Competent Data Management – a key component

Part I – Elements of a Good Data Entry System

April 2009

University of Reading
Statistical Services Centre

Data Management
Support to RIU Projects

Background

This booklet, and an accompanying parallel booklet entitled *"Part II – Illustrating the Process using CS-Pro"* aim to provide guidance on how data entry systems can be set up and managed. The first part in here is aimed at the project Data Manager and focuses on discussing the components of a good data entry system that will help in ensuring quality data. This first booklet can be used as an aid to development of the data management guidelines for the project.

In the second part we show how to implement a data entry system. We use CS-Pro in this second guide, but the ideas could be implemented equally well in other packages such as Epi-Info and MS-Access®. The exact method of implementation will of course differ, but the concepts remain the same.

Disclaimer

This document is an output from the Research into Use Programme, managed by NR International, a project funded by the UK Department for International Development (DFID) for the benefit of developing countries. The views expressed are not necessarily those of DFID or NR International.

Acknowledgements

We are very grateful to Dr. Parvin Sultana, Middlesex University Flood Hazard Research Centre, for granting permission to use an example from her research work to illustrate good features of a data entry system. We also thank Dr. Frances Kimmins, RIU Management Team for funding the preparation of this guide.

Contents

Introduction

Organising a data entry system that will minimise data entry errors and enhance data validation procedures, is one component of a chain of activities that should be followed to generate reliable and good quality data. Research and development (R&D) workers need to pay attention to this component to lend credibility to the results they generate from their activities.

This guide is divided into three sub-sections. In the first of these we consider the structure of the data and how individual pieces of data fit into the whole structure.

In the second section we look at techniques for ensuring good quality data, emphasising data checking as a means of minimising errors.

In section 3 we discuss the mechanisms for data entry in terms of organising the data entry team.

For illustration purposes throughout this booklet we will use the questionnaire in Figure 1 which has been adapted and much shortened from a questionnaire used in a study in Bangladesh in 2004 aimed at determining whether Participatory Action Plan Development (PAPD) within phase two of the Community Based Fisheries Management project (CBFM) has been effective.[1]

[1] This adapted and shortened questionnaire is being used with the kind permission of Parvin Sultana, leader of the PAPD project.

Figure 1 – Adapted questionnaire from focus group discussions for PAPD project

The Effectiveness of the PAPD Method

Focus Group Discussions in CBFM Sites

Identification Number: [ ][ ][ ]

1. Waterbody Type: [__]
   (1=Open Beel, 2=Flood Plain Beel, 3=River, 4=River + Open Beel)

2. Area of Waterbody (hectares): [__ __ __ __ . __]

3. Is this a PAPD site?: [__]
   (0=No, 1=Yes)

   3.1. If yes, please state date PAPD started: [__ __ / __ __ / __ __ __ __] *(dd/mm/yyyy)*

4. Partner NGO: [__ __]
   (1=BRAC, 2=Proshika, 3=Caritas, 4=Banchte Shekha, 5=CNRS, 6=CRED, 7=SUJON, 8=ERA)

   4.1. Date of NGO contract: [__ __ / __ __ / __ __ __ __] (dd/mm/yyyy)

5. Date of start of CBFM activities: [__ __ / __ __ / __ __ __ __] (dd/mm/yyyy)

6. Date of formation of CBO: [__ __ / __ __ / __ __ __ __] (dd/mm/yyyy)

7. Please enter the following:

   7.1. Number of poor fishers: [__ __ __]

   7.2. Number of fishers who fish for income: [__ __ __]

   7.3. Number of fishers who fish for food: [__ __ __]

   7.4. Number of persons who do not fish: [__ __ __]

8. Number of awareness training activities: [__ __]

9. Number of people in Focus Group discussion: [__ __]

10. Basic information concerning respondents:

No.	Type of person *(1=Full-time fisher; 2=Part-time fisher; 3=Subsistence fisher; 4=Fish trader; 5=Fish processor*	Gender *(1=Male, 2=Female)*	Involvement in CBFM activities *(1=Lots; 2=Some; 3=Little; 4=None)*	Perception of changes in social cohesion since start of CBFM activities – scale of +5 to -5
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

1. Understanding the Structure of the Data

The first thing you need to think about is the number of *levels* in your data where, in general, a *level* equates to a questionnaire in your survey. For example a survey may include a Community *level* questionnaire and a Household *level* questionnaire thus giving you two *levels* of data. The data at the lower *level* (i.e. the Household *level*) must include a link to the data at the higher *level* (i.e. a way of linking the two *levels* of data). We will expand on this point in our discussion of unique identifiers.

