# Case Study No. 6

## Good practice in data management

This Case Study is drawn from the DFID-funded Farming Systems Integrated Pest Management (FSIPM) Project conducted between 1996 and 1999 in Blantyre-Shire Highlands in Southern Malawi. It describes procedures for data management that are designed to achieve project outputs based on high quality data. The case study also highlights resources needed for this purpose.

### 1. Background

A large number of information collection exercises took place during the FSIPM Project. They included diagnostic surveys, a baseline survey, meetings with farmers, transect surveys, scoring and ranking exercises, monitoring exercises and numerous measurements on research plots in farmers' fields to evaluate pest attack and crop yields.

Managing this large volume of data through a well defined data management system was crucial to the success of the project. Methodological aspects of the data management process are described below using examples drawn from the FSIPM project's activities. These relate largely to the experimental trials conducted on farmers' fields but the general concepts are equally applicable to other data collection exercises.

### 2. Steps towards good data management practices

### *2.1 Preparing forms for data collection*

Different types of data collection forms were constructed each activity. For example,

- Pre-coded questionnaires were used in formal survey work, (e.g. the Baseline Survey).

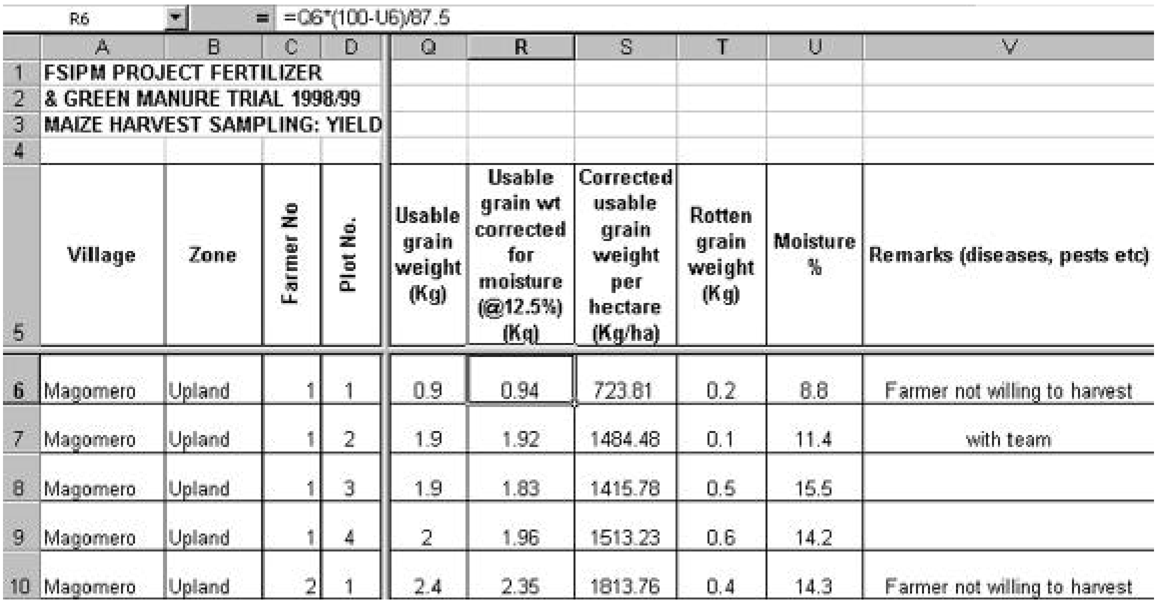
- Where open-ended questionning was envisaged (e.g. study of networks of communication), the form had broad discussion headings with space between to enter notes.

- In recording pest damage and crop yield measurements on experimental plots in farmers' fields, the data collection form typically contained background information to identify the plot location (e.g. village, farm identification, field number, plot number) and blank columns with headings to enter the measurements. Additional "check" columns were included where necessary, e.g. total grain weight was recorded in addition to damaged grain weight and usable grain weight; the number of plants with pods was used as a check on grain yields. Units of measurement were made clear on the recording sheet. Space was included for comments so that unusual occurrences in the field could be recorded (e.g. waterlogging causing stunting of plants).

