problems

fm

research objectives

dv

protocol

ds

observation units

coll

data sheet

cm

data set

qy

data selection

as

results

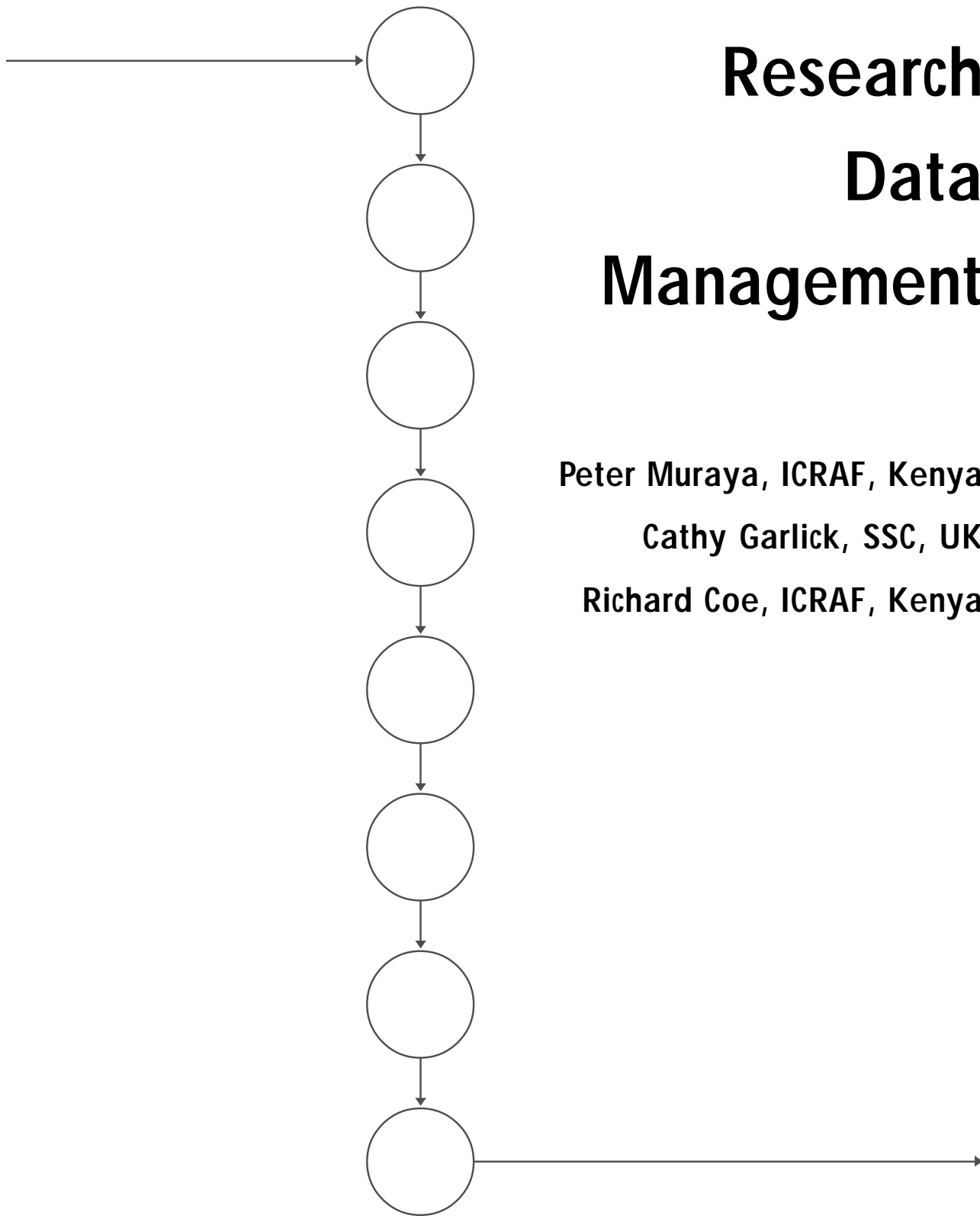pb

knowledge

Research Data Management

Peter Muraya

Cathy Garlick

Richard Coe

World Agroforestry Centre
TRANSFORMING LIVES AND LANDSCAPES

Research
Data
Management

Peter Muraya, ICRAF, Kenya

Cathy Garlick, SSC, UK

Richard Coe, ICRAF, Kenya

World Agroforestry Centre

TRANSFORMING LIVES AND LANDSCAPES

STATISTICAL SERVICES CENTRE

The World Agroforestry Centre (ICRAF) is the international leader in Agroforestry - the science and practice of integrating 'working trees' on smallholder farms and in rural landscapes. Agroforestry is an effective and innovative means to reduce poverty, create food security, and improve the environment. The Centre and its many partners provide improved, high quality tree seeds and seedlings, and the knowledge needed to use them effectively. We combine excellence in scientific research and development to address poverty, hunger and environmental needs through collaborative programs and partnerships that transform lives and landscapes, both locally and globally.

Founded in 1983, the Statistical Services Centre (SSC) is a not-for-profit body within the School of Applied Statistics at The University of Reading, UK. The SSC provides training and consultancy in both statistics and data management in the international arena. We aim to encourage good statistical practice, and the use of modern statistical methods in applied problems.

The SSC currently has nine statisticians, plus computing professionals and administrative staff.

Contents

About the course

Introduction

This document describes the Research Data Management training course organized by the World Agroforestry Centre. Data management is a key area in any research and if not done well can limit the usefulness of the data.

The numbers we quote and use to reach conclusions must be correct. If research is worth doing it is because we need answers. Erroneous answers will not only damage the reputation of the institute and scientists, but will not help us solve the development problems we are working on. We should therefore be concerned about the validity of every number. Validity of data is not ensured simply by getting the numbers correct, but also by getting the context of the numbers correct. We need to know how and why data were collected in order to make valid interpretations.

Sound management of research data is important for two other reasons in addition to ensuring validity. Well-managed datasets will be easy to process, so that the turning of raw data into useful information can be done efficiently. Well-managed datasets will also be accessible in the future, increasing their half-life and adding value.

The main aim of this training is to encourage research scientists to allocate necessary resources to data management. These resources include both time and skilled personnel. The course also aims to give participants the necessary skills to handle their data in a systematic and organized way and to preserve their data for future use. Implementing the ideas introduced in the course should lead to:

1.	Improved processing efficiency.

2.	Improved data quality.

3.	Improved meaningfulness of the data.

Resource persons

The resource persons should be skilled in scientific data management. They must be familiar with MS-Excel and MS-Access and the general principles of database design and data modelling. They should also have some familiarity with the types of data being handled by the participants.

Audience

The course is intended primarily for scientists undertaking agroforestry research. Materials will also be of interest to other research scientists doing similar field research. Session 7 (Building a data management strategy) can also be used as a stand-alone session for project managers.

We assume participants are regular users of software in the Microsoft Windows environment. We also assume they have used a spreadsheet package to enter data, perform calculations using formulae based on cell references, and produce simple charts such as bar charts and scatter plots. Most will be familiar with MS-Excel and we will be using that spreadsheet package as an example. However, the notes are designed such that the concepts should apply to any similar spreadsheet package and thus participants should be able to substitute the package of their choice. Scientists who do not have regular access to a computer and who are not comfortable with the use of a word processor and a spreadsheet should not take this course.

We assume that participants are in posts where they are able to use the materials that are covered here regularly in their work. In the final session of the course we will be encouraging users to think about how they can improve their own data management immediately, in the near future, and in the more distant future. Participants should go home with a phased action plan. We believe that follow-up, say in 2 or 3 months time, to check how data management has improved, is an important part of effective training.

Datasets

A requirement of the course is that participants bring with them a set of data either from an experiment or from a survey that they have been involved in. The data can be in the form of a computer file or on paper. In either case participants should also bring the protocol or documentation that describes how and why the data were collected.

The materials presented here include datasets from ICRAF experiments and surveys. Participants will first work with these example datasets and then repeat the process with their own data. These examples should be substituted with others more relevant for participants working in different subject areas.

Duration

The course is intended to last a week. Some of the sessions can be covered in just half a day whereas others are likely to take longer. The time taken for any one session will depend in part on the previous experience of the participants in the subject area.

The speed with which participants are able to cover the material depends a lot on their general computing competence and experience.

Teaching style

Throughout the course lecturing is kept to a minimum.

It is envisaged that a typical half-day session would start with a lecture/demonstration of between 20 and 45 minutes. The main part of the session would then be devoted to practical and discussion work. Participants would usually divide into groups for this work, with each group having a set of tasks.

This training requires the use of computers and there must be sufficient for participants to have considerable hands-on practice. One computer between 2 participants is a minimum. Often we find that one between 2 is better than one each, because it encourages more discussion during practical work.

There should also be a projection device that allows the whole group to view a computer screen, so data management techniques can be discussed jointly.

Presentations can use an overhead projector. They can be prepared using presentation software such as MS-PowerPoint, or they can use the spreadsheet or database package directly. Ideally, the computers would be networked so the presentations that use a computer can use the same computer on which the materials were prepared.

If none of this is available, the resource person could still put presentations and other background material on a CD-ROM and give this to each participant. This way, participants cannot only follow the presentations on their screen, but have all background material for later reference.

Software

A spreadsheet package and a database package are required. The aim is to teach the concepts and practice of data management, not the skills of using a particular package.

We currently use MS-Excel as the most widely used spreadsheet package available today. The current version at the time of writing is MS-Excel 2000 (part of the MS-Office 2000 suite of packages). As an example of a database package we are using MS-Access. The current version at the time of writing is MS-Access 2000 (also part of the MS-Office 2000 suite).

In addition to the spreadsheet and database packages a word processor (MS-Word) and software for presentations (MS-PowerPoint) are also needed.

Course content

This training course will consider the life cycle of the project from formulating the research objectives through to publishing the results. The project life cycle is shown in Figure 0.1.
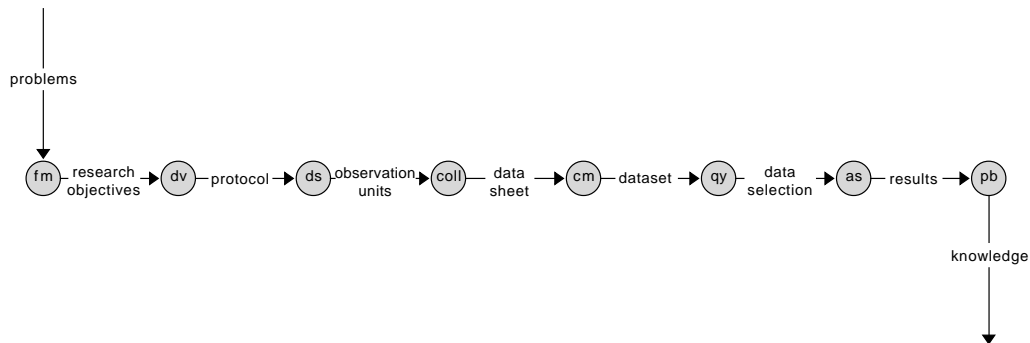
Figure 0.1 - **From problems to knowledge, through data transformations**

Going through the cycle we can imagine **formulating** (fm) the research objectives to address particular problems. Once we have the research objectives we **develop** (dv) the protocol from which we **design** (ds) the observation units. We can then go into the field to **collect** (coll) the data. These are **compiled** (cm) into a well-structured dataset which we can **query** (qy) and select subsets of the data to **analyse** (as). The results of our analyses are hopefully **published** (pb) leading to knowledge.

The main scope of this training covers the processes from the collection of data and production of the data sheets, through to selection of data for analysis. We will however, consider the entire life cycle as the processes and transformations are linked in such a way that a weak point anywhere in the chain can cause the entire process to collapse.

Throughout the lecture notes you will notice references back to the project life cycle so that you can see how each topic links into the entire research process.

Training strategy

The course comprises 7 sessions. For each session you will receive:

- **Summary** - Indicating the topic and the teaching method. The summary also lists any required skills, equipment and data files needed for the session.

- **Lecture note** - As far as possible the lecture notes are independent of the exercises. Concepts rather than tools are emphasized so that the notes are more flexible. Where we have referred to particular packages we have used MS-Excel and MS-Access.

- **Exercises** - Practical methods using MS-Excel and MS-Access. As indicated above MS-Excel and MS-Access are used merely as examples. It should be possible to alter these exercises to use other spreadsheet and database packages. For most of the exercises you will work in small groups. You will first carry out the exercises on datasets that we provide and then repeat the exercises using your own dataset(s). Please bring with you a dataset together with as much information about the dataset as possible (the metadata).

The sessions

1. **Why worry about data management?**

   In this session we will explain what we mean by data management and describe the main steps in the data management process.

2. **Designing a spreadsheet for research data**

   Here we encourage participants to consider the structure of their data and construct spreadsheets prior to data entry. These spreadsheets will contain all the information about the data (i.e. metadata).

3. **Spreadsheet data entry and checking**

   In this session we will look at a disciplined approach to using spreadsheets for data entry. We will also demonstrate techniques for data checking including pivot tables and graphs.

4. **Why go beyond the spreadsheet?**

   The main objective of this session is to alert participants to the limitations of using spreadsheets for data management.

5. **Improving data querying efficiency**

   In this session we show how the database offers a more elegant solution to many of the data querying problems encountered in spreadsheet packages. We look at row and column selections, calculations and summaries using data queries.

6. **Data modelling: Organizing data for easier querying**

   In this session we look at building a model of the data. A well-defined database structure facilitates the querying process and helps preserve data integrity.

7. **Building a research data management strategy**

   This session is designed to pull together topics covered in the course and to help you develop a Data Management Strategy for your own research. At the end of the course you will take away a phased **action plan** to help you set milestones for the implementation of your strategy.

Preparing for the training

This booklet of training materials is not complete without the accompanying CD. It consists of the lecture notes, summaries and instructions for exercises only. We have not printed documents needed for practical exercises, expecting that course organizers will want to replace them with others that may be more relevant for their situations than the ones we have provided. The file name for each document is shown in grayscale, e.g. **Sheet 1** (*Root.xls*). In preparation for this course, the resource person will copy the data files from the CD to an appropriate location, say DataFiles, in the participants' computers. He/she will need to announce this location to the participants well before the start of the exercises. For each session, the files to copy to DataFiles are listed in the sub-section **'Files needed for Practical Exercises'** in the Summary section.

To keep the size of the booklet manageable, we have opted not to print the support documentation. Instead we have provided hyper-links on the CD, so that the documents are easily located (and printed, if required).

Acknowledgements

Why worry about data management?

Objectives

By the end of this session participants should:

- Understand what is meant by 'data management' in agroforestry research.

- Describe the main steps in the data management process.

- Understand why data management must be a concern of all agroforestry researchers.

- Be motivated to benefit from the remainder of the course.

Summary

In the context of this course, **data management** refers to any activity concerned with processing or looking after data generated as part of research activities.

Attention to data management is important for three reasons:

1. We must be sure that data used to reach conclusions is of the highest quality.

2. Data is expensive to collect. It should therefore be looked after and archived in such a way that it can still be retrieved and interpreted in the future.

3. Data processing and analysis proceeds most efficiently if the data have been managed well.

Data management problems may be technical, organizational or conceptual in origin.

Strategy

1. Short introduction on the meaning of 'data management' in the context of agroforestry research and why it is important.

2. An exercise based on a simple agroforestry trial. Ideally the trial would be briefly visited, but an introduction to it with slides is sufficient. The exercise aims to raise awareness of common and simple problems in data recording and data entry.

3. Discussion as follow up to the exercise, focusing on how common these problems are within participants own work, and encouraging them to raise other data management problems of which they are aware.