Generally you could consider a relationship between the *levels*; e.g. one community is likely to include many households so this would be a *one-to-many relationship*.

Our example survey – see Figure 1 – comprises a single questionnaire and therefore has just one *level* of data.

1.2 Records

Within a *level* (questionnaire) there are likely to be one or more tables of information in which the number of rows is variable. For example in a household questionnaire there is likely to be a household roster containing information about individuals within the household. The number of rows in the roster will vary between households. Thus within the household *level* there will be two types of *record* – the household *record* and the individual *record*.

Here again there would be a relationship between the *records*; e.g. one household and many individuals within the household. The *record* for the individuals would include a link to the household *record*.

When considering the structure of your data there is little to distinguish between *levels* and *records*. The difference is really in the application and how the data are entered. Different *levels* equate to different questionnaires (e.g. Community and Household) and in data entry you are likely to enter the Community data before entering any of the Household data. *Records*, on the other hand, are within the same questionnaire. Data in a household roster, for example, would be entered at the same time as the corresponding household data.

Our example survey – shown in Figure 1 – has just one questionnaire and thus one *level*, but has two *records*. Questions 1 to 9 make up the background information for the focus group discussion and question 10 is a roster table for recording data about respondents.

1.3 Items (Variables)

The next step is to consider the *items* or variables that are needed within each *record*. An *item* describes a single piece of information given as a response to a question. Note that some questions generate many *items*. For example a question may include the instruction "Tick all that apply". This is known as a *multiple response* question and you must have enough *items* in your data entry system to allow for a respondent ticking all the options. For example you would need 4 *items* for the following question:

> Which of the following crops have you grown in the last 12 months? *(tick all that apply)*
>
> [ ] Maize
> [ ] Rice
> [ ] Sugar beet
> [ ] Potatoes

Pieces of information that are often neglected when considering *items* are the responses to "Other, specify" questions. Consider the following question:

> What is your primary occupation? *(tick one box only)*
>
> [ ]$_1$ Farmer
> [ ]$_2$ Shopkeeper
> [ ]$_3$ Teacher
> [ ]$_4$ Health Professional
> [ ]$_5$ Unemployed
> [ ]$_6$ Retired
> [ ]$_7$ Other, please specify: _____

There must be two *items* for this question – the first will store the numeric code (1 to 7) and the second will store the "Other, specify" text response. This second *item* will be blank for all cases where the occupation is between 1 and 6. Note you must not attempt to mix numeric codes and text in the same *item*. We will discuss data types in the next section.

1.4 Data types

When considering the *items* needed for your data, you must also consider the *type* of data to be stored. Most data will fit into one of four data types:

> Number
> Numeric Code
> Text (alpha)
> Date

Numbers and *Numeric Codes* are generally treated in the same way by data entry software but we have separated them here as there are important differences in the interpretation of the data between these two *data types*.

1.4.1 Number

The *Number* data type is used for questions that require a real number as a response. Examples include "Age of respondent", "Number of children", "Rice Yield", "Monthly income", etc. These are often referred to as *quantitative* data items. Some of these *items* should only take whole numbers – e.g. Number of children; others, such as Rice Yield, can and should allow for decimal places.

In many data entry applications there are properties for the items that determine whether or not decimal places are allowed and determine the range of possible values. This may include fixing the number of digits allowed for the *item*. For each *number item* you need to know the range of possible values so that you set the properties accordingly. For example, for Age of respondent, two digits would allow respondents up to 99 years of age.

1.4.2 Numeric Code

Many questions in surveys are pre-coded and numeric codes are entered into the data file. An example is Sex of respondent with codes 1=Male, 2=Female. These are referred to as *categorical* data items. As with the *number* data type, you will need to consider the number of digits needed to store the numeric code – if you have less than 10 codes then a single digit is sufficient; if you have 10 or more codes (but less than 100) then you will need two digits.