Pre-testing of recording forms in the field was important and the forms were modified where needed. Separate recording forms or questionnaires were used for collecting plot level information (e.g. crop yields, pest damage), field level information (e.g. soil type) and farm level information (e.g. socio-economic variables).

Recording sheets were set up in a way that allowed the data to be entered directly into a computer. Designing the data entry form required a skilled person with a good awareness of the objectives of the data collection process. This is rarely the data entry person. Figure 1 shows a typical example.

**Figure 1**



R6 = =Q6*(100-U6)/87.5

| | A | B | C | D | Q | R | S | T | U | V |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | FSIPM PROJECT FERTILIZER | | | | | | | | | |
| 2 | & GREEN MANURE TRIAL 1998/99 | | | | | | | | | |
| 3 | MAIZE HARVEST SAMPLING: YIELD | | | | | | | | | |
| 4 | | | | | | | | | | |
| 5 | Village | Zone | Farmer No | Plot No. | Usable grain weight (Kg) | Usable grain wt corrected for moisture (@12.5%) (Kg) | Corrected usable grain weight per hectare (Kg/ha) | Rotten grain weight (Kg) | Moisture % | Remarks (diseases, pests etc) |
| 6 | Magomero | Upland | 1 | 1 | 0.9 | 0.94 | 723.81 | 0.2 | 8.8 | Farmer not willing to harvest |
| 7 | Magomero | Upland | 1 | 2 | 1.9 | 1.92 | 1484.48 | 0.1 | 11.4 | with team |
| 8 | Magomero | Upland | 1 | 3 | 1.9 | 1.83 | 1415.78 | 0.5 | 15.5 | |
| 9 | Magomero | Upland | 1 | 4 | 2 | 1.96 | 1513.23 | 0.6 | 14.2 | |
| 10 | Magomero | Upland | 2 | 1 | 2.4 | 2.35 | 1813.76 | 0.4 | 14.3 | Farmer not willing to harvest |

### *2.2 Data collection*

Training field staff prior to data collection is always important. In the first two seasons, planning meetings were held to discuss the format and other details concerning the data collection process. Some problems were encountered in the first two years and to improve the data quality, it was decided in the third year to give field staff the responsibility for the data entry work. Six Technical Assistants were therefore provided with some formal training on field methods, data collection and data entry and checking procedures (Ross, 1999). The training was facilitated by a Professional Officer, who had been responsible for data management activities in the previous season.

The training was particularly useful for the following activities:

- in the identification of disease symptoms and attributable causes of plant deaths;
- for ensuring uniformity in subjective assessments made by field assistants when scoring scales were used, e.g. in assessing the level of plant damage within a plot or in farmer participatory exercises to elicit farmers' opinions about particular pest management strategies;
- to raise field staffs' awareness of the importance of collecting reliable field measurements. For instance, ambiguities in the measurements recorded should be spotted, e.g. pod weight lower than grain weight; and inconsistencies noted and checked, e.g. many plants with pods, but low yields.
- to highlight the importance of recognising the difference between a genuine zero (no pod yields because of high damage attack) and a missing value (yield data unavailable because the farmer harvested the crop early);
- to draw attention to the units of measurement and ensuring measurement scales have the desired level of accuracy; and
- to emphasise the need to note down any unusual occurrences (e.g. a sudden high attack of *Sclerotium* on pigeonpea plants).

If labourers are employed to help with more straightforward data collection activities (e.g. harvesting the crop), then one or more trained field assistants are needed in a supervisory role to ensure that the correct data are collected e.g. that net plot sizes are correctly identified; all seed are placed in the appropriately labelled container, etc.