4. Exercise based on participants own data. This follows a similar format to the previous exercise, with participants discussing data files (paper and electronic) that they have brought with them. The main aim is to get participants to think critically of their own data management problems. If the course is progressing to data analysis, then the exercise also starts the process of cleaning up participants' files so that analysis can proceed smoothly.

Required skills

Participants must be regular users of software in the Microsoft Windows environment. They must also be familiar with the use of spreadsheet packages.

Files needed for practical exercises

- *Leucaena diversifolia Progeny Trial.doc* - Protocol for the 'Leucaena diversifolia Progeny Trial - cum - Seed Orchard'.

- *ROOTCOLL.XLS* - containing Sheet 1, Sheet 2 and Sheet 3.

Support documentation

- *The sock under the bed* - Coe R and Muraya P, ICRAF (2000), 6pp.

- *Data management guidelines for experimental projects* - Statistical Services Centre (1998).

- *Project Data Archiving* - Lessons from a Case Study - Statistical Services Centre (1998).

- *Audit Trail in Research Data Management* - ICRAF.

- *Data Collection Form* - ICRAF.

Why worry about data management?

What is data management?

The 'raw materials' of a field or laboratory research activity are data. These are the numbers resulting from measurement, together with information about where they came from. 'Data Management' refers to any activity concerned with looking after and processing this information. It includes:

- Looking after field data sheets.

- Entering data into computer files.

- Checking data and preparing for analysis.

> *Project Life Cycle Reference*
>
> *Referring back to the project life cycle in Figure 0.1 in 'About the course', these activities cover the processes of collecting the data, compiling the dataset, and querying for data selections.*

- Maintaining records of the processing steps.

- Archiving the data for future use.

Why is data management important?

Quality of data

Data used for analysis and reaching conclusions, must be correct. If proper attention is not paid to data management it is too easy to make a mistake in the processing, resulting in incorrect conclusions.

Common errors that go undetected include:

• Data entry errors.

• Incorrect methods of conversion and combining numbers.

• Confusing variables and datasets.

Documenting and archiving

Data has to be documented and described so that it makes sense, not just now to the researcher who collected it, but in the future and to others. Field data is expensive to collect so must be considered as valuable, but its value is maintained only if it can be used in the future. A computer file full of numbers, the exact meaning or origin of which is not known, is worthless.

Efficient data processing

Most scientists analysing data from a research project find that much time is taken in preparing the data for the statistics, modelling, mapping etc. that they want to complete. This preparation includes converting the data to suitable formats, merging data originally entered in different files and producing various summaries and conversions from raw field measurements. Careful planning of the data management before starting the project can make a great difference to the efficiency of this data processing step.

Project Life Cycle Reference

Efficient data processing follows the transformation processes of compiling the dataset and querying for data selection.

It is helpful to think of comparisons with financial data management. Any organization will use a financial and accounting system to keep track of money. The system will be designed to ensure, among other things:

• Each transaction is correct and legitimate (quality).

• Records are kept so that information can be found and understood long after transactions have completed (documenting and achiving).

• Invoices, payments, summary accounts and so on can be prepared quickly and accurately (efficient data processing).

Accountants are trained for years to become proficient at financial data management. We often expect scientists to do a similar job with no training at all.

Some steps in data management

(Steps taken remembering the aims of validity, quality and efficiency.)

- Planning data management for a project, taking into account the objectives and planned outputs, the resources and skills available.

- Designing field data recording sheets.

- Collection of data, with appropriate quality control.

- Checking of raw data.

- Data entry and organization of computer files.

- Backup of data files.

- Processing of data for analysis.

- Checking of processed data.

- Maintenance of a data processing log.

- Archiving data for future use.

Data management problems

Problems in data management can be very diverse. Many will be exposed during the workshop. They can usefully be classified into the following areas:

Area	Examples
Technical	• Not being able to use software. • Not being able to set up data checking procedures. • Organizing data in ways that are not compatible with some of the required uses.
Organizational	• Multiple copies of files. • No one with responsibility for checking data. • No feedback on quality to technicians. • No clear policy on archiving data and making it available.
Conceptual	• Multiple entry of the same data or hand pre-processing of data. • Links between the numbers and information on the source of the numbers. • Audit trail.

Suggestions for diagnosing and overcoming some of these will be presented later in the course.

Why worry about data management?

Exercise 1 - Spotting problems

In this activity you will be considering some of the steps in managing data from a simple trial.

Step 1: Read the protocol **Leucaena diversifolia Progeny Trial - cum - Seed Orchard** (*Leucaena diversifolia Progeny trial.doc*). If you are not able to visualize the trial, ask for clarification.

Step 2: Work on your own for this step.

You are going to the field to measure the root collar diameter of the trees. **Sheet 1** (*ROOTCOLL.XLS*) is the form on which the data will be recorded. Prepare the sheet for data recording by adding information to the header area and the columns. Make up typical information if you do not know the actual facts. You may not need all the columns.

Step 3: Work in pairs for this step.

Sheet 2 (*ROOTCOLL.XLS*) is the actual recording sheet brought from the field. Look through it and identify:

• Aspects that you do not understand.

• Points that might be ambiguous or lead to misunderstanding later.

• Obvious mistakes.

Step 4: Also in pairs:

Sheet 3 (*ROOTCOLL.XLS*) is a copy of the computer file into which the data was entered.

Exercise 2: Your own data

For this exercise, work in pairs or small groups.

You will need the data brought by each participant. If the data is on paper, you can work from the paper copy. If it is on a computer file then print out a copy. If it is very extensive, print enough to show the nature of the data and the layout (say one or two pages).

Choose one member of the group whose data is to be discussed. The other members of the group raise questions about the data, highlighting any aspect of the data that seems incomplete or unclear. Repeat for each of the group members.

Try to compile a brief report on the common problems.

The following list of questions may be helpful:

Study

- Is the study clearly identified?

- Is there information that may help one find related studies?

- Are the objectives well stated?

- Is the location of the trial recorded?

- Is the principle investigator clearly shown?

Measurements

- Are all the measurements well described?

- Do they have short (<=8 characters) names?

- Is it clear what the measurement units are?

- Is it clear exactly what object has been measured (e.g. the sample or plot size)?

Field layout

- Is the field plan discernible from the data set?

- Has the type of design been stated?

- Are the treatments clearly defined?

- Is it clear what each level of a treatment means?

Data

- Are these original values or just a summary? If not, where are the originals?

- For each recorded value, is it clear exactly what measurement it represents?

- Can you tell the design, field layout or treatment level for each data value?

- When were the values recorded?

- How were missing values recorded?

- If the values are discrete measurements, is each level described clearly?

1. why worry about data management

exercises

Designing a spreadsheet for research data

Objectives

- To construct spreadsheets in MS-Excel, prior to data entry, which will contain all relevant information necessary for efficient analysis of the data.

- To carry out checks on these spreadsheets to ensure correct construction of data. For 'advanced' participants we will also set up validation checks for subsequent data entry.

Summary

During Session 1 we discussed the importance of good data management in agroforestry research. Electronic spreadsheets such as MS-Excel are widely used for entering and processing data. However, their use requires discipline on the part of the researcher. Spreadsheets have the advantage that several data checks and calculations can be carried out within the spreadsheet. These spreadsheets can then be easily linked to both database and statistical analysis software packages for further management and analysis.

Spreadsheets must be set up so that they contain all information about the data, not just the raw numbers. The spreadsheet can be used for data entry and data checking. The design of the spreadsheet determines how efficient it will be, both for data entry and subsequent data analysis.

The setting up of a workbook, containing our designed spreadsheets, can be done before the data have been collected. These can then be printed as blank data collection forms. This would help the data entry work, which could then be started immediately the data are available.

Designing the spreadsheet is complicated for some experiments and surveys. Complex designs include those with large numbers of measured variables, more than one type of unit, and data that are measured at a 'low' level but summarized to a higher one before analysis.

Strategy

Lecture - Part 1

Good design of a spreadsheet for research data, using MS-Excel as an example.

Exercise 1 (in 2 parts)

Part 1 - Interactive session with resource person, using the example protocol and raw data collection sheet to design a spreadsheet in MS-Excel.

[Resource person will take the participants through the steps of designing the spreadsheet, including explanations of various MS-Excel formatting and validation tools.]

Part 2 - Participants take one of their own protocols and datasets and design an appropriate spreadsheet in MS-Excel for their data.

At the end of this session one or two participants could take 2 - 3 minutes to present their spreadsheet designs for others to comment upon.

Lecture - Part 2

Checking the design information in a data spreadsheet.

Exercise 2 (in 2 parts)

Part 1 - Interactive session with resource person, using the example spreadsheet to perform checks on the design information.

[Resource person will take the participants through the practical steps for checking design information, including explanation of Pivot tables in MS-Excel.]

Part 2 - Participants will use the designed spreadsheet for their own data and check the design information.

Lecture - Part 3

- Complicated designs

- Data collection forms

- File names and saving

[Note to Resource people:

In the exercises for this session participants use their own data to design a spreadsheet. Their trials may have 'complicated designs', as described in Part 3 of the lecture. It may make sense therefore, to present Parts 2 and 3 of the lecture together, and then go on to Exercise 2.]

Required skills

Participants should be familiar with the material covered in Session 1. In addition they should be familiar with the use of spreadsheet packages for data entry.

Files needed for practical exercises

*   *Influence of Improved Fallows.doc* - Protocol for the trial 'The influence of improved fallows on soil phosphorus fractions - an on-farm trial'.

*   *Improved fallows - raw data sheet.jpg* - Raw data sheet for above trial.

Support documentation

*   *Disciplined use of spreadsheet packages for data entry* - Statistical Services Centre (2000).

*   *Session 2: Interactive teaching support (2000).doc* - Teaching support document.

2 4

Designing a spreadsheet for research data

Part 1

Introduction

Session 1 'opened our eyes' to the importance of good data management for our research data. We saw that in order to produce high-quality, valid results from our trials we must ensure that our data are collected, documented and archived to the highest standard. This can be achieved by careful planning of your data management before the start of a project.

In this session we are looking at the design and organisation of our computer files, to store all of the relevant trial information. This does not just mean the 'numbers' that we have measured and collected in the field. We want to include the design information, any additional field observations and basic data summaries. Anyone accessing the files from a particular trial should be able to find and understand almost everything there is to know about that trial.

Project Life Cycle Reference

Referring back to our project life cycle diagram (Figure 0.1) this covers the processes of formulating the research objectives, developing the protocol and designing the observation units.

We are using MS-Excel to design a spreadsheet ready for our data to be entered. MS-Excel is probably the simplest and most familiar spreadsheet package to use for storage of our data. It is a flexible package, which allows us to manipulate our data easily. However, its very flexibility means that, if we do not use it with great care, the end result is poor data entry and management. We need to apply a great deal of rigor and discipline to ensure that this does not happen.

The basic spreadsheet

The recommended layout for experimental data in a spreadsheet has 3 components; **experiment details, design factors and measurement variables** (Table 2.1). There is also a row where we will put short, but relevantly named column titles. This will make it easier for future export to a statistical package.

EXPERIMENTAL DETAILS	
DESIGN FACTORS	MEASUREMENT VARIABLES
COLUMN TITLES/CODES	
FACTOR LEVELS	OBSERVATIONS

Table 2.1 - Spreadsheet layout for Experimental data

Table 2.2 shows a completed spreadsheet that was designed using the template shown in Table 2.1.

Table 2.2 - Completed Spreadsheet

program	4	Systems evaluation and dissemination
project	4.1	Developing choices for farmers
experiment	glm	Gliricidia leucaena mulch trial
Scientist	amh	A. M. Halsman
Location	maseno	Maseno, Western Kenya
Design	rcbd	Randomised complete block design

EXPERIMENT DETAILS

		Type of mulch (leu=Leucaena, gli=Gliricidia, con=Control)	Mulch intensity (t/ha)	Sample air dry cob weight (kg)	Oven dry cob weight (g)	Total air dry cob weight (kg)
DESIGN FACTORS				**MEASUREMENT VARIABLES**		
block#	plot#	type#	intensity#	airdry	ovendry	totaldry
1	1	leu	5	1.27	965	17.7
1	2	con	0	1.95	181	19.9
1	3	gli	10	2.6	208	28.3
1	4	leu	10	2.7	220	25.7
1	5	gli	5	2.42	187	21.08
2	1	leu	10	2.3	215	24.21
2	2	gli	5	2.59	202	24.88
2	3	con	0	2.3	186	25.46
2	4	gli	10	2.27	179	12.5
2	5	leu	5			
3	1	con	0	2.39	198	16.06
FACTOR LEVELS				**OBSERVATIONS**		
3	2	gli	5	1.37	112	16.26
3	3	gli	10	2.45	185	25.56
3	4	leu	10	2.48	196	26.61
3	5	leu	5	2.34	187	22.01
4	1	gli	5	2.5	203	25.31
4	2	leu	5	2.4	195	18.91
4	3	leu	10	2.14	168	24.73
4	4	con	0	2.27	183	23.78
4	5	gli	10	2.49	195	22.87

JPoole: Plot was by gate and maize was stolen before harvest

Experimental details

The top section of the worksheet contains information relating to the whole trial - project name and description, experiment name, leading scientist, collaborators, location, design, and other useful information. It may be easier to allocate short codes to these descriptions (as in Table 2.2) if there are several worksheets in the same file, to save repetition. This information is vital to anyone opening the file as it provides the background to the data.

It is also a good idea to include the names of the data collector, entry technician and data checker in case of further need for data verification and/or validation. This type of information can be recorded in the properties of the workbook.

Use *File* ⟶ Properties ⟶ *Custom* to enter this type of information

Design factors

Before the collected data can be entered into the computer, the design factor columns need to be allocated and labelled. These are columns that identify the rows (block number, plot number) and the treatments applied to them. Every row must be uniquely identified. This unique identity could either be formed by a single column (i.e. 1,2,3...n) or by using a combination of two or more columns. For example, plots in a randomized block design need two columns, one showing the block and one showing the plot. Plot numbers may also show the block (e.g. 101, 102, 103, 104, ...201, 202...) but block identification will be needed as a separate column for analysis.

> *Project Life Cycle Reference*
>
> *In the project life cycle we are here considering the process of designing the observation units.*

Measurement variables

These are columns of the measured data. Each column has a title section at the top that describes the contents of the column in detail, including the units of measurement. All statistical packages read short names for each of the columns, so below the variable description there is a 'name' row with a short code describing the column. In Table 2.2, for example, the 'name' for oven dry cob weight is ovendry. Most statistical software requires variable names that are 8 characters or less, starting with a letter and containing no 'odd' characters or symbols.

Additional notes on spreadsheet design

Rows

The worksheet has one row for each unit (plot, subplot) that is measured. For example, an experiment to compare 5 provenances of sesbania has 4 replicates, giving 20 plots. Each plot has 100 trees of which the height of 15 is measured. Height data will be recorded on a sheet with 20 plots x 15 trees = 300 rows. If the total biomass of all 15 trees is measured, this will require a worksheet with 20 data rows - one for each plot.

Rows are arranged in field order, NOT treatment order. The order of the rows is the order data has been collected in the field, and logically the order of data on the raw data collection form.

Comments

Comments are useful for adding non-numerical information to the spreadsheet. In this way all of the experimental information can be stored in the workbook. Additional information from the field collection can be included here. In Table 2.2, one comment has been added to the spreadsheet to explain why a row of observations is missing. Under no circumstances should you type the comment directly into the cell of the spreadsheet, as this will cause major problems if and when the data are transferred to a statistical package for analysis.