1.4.3 Text (Alpha)

Text items relate to descriptive responses that have not been pre-coded. Responses to "Other, specify" questions would fit into this category as would responses from open-ended, descriptive questions – e.g. "Please comment on the planning process". As with number data items you generally need to specify the number of character spaces to be allocated to the *text item*. Note that in some data entry software there is a limit of 255 characters for a *text item* so you may need to consider alternative methods for capturing lengthy comments.

1.4.4 Dates

Some data entry software have special data types to deal with dates, but others do not. Regardless of the situation dates can cause problems because of the different formats used in different countries – for example some use dd/mm/yyyy while others specify the month before the day. The recommended method for dealing with dates, regardless of the software used, is to separate the date into the three components of

day, month and year. These would be set up as three *number items* (though month could be considered to be a coded item). For day and month you must allow for two digits and for year allow four digits – we strongly recommend always using four digits for the year.

Most data entry software will allow you to set ranges for the items (1 to 31 for day, 1 to 12 for month) and to run consistency checks between items, e.g. to check for invalid dates such as 31st February.

1.5 Unique Identifier

For every *record* in your data there must be a *unique identifier*. This is an *item*, or combination of *items* that uniquely identifies an individual case. You may sometimes hear this referred to as the *primary key*. The *unique identifier* cannot be blank and cannot be duplicated – i.e. two or more cases cannot have the same value for the *unique identifier*.

In our focus group example in Figure 1 the **Identification Number** is the *unique identifier* for the focus group *record*. The *record* for the respondents (question 10) would have, as its *unique identifier*, the combination of **Identification Number** and **Respondent Number**; the latter being the row number from the respondent table in question 10. The *unique identifier* has to be the combination of the two *items* in this instance, because respondents are numbered from 1 for each focus group so **Respondent Number** by itself is not unique.

For each *level* and *record* in your data you must consider which *item* or combination of *items* is unique.

In multi-level surveys, the data file at the lower level often includes a link to the data at the higher level. For example in a survey with community *level* and household *level* data the household *level* would include a link to the community *level*. Whether or not the community level identifier forms part of the identifier at the household level depends on how the households have been labelled. Consider the data in Figure 2.

Figure 2 – Household Level Data

HHID	COMMID	OWNHOUSE	NUMROOM	ELEC
1	1	1	2	1
2	1	1	3	1
3	1	2	3	1
4	1	1	2	1
5	1	1	1	1
6	1	1	1	1
7	1	2	2	1
8	1	2	1	1
9	1	1	2	1
10	2	2	2	1
11	2	2	1	2
12	2	2	2	1
13	2	1	3	1
14	2	2	2	1
15	2	1	2	1
16	2	2	3	1
17	2	2	4	1

Here the households are numbered from 1 to 17 regardless of the community ID (**COMMID**). The *unique identifier* for the household in this scenario is **HHID**. Now consider the data in Figure 3.

Figure 3 – Households numbered within Communities

HHID	COMMID	OWNHOUSE	NUMROOM	ELEC
1	1	1	2	1
2	1	1	3	1
3	1	2	3	1
4	1	1	2	1
5	1	1	1	1
6	1	1	1	1
7	1	2	2	1
8	1	2	1	1
9	1	1	2	1
1	2	2	2	1
2	2	2	1	2
3	2	2	2	1
4	2	1	3	1
5	2	2	2	1
6	2	1	2	1
7	2	2	3	1
8	2	2	4	1

Here the households are numbered within communities so **COMMID** must be part of the *unique identifier* for the household.

1.6 Conclusion

In conclusion you must determine where each *item* of data fits into the structure of *levels* and *records*. Each *item* must be assigned a *data type* and each *level* and *record* must have a *unique identifier.*

2. Techniques for Ensuring Quality Data

2.1 Checks prior to Data Entry

It is recommended that you minimise the time between data collection and data entry. However, it is also recommended that you carry out various checks prior to data entry so that the process runs more smoothly. All completed questionnaires should be manually checked for completeness, clarity and consistency before being passed to the data entry staff.