### *2.3 Data entry and validation*

Data were entered in their “raw” form, using the same software that generated the data collection form. Data from most studies within the FSIPM Project were entered using the spreadsheet package Excel. Direct entry into the statistics software package SPSS was used for others, e.g. survey data. The advantages of using a relational database packages such as Access for hierarchical and complex data structures were recognized (e.g. data from socio-anthropological work concerning a case study of activity diaries (Abeyasekera and Lawson-McDowall (2000)), but not used because the study was relatively small.

Data entry and validation procedures within the FSIPM Project varied during the three crop seasons. In the 1996/97 season, this was the responsibility of the Senior Technical Officer (STO) who was to have received 5-weeks of UK based training on data validation and management issues. Unfortunately the STO passed away shortly before the training. Data queries were therefore usually resolved at the time of analysis in consultation with FSIPM team members.

In the 1997//98 season, an ex-Associate Professional Officer, an Entomologist with M.Sc. level training, was employed on a full-time basis for two months to carry out the data entry and validation task. Some data management tasks were also included. Validation was by visual comparison of computerised sheets with the data collection sheets although current versions of Excel have good facilities for some data validation at the time of data entry (see SSC Biometric Guideline Series, 2000). Suspicious data recordings identified during validation were checked with the data collection team and the computer files updated where needed.

At a training course given to field assistants in the 1998/99 season, the following actions were taken.

- A schedule of agreed responsibilities for data entry and quality control was drawn up with respect to each of the experimental trials.
- Training in Excel was given.
- A senior staff member was assigned to conduct regular quality checks on the entered data.

This allocation of responsibilities gave field staff a sense of ownership and led to a considerable improvement in the quality of the data collected and a reduction in data entry errors. As a result there were very few data queries at the analysis stage in the final year of the project.

### *2.4 Organisation of the data files*

Following data entry and checking procedures, a system was needed to organize the data files in a systematic way and to have an appropriate format for data storage.

- A separate directory was created for data from each season (see Figure 2). Next sub-directories were created within each *season* directory to store data files from each of the on-farm experimental studies of which there were about ten in each season.

- Within each sub-directory, further sub-sub-directories were created (see Figure 2). One was to keep the ***original set*** of checked data files ***in their initial format*** and included formulae used for standard data calculations (e.g. original plot yields together with yields which had been corrected for moisture content and converted to kilogrammes per hectare). These files were retained as part of the historical archive showing the data as it was originally created.

- Copies of the above were made for subsequent analysis and were referred to as ***working Excel files.*** They formed the *Master Copy*. They were stored in a separate sub-directory and any subsequent data corrections were made in this master copy.

- The Excel workbooks used had multiple worksheets. This enabled each worksheet to be simple and named appropriately. For example, an Excel workbook could include several Excel worksheets corresponding to different sampling occasions, e.g. 8-16 occasions for damage data and 1-3 occasions for harvest data. Copies of these Excel workbooks were used for analysis purposes, with additional worksheets created where necessary to re-structure the data into a suitable form for the type of analysis envisaged. For instance, damage data that had been entered in several sheets were collated together into a single sheet for analysis purposes.

- Worksheets or workbooks included all the necessary background information (e.g. farmer's name, village, etc) entered prior to data collection, always in the same order so that data recording and data entry errors would be minimal.

- For some of the ***analysis***, small command files were written, mostly using the software package Genstat. The statistical software has a Windows interface, so dialogues could be used instead for the analysis. In this case the running of the dialogues creates a 'log file' of the equivalent commands, and this should be kept as the record.

- Reports were prepared as ***WORD documents*** . In this project, they were stored as separate directories to the command files.