Coding factors and explanation

In the spreadsheet it may be more practical to introduce codes to identify treatments, as in Table 2.2 where the mulches have been coded. Although codes are useful, it is essential that explanations of the codes also be maintained in the workbook. It may be easy to put it above the data as shown in Table 2.2, or if more space is needed you can insert a section into the experimental details section. If there are many factors that are coded and many treatments, it may be more practical to place a table(s) in a separate spreadsheet (but same workbook) to give the details.

Survey data

With survey data you can use a similar template although some of the components will be slightly different. Table 2.3 shows the recommended layout. This is the equivalent of Table 2.1 for survey data.

SURVEY DETAILS	
SURVEY DESCRIPTORS	MEASUREMENT VARIABLES
COLUMN TITLES/CODES	
SURVEY DEFINED VARIABLES	OBSERVATIONS

Table 2.3 - Recommended spreadsheet layout for survey data

Survey details

The top section of the worksheet is similar to that used for experimental details in Table 2.1; it contains information about the survey as a whole. This information provides the background to the survey.

Survey descriptors

In a survey there are no **design factors** as there are in experiments. Instead we are suggesting allocating this section to variables defined by the survey design. These could include the sample site details and/or the unique identifiers for each interview - the interview ID. In Table 2.4, which shows a completed worksheet based on the template in Table 2.3, we have included the name of the enumerator and the date and time of the interview as the data for these variables are determined by the survey itself and not from the responses made to the questionnaire. Unlike with experiments though, these descriptors cannot be entered prior to data entry.

Study	Timber business census			SURVEY DETAILS					
Scientist	C. Holding								
Interview ID	Date of Interview	Start time of interview	Enumerator	Name of business		Name of owner	Date business started	Reason business started 1=saw good opportunity in timber 2=made redundant 3=retrenchment 4=skill known by respondent 5=other	Size of business 1=small 2=medium 3=big
	SURVEY DESCRIPTORS			MEASUREMENT VARIABLES					
IntID	IntDate	StTime	Enum	BusName	Town	Owner	DtStart	WhyStart	BusSize
26	19/06/2001	0.52	P.N. MuigaiHold	next to Kirun Pi	Kirua	Samuel	1994		1 medium
31	19/06/2001	0.17	P.N. MuigaiHold	Michoni w/shop	Kbirichia	Micheni Mbog	1991		4 large
44	20/06/2001	1.50	P.N. MuigaiHold	Nkubu Builder v	Nkubu	Mwongera	1996		4 small
52	20/06/2001	12.50	P.N. Muigai	Mikumbune G. F	Nkuene	Kimathi	1998		1 medium
57	23/06/2001	11.00	P.N. Muigai	Ribui Holdings	Ruiri	Charles Ribui	1982		medium
61	23/06/2001	12.30	P.N. Muigai	Ntarungwi T. d	Ruiri	Ntarangwi	1999		1 small
63	23/06/2001	14.15	P.N. Muigai	Napolian F. Mar	Kanjai	Napolian		No answer	1 small
66	23/06/2001	15.10	P.N. Muigai	Japoto F. shop	Kanjai	Jason Mururi	1986		1 small
67	23/06/2001	15.35	P.N. Muigai	Mutdthia F shop	Kanjai	jackson Itari	1993		1 small
69	23/06/2001	16.20	P.N. Muigai	Franco Furnitur	Kanjai	Franco			1 small
	SURVEY DEFINED VARIABLES			OBSERVATIONS					
70	23/06/2001	16.40	P.N. Muigai	Umoja T.sale	Kanjai	James Muturi	1993		1 small
71	25/06/2001	8.30	P.N. Muigai	sunrise furnitur	Chuka	Kariuki	1997		4 medium
76	25/06/2001	11.00	P.N. Muigai	Chuka youth po	Ndagani/Chu	Govt.	1958		4 small
77	25/06/2001	11.30	P.N. Muigai	Ndagani F shop	Ndagani/Chu	Munithi	2000		small
79	25/06/2001	12.00	P.N. Muigai	kamokia Marima	Marima	Njaga	2001		1 medium
80	25/06/2001	12.20	P.N. Muigai	Muchai njagi F s	Marima	Muchia Njagi	1949		4 small
82	25/06/2001	14.20	P.N. Muigai	Amani wood w	Chogoria	Kirimi	1999		1 medium
83	25/06/2001	14.40	P.N. Muigai	Chogoria furnit	Chogoria	Mitu	1996		1 medium

Table 2.4 - Completed survey spreadsheet

One answer per cell

When storing survey data you need to consider the 'one answer per cell' rule. Some survey questions generate more than one answer, so you must think carefully about how to store these answers. For instance in the Timber Business Census some respondents may give two or more reasons for the business starting up. You should not store data such as **1,4** in a single cell in the worksheet. This is impossible to analyse in this format. In Session 6 we consider this situation in more detail when we talk about **Data Modelling**.

Final tip

Do not leave blank rows either, between the column title row and the data, or within the data. Statistical packages will not accept these.

Part 2

Checking design information

After setting up the **experimental details** and **design factors** in the spreadsheet it is necessary to check that the structure is correct. These checks should be carried out prior to data entry to ensure that no design information is missing. The levels of factors should be checked, as well as the numbers of rows for each factor. For surveys the Interview ID must be unique for each questionnaire.

'Unique' levels of factors

For factors containing text levels (e.g. species names) it is easy to have typed in the wrong spelling of a word into a particular cell. MS-Excel will then have added this label as an extra level to the factor. Note however, that if drop-down menus have previously been used to set up the spreadsheet this problem should not occur.

Number of 'units' for each factor

The number of replicates of each treatment, occurrence of treatments in blocks, etc. should be checked. The number of rows should be equal to the number of units (e.g. plots). The number of units per block, treatment, etc. should also be checked.

We can use **pivot** tables in MS-Excel to carry out these checks.

Part 3

More complicated designs

We have designed a spreadsheet for a fairly simple experiment. It was easy to work out the most appropriate design for the data that had been collected. However, in other experiments it may not be so clear.

The simple rule of a row for each 'unit' (e.g. plot) and a column for each variable, is sometimes difficult to follow. Here are some examples of complications that might arise and how to adapt the basic spreadsheet design to handle these.

More than one type of unit of measurement

It is common to have several types of experimental unit within a single experiment. For example, in a tree variety trial, treatments were applied to plots; biomass yield at final harvest was taken over a whole plot. Hence a plot is a unit and a worksheet must have one row per plot to enter the data. However, tree height and diameter was measured on 9 individual trees in each plot. These data will require a worksheet with a row for each measured tree (i.e. 9 rows per plot). It is easiest to use separate sheets within the same workbook for these two data types.

Project Life Cycle Reference

Referring back to our project life cycle diagram we are talking here about compiling the dataset.

Some analyses of data measured on individual trees will be carried out at the tree level (e.g. relationship between height and diameter). However, other analyses will be done at the 'plot level' (e.g. comparison of the height of different varieties). The tree level data must therefore be summarized to the plot level. This can be done using pivot tables in MS-Excel. See Table 2.5 below which shows the example of soil nitrate measured in the field at three depths per plot; for analysis only an average soil nitrate is required per plot.

Table (a)   Pivot Table (b)

NM-xx 97	Soil Nitrate Measurements		Depth 1=0-15cm
Aug-97			Soil Nitrate 2=15-30cm
			mg/kg 3=30-60cm
Blockl	Plot	Depth	nitrate
1	1	1	9
1	1	2	6
1	1	3	1
1	2	1	18
1	2	2	7
1	2	3	1
1	3	1	22
1	3	2	10
1	3	3	1
1	4	1	19
1	4	2	12
1	4	3	3
2	1	1	13
2	1	2	9
2	1	3	5

Average of nitrate		
Blockl	Plot	Total
1	1	5.3
	2	8.7
	3	11.0
	4	11.3
2	1	9.0
	2	7.3
	3	9.3
	4	9.3
3	1	11.3

The data in (a), which is measured at 3 depths per plot, is summarized using a pivot table (b) to give an average nitrate level per plot.

Table 2.5 - Soil Nitrate example

> *Project Life Cycle Reference*
>
> *In the project life cycle this task is Querying for data selections.*

It is also common in surveys to have several levels of data. For instance a household survey may well ask questions of the individuals within the household thus generating data at both the household and individual levels. As with experimental data in different units, it is recommended that you store different levels of survey data in separate worksheets.

Large numbers of variables

If a large number of variables have been measured or calculated, then the worksheet will get wide and become difficult to navigate around. One idea is to divide it into several sheets within the same workbook. You can copy the **experimental details** and **design factors** from one sheet to the next, so that these are consistent between sheets. It may make sense to group certain variables together within a sheet. For example, for data collected in a multi-season trial, each season of data could be placed on a separate sheet.

Naming each sheet appropriately will make the required worksheet easier to find.

Awkward cases

1. If the data are measured at a 'low' level but will always be summarized to a higher one before analysis, then it is easiest to enter the data as a row for each higher-level unit and a column for each of the lower levels. Summarizing is then a simple matter of a formula across columns.

    For example, tree diameter is measured twice in perpendicular directions. The individual measurements are rarely used in the analysis, but usually averaged. Therefore, we can enter the diameter measurements as two variables (columns) and calculate the average as a third.

2. The number of measurements may not be known until the time the data are collected. For these data it is difficult to design the spreadsheet before data collection.

    For example (see Table 2.6), the lengths of all shoots on a tree are measured. The number of shoots will not be known until the measurements are done. If we enter the data as a row per shoot then we cannot label all of our rows. Conversely, if we enter each shoot as a column (as in Table 2.6) then we may end up with many columns and lots of empty cells, as the tree with the most number of shoots may have many more than a typical tree. The advantage of placing shoots in separate columns is that we can easily summarize the data with a formula across columns.

Experiment: Gliricidia provenance trial
Location: Msekera Research Station, Chipata, Zambia
Scientist: Dr F Kwesiga
Started: 12/1/91

Replicate	Blocks within replicates	plot number	provenance	tree number (the central 9 trees are measured)	Diameter of regrowth shoots, 10cm from stump, 6 months after cutting (cm)						Average diameter of all shoots measured (cm)	Square root of total squared shoot diameter (cm)
rep	block	plot	prov	tree	shoot1	shoot2	shoot3	shoot4	shoot5	shoot6	meandiam	rtotsqr
north	1	1	br12-64	1	1.30	1.50	0.60	0.40			0.95	2.11
north	1	1	br12-64	2	2.40	1.00	1.10	0.30	0.30		1.02	2.88
north	1	1	br12-64	3	0.80	0.90					0.85	1.20
north	1	1	br12-64	4	1.30	1.80	0.50				1.13	2.12
north	1	1	br12-64	5								0.00
north	1	1	br12-64	6	2.60	1.90	1.50	1.20	0.70	0.30	1.40	3.96
north	1	1	br12-64	7	3.50						3.50	3.50
north	1	1	br12-64	8	1.60	1.30	0.80				1.23	2.21
north	1	1	br12-64	9	1.80	1.70	0.60	0.50			1.15	2.60
north	1	2	kr-16-84	1	3.15	2.22	1.53	0.08			1.74	4.15
north	1	2	kr-16-84	2	2.48	1.34					1.91	2.82
north	1	2	kr-16-84	3	3.91	3.07	2.61	1.72	2.58	1.78	2.54	6.51
north	1	2	kr-16-84	4	1.13						1.13	1.13
north	1	2	kr-16-84	5	2.16	1.99	1.29	1.19	0.85	0.87	1.39	3.63
north	1	2	kr-16-84	6	2.82	2.08					2.33	3.27
north	1	2	kr-16-84	7	3.86	3.43	2.87	2.30	2.57	1.27	2.83	6.76
north	1	2	kr-16-84	8	2.91	2.93	1.04	1.87	2.79		2.31	5.42
north	1	2	kr-16-84	9	1.10	1.94	2.47	1.76	1.63		1.78	4.10
north	1	3	kr-23-84	1	2.98	2.69	2.00				2.56	4.49
north	1	3	kr-23-84	2	3.53	1.81	2.02	2.97	0.33	0.57	1.87	5.39
north	1	3	kr-23-84	3	3.00	2.11	2.00	1.65	1.74		2.10	4.82
north	1	3	kr-23-84	4	3.79	2.81	1.61	1.25	0.59	0.53	1.76	5.20
north	1	3	kr-23-84	5	3.19	2.66	1.99	1.73	1.15		2.14	5.05
north	1	3	kr-23-84	6	2.86						2.86	2.86
north	1	3	kr-23-84	7	2.90	1.34					2.12	3.19
north	1	3	kr-23-84	8	3.55	2.57	2.20	1.23	0.41		1.99	5.07
north	1	3	kr-23-84	9	3.03	2.11	1.90	1.30	0.96	0.82	1.70	4.55

Table 2.6 - Shoot level data stored as columns

However, it is harder to analyse the data at the shoot level. Perhaps it would have made more sense to have only measured the 5 - 6 longest shoots! In Session 6 we discuss an alternative method for dealing with these data.

Project Life Cycle Reference

Again referring back to our project life cycle, the process described here is compiling the dataset.

What is a row and what is a column?

Deciding what 'unit' constitutes a row is not a trivial task. In the next example (Table 2.7) it is difficult to say which is the superior design, the 'long' version or the 'short' version.

The type of analysis to be used will dictate which version is required for the data. It is possible to manipulate the data in MS-Excel between the two versions. However, although pivot tables are powerful tools for summarizing data to a higher level and for turning data entered as rows into data as columns, the reverse (i.e. from 'short' version to the 'long' version) is NOT as straightforward.

Short version

Block!	Treatment!	number damaged leaves (harvest 1) harv1dam	number marketable leaves (harvest 1) harv1mark	number damaged leaves (harvest 2) harv2dam	number marketable leaves (harvest 2) harv2mark
1	1	182	250	103	234
1	2	249	255	182	245
1	3	246	273	154	294
2	1	240	200	99	200
2	2	281	261	95	228
2	3	503	125	161	196
3	1	375	265	192	235
3	2	413	107	339	149
3	3	364	105	316	126

Long version

Harvest date Harvest!	Block!	Treatment!	number of damaged leaves numdam	number of marketable leaves nummark
1	1	1	182	250
1	1	2	249	255
1	1	3	246	273
1	2	1	240	200
1	2	2	281	261
1	2	3	503	125
1	3	1	375	265
1	3	2	413	107
1	3	3	364	105
2	1	1	103	234
2	1	2	182	245
2	1	3	154	294

Table 2.7 - Two displays of the same data

Data collection forms

Looking back at the spreadsheet we designed earlier we can see that the spreadsheet with the **experiment details** and **design factors** filled in would make an ideal raw data collection form. It contains all of the information that a technician might reasonably need to include on their form.

If the spreadsheet has been designed at the start of an experiment, as is recommended, then we could just print off the spreadsheet and use that as our data collection form. The added advantage of doing this is that data entry into the computer after collection should be more efficient as the layout and format of the forms is identical.

Project Life Cycle Reference

Obviously the process we are talking about here is that of collecting the data.

Saving and file protection

Ideally only one workbook per experiment or survey is required for keeping all of our data. However, large datasets and experiments or surveys with several different aspects of research may justify the use of more than one workbook.

Naming

Keep data files in directories for data, NOT in directories where programs are stored (e.g. MS-Excel worksheets, with extension .XLS, should not be stored in the directory where the MS-Excel program is stored, such as \EXCEL).

Give files and directories meaningful names. Organize directories in a way that makes sense so that the researcher (and anyone else) can find the files easily. Researchers often keep all data files in a directory called \USER. Individual research locations have a subdirectory, and each trial is a subdirectory of the location. Thus the 'Prototype hedgerow intercropping trial' at Machakos, Kenya has data stored in a directory called \USER\MACHAKOS\PROTO. See Figure 2.1 for an example.

Project Life Cycle Reference

This aspect of data management is to do with compiling the dataset.