By **completeness** we mean that all questions should be filled in. If a question is not applicable – for example a question about children in a household with no children – then established procedures should have been followed; e.g. a code for "Not applicable" used.

All responses given on the form must be **clear** enough for the data entry staff to read. Anything that is not clear should be referred back to the fieldworkers.

Data must be internally **consistent** and the sooner inconsistencies are found and resolved the better. Prior to data collection we recommend you list all possible checks for your questionnaire – for example check age against marital status (you would not expect a 6 year old to be married). With a comprehensive checklist you can catch many errors before they reach the computer.

2.2 Checks during Data Entry

2.2.1 Data Entry Forms

When the on-screen data entry form matches the layout of the paper questionnaire, data entry is easier as the data entry staff know exactly where to enter each value – they are able to match the position of a value on the questionnaire with the data entry box on the screen. Creating the forms is a time-consuming activity but the benefits in terms of data quality, outweigh the cost.

Text on the screen can be in the same language as your questionnaire though this would depend on the character set used. With some software it is possible to have more than one set of data entry forms – each in a different language. This should be a consideration for surveys conducted in several languages.

2.2.2 Range Checks

A good data entry system will include *range checks* or lists of valid values for each *item*. Depending on your system, values outside the valid range can either be disallowed or simple flagged as potential problems. For numeric codes a range check is straight-forward – if you have options 1 to 10 then your range of valid values is clearly 1 to 10. Real numbers need more considered thought – what, for example, is a feasible range for monthly income? It is useful to at least have a list of what are considered to be valid values so that extreme values can be flagged and investigated.

2.2.3 Labels for Numeric Codes

Your data entry system should include a method of labelling your numeric codes – for example 1=Male, 2=Female; 1=Yes, 0=No. The labels make your data more easy to understand and it is easier for the user if they are stored with the data.

2.2.4 Missing Values

Missing values are a fact of life in surveys. There are a number of reasons for a value being missing – the respondent didn't know the answer; the respondent refused to answer; the interviewer missed the question or didn't enter the response clearly enough. You need to consider whether you want to distinguish between different types of missing value or want to treat all missing values in the same way. In either case you need to allocate missing value codes for each *item* in your data. Some people choose to simply leave the *item* blank if it is missing but we would not recommend this course of action as there would be no way of telling whether this was a genuine missing value or something that was simply omitted by the data entry staff.

The missing value code should be a value that is not a feasible value for the *item*. In other words it must be outside the range of valid values. For numeric codes this is quite straight-forward. For example consider the question:

What is your primary occupation? *(tick one box only)*

[ ]$_1$ Farmer
[ ]$_2$ Shopkeeper
[ ]$_3$ Teacher
[ ]$_4$ Health Professional
[ ]$_5$ Unemployed
[ ]$_6$ Retired
[ ]$_7$ Other

Here any number below 0 or above 7 could be assigned as the missing value code (Note that 0 itself is not a good choice for a missing value code and should never be used as such). However, there are likely to be other coded questions in your survey that have more than 7 options. It is recommended you use the same missing values

codes for all your coded questions so choose a code that is not used for any of your *items*. Assuming no question has more than about 80 codes, 99 is often a good choice for the missing value code.

For real numbers the choice of code is not so obvious. For example, 99 could easily be the age of the respondent. A popular choice for real numbers is to use a negative number as the missing value code. To keep some degree of consistency you could consider using 99 for numeric codes and -99 for real numbers. However, you must make sure that the value you choose is not a feasible value for the *item* - age cannot be negative but what about temperature?

A special type of missing value is generated when a question is "not applicable" – i.e. did not apply to a particular respondent because of the response to a previous question. For example in the question sequence:

> 3. Do you have any school aged children? Y/N
> 3.1 If yes, do they attend the local school? Y/N

question 3.1 is clearly not applicable for respondents who do not have school aged children. It is useful to distinguish between values that are missing because the question was not applicable and those that are missing for other reasons. This is usually done by assigning a different code for "not applicable" – e.g. 88=Not applicable, 99=Missing.