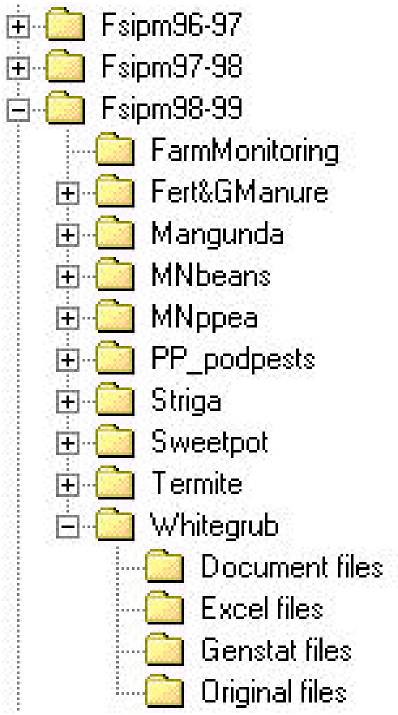
- A ***standard notation*** was used to facilitate the process of locating relevant files when needed. We chose to use short names of eight characters each. The first two characters of the filename identified the trial (MN for the Main Intercrop Trial, ST for *Striga* Trial, etc), the next two characters identified the season/year (97, 98, 99); the next two characters identified the crop (MZ, BN, PP, etc) while the last two characters identified the type of data (H for harvest data, D for damage data) and gave a sequential number for different versions of the data (1, 2, etc).

- Data from each study were stored on separate floppy disks with ***back-up copies of these disks kept up-to-date*** throughout project activities.

### *2.5 Data archive*

The archiving of all information collected during the FSIPM study was made available to relevant parties within Malawi on a CD ROM. Copies were also lodged with the DFID Natural Resources Adviser in Malawi from whom further copies could be obtained on request. The archive included all the trial and survey information collected by the FSIPM project and key reports that were written during the course of the project cycle.

**Figure 2. Data storage structure of FSIPM experimental trials**



```
Fsipm96-97
Fsipm97-98
Fsipm98-99
    FarmMonitoring
    Fert&GManure
    Mangunda
    MNbeans
    MNppea
    PP_podpests
    Striga
    Sweetpot
    Termite
    Whitegrub
        Document files
        Excel files
        Genstat files
        Original files
```

***Point No. 1: Allow sufficient time, perhaps six months, to procure the necessary computer hardware and software.***

***Point No. 2: Plan a practically feasible data management system before starting the data collection activities. Make sure the appropriate software for data entry, validation, management and analysis are available.***

***Point No. 3: Keep a record of planned net plot sizes and check that field assistants are aware of this before data collection commences.***

***Point No. 4: Ensure a relatively senior staff member is available to supervise the field data collection activities and the data entry and validation procedures.***

***Point No. 5: Ensure that data entry and validation activities take place as soon as possible after data collection, to facilitate timely checks on any records that are found unusual.***

***Point No. 6: Ensure the following resources are available.***

- Computer(s) with adequate power and capacity for data storage. These should have facilities for regular back-ups (e.g. a built in CD writer). Alternatively, a zip-drive or a large number of good quality diskettes are needed to allow at least two back-up copies of data files.
- Appropriate software for data entry and management (e.g. Excel, Access) and for data analysis (e.g. Genstat, SPSS).

For projects that have large data collection activities, staff resources are also needed, such as

- a full-time technical officer (TO) given training in data management and with sole responsibility for ensuring data quality and supervising field data collection; and
- a senior level staff member responsible for over-seeing the work of the TO.

## 5. Acknowledgement

We are very grateful to Dr Mark Ritchie, leader of the FSIPM project, for his valuable comments on initial drafts of this case study. We are also grateful for his permission to use this example as part of our case study series. A similar version is in the *Methods Manual* published by the Natural Resources Institute as an output of the FSIPM Project.

## 6. References

Ross, S. (1999) *Training workshop in field methods, data collection and data processing for FSIPM Project Field Assistants.* Consultant's report. Mimeo.

Abeyasekera S. (2000) *FSIPM 1998-99 On-farm experimental trials*. Statistical Analysis Reports. Mimeo.

SSC Biometric Guideline Series (2000) *Disciplined use of spreadsheets for data entry*. Publication of the Statistical Services Centre, The University of Reading. (Available at http://www.reading,ac.uk/ssc/).