All files related to the trial are kept in the same directory. Thus raw data, reports, statistical programs to do the analysis, graphics etc. for the 'Prototype Hedgerow Intercropping trial' are all found in \USER\MACHAKOS\PROTO.

MS-Excel provides an opportunity to record **summary information** (title, subject, keywords) about a workbook. Use these if you are going to end up with many workbooks. Take advantage of the long file names that are supported by modern version of the Windows operating system (Windows 98, Windows 2000, Windows NT, etc.). These names can have as many as 256 characters and can include spaces.

To record summary information use *File* ➡ *Properties*
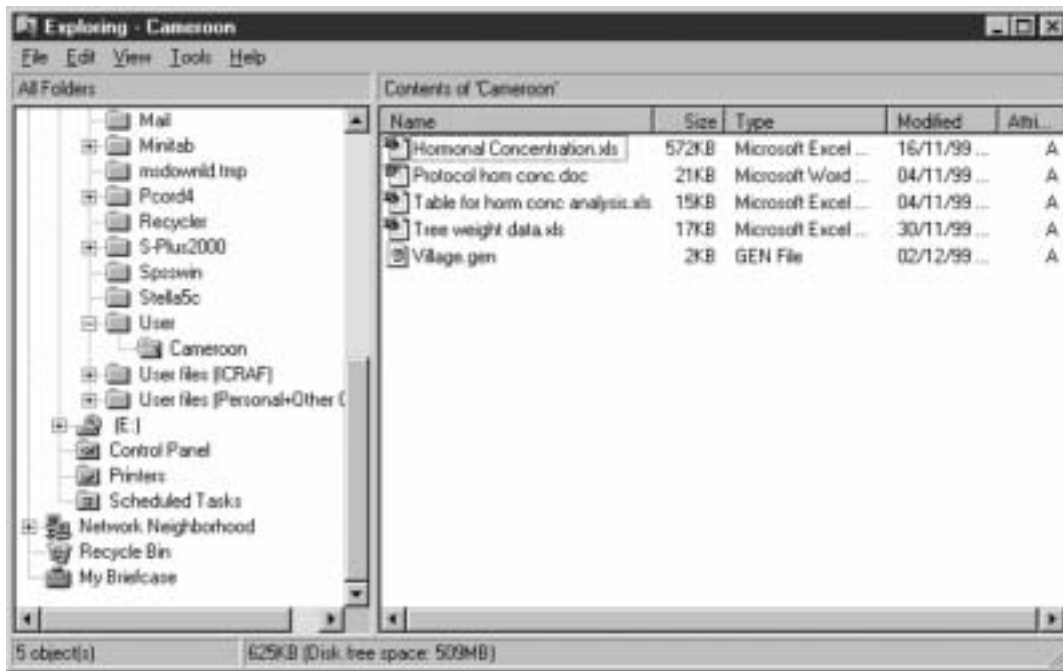
Figure 2.1 - PC File system structure

Saving

Do it!

Every day that goes by without a backup of your files, is one day too many. You are gambling that your hard disk won't 'crash', that someone won't accidentally erase your files, that a disaster won't occur, and that a virus won't infect your data. Hard disks, floppy disks and disk drives can fail at any time. They can also be stolen, burnt down, lost, abandoned during emergencies or accidentally overwritten. At least two copies of data files, in addition to the working file on the hard disk, or floppy disk, should be kept. Keep one in a different location to the computer (e.g. at home) and keep them up-to-date.

Backup method

The only way you can really protect your data is to keep it in more that one place. Depending on the operating system, there are many ways to backup files/directories. Using a floppy disk is one choice, if one floppy is not enough then you can use several disks. You could also back up to tape and keep the tape in a safe place (preferably 'off-site'). Many computers also now have CD-RW writer drives and one CD-ROM can usually hold all of your files.

Master Copies

Sometimes when using workbooks scientists save each updated version under a new file name. This is NOT advisable as the end result is several files containing similar information and difficulty in locating the most updated edition. It is preferable to maintain just one copy of the workbook (Master Copy). Adjustments to the data within the file can then be highlighted and/or comments added to explain changes.

File Protection

It is also advisable to add protection to the workbook so that any spreadsheet changes can be either highlighted or actually prevented.

To set up rules for file protection use

Tools ⟶ *Track Changes or Tools* ⟶ *Protection.*

Designing a spreadsheet for research data

Exercise 1 - Part 1 (Example protocol and data)

You will complete this exercise under the guidance of the resource person. Working in pairs, read through the protocol **The influence of improved fallows on soil phosphorus fractions - an on-farm trial** (*Influence of Improved Fallows.doc*).   Now look at the **raw data collection sheet** (*Improved fallows - raw data sheet.jpg*) which has been made for this trial. We will now design a spreadsheet ready for import of these data. (N.B. we are only looking at 4 of the farms).

- Work out what Experimental Details information needs to be put on the spreadsheet.

- Decide what Design Factor details are going to be required. You need to think about column headings/titles and what 'unit' constitutes a row in the spreadsheet.

- Code your factors and make sure that an explanation of the codes is included in the spreadsheet.

- With the help of the resource person, design the spreadsheet.

Exercise 1 - Part 2 (Participants' data)

Now repeat the task of Exercise 1 - Part 1, but using a trial protocol provided by one participant in each pair. If you have time, you can also do your partner's design. If your design is for a survey you can follow much the same procedure. We have demonstrated the design of a spreadsheet for experimental data, as that is the more complicated of the two.

If your data is already entered onto an MS-Excel spreadsheet, examine it; decide what experimental information is missing and whether the design factors are constructed correctly. If your data are from a survey ensure that each questionnaire has a unique identifier. Identify the survey descriptors and move those columns to the left of the measurement variables. Edit your spreadsheet until it is 'well designed'.

Exercise 2 - Part 1 (Example protocol and data)

We now need to check the design information that we put into our spreadsheet in Exercise 1 - Part 1. With the help of the resource person you need to:

*   Check the levels of each factor. Do you have the correct number of levels and the correct titles for each?

*   Use a pivot table (the resource person will help you with this) to check the design layout. You are looking to see if the number of occurrences of each treatment, in each block, is correct. The design is split-plot with 4 x 2 treatments in total. We are looking at data from 4 farms only. Therefore, we should have each treatment occurring 4 times, in total, and once on each farm.

Exercise 2 - Part 2 (Participants' data)

Now repeat the first part of Exercise 2, using the spreadsheet you designed earlier for your (or your partner's) data. For survey data, the only one of the survey descriptors that you are likely to know in advance is the interview ID.

Spreadsheet data entry and checking

Objectives

- To form a strategy for data entry into a spreadsheet which allows for validation of the data, during and after entry.

- To learn various techniques for data checking, after entry into a spreadsheet. This includes the use of pivot tables, graphs and validation checks.

Summary

Spreadsheets such as MS-Excel are widely used for data entry. To be effective for this they have to be used with discipline.

Data should be transferred directly from the raw collection forms to the MS-Excel spreadsheet. All of our necessary calculations (e.g. converting yield per plot to yield per hectare) should be carried out in MS-Excel. Doing calculations and conversions 'by hand' will very likely result in errors and therefore require more data checking once it is in MS-Excel. Data entry should be done as soon after collection as possible so that queries can be followed up. Similarly a data summary should be given to a supervisor or scientist responsible for a trial, to check that the numbers look reasonable.

There are a number of techniques for checking data in MS-Excel. Most of the techniques require an understanding of the context of the data and the sort of numbers and patterns that are expected.

Strategy

Lecture - Part 1

Entering data into a spreadsheet.

Exercise 1 (in 2 parts) - Optional

Part 1 - Interactive session with resource person, entering the data into the spreadsheet designed in Session 2: Exercise 1 - Part 1.

[Note to Resource people:

If the design of the spreadsheet was not completed in Session 2, then you need to provide participants with the designed example spreadsheet. This file named *Improved fallows designed sheet.xls* is provided on the accompanying CD.]

(Resource person will take the participants through the steps of entering data into the spreadsheet, including explanations of MS-Excel formatting and validation tools.)

Part 2 - Taking the spreadsheet participants' designed for their own data, and inputting the data.

[Note to Resource people:

This exercise may become lengthy or/and be unnecessary to participants learning. Part 1 of the exercise is useful for advanced level participants who can start using the MS-Excel validation tools. Part 2 of the exercise must be completed if participants do not have their data already in MS-Excel.]

Lecture - Part 2 & 3

Adding calculations and conversions to the spreadsheet. Data checking using pivot tables and graphs. Preparing the data for import to a statistical package and MS-Excel Help.

Exercise 2 (in 2 parts)

Part 1 - Interactive session with resource person. Carrying out checks on the data entered, adding calculations etc. as instructed on the Exercise sheet.

(Resource person will take the participants through the steps of checking data in the spreadsheet, including explanations of how to construct graphs, how to design pivot tables and using the autofilter MS-Excel tool.)

Part 2 - Checking the data from participants' own spreadsheets.

[Note to Resource people:

If Exercise 1 - Part 1 has not been completed, then you need to provide participants with the example data MS-Excel spreadsheet. This Data Sheet is provided in the accompanying CD in the file named *Improved fallows complete sheet.xls*.]

Required skills

Participants must be familiar with the material covered in Session 2. They should be comfortable using spreadsheet packages for data entry and calculations.

Files needed for practical exercises

- *Improved fallows designed sheet.xls* - Designed MS-Excel spreadsheet for the trial 'Influence of improved fallows on soil phosphorus fractions'.

- *Improved fallows complete sheet.xls* - Complete MS-Excel spreadsheet for the trial 'Influence of improved fallows on soil phosphorus fractions'.

- *Improved fallows - raw data sheet.jpg* - Data collection sheet from the trial 'Influence of improved fallows on soil phosphorus fractions'.

Support documentation

- *Interactive teaching support.doc* - 'Session 3 - Practical details required for interactive exercises'.

- *Disciplined use of spreadsheet packages for data entry* - Statistical Services Centre (2000).

4 4

Spreadsheet data entry and checking

Part 1 - Data entry

Data entry

Data entry should be started as soon after collection in the field as possible. If the spreadsheets have been designed before the data were collected, then the job of inputting the field observations can begin immediately. One advantage of prompt data entry is that the data collection is still 'fresh' in the data collector's mind. The observations made in the field and any notes made on the data collection form can still be explained easily. How much more difficult is it if the data are input into the computer, say, a year later? Memories have faded and perhaps the technicians have even changed jobs!

After initial data checks by the technician who carried out the data entry, a data summary should be given to the supervisor or scientist responsible for the trial. They should check that the numbers look 'reasonable'.

Another important aspect of data entry is that the raw data collected should be entered directly into a computer. You should NOT be carrying out calculations (e.g. % dry matter) or conversions (e.g. kg/ha to t/ha) by hand. Doing calculations and conversions by hand will very likely result in errors and therefore require more data checking once the data are in MS-Excel. We will see later in this session that we can write formulae in MS-Excel to produce these calculations and conversions. In MS-Excel, as long as the formula is written correctly, we will produce no errors. It is easy to update the formulae if corrections to the data are made and it is also easier for you or anyone else to trace the operations in data processing.

> *Project Life Cycle Reference*
>
> *In our project life cycle these points are concerned with the processes of collecting the data and compiling the dataset.*

Part 2 - Data checking

Adding calculations and conversions

After data entry, it is likely that we will want to perform certain calculations on our data (e.g. sum of wood biomass and leaf biomass to give total biomass). Calculations can be written in MS-Excel using formulae. We will look at the practical details of writing these formulae in the next practical session (Exercise 2).

We can use our calculations and conversions for data checking. For example, if we have collected grain yield per plot, it may be difficult to see whether the values are reasonable. However, if we convert the yield to yield per hectare, then we can compare the numbers with our scientific knowledge of grain yields.

We can also write simple formulae to check for consistency in the data. For example, if we have measured tree height at 3 times in the year we can write a formula that subtracts 'tree height 1' from 'tree height 2'. The numbers in the resulting column should all be positive…we cannot have a shrinking tree!

Project Life Cycle Reference

In our life cycle this is the process of querying for data selection.

For new columns of calculated or converted data we need to remember to include suitable header information (what the new column is, units and short name) at the top of the data.

Missing values

We should check in the spreadsheet that the only missing values (i.e. empty cells within the Measurement Variable columns) are actual 'missing values'. In MS-Excel the missing values are BLANK cells. It is useful to know this when calculating formulae and summaries of the data. For example, when calculating the average of a number of cells, if one cell is blank MS-Excel ignores this as an observation (i.e. the average is the sum/number of non-blank cells). But if the cell contains a '0' then this is included in the calculation (i.e. the average is the sum/number of cells). The scientist and technicians should make sure that the data entry distinguishes where necessary between the reasons for a missing value.

In a column of 'number of fruit per plot', a missing value could signify zero (tree is there but no fruit), dead (tree was there but died so no fruit), lost (measurement was lost, illegible…) or not representative (tree had been browsed severely by goats). In this example, depending on the objectives of the trial, the scientist might choose to put a '0' in the cells of trees with no fruit and leave blank (but add comments) for the other 'missing values'.

Note that there should be NO missing values in the Design Factor columns.

Project Life Cycle Reference

These points come into the processes of developing the protocol and designing the observation units.

Dates

The storage of dates in computer files should be straightforward, but is often a nightmare. Often people type **2/5/01** to mean **2nd May 2001**, but the computer records this as **5th Feb. 2001**. The problem is with the regional settings on the computer. In many cases the default date format is **mm/dd/yy** which is the format used in the USA. To change this default you must change the settings on your computer.

Go to **Start** ➤ **Settings** ➤ **Control Panel** ➤ **Regional Settings**.

Go to the **Date** tab and ensure the date styles are **dd/MM/yyyy** for short date

and **dd MMMM yyyy** for long date.

In many applications there is the possibility of formatting date cells to display the month as letters (e.g. Jan., Feb., etc.). This can be used as a check to ensure you are entering the date and month in the correct order.

We recommend you display the year with all 4 digits to avoid any problems. Although the so-called 'Millennium Bug' did not really materialize, it must be remembered that software applications interpret two-digit years as lying within a particular range.

In MS-Excel for instance, a two-digit year is assumed to be between 1930 and 2029. Generally this range is fine but what about survey data for instance when you are asking for dates of birth? According to MS-Excel someone with the birth date of 6/9/23 has not yet been born!

Techniques for data checking

Pivot Tables (to check consistency between replicates)

Variation between replicates is expected, but some level of consistency is also usual. We can use pivot tables to look at the data. We can use our knowledge of what happened in the field to spot errors. In Table 3.1 below, the grain yield of 21 for replicate 2 of treatment 5 is unusually high compared to the other two replicates of treatment 5. A check of the original data will show whether this plot really was this high yielding or whether a recording error has been made. In Table 3.1, average of grain yield within each treatment and block are shown.