2.2.5 Automatic Skips in Data Entry

The discussion on *not applicable* values leads naturally into a discussion on programming automatic skips into the data entry system. Questionnaires often include instructions to skip questions that are not relevant. The skip is sometimes explicit as in:

> if yes, skip to Q12

and sometimes implicit as in:

> if yes, please state the date

Here the implication is:

> if no, skip this question

Whether or not to program automatic skips into the data entry system is an issue that provokes much debate. There are those who argue that the data entry clerk should enter data exactly as it is on the questionnaire, manually skipping not applicable questions if they are blank on the questionnaire. They argue that this is less confusing for the data entry staff and therefore data entry is quicker. The idea is that errors and inconsistencies are sorted out after data entry.

Others prefer to have automatic skips matching those on the questionnaire. Not applicable codes can automatically be inserted where appropriate. Inconsistencies on the questionnaire – for example answering No to the question *"Are there school going children in the household?"* and Yes to the question *"Do they participate in farm activities?"* – need to be dealt with during data entry (or preferably before) as the data entry system would skip the second question.

The debate then becomes a case of speed versus consistency. In our experience we have found post-data entry checks to be sadly lacking and as such would recommend taking steps to ensure that the data entered into the computer are as accurate and consistent as possible even if this means taking more time over data entry.

2.2.6 Consistency Checks

A good data entry system will include consistency checks between items. For example if the month is entered as 11 (November) it should be possible to check that the day is not 31. Many of the consistency checks done prior to data entry can and should be programmed into the data entry system. If they are missed in the manual checks they can then be captured during data entry. Note this does not negate the idea of the manual checks prior to data entry. There will always be errors in your data but your aim should be to catch as many as possible as soon as possible.

2.3 Checks after Data Entry

Two techniques we would recommend for checks after data entry are *Double Data Entry* and *Exploratory Data Analysis*.

2.3.1 Double Data Entry

Double data entry (DDE) is another topic that can be controversial. It can be expensive and time-consuming. However, we would recommend using it in most of your projects and suggest you budget for it in terms of both time and money at the very start of your project.

DDE works as follows:

> Two data entry clerks enter the same data using separate copies of the data entry program. The resulting files are then compared and any discrepancies checked against the original questionnaires. If necessary corrections are made to both files so, by the end of the whole procedure, both files should be identical.

The idea behind DDE is that two separate individuals are highly unlikely to make the same mistake in data entry. Remember that even the most experienced and careful data entry clerk can and will make mistakes. DDE is an excellent tool for helping to ensure good quality data.

2.3.2 Exploratory Data Analysis (EDA)

Exploratory data analysis for data checking means using summary statistics to explore the quality of your data. We would recommend producing frequency tables for all categorical items (i.e. those with numeric codes). For real numbers produce summary statistics such as mean, maximum, minimum, inter-quartile range, etc. Graphical summaries are also very useful – for example a scatter plot of one *item* against another can often highlight cases that appear not to follow the general pattern of the graphical display. Box-plots are very good at showing outliers in the data. Some of the tabulations can often be done within the data entry system. For other analyses, particularly graphical summaries, you may need to export some of the data into a statistics package. If errors are found in the exported data, corrections must be made in the original data file which you should treat as your master database.

2.4 Conclusions

The key point being made in this section is that the only way to ensure high quality, reliable data is to check the data at every stage of the process. We strongly recommend documenting your checking procedure and any corrections you subsequently make to your data, as this demonstrates that you have taken steps to ensure good quality data. It also adds weight to any conclusions you draw from your analyses.

3. Organising the Data Entry Process

In addition to preparing the data entry system, you need to consider the *Data Entry Process.* This includes deciding who is going to do the data entry; where the data entry will be done; how it will be done and when it will be done.

3.1 The Data Management Team

The project Data Manager should have overall responsibility for the data but, on larger projects in particular, is likely to be just one member of a substantial team which may comprise one of more of the following:

Questionnaire Archivist/Logger:

responsible for logging questionnaires as they are returned from the field and for keeping track of which questionnaires have been entered;

Data Entry Supervisor:

responsible for training data entry staff; allocating questionnaires to data entry clerks; keeping regular backups; dealing with problems with the data entry system;

Data Organiser:

responsible for running data checks including double data entry comparisons; merging files; preparing cleaned data sets for researchers;

Data Entry Clerk:

these are the people who do the bulk of the data entry; there may be a team of such persons.

