time for reentering households whose data entry was interrupted before the information was recorded.

➔ Further considerations

hhEntry is a program whose sophistication corresponds to the standards accepted worldwide for this kind of surveys.

Possibly the close attention paid by the program to the quality of the data entry job made by the operator does exceed what has been the standard up to date in the NIS. In many

aspects, this program reacts and behaves very differently than the remaining data entry applications currently in use in the NIS.

For this reason, the data entry staff must be carefully trained by the Data Entry Supervisor in order to achieve the best performance of all the operators. One thing to always keep in mind is that "best" means best quality, not necessarily high speed. Forcing the data entry operators to work faster than advisable may seriously compromise the required quality levels for the CSES. In worst of cases, "best" will be a careful operator, rather than a "fast" one.

The consistency program (*hhCheck*)

The pace of data entry job in the CSES is strongly determined by the delivery dates required for each month. The data entry module itself has a rather relaxed set of controls (i.e. allowing the operator to enter suspect data, just with warnings issued by the module, but never blocking the operation), so the information collected in the primary data entry likely will contain a significant rate of errors.

For this reason, the data management component of the CSES includes a batch-consistency program stronger than usual. This module is designed for checking at least following:

- Thorough check of the demographic structure of the household: presence of head of the household, inter-generation gaps, reliability of relationships, checking on ages versus dates of birth, opposed sex of spouses and parents, etc.
- Fully-compliant checking of children measurement using WHO standard tables, mainly to correct common inversion of height/weight measurements
- *<not implemented>* Balance of calories-per-capita (BCPC)
- *<not implemented>* Balance of incomes/expenditures: *this needs a model based on the diary, and a set of boundaries for allowable ratios*
- *<not implemented>* Close-up on food groups, to check for improper assignment of "household use" instead of "business utilization"
- Similar structure of messages as in the data entry program: both programs use the same numbered-messaging approach and the same messages' file.
- Strong control of the datafile contents, dates of activities and staff involved in the fieldwork, using the background information stored in the control catalog.
- The program reutilizes most of CSPro' user-defined functions of the data entry program, so sharing most of the controls implemented in the data entry.

hhCheck produces one listing for each PSU containing all the detected errors and problems found in the information. The listing is normally available at the end of any data entry session, and it should be delivered by the Data Entry Supervisor to the Quality Control Supervisor (along with the original questionnaires in the PSU-packet and a safe copy of the PS-file) once the entry of the PSU finished.

The Editor in charge of the PSU must analyze the problems reported in the listing, check that they are not originated in a data entry error, indicate the corrections to be made over the same listing, and marking his approval when the situation reported cannot be fixed by any reasonable mean. Once the analysis ended, the listing will be returned to the Data Entry room in order to fix the mistakes.

Once all the possible corrections have been made, the PSU reached its final status. At this moment, it will be delivered to the Data Management Chief to reset it to "finally approved".

An approved PSU is to be stored forever. Its PSU-packet will be sent to a separate warehouse, and the PSU-file will be transferred to a privileged folder. None of them will never again be placed back to the production activities.

The Editors and the Quality Control Supervisor *should be aware that the approved PSUs may be submitted to further analysis by this or other advisors*. All of the source materials (PSU-packet and PSU-file) must be available for auditing purposes. By this time there should not be any surprises regarding the contents of the files (i.e. corrections still not included in the data file, problems deserving a solution that have been deliberately ignored by the Editor, and so far).

Checking database integrity, parametric selection tool (*Select*)

It is absolutely required to set up a separated tool to control the overall integrity of the database, capable to detect any possible missing or duplicate PSU-files.

The same piece should be able to deliver parametric selections of the database. For instance, to produce a selection for three arbitrary months (including formal quarters), or just for Urban or Rural, of for predefined sets of provinces, etc.

Selected databases resulting from any selection performed should be always submitted to the consistency program, primary production of basic frequencies and quality fieldwork tables. The resulting reports should be attached to any database delivered to the analysis.

Of course, the databases cannot be delivered in their original CSPro format. They have to be filtered by a CSPro-export module to SPSS (and SAS or Stata or even dbf) according to the requirements of the analysis staff. *This issue is currently being discussed with the Statistics Sweden consulting team providing advisory to the project*.

The collection of CSPro' applications included in this tool are to be packaged in suitable menus to be operated with minimum effort.

The data entry program for the Village questionnaire (*viEntry*)

This piece is to be entirely developed by Yip Thavrin and Mam Manith. They will follow, as carefully as possible, the conventions and programming styles of ***hhEntry***.

This advisor will check the quality of the implementation of ***viEntry***, and will provide all the needed support by email as far as it may be required.

It is advisable to train a couple of operators for entering these questionnaires. The job should be made exclusively by this specialized staff, avoiding the intervention of different operators in the task.

The final data files delivered by ***viEntry*** must smoothly fit in the global data management system. This assumption will be checked when exporting the first database for analysis, and any remaining mistakes will have to be fixed by the involved staff.

Other supplementary tools

The design of the data management system includes a number of small pieces and menus. Depending of the time available and the possible contribution of the NIS' technical staff, these

pieces may be solved by means of a variety of choices, ranging from elementary bat-files written in DOS scripts, up to standard services written in a higher-level language managed by the NIS.

Advisable features to be implemented by the data management team

It is strongly suggested to randomly pre-select a core set of PSUs for double, independent data entry, for not less than 5% of the PSUs (that means 45 PSUs in total). The product of both independent data entry operators is to be checked using the standard “compare” support embedded in CSPro prior to any other correction activity with the involved datafile.

This type of comparisons usually uncovers a significant rate of interpretation mistakes and errors of one, and sometimes tow, of the involved operators, allows both the Data Entry Supervisor and the Data Management Chief to produce an accurate assessment of the quality of the job made by the Data Entry Operators, and permits to adopt corrective measures and to focus on specific retraining before the misconducts spread across the whole staff.