Average of Grain	Block			
Treatment	1	2	3	Grand Total
1	0.2	0.5	0.6	0.4
2	0.7	0.8	1.0	0.8
3	1.1	1.3	1.0	1.1
4	1.6	1.5	1.0	1.4
5	1.5	21.0	3.4	8.6
6	7.2	4.3	8.3	6.6
7	6.5	4.4	7.6	6.2
8	5.3	2.8	4.8	4.3
9	6.4	4.9	7.9	6.4
10	7.9	6.5	5.9	6.8
Grand Total	3.8	4.8	4.2	4.3

Table 3.1 - Pivot table showing average grain yield within each treatment and block

Scatter Plots (to check consistency between variates)

We can often expect two measured variables to have a fairly consistent relationship with each other. For example, 'maize grain yield' with 'maize stover yield' and 'number of

fruits' with 'weight of fruits'. To look for odd values we could plot one against the other in a scatter plot. Scatter plots are useful tools for helping to spot outliers. Graph 3.1 shows a scatter plot of stover plotted against grain yield. This plot shows that data lie from 0.0 to 10 units, with one suspicious observation with high grain yield (>20). Such errors could be trapped at data-entry if validation rules are set up.
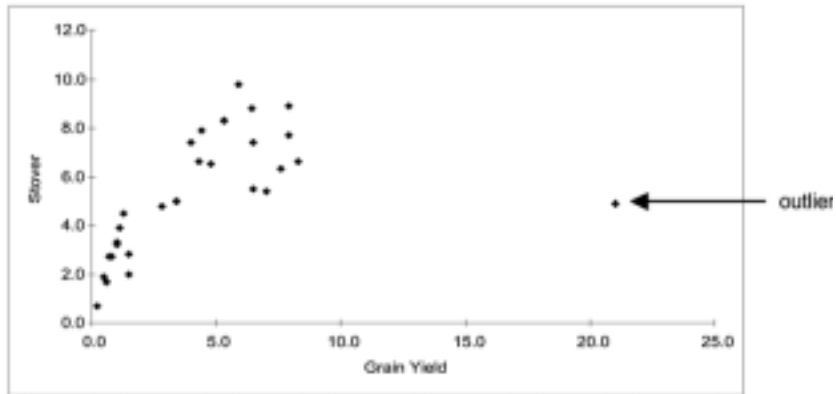
Graph 3.1 - Stover against grain yield

Line Plots (to examine changes over time)

Where measurements on a 'unit' are taken on several occasions over a period of time it may be possible to check that the changes are realistic. For example, Graph 3.2 shows the changes over time for measurements of plant infection by a particular insect. We can see clearly that there is a gradual increase in the percentage of plant infection over time. However, on day 11, variety C records a much lower level. This suggests a problem in the data for that day and variety. A check back at the data will identify the doubtful value.
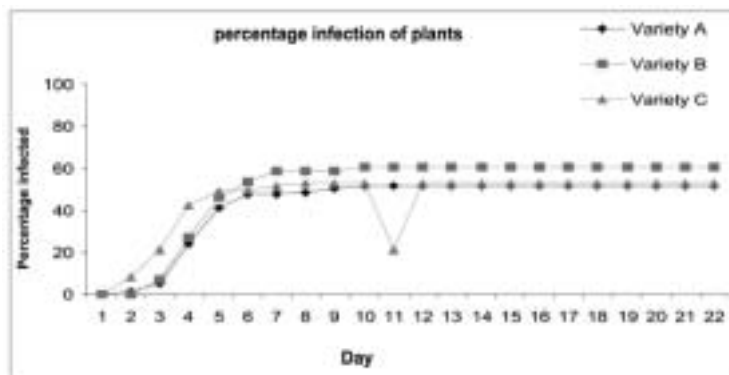
Graph 3.2 - Line plot to show infection of plants

Double data entry

One effective, although not always practical, way of checking for errors caused by data entry mistakes is double entry. The data are entered by two individuals onto separate sheets that have the same design structure. The sheets are then compared and any inconsistencies are checked with the original data. It is assumed that the two data entry operators will not make the same errors.

There is no 'built-in' system for double entry in MS-Excel. However, there are some functions (which we talked about in the calculations/conversions section) that can be used to compare the two copies. An example is the DELTA function that compares two values and returns a 1 if they are the same and a 0 otherwise. To use this function we would set up a third worksheet and input a formula into each cell that compares the two identical cells in the other two worksheets. The 0's on the third worksheet will therefore identify the contradictions between the two sets of data.

This method can also be used to check survey data but for the process to work the records must be entered in exactly the same order in both sheets. If a section at the bottom of the third worksheet contains mostly 0's, this could indicate that you have omitted a record in one of the other sheets.

Final tip

When you've found errors or values that require checking it is a good idea to highlight them (using one of the symbols shown below) so they can be easily referred to. 'Odd' values that are consistent with values on the raw data collection form must be referred to the data collector.

Part 3

Preparing data for export to a statistical package

Statistical analysis of research data usually involves exporting the data into a statistical package such as GenStat, SAS or SPSS. These packages require you to give the MS-Excel cell range from which data are to be taken. In the latest editions of MS-Excel we can mark these ranges within MS-Excel and then transfer them directly into the statistical packages.

Project Life Cycle Reference

Here we are talking about selecting data so this fits into the process of querying for data selection.

Resource person to demonstrate:

- **Highlight the data you require including the column titles (the codes which have been used to label the factors and variables).**

- **Go to the Name Box, an empty white box at the top left of the spreadsheet (shown below). Click in this box and type a name for the highlighted range (e.g. Data).**

- **Press Enter.**

- **From now on, when you want to select your data to export go to the Name Box and select that name (e.g. Data). The relevant data will then be highlighted.**

Note that you can also use this method of range setting to quickly select data for other uses. For example, if you want to create several pivot tables, rather than having to highlight the data for each new table you can just click on the range name and the data will be immediately highlighted.

MS-Excel help

If you get stuck on any aspect of MS-Excel then use the Help facility. It contains extensive topics and by typing in a question you can extract the required information. See the picture below for an example.
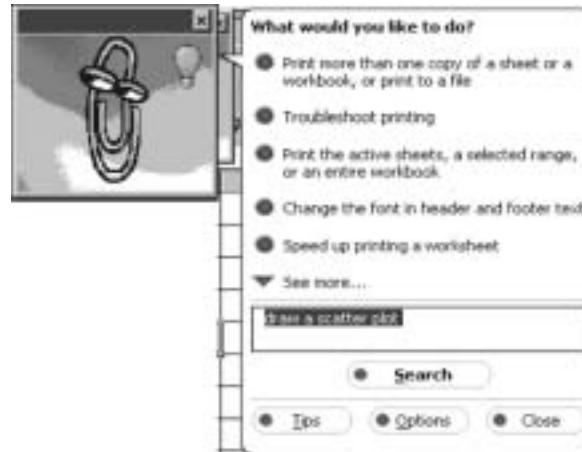
Exercise 1 - Part 1 (Example data) Optional

Working in pairs, and with the help of the resource person, enter the data from the collection sheet for the 'Influence of improved fallows on soil phosphorus fractions trial' (*Improved fallows - raw data sheet.jpg*) into the spreadsheet we designed in Session 2.

If you did not complete the design of the spreadsheet in Session 2, then the resource people will provide you with a fully designed MS-Excel spreadsheet from the accompanying CD with the file named *Improved fallows designed sheet.xls*.

If some values are difficult to read, input the data you think are correct. If there are problems we will find them during our data checking.

The resource person will show you how to add comments to your spreadsheet and set up validation checks before you enter the data.

Exercise 1 - Part 2 (Participants' data) Optional

Repeat Part 1 of the exercise, using your own (or your partner's) data and the spreadsheet you designed in Session 2.

Exercise 2 - Part 1 (Example data)

With the help of the resource person, carry out checks on the example data set you entered in Exercise 1 - Part 1. Try to carry out the following instructions to perform checks on the data. If you find any curious or incorrect values, highlight them and decide how to check and correct them (e.g. speak to the data collector, see the raw data sheets etc.).

- Calculate percentage dry matter = dry sub-sample / fresh sub-sample 100. Any interesting values?

- Convert the fresh weight of cobs from kg/plot to t/ha (= fresh weight per plot / plot area 10). Any interesting values?

- Calculate the dry weight of cobs in t/ha = fresh weight cobs t/ha   (% dry matter/100).

- Calculate cob sub-sample fresh weight   cob sub-sample dry weight. What type of numbers do you expect in the resulting column? Are there any problems?

- Produce a pivot table with Farmers as the columns, Fallow and Nitrogen treatments as the rows and dry cob weight t/ha as the cell numbers. The resource person will show you how to create the pivot table. Look at the pattern across farms and within farms. Are there any unusual cob yields?

- Produce a scatter plot of cob weight per plot versus stover weight per plot. What pattern do you see? Are there any data points that do not fit into this pattern?

- Finally, set a range to include all the data, ready for import into a statistics package.

If you did not complete Exercise 1 - Part 1 in this session, the resource people will provide you with an MS-Excel spreadsheet (*Improved fallows complete sheet.xls*) containing the raw data. Carry out your checks and answer the questions using this spreadsheet.

Exercise 2 - Part 2 (Participants' data)

Carry out similar checks and calculations on your own (or your partner's) data.

summary 4

Objectives

The main objective of this training session is to **alert** participants to the limitation of using spreadsheets as a base for building a research data management system, and to **introduce** databases as a more elegant alternative that goes beyond spreadsheets.

Summary

Simple datasets are easy to manage with a spreadsheet, but many interrelated data are not. They require creating multiple copies of the same data to produce the intended result in an intuitive way. The separate copies are difficult to keep consistent for an active dataset. The appropriate way of using spreadsheets that is efficient and less error-prone is not intuitive at all. It requires deep understanding of advanced functions that are very spreadsheet-specific and a lot of self-discipline in avoiding shortcuts. Data managed as relational databases, on the other hand, are more likely to produce accurate results because they allow integrity rules to be enforced and are efficiently organized for querying when the volume is large and the correct relationships between them are well defined and implemented.

Strategy

Œ   Short introduction to this session (30 minutes).

Œ   Practical exercises, in groups, that are designed to lead users into experiencing common problems of data integrity, organization and query when they try to perform simple tasks on data that have been managed using spreadsheets (6 exercises of 30 minutes each).

Œ   A plenary session to demonstrate appropriate spreadsheet solutions to some of the problems experienced during the exercise, as well as the database approach of addressing the same problems (60 minutes).

Required skills

Participants will be familiar with the material covered in Sessions 2 and 3. In particular they should know how to:

Œ   Create a well-designed spreadsheet for data entry.

Œ   Use pivot tables for summaries and data checking.

Œ   Use formulae to carry out calculations in spreadsheets.

Files needed for practical exercises

Œ   *Onfarm intercropping.xls* - Data from the experiment 'Mixed intercropping with *Gliricidia sepium* and relay cropping with *Sesbania sesban*'.

Œ   *Soil conservation in Embu.xls* - Data from the experiment 'Soil and water conservation using contour hedgerows of fodder grass and trees'.

Œ   *Mulch Trial.xls* - Data from the experiment 'The effect of mulch on crop performance'.

Why go beyond the spreadsheet

Introduction

A good research data management system will, among other properties, allow data to be processed efficiently and accurate results to be obtained. For data recorded manually in the field, efficient processing of most data requires that they be transcribed from paper forms into electronic media. Most researchers manage their data using a system based entirely on spreadsheets. This is often efficient for some very common processing tasks that need to operate on one data sheet at a time. However, the systems become poor for data management (i.e. inefficient and error-prone) when there is a need to query related data that are held as separate sheets in the same (or different) files. This lecture note describes six simple tasks. They are chosen to illustrate that the intuitive spreadsheet-style of querying data is inefficient and error-prone for handling some common situations. All these tasks have a more sophisticated spreadsheet function of achieving better results – but they are not intuitive and very much depend on detailed specific knowledge of the spreadsheet software used. Solutions based on a database system will be presented, highlighting the features of a database that help to achieve some level of processing efficiency beyond that of a spreadsheet.

Data integrity

This task is designed to let participants experience the limitations of a spreadsheet in registering simple data integrity rules and enforcing them during data entry to ensure accurate datasets. One example that could illustrate this point is when different treatment levels of the same factor are applied to a plot. The dataset used in this exercise was obtained when 41 farms were assessed for slope, weeding and fertilizer application. One record per farm is expected. However, the entries in the dataset are more than expected, because of duplicate entries that have different assessment for the same farm. Participants will be asked to suggest ways of preventing this simple violation of data integrity – i.e. no more than one record per farm. Those with more knowledge of MS-Excel might suggest the use of the Data/validation option, but will realize that it is limited to simple range checks. In the plenary session the instructor will demonstrate how to define a data entry table in MS-Access that has the farm identifier designated as the primary key, thus registering this integrity rule. The instructor will attempt to input

duplicate farm data, raising an error message that disallows the input – thus demonstrating that the database system is indeed enforcing the constraint.

Data organization

This task is designed to alert users to the fact that spreadsheets do not recognize column names, but column positions. This is a headache when you want to append one dataset to the bottom of another, to form a combined dataset that is suitable for analysis as a single unit – otherwise you end up with a dataset that mixes oranges with apples due to columns not identically positioned in the different sheets – even though they are labeled appropriately. The data for this exercise are maize yields over 4 seasons – all recorded on separate sheets. The columns are headed appropriately, but are not in identical positions. If the participants do not notice this they will certainly mix apples and oranges, ending up with an invalid result. If they do (notice) then they will go through the tedious operation of cutting and pasting to align columns – also an error-prone operation that is likely to produce invalid results.

> *Project life cycle Reference*
>
> *Here we are referring to the process of compiling the dataset.*

In the plenary session, the instructor will demonstrate how to combine the datasets seamlessly and efficiently by importing the first data sheet into MS-Access. The subsequent data sheets will also be imported and appended to the first data sheet. This will work if the columns are named identically and if the appropriate range for import has been defined. It will be difficult to define the import range if some of the recommendations for data organization in a spreadsheet (Session 3) have not been observed.

The instructor will demonstrate how simple calculations are done in a database environment.

> *Project life cycle Reference*
>
> *This uses the process of querying for data selection.*

Simple query

This task is designed mainly as an input to the more advanced query presented in the next section. It may or may not be simple to do, depending on the participants' level of knowledge of MS-Excel. The exercise is on selecting plots on farms, based on the tree

species growing in the plots. The ordinary users will most likely select by inspection – which works for few entries, but is less efficient when there are many plots. Some of the more knowledgeable users might sort the data to isolate the required plots; others might use the Autofilter option in MS-Excel. Both of these require that the data follow many of the recommendations of Session 3.

> *Project life cycle Reference*
>
> *This task also uses the process of querying for data selection.*

The main lesson from this exercise is that when there is an option that does just exactly what you want to do, then the spreadsheet is a great attraction. Sometimes there is not, and then you have to work out your own unique way of achieving your objectives. Different people will do it differently often with lots of cut-and-paste operations that are difficult to define clearly. Errors can be readily introduced, and it is extremely difficult to share experiences among people wanting to do data management the same way. Standardization is difficult.

The equivalent database style of selecting datasets will be demonstrated. It involves looking at the data in the spreadsheet through a database front end. This is achieved by creating both a database file and a table that is linked to the data in the spreadsheet. You demonstrate the concept of a query (as an object that is distinct from a query result) by creating one, naming it and saving it for later use. It should be stressed that queries are fairly standard objects in a database, and are extremely valuable for data management. In Session 5 we will learn about data querying in more detail.

Advanced query

This task will let participants experience that in spreadsheets it is difficult to perform complex operations without duplicating the data. The exercise requires the participants to use the results of the previous task to calculate some simple results. They will not find it easy to do without copying the same data elsewhere – thus introducing problems of consistency often associated with multiple occurrences of a dataset that is likely to be updated.

There are ways to get round the problem (e.g. use of pivot tables, use of paste-link – not just paste etc.) but they are not well known to ordinary users. These will be demonstrated in the plenary session, together with the database approach of achieving the same results.

The database approach will be saving the query formulated in the previous task; then use the query and one of the aggregate functions to summarize the data.

Related data

This task is about working with two data sheets that have a one-to-many relationship between them. Participants will be asked to perform exercises that will be difficult without joining the data sheets, manually by cutting and pasting or automatically by use of specialised formulae. They will appreciate that the manual join is tedious and error-prone; it is equivalent to entering the data a second time. It is not efficient. The better-automated way uses the specialized Vlookup function, which is not so intuitive to use, and it is expected that most participants will not be familiar with it. The instructor would demonstrate its use in a plenary session after the participants' exercise.