You should have clearly defined roles and responsibilities for each team member so that everyone knows what is expected of them and no task is missed out. These roles and responsibilities should be documented as "Terms of Reference" for each person.

3.2 Training

The data entry supervisor, the data organiser and the data manager must be familiar with the data entry system and should each have fully tested the system by entering a number of questionnaires themselves. Training must be provided to the team of data entry clerks on how to use the data entry system – the training sessions would generally be run by the data entry supervisor.

A data entry manual should be written as the system is developed and this can be used for the training. Training should include the procedure to follow when problems are found with the data entry system and with the data (e.g. values out of range). A log book should be kept detailing problems found and how each problem was solved.

3.3 Signing the Questionnaires

To establish an audit trail for questionnaire data we suggest having space at the front of the questionnaire that can be signed and dated by project staff as they complete their tasks. Figure 4 shows an example:

Figure 4: Space for project staff to sign the questionnaire

	Name	Signature	Date
Fieldworker			
Fieldwork Supervisor			
First Data entry clerk			
Second Data entry clerk			

The fieldworker signs and dates the form when he/she has completed the interview and checked the data for consistency. The supervisor then manually checks the data and signs the form when that task is complete. Assuming double data entry is to be

used, allow space for two data entry clerks to sign and date the form when they have each completed the data entry.

It must be made clear that in signing the questionnaire, the team members are taking responsibility for the quality of data on the questionnaire and/or in the data file.

3.4 Backups

Backups must be carried out at least once a day during the data entry phase. A range of media are available for backups including CDs, DVDs and memory sticks. You should have at least two sets of the media you choose for your backups and alternate between them. For example use the first set on a Monday, the second set on Tuesday, then re-use the first set on the Wednesday, etc. With this method, should a problem occur during the backup in which both the file on the computer and the file on the backup media become corrupted, you still have the previous day's backup file to return to. The backup media should be stored in a safe location away from the computers being used for data entry – i.e. you should not keep your backups in your computer room.

Someone in the team – generally the data entry supervisor – must have responsibility for creating the backups and for their safe storage.

3.5 Merging Data Files

In a large survey there is likely to be a team of data entry clerks with each person entering a subset of the completed questionnaires. The data files thus produced will need to be merged once data entry and checking is complete. We would recommend checking and cleaning the smaller subset files before merging into a master data file.

3.6 Example Scenario for a Data Entry Process

Let's imagine we have a household survey with 2,000 households numbered 0001 through to 2000. We have a team of 10 data entry clerks divided into five teams of two – the teams are numbered 1 to 5 and individuals within the team are labelled as A or B. Each team is allocated 400 completed questionnaires to enter. We would need to document which questionnaires have been allocated to which team –

for example:

Team Number	Households
1	0001 to 0400
2	0401 to 0800
3	0801 to 1200
4	1201 to 1600
5	1601 to 2000

Within each team person A is allocated the first 200 questionnaires and person B is allocated the second 200. Once these have been entered and signed and dated, the team members swap batches so person A then enters the second batch of 200 and person B enters the first batch. By the end of the process both team members should have identical data files and overall you will have 5 pairs of data files on which to run the double data entry comparisons. As mentioned earlier it is better to run the comparisons and do the necessary corrections on each of the 5 pairs of data files before merging to a master data file – it is easier to look through 400 questionnaires than to look through 2000 when you find discrepancies.

Depending on the software used, it is usually possible to save a report of the results of the DDE comparisons. These should form part of your data management report together with details of corrections made to the data.

3.7 Conclusion

The key messages from this section are (1) to document your procedures and (2) to make sure all team members have clearly defined roles and responsibilities.

The Statistical Services Centre (SSC), University of Reading, UK, undertakes training and consultancy work on a non-profit-making basis for clients within and outside the University. This guide has been produced by the SSC for the DFID funded Research Into Use (RIU) programme, which is being managed by Natural Resources International Limited.

Statistical Services Centre, University of Reading
P.O. Box 240, Reading RG6 6FN, UK
Tel: +44 118 931 8025
Fax: +44 118 931 8458
web: http://www.reading.ac.uk/ssc/