> *Project life cycle Reference*
>
> *This is related to the process of compiling the dataset.*

The database approach to working with related tables will also be demonstrated in the plenary session. A database file will be created. Using the File/Get External data/Link Tables option the instructor will create 2 tables – one linked to each data sheet. Then a query built with fields from the 2 joined tables will be formulated and saved. It will be stressed that, in a database environment, it is very easy to specify a join between tables using a common field.

Real situation

The dataset to be used for this exercise has been captured by a different set of people to those who will analyse it. They are captured on paper to follow the field plan naturally, and to minimise errors of transcription from paper to spreadsheet, they are entered to match the paper forms. Participants will experience the difficulty of using this data in its current form to answer a simple question that is designed so that contributions from all the data sheets are required. Even the database environment will not offer anything more to simplify working with the datasets, without re-organizing the data in a drastic way or resorting to advanced programming techniques for large datasets.

The Logbook system will be introduced to demonstrate how it increases data processing efficiency and reduces errors associated with data organization in situations similar to this. The system has two components: a predefined database designed for experimental data, and software that simplifies working with the database. In the plenary session the process of loading the data sheets into the empty database will be demonstrated. Following that, the data will be retrieved, and participants should witness that the results of the output query are exactly what they may have tried to generate manually.

Why go beyond the spreadsheet

Following, are six tasks, which are to be performed with data generated from 3 experiments. The first section of this document presents a very brief description of each experiment (in terms of objectives, experimental design, soil and crop measurements) and the data files associated with it. Before you start, ensure that the data files have already been copied to your computer's desktop.

For each task, plan what you are going to do, writing the steps down carefully. Then do it. Record your experiences, noting especially, actions that are likely to introduce errors in to your data as well as operations that you find tedious and repetitive.

Experiments

1. Mixed intercropping with *Gliricidia sepium* and relay cropping with *Sesbania sesban*

Objectives

1. To collect biophysical data to assess the effect of green manure of *Sesbania sesban* and *Gliricidia sepium* on maize yield.

2. To assess the compatibility of these technologies on maize yield.

Observations

Maize yield every season.

Initial soil samples were taken in October 1994 and analyzed on pH, total N, available P, Mg, K and Ca. Soil samples will be taken again at the end of the trial. At the end of the trial the bulk density of the experimental plots will be measured.

Data file

Onfarm intercropping.xls

This data file contains 2 data sheets: one for data taken at the plot level, the other for data collected at the farm level. The common link between them is the **farmid**.

2. Soil and water conservation using contour hedgerows of fodder grass and trees

Objectives

1. To determine the degree to which grasses and trees, in combination and alone, can reduce soil erosion.

2. To determine the amount and quality of fodder these various combinations can produce over time.

3. To determine the degree of competition.

Treatments

1. Control: No grass or trees.

2. Grass hedge: *Pennisetum purpureum*.

Observations

Grain and total yield on a per-row basis, to determine competitive effects.

Total soil runoff, nutrient redistribution, particle size distribution, soil deposition up-slope from the hedgerows for terrace formation as determined by survey.

Data file

Soil conservation in Embu.xls

This is a carefully prepared dataset; the sheets are not identical, but they generally contain data for maize production for 4 years.

3. The effect of mulch on crop performance

Treatments

Two tree species and two intensities of mulching are used in a factorial combination. The treatments are mulch type and mulch intensity. Mulch type has two levels, *Leucaena leucocephala* and *Gliricidia sepium* (2 species of nitrogen-fixing tree) but there is also a control plot to which no mulch is applied. There are three possible levels of mulching 0 (the control), 5 and 10 tons per hectare (t ha-1). Since the 0 level of mulch application will not involve the mulch or any species difference, there are 5 possible combinations of treatments. Each block has 5 plots: one control plot with no mulch and 4 mulched plots.

Observations

The level of soil carbon is measured both before the experiment starts and at subsequent intervals after applications of the treatment, in successive seasons, to see whether the mulch has an effect on the level of soil carbon.

The other observation is the grain produced by the maize. The total weight of cobs from each plot is recorded. Twenty cobs are selected at random from each plot and weighed in the field. These are oven dried and weighed after drying. They are then shelled and the grain is weighed. Thus 4 different maize measurements are recorded at the end of the cropping season.

Data file

Mulch trial.xls - There are three data sheets in this file:

1.   A field plan showing the allocation of treatments to the experimental units.

2.   Initial soil carbon measurements.

3.   Maize harvest for the first cropping season.

Exercise 1 – Data integrity

The data required are from Experiment 1, and are found in the second sheet named **farm level** data in the file *Onfarm intercropping.xls.* There were 41 farms in the experiment.

Œ   How many farms does the data actually show?

Œ   Why is there a discrepancy between the expected and actual number of farm entries?

Œ   Use a pivot table (or any other method) to isolate the entries that are likely to be wrong.

Œ   Design a data entry system that can trap this type of error whenever it occurs during data entry.

Exercise 2 - Data organization

The data you need are from Experiment 2 and are found in the file named *Soil conservation in Embu.xls.*

Organize these data so that seasonal trends can be investigated. Specifically, you need to:

Œ   Add a column to each data sheet to record the year for each dataset.

Œ   Combine the data sheets by appending one sheet to the bottom of the other, so that the resulting sheet has all the data in one sheet.

If there is time you may carry on to:

Œ   Add a column to convert fresh weights (kg/row) to dry weights (t/ha) at 15% moisture content.

Œ   Use pivot table (or any other method) to obtain a table of grain yield for each treatment each year.

Exercise 3 - Simple query

The data required for Exercise 3 are the **plot level** data found in the file *Onfarm intercropping.xls.*

Œ   Select all *Gliricidia* plots for output to a printer. Test the output by doing a print preview.

Exercise 4 - Advanced query

This task requires the output from Exercise 3.

Œ   Calculate the total harvest from *Gliricidia* plots for 1998.

Exercise 5 - Related tables

This task requires the **plot and farm level** data in the file *Onfarm intercropping.xls* from Experiment 1.

Œ   How many plots do you have with *Sesbania* on steep slopes?

Œ   Select them and print preview them.

Exercise 6 - Real situations

The data required for this task is in the file *Mulch Trial.xls* from Experiment 3.

Œ   Tabulate the treatment means for change in soil carbon.

Improving data querying efficiency

Objectives

In this session you will learn how to:

Œ   Use a database management system to overcome spreadsheet limitations.

Œ   Move data between spreadsheet and database applications.

Œ   Formulate a query.

Œ   Link tables in a query.

Summary

This session focuses on overcoming some of the spreadsheet limitations highlighted in Session 4. For this we introduce data querying in a database package. Data querying allows the user to:

Œ   Perform calculations on the data, in general only the **raw** data should be stored, new variables are created as and when they are needed using the power of the computer.

Œ   Select data, both row-wise (records) and column-wise (fields).

Œ   Create summaries.

The session also looks at using queries as a means of data checking.

Strategy

Œ   Brief introduction to the concept of data exchange between applications.

Œ   Introduction to query design, using queries to ask particular questions of the data.

Œ   Exercises based on data from agroforestry trials and surveys. The exercises will include using queries for calculations, selections, summaries and data checking. There are 8 exercises in total.

Required skills

It is assumed that participants have a basic knowledge of spreadsheets as a tool for data management. They should be familiar with the concept of filtering data in the spreadsheet package.

Files needed for practical exercises

For this session you will need the same 3 datasets that were used in Session 4. These are the three MS-Excel files:

Œ   *Onfarm intercropping.xls*

Œ   *Soil conservation in Embu.xls*

Œ   *Mulch trial.xls.*

Improving data querying efficiency

Introduction

> *Project Life Cycle Reference*
>
> *The majority of this session is concerned with the querying for data selection part of the project life cycle (Figure 0.1 of About the course).*

What is data querying?

Specifically data querying is used for selections, calculations and summaries. It involves two activities; (1) the specification of what to do (the command); and (2) carrying out the command (the execution). Data querying can also help you to track errors in the data. This information can be fed back into the data cycle in the form of corrections ensuring a cleaner dataset at the end of the cycle.

Why is data querying important?

Raw data is rarely in a form that is ready for graphing, analysis etc. It often needs some degree of massaging. This operation can take a lot of time and energy and is often the cause for data lying around unanalysed for much longer than it should be. It can also be the cause of errors lurking in the data, undetected until late in the data cycle.

Why is it inefficient to use spreadsheets for data querying?

Generally in spreadsheets, users mostly do the 'command' and the 'execution' together. For example to filter data in a spreadsheet you need to define the filter each time you want to use it. This process is error-prone and tedious as was demonstrated in the exercises for Session 4 – 'Why go beyond the spreadsheet'.

What is different about a database?

With a database you build queries to specify just the command, the computer does the execution. The specification of the command is saved with the database and can be rerun at any time – you do not need to specify it each time. It is this that makes the

difference between using a Pentium IV and a 386 computer for querying your data. To do proper querying in a spreadsheet you tend to use brute force (or highly specialized functions) and therefore computer speed is irrelevant. Such tasks are rarely repeatable.

What's required?

In the project life cycle we show the collection of data, which often means capturing the data in spreadsheets, followed by the compilation of these data sheets into a structured dataset. In our examples we show the 'compiling' (or linking), of MS-Excel data sheets to a structured dataset (i.e. database). The session then looks at the command specification facilities (i.e. queries) of a database.

What is this session about?

This session gives you the skills to meet these requirements. The main objective is to improve your querying efficiency. Ideally this should speed up the feedbacks to the earlier data management operations in the life cycle. For instance, detecting and fixing errors as the data are being collected.

Data exchange between applications

There is often a need to transfer data between applications. This is so that you can take advantage of the strengths of each application type. Typically you might want to use data from a spreadsheet in a database or vice-versa for various reasons. For example, when:

1. Data size exceeds spreadsheet limits hence requiring the advantage of a database.

2. Spreadsheet data exist in separate files that should be linked together.

3. Ease of querying in the database (this will be covered in more detail later in this session).

4. Easier charts and tables from spreadsheet packages.

We might also like to transfer subsets of the data to statistics packages, graphics software, etc. There are many different tasks in the life cycle of a research project and you should try to use the most suitable software for each task. You wouldn't, for instance, try to write a project proposal using just a statistics package!

Transfer of data between applications is now very flexible and much easier than in the past. However, it is essential that you always check your data in the target application to ensure it has transferred correctly; errors can occur in any manipulation of data and it is important to find such errors as soon as possible.

The data exchange process involves importing or linking data. **Importing** data from a spreadsheet into a database means you end up with two copies of the data; one in the spreadsheet file and one in the database. From the data management viewpoint two or more copies of the data can lead to major problems of data integrity – if corrections need to be made, they must be made in all copies of the data. This is easily forgotten, thus the copies of the data become unsynchronised and it is easy to forget which is the correct version.

If you need to use the power of both databases and spreadsheets we suggest you designate one copy of the data (either the spreadsheet file or the database) as the definitive **master** copy of the data. From this master copy you can extract subsets into other applications for analysis etc, but any corrections must be made to the **master** copy and subsets extracted again if you need to repeat any analyses.

An alternative to importing from spreadsheet to database is linking. **Linking** leaves the source file data in its current location and stores a link to that data in the database file. Changes made to the linked data are automatically reflected in the source file. You have just one copy of the data to worry about. With linked data you will need to ensure the links are up-to-date and if you move the data into a different folder on your computer (or send the data to someone else) you will need to modify the links.

From the data management viewpoint, linking solves the problems with data integrity mentioned earlier. However, if your data exceed the spreadsheet limits of 65 000 records then you must import it using the database as the master copy of the data. It is not advisable to store your data in lots of separate spreadsheet files, as it becomes a major task keeping these files organized. The number of data files in a study is an indication of the level of data integration and the efficiency of data analysis. Fewer data files are easier to manage. Also databases generally work faster when the data are stored within the database itself rather than linked from other files.

> *Project Life Cycle Reference*
>
> *Linking or importing data describes part of the process of compiling the dataset.*

Based on your own situation you need to decide whether importing or linking is more appropriate for you.

Note that when we talk about having a single master copy of the data, we are of course

referring to the active copies of the data. It is essential that you have at least 2 backup copies of all your important data on removable media such as CD or zip disk. These backup copies should be stored in separate locations and should be tested regularly to ensure they are still readable. **If you do not keep backups of your important data and other files, then one day you will lose some very important work.**

Using queries for calculations

A basic principle of data management is that you store only the raw data (and the metadata). You should not store calculated data that can be derived from other data. Instead you would store the formulae and let the application carry out the calculations as and when required. Most of you will be familiar with the idea of formulae in spreadsheet packages. For example if you have the number of cobs per plot in cell D2 and the total weight of cobs in cell E2, then to calculate the average weight of each cob you would use the formula =E2/D2. In this case you store the formula, and the results are automatically displayed. Should the value in one or both of D2 and E2 change, then the displayed result from the formula will automatically change.

If you have your data in a database such as MS-Access, calculations are done using **queries**. Again it is the formula or expression that is saved to the database, not the results of the calculation. Using the same example as above let's assume cobs per plot are stored in the field **cobs**, and total weights are stored in a field called **totweigh**; to calculate the average weight per cob we would use the expression **avweight: [totweigh]/[cobs]**.

> *Project Life Cycle Reference*
>
> *Here we are referring to the querying for data selection part of the cycle.*

One advantage the database query has over the formula in the spreadsheet is that you cannot enter data into the calculated field in the database; in the spreadsheet it is easy to accidentally type a value in the cell containing the formula.

Using queries for selecting records

In the spreadsheet package, selection of records or rows can be done using **filters**. The process is quick and relatively straightforward. It could be used for example to carry out task 3 in the exercises for Session 4 where you were asked to select all *Gliricidia* plots.

There are, however, several disadvantages to this method of selection:

1.   The filter needs to be done manually each time (unless you are prepared to start writing macros – a subject not covered by this training). Manual processes are tedious and prone to errors.

2.   The filtered data need to be copied and pasted into a separate worksheet before they can be used in any analysis or tabulation. The non-selected rows are merely hidden from view but they are included in any calculations done on the worksheet. The process of copying and pasting is itself error-prone, but also results in duplicate copies of the data and all the problems that entails.

NB: Duplicates are bad; backups are good.

An alternative is to create a query in the database to select the records you want to see. In a query you specify a criteria. This is similar to defining the filter. However, in the database the criteria are saved with the query and to redo the same selection it is simply a case of re-running the query. The results of the query can be used in a report, exported to another application for analysis, etc. As with calculated fields, the subset of data selected by your criteria is not stored as a separate entity in the database. The original raw data remain unchanged and the subset is produced each time the query is run.

Criteria can be simple such as selecting all records where a particular field contains a specified value (all *Gliricidia* plots for instance) or much more complex involving more than one field and criteria joined with **and** or **or**. (For example select all records where the cropping system is either **napier** or **combi** and the total cob weight is greater than 200g.) This sort of criteria is relatively easy to specify in the database.

Project Life Cycle Reference

Again this refers to the process in the cycle of querying for data selection.

Another type of query also used for record selection is a **parameter** query. Here you specify a parameter as your criteria and the database asks you for a value whenever you run the query. Thus the same query can produce different subsets depending on the value you enter.

Using queries for selecting fields

Selecting a subset of the columns in a spreadsheet is not automatic. If you wanted to print out just a few of the columns or fields, you would either have to hide unwanted columns or copy the desired columns pasting them into a separate sheet. Either method is tedious and therefore prone to errors.

With a query you can just select the fields you want to see. MS-Access for example, has a design grid for queries to which you can simply drag the fields you want in your selection.

Using queries for data summaries

Summarizing over particular fields is not at all straightforward in a spreadsheet without using data analysis tools. On the other hand in a database such as MS-Access you can use queries to find averages, totals etc. grouped by one or more fields. For example if we have maize yield figures for individual *Gliricidia* plots and yield figures for individual *Sesbania* plots, then we can use a query to calculate the average yield per plot for each species and/or the total yield for each species. Note this is not the same as when we mentioned calculated fields earlier in this session. Then we were calculating over columns (fields), here we are summarizing over rows (records).

Using queries to link data

Consider task 5 in Session 4 where you were asked for the number of plots with *Sesbania* on steep slopes. The data required for this exercise were in two separate data sheets in the spreadsheet file. The exercise is difficult without joining the data sheets either manually by cutting and pasting or automatically by using specialized formulae. The manual join is tedious and error-prone. The automatic formula using the **VLOOKUP** function is not intuitive and needs a great deal of care to use correctly.

In the database approach, data from the two data sheets are stored as two separate tables but with a common field. The common field is generally unique in one of the tables but not in the other, thus we have a one-to-many relationship between the tables. We will talk more about relationships in the next session. For now you will see that both these tables can be taken into the query design grid and the link made between the common fields. It is then easy to select fields from both of the tables. Data from the one side of the relationship will appear to be repeated in the query results but it should be noted that although displayed more than once, these data are only stored once, thus we do not have the problem of compromising data integrity that we would be likely to come across in the spreadsheet.

Using queries for checking data

Queries can be designed to help check for inconsistencies in the data. Take as an example a household survey in which you have recorded the total number of people in the household and the number of children in the household. Obviously there cannot be more children than there are people in any one household. You can easily design a query to show all the households (records) where **[children] > [people]**. In a similar way you can calculate the number of adults as being **[people]-[children]**, then design another query to show all the households where there are no adults. The majority of checks you want to make on your data you should be able to turn into a query. The skill comes in the phrasing of the question and this comes with practice.

Some database packages such as MS-Access have special queries that can help you to find errors. MS-Access for instance has a **Find Duplicates Query** in which you can check for duplicate entries in particular fields. There is also a **Find Unmatched Query** that allows you to find records in one table, which have no linked records in another table. This can be useful for finding redundant data records.

Structured Query Language (SQL)

We have continuously referred to the query being stored but not the results of the query. Most databases use SQL (Structured Query Language) to formulate queries. This is a universal language that most modern database packages understand. However, you do not need to know SQL in order to use queries. We have already mentioned that MS-Access uses a design grid as a user interface. In the design grid you can select the tables you want to use, link the tables where appropriate, select the fields you want to see in the results and set any necessary criteria. MS-Access will translate what you have put into the grid into the corresponding SQL statement and it is this SQL statement that is saved with the database. You have the facility for viewing this statement and for editing it directly but it is not necessary. The vast majority of queries can be created using the grid.

Exercise 1 – Linking data into MS-Access

The first task is to link data from the spreadsheet into the database. Note you can either link an entire worksheet or a named range. Data to be imported or linked should contain just the short variable name and the data. In other words, a neat rectangular block of rows (records) and columns (fields). For this task you will be using data from the trial 'Mixed intercropping with *Gliricidia sepium* and relay cropping with *Sesbania sesban*'.

If you have followed Sessions 2 and 3 of this course you would have metadata in the first few rows of the data sheet. The whole sheet cannot therefore be linked into MS-Access. In Part 3 of Session 3, we showed you how to name a range of cells in the spreadsheet to facilitate transfer to other packages. For this task you will need to ensure that you have a named range.

Œ   In MS-Excel open the file *Onfarm intercropping.xls* and go to the sheet called **Plotlevel** data. You want to use the range A5:K113 that should already have been named.

Œ   In the **Name** box click the down arrow and select **data_at_plot_level**. The correct range should be selected. If not then ask the resource person for help.

Œ   Open a new, blank database in MS-Access and from the **File** menu choose Get external **data** followed by **Link Tables**.

Œ   In the **Link** dialog box, go to the **Files of Type** box and select **Microsof**t **Excel**.

Œ   Click the **down arrow** to the right of the **Look in** box and select the drive and folder where the spreadsheet file is located. Double-click on the file icon.

Œ   In the **link spreadsheet wizard** select the option to **Show named ranges** and select **data_at_plot_level**. Click Next.

Œ   Ensure that **First row** contains column headings is ticked and click Next. The default table name will match the name of the range. Click **Finish.**

NB: MS-Access uses different icons to represent linked tables and tables that are stored

in the current database. If you delete the icon for a linked table, you delete the link, not the external table itself.

Œ   Take a look at the data in the linked table within MS-Access and make a note of anything that might cause a problem later.

Exercise 2 – Simple query

The next task uses the database created in Exercise 1. The objective is to design a query to select all *Gliricidia* plots.

Œ   In the database window go to the **Queries** section and click **Create Query in Design View.**

Œ   Add the table **data_at_plot_level.**

Œ   Select all the fields and drag them to the grid.

Œ   In the grid go to the **Criteria** row for the field **species** and type **g**. NB: the linked data has one-letter abbreviations for the species names. In the next session we will show how to link these abbreviations to the full species names.

Œ   Click the **Run** button. ![!]

Œ   Save the query giving it a suitable name (*Gliricidia* **plots** for example)

Exercise 3 – Summary query

This exercise uses the same database as Exercise 2. In this exercise you will use a summary query to calculate the total harvest for each species for 1998.

Œ   Follow the first 2 steps of the above procedure.

Œ   Transfer species and F7 to the grid.

You should have noticed in Exercise 1 above that the columns headed 1995, 1996, 1997 and 1998 in MS-Excel, have linked into MS-Access with fieldnames F4, F5, F6 and F7 respectively. Although MS-Access is happy to have field names consisting entirely of digits when you create the tables in MS-Access itself or when you import data from MS-Excel, it is not happy for linked tables to have such fieldnames and instead labels the fields as F followed by a digit representing the column position. Therefore F7 holds data for 1998.

Œ   Click the **Totals** button and in the Total row for F7 change the value to **Sum**. Change the fieldname to Total Yield: F7

Œ   Click the **Run** button.

Exercise 4 – Related tables

This task requires the plot and farm level worksheets from *Onfarm intercropping.xls*. You need to find out how many plots have Sesbania on steep slopes. You have already linked the plot level data to MS-Access, now you need to link the farm level data.

Œ   Similar to the way you carried out Exercise 1, link the farm level data from the MS-Excel file.

Œ   Now start a **new query** this time including both tables in the query grid.

Œ   The field **farmid** is the common field in both these tables and we need to use that to form the link. Drag **farmid** from the farm level table to **farmid** in the plot level table. This should produce a line joining the two tables.

Œ   Drag the fields **slope**, **species** and **farmid** (from either table) to the grid.

Œ   In the **criteria** row for slope type **2** (2 is the code for steep slopes); in the **criteria** row for species type **s**. (In Session 6 we will see how to include the text value matching the code.)

Œ   **Run** the query to see how many plots have *Sesbania* on steep slopes and which farms these are on.

Exercise 5 – Data integrity

For this exercise we will use the data linked from the farm level details of the MS-Excel file *Onfarm intercropping.xls*. We saw in the previous session how there were 41 farms in the experiment but there were 42 records in the file.

Œ View the **farm level data** in your MS-Access database to confirm that there are indeed 42 records.

We will run a query to find the duplicate farm.

Œ On the **Queries** section click **New** and choose **Find Duplicates Query Wizard**.

Œ Select the **data_at_farm_level** table and look for duplicates in the **farmid** field.

Œ What is the ID of the duplicated farm and what differences are there between the two records for this farm?

In the next session we will look at ways to stop this sort of duplication occurring at data entry.

Exercise 6 – Importing and appending

This exercise uses data from the trial 'Soil and water conservation using contour hedgerows of fodder grass and trees'' stored in the file *Soil conservation in Embu.xls*. Exercise 2 in Session 4 on data organization, involved a great deal of cutting and pasting to append each sheet onto the bottom of another so that all the data were in one sheet. One problem you probably found in this was that the columns were not in the same order in every sheet making this a very tedious and error-prone activity.

Now you will import these data into MS-Access.

Œ If not already done add a column to each sheet in the MS-Excel file to record the year. Make sure this column is not the last column otherwise the named ranges that have already been set up will not include it.

Œ In MS-Access create a new blank database and import the data from the named range **Ir93**. Import these data to a new table calling the new table **Soil Conservation**.

Œ Now import each of **Ir94**, **Ir95**, **Ir96** and **Ir97** each time appending the data to the existing table.

Œ   What errors do you get and what do you think might cause these errors? Look
     back at the MS-Excel file and, comparing the worksheets, try to work out a way
     round the problem.

Exercise 7 – Calculated fields

In this exercise you will use the database created in Exercise 6 to convert the fresh weights
to dry weights based on the given percentage moisture content.

Œ   Once again go to the queries section of MS-Access and start a new query based on
     the table Soil Conservation. Take all fields down to the grid.

Œ   In the next column in the grid enter the field name as **drygrain:[tfgrainwt]*(100-
     [mc])/100.** The square brackets are used to enclose fieldnames.

Œ   Run the query. You should have an additional column containing the dry weights
     of grain.

Exercise 8 – Find unmatched query

In this exercise we will look at the farm level data and the plot level data from the file
Onfarm intercropping.xls. By this stage the two sheets in this file should already be linked
into an MS-Access database. First of all we are going to check whether we have farms at
the farm level for which we have no plot data.

Œ   Open the database with the correct data and go to the queries section.

Œ   Click **New** and choose **Find Unmatched Query Wizard.**

Œ   Choose **data_at_farm_level** as the first table and **data_at_plot_level** as the related
     table.

Œ   Choose **farmid** as the matching field from both tables.

Œ   Select all remaining fields to appear in the query results then click **Finish**.

Œ   Are there any farms for which we have no plot data?

Œ   In the same way check that every farm mentioned at the plot level has a
     corresponding record at the farm level.

exercises

5. improving data querying efficiency

Data modelling:
organizing data for easier querying

Objectives

The aim of this session is to introduce hierarchical data structures and methods to help preserve the integrity of your data. By the end of the session you will be able to recognize structures in your data and create your data model accordingly.

Summary

During the session we will look at patterns frequently found in research data and demonstrate ways of structuring these data so as to improve data integrity and eradicate unnecessary duplication. You will see how a well-designed database makes for easier querying. During the session we will be introducing some of the key database concepts. We will use MS-Access by way of example although the concepts should hold for any database management system.

Strategy

1. Introduction to the concept of data modelling demonstrating techniques for recognizing patterns in the data.

2. Building a logical diagram of the data structure.

3. Participants work with example datasets designing the structure and transferring this structure to MS-Access.

4. Participants will work with their own datasets to develop logical and physical designs.

Required skills

It is assumed that participants are fluent in a spreadsheet application and have used a database package for data querying.

Files needed for practical exercises

Œ   *Timber business census questionnaire.doc* - Blank questionnaire for survey on timber businesses.

Œ   *Improved fallows complete sheet.xls* - Data file for the trial 'Influence of improved fallows on soil phosphorous fractions'.

Support documentation

Œ   *The role of a database package for research projects* - Statistical Services Centre (2000).

Introduction

Most researchers in the target audience for which this training is designed will have come across the term modelling in the usual sense, but not **data modelling**; so it is worth spending a little bit more time to explain what it is, why it is important, how it fits in with the transformation steps in the project life cycle (Figure 0.1 in the introductory notes) and how it's done.

What is it?

In layman terms it is deciding how to organise your data (in terms of tables) BEFORE YOU COLLECT IT. This is motivated by the need to be sure that data integrity is enforced throughout the life of the data; there should be no redundancy; querying is intuitive and the data organisation transcends any application.

Why is it important?

Without it we will not be able to go beyond the spreadsheet (Session 4) even if we now have the new skills of querying a database (Session 5).

For example:

1. In Sessions 4 and 5 we linked two worksheets from the MS-Excel file *Onfarm Intercropping.xls* into an MS-Access database. We were then easily able to design a query linking data at the farm level and at the plot level, and select plots based on criteria at the farm level. However, we paid no attention at that time to data modelling and thus whatever errors were in the spreadsheets were also in the database; for instance the duplicate farm record. Also, although we could check that every farm mentioned at the plot level was also mentioned at the farm level, we had no way of enforcing this condition.

2. Table 2.7 in Session 2 (also shown below as Table 6.1) shows two ways of displaying the same data. Although the short version is more compact and perhaps easier to

read for the layman, it is not as straightforward to query data in this format. Imagine for example you wanted to summarize these data over treatments. From the long version it is just a matter of summarizing over rows (records). From the short version you would need to summarize over rows and columns. Adding data from a third harvest would involve adding extra fields (columns) in the short version, which is not as easy as adding extra records (rows), which is all you would need to do with the long version. Adding fields involves changing the structure of the data (the data model) and ideally you want this to be complete prior to data entry.

As mentioned in Session 2, it is relatively easy to change data stored as the long version into the short version using pivot tables; the reverse however, is not at all straightforward.

Short version

Block!	Treatment!	number damaged leaves (harvest 1) harv1dam	number marketable leaves (harvest 1) harv1mark	number damaged leaves (harvest 2) harv2dam	number marketable leaves (harvest 2) harv2mark
1	1	182	250	103	234
1	2	249	255	182	245
1	3	246	273	154	294
2	1	240	200	99	200
2	2	281	261	95	228
2	3	503	125	161	196
3	1	375	265	192	235
3	2	413	107	339	149
3	3	364	105	316	126

Long version

Harvest date Harvest!	Block!	Treatment!	number of damaged leaves numdam	number of marketable leaves nummark
1	1	1	182	250
1	1	2	249	255
1	1	3	246	273
1	2	1	240	200
1	2	2	281	261
1	2	3	503	125
1	3	1	375	265
1	3	2	413	107
1	3	3	364	105
2	1	1	103	234
2	1	2	182	245

Tables 6.1 - Two ways of displaying the same data

3. Also in Session 2 we looked at the case of measuring lengths of shoots on trees where the number of shoots varied. We noted that storing the data with each shoot in a separate column made it difficult to analyse the data at shoot level. Re-organising the data (part of data modelling) makes analysis and querying easier.

The end product of data modelling is an EMPTY database waiting to be POPULATED with data. An empty database is different to an empty spreadsheet. The former is a series of tables with enforced relationships and the data in there is guaranteed to obey the design rules; the latter is a free field where anything goes.

What you will learn from this session?

Œ   How to represent a data model (Entity/Attributes/Relation).

Œ   How to translate a data model into a database (i.e. table definition, fields and data types, primary key fields and foreign keys for setting up relationships, etc.).

Œ   How to design a data model for data that already exist. We do not go into the formal normalization process, but instead provide common indicators in research datasets when data modelling is needed. We discuss 4 such indicators:

1. Numeric suffixes to field names.

2. One item per cell.

3. Row and column order independence.

4. Repetition and consistency of data values.

Data modelling involves structuring data to meet the needs of one or many users of the data. A data model should be developed before designing and developing the actual tables used to hold the data.

Numeric suffix to field names

Consider a household survey where among other information you are asked to record the age and gender of each child. How would you store these data at the household level? The number of children in each household is not know prior to the survey so there is no way of knowing how many columns to allow. Often a maximum household size is assumed and smaller households have lots of empty cells in their record. Often these columns are headed age1, gender1, age2, gender2, ..., age#, gender#, where # is the maximum number in the household. Table 6.2 shows part of such survey data taken

from a spreadsheet. The spreadsheet actually allows for a total of 12 children in each household.

hhid	date	namehead	agekid1	yearkid1	genkid2	agekid2	yearkid2	genkid3	agekid3	yearkid3	genkid4	agekid4	yearkid4	genkid5	agekid5	yearkid5
101	17-Feb-98	Sabwa, Rosia	17		0	13		0	10		0	3		0		
102	18-Feb-98	Makutsa, David	13	85	0	9	89	0	6	92	0	4	94	0	1	97
103	18-Feb-98	Amaloba, Esther	20	78	0	25	73	0	18	80	0	16	82	0	14	84
104	19-Feb-98	Onutsko, Samson	19	79	0	18	80	0	16	82	0					
105	20-Feb-98	Ndele, Samuel Musenda	17		0	15		0	15		0	13		0	10	
106	20-Feb-98	Ochwa, David	18		0	17		0	11		0	9		0	1	
107	23-Feb-98	Eskangwa, Salmon	24	74	1	22	76	1	20	78	2	19	79	1	13	85
108	23-Feb-98	Oguang, Florence Ayuma	26	72	1	25	73	2	21	77	2	16	82	1	16	82
109	24-Feb-98	Kitwa, Andrew Otocho	26	79	2	25	73	2	23	75	2	21	77	1	18	80
110	25-Feb-98	Ndele, Julius Omulama	45	53	1	41	57	1	36	62	1	33	65	1	36	65
111	27-Feb-98	Sabwa, Livingston			1			1			1	36	62	1	36	66
112	27-Feb-98	Okonda, Rebecca Nyakowa	10	88	1	8	90	2	7	91	2	5	93	2		
113	2-Mar-98	Ngele, Jane	14	84	1	15	88	1	6	92	2	2	96	2		
114	3-Mar-98	Manjichi, Daniel														
115	3-Mar-98	Shiundu, Mary	26	72	2	21	77	2	19	79	2	17	81	2	13	85
116	5-Mar-98	Mukaka, Simon Abilo														
117	6-Mar-98	Sabwa, Mary	35	63	1	33	65	1	29	69	2	26	72	2	24	74
118	9-Mar-98	Kitwa, Samuel Obuwa	31	67	1	28	70	2	23	75	2	18	80	1	16	82
119	11-Mar-98	Mukaka, Samuel	24	74	2	22	76	3	19	79	1	16	82	1	12	86
120	11-Mar-98	Oponyo, Saloma Nyakeso	7	91	1	5	93	1	2	96	2					
121	13-Mar-98	Anyimbwa, Zedekiah Sakhunya	36		1	36		2	29		2	28		1	27	
122	13-Mar-98	Malanda, Leonidah	31	67	2	29	69	2	26	72	2	24	74	2	23	75
123	16-Mar-98	Okonda, Josephat	4	94	1											
124	17-Mar-98	Mukaka, Patrick Ondeve	30	68	1	28	70	2	24	74	2	22	76	2	20	78
125	19-Mar-98	Okwaro, Enos Kitwa	44	54	1	41	57	1	40	58	1	35	63	1	35	63
126	19-Mar-98	Ndele, Helen	41	57	1	36	62	2	33	65	2	30	68	2	29	69

Table 6.2 - Household survey data showing child data in seperate colums

How would you find the age of the youngest child in each household?

Table 6.3 shows the children data as one row per child. This method of storing the data is preferable as it is much easier to summarize over children. Notice that we have included the field **hhid**. This is in order to link the child data to the household data shown separately in Table 6.4; **hhid** is a unique identify for the household. In database terms it is known as the **primary key**. The primary key cannot be left empty and cannot have duplicates. Thus by setting a primary key we ensure we have only one record per household. In the child table **hhid** is known as a **foreign key** as it is the primary key from another table. The relationship between the two tables is known as a **one-to-many** relationship as a single household can have many children, but a child can be a member of only one household. Note there may be households with no children.

In the child table we have also added a field called **kidno**. This is to form a unique identifier for the child table. Note that **kidno** is not unique by itself but combined with **hhid** it is thus the combination of these fields from the **primary key** for the child table.

hhid	kidno	age	year	gender
101	1	17		0
101	2	13		0
101	3	10		0
101	4	3		0
102	1	13	85	0
102	2	9	89	0
102	3	6	92	0
102	4	4	94	0
102	5	1	97	0
103	1	20	78	0
103	2	25	73	0
103	3	18	80	0
103	4	16	82	0
103	5	14	84	0
103	6	6	92	0
104	1	19	79	0
104	2	18	80	0
104	3	16	82	0
105	1	17		0
105	2	15		0
105	3	15		0
105	4	13		0
105	5	10		0
105	6	9		0
106	1	18		0
106	2	17		0
106	3	11		0
106	4	9		0
106	5	1		0
107	1	24	74	1

Table 6.3 - One row per child

| hhid	date	namehead
101	17-Feb-98	Sabwa, Roda
102	18-Feb-98	Makutwa, David
103	19-Feb-98	Amakobe, Esther
104	19-Feb-98	Omutoko, Samson
105	20-Feb-98	Ndele, Samuel Musanda
106	20-Feb-98	Ochwa, Daudi
107	23-Feb-98	Esikangwa, Salmon
108	23-Feb-98	Ojuang, Florence Ayuma
109	24-Feb-98	Kitwa, Andrew Olocho
110	26-Feb-98	Ndele, Julius Omulama
111	27-Feb-98	Sabwa, Livingston
112	27-Feb-98	Okonda, Rebecca Nyakowa
113	2-Mar-98	Ngala, Jane
114	3-Mar-98	Manjichi, Daniel
115	3-Mar-98	Shiundu, Mary
116	5-Mar-98	Mukaka, Simon Atulo
117	6-Mar-98	Sabwa, Mary
118	9-Mar-98	Kitwa, Samuel Obuwo
119	11-Mar-98	Mukaka, Samuel
120	11-Mar-98	Oponyo, Salome Nyaleso
121	13-Mar-98	Anumbwe, Zedekiah Sakhunya
122	13-Mar-98	Malanda, Leonidah
123	16-Mar-98	Okonda, Josephat

Table 6.4 - Data at household level
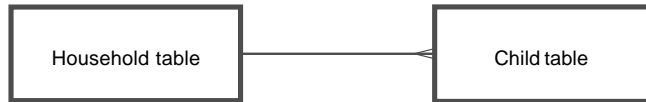
Figure 6.1 - One-to-many relationship between Household and Child tables

Another example was used in Session 2 where the lengths of all shoots on a tree were measured. Entering each shoot as a column so that the row represents a single tree gives us many columns and lots of empty cells as the tree with the most shoots dictates the number of columns and may have more than a typical tree.

Alternatively if we enter the data as a row for each shoot then the tree information is repeated for each shoot. Such duplication is strongly discouraged for reasons of data integrity. These data as they were entered into the spreadsheet, are shown below in Table 6.5.

Experiment: Gliricidia provenance trial
Location: Msekera Research Station, Chipata, Zambia
Scientist: Dr F.Kwesiga
Started: 12/1/91

Replicate	Blocks within replicates	plot number	provenance	tree number (the central 9 trees are measured)	Diameter of regrowth shoots, 10cm from stump, 6 months after cutting						Average diameter of all shoots measured	Square root of total squared shoot diameter
rep	block	plot	prov	tree	cm shoot1	cm shoot2	cm shoot3	cm shoot4	cm shoot5	cm shoot6	cm meandiam	cm rtotsqr
north	1	1	br12-84	1	1.30	1.50	0.60	0.40			0.95	2.11
north	1	1	br12-84	2	2.40	1.00	1.10	0.30	0.30		1.02	2.66
north	1	1	br12-84	3	0.60	0.90					0.65	1.20
north	1	1	br12-84	4	1.30	1.80	0.50				1.13	2.12
north	1	1	br12-84	5								0.00
north	1	1	br12-84	6	2.80	1.90	1.50	1.20	0.70	0.30	1.40	3.96
north	1	1	br12-84	7	3.50						3.50	3.50
north	1	1	br12-84	8	1.60	1.30	0.80				1.23	2.21
north	1	1	br12-84	9	1.80	1.75	0.60	0.50			1.15	2.60
north	1	2	kr-18-84	1	3.15	2.22	1.53	0.06			1.74	4.15
north	1	2	kr-18-84	2	2.48	1.34					1.91	2.82
north	1	2	kr-18-84	3	3.91	3.87	2.61	1.72	2.18	1.76	2.54	6.51
north	1	2	kr-18-84	4	1.13						1.13	1.13
north	1	2	kr-18-84	5	2.16	1.99	1.29	1.19	0.85	0.87	1.39	3.63
north	1	2	kr-18-84	6	2.52	2.06					2.30	3.27
north	1	2	kr-18-84	7	3.66	3.43	2.87	2.30	2.07	1.27	2.60	6.79
north	1	2	kr-18-84	8	2.91	2.93	1.04	1.87	2.79		2.31	5.42
north	1	2	kr-18-84	9	1.10	1.94	2.47	1.76	1.63		1.78	4.10
north	1	3	kr-23-84	1	2.98	2.89	2.00				2.58	4.49
north	1	3	kr-23-84	2	1.53	1.81	2.02	2.97	0.33	0.57	1.87	5.39
north	1	3	kr-23-84	3	3.00	2.11	2.00	1.65	1.74		2.10	4.82
north	1	3	kr-23-84	4	3.79	2.81	1.61	1.25	0.59	0.53	1.76	5.20
north	1	3	kr-23-84	5	3.19	2.68	1.99	1.73	1.16		2.14	5.05
north	1	3	kr-23-84	6	2.86						2.86	2.86
north	1	3	kr-23-84	7	2.90	1.34					2.12	3.19
north	1	3	kr-23-84	8	3.55	2.57	2.30	1.23	0.41		1.99	5.07
north	1	3	kr-23-84	9	3.03	2.11	1.90	1.30	0.86	0.92	1.70	4.86

Table 6.5 - Seperate columns for each shoot

As with the many children in each household the solution is to divide the data into two tables, having one table for data at the tree level and one for data at the shoot level. A unique identifier is created for each tree and this identifier is used in both tables to act as a linking field. Figure 6.2 shows the one-to-many relationship between these tables. In this example we have included the field names for each table. The underlined fields are the primary key fields. Note that as well as being part of the primary key in the Shoot table, **TreeID** is also a foreign key as it is the primary key for the Tree table. The relationship links TreeID in both tables. In database terminology you may hear the tables referred to as **entities** and the fields within the tables as **attributes** of those entities.
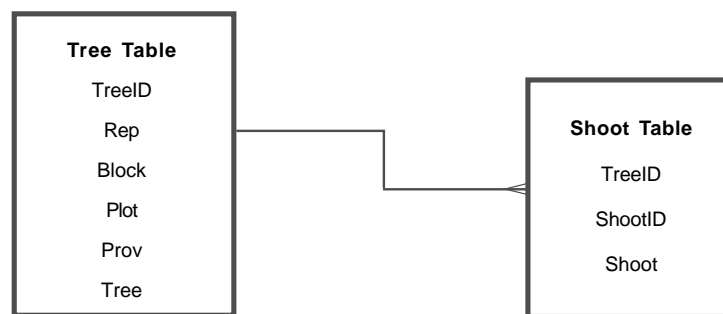
Figure 6.2 - Relationship between Trees and Shoots

With this model of the data it does not matter how many shoots each tree has, or, in the children in the household example how many children each household has. Adding an extra child or an extra shoot is simply a case of adding an extra record to the shoot or child table. Data in the related table remains unchanged. Remember that adding extra records is trivial, adding extra columns/fields is not.

One item per cell

In a typical survey you often have questions that generate more than one answer. For example, imagine a timber census survey in which one of the questions was 'Who are your customers?' Typically a business would have many customers so there are many answers to this question. A well-designed dataset – whether it be stored in a spreadsheet or in a database – would have just one item per cell and in a survey this would generally be the answer to a single question. So how do we cope with the situation where we have many answers to a single question?

The solution is to store the multiple answers as records in a separate table or worksheet. At the questionnaire level you store a link to this extra table. Thus the 'single item' in the cell for this particular question is the link to the other table of data.

This could be seen as similar to our example in the previous section where we had many shoots on each tree, or many children in each household. Here we have many customers for each business (hopefully!). Thus we have a **one-to-many** relationship between business and customer.
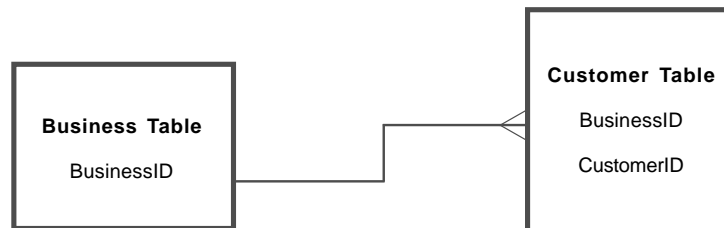

Figure 6.3 - Relationship between Business and Customers

In summary, when a question in your survey is generating multiple answers you should consider dividing the data into related tables.

Row and column order independence

In Sessions 2 and 3 we encouraged you to include details about the survey or experiment in the first few rows of the spreadsheet. This was so that all relevant information is kept together and it is easier for others to understand the data (It is also easier for yourself to understand when you come across the file again 6 months later having forgotten all about it!).

In Session 4 we saw some of the limitations of spreadsheets in the field of data management and in Session 5 we showed you how to overcome some of these problems by using a database package. However, our well-designed spreadsheet, like the one shown below in Table 6.6, cannot as a whole be imported into a single database table.

Title:	On-farm cropping with sesbania and gliricidia								
ExptID	Mak20								
Program:	4								
Project	4.2								
Date	Nov-94								
Scientists:	Prof J. Maghembe								
	Species Name								
	g gliricidia			Maize Yield					
	s sesbania			Dry maize weight (t/ha)			(%)	(ppm)	
	c control								
farmer	farmid	SpeciesId	harv95	harv96	harv97	harv98	N	P	pH	cec
Chakama	1	g		1.97	0.34	3.38	0.065	7.85	5.8	3.79
Chakama	1	s		1.91	0.65	2.14	0.055	11.71	6.1	5.33
Chakama	1	c		2.25	0.49	0.76	0.06	8.9	6	4.14
Thobola	2	g	2	0.26	0.50	1.86	0.07	26.82	5.75	
Thobola	2	s	3.28	0.27		0.51	0.055	33.79	6	
Thobola	2	c	4.68	0.62	0.39	1.02	0.07	28.07	5.95	7.06
Adisani	3	g		0.60	0.86	1.45	0.075	5.67	6.05	6.59
Adisani	3	c	2.63	1.61	0.28	0.80	0.08	2.38	5.95	5.11
Majoni	4	g			0.24	4.35	0.06	7.21	6.25	4.05

Table 6.6 - Designed spreadsheet

A general rule to remember is that data in a database table must be row and column independent. In other words re-ordering the rows, or re-ordering the columns has no effect on the data. Looking at the data in Figure 6.6 we can see that the section of the worksheet from the column headers onwards (farmer, farmid, etc.) is indeed row and column independent; however, the worksheet as a whole is not. To transfer these data into a database structure we need to separate out the basic experimental details and the plot level data. Thus we have one table at the experiment level and one at the plot level. The tables are shown below:

Title	ExptID	Program	Project	Date	Scientist
On-farm cropping with sesbania and gliricidia	Mak20	4	4.2	Nov-94	Prof J. Maghembe

Table 6.7 - Experiment level data

ExptID	farmer	farmId	speciesId	harv95	harv96	harv97	harv98	N	P	pH	cec
Mak20	Chakame	1	g		1.97	0.34	3.38	0.065	7.85	5.8	3.79
Mak20	Chakame	1	s		1.91	0.65	2.14	0.055	11.71	6.1	5.33
Mak20	Chakame	1	c		2.25	0.49	0.76	0.06	8.9	6	4.14
Mak20	Thobola	2	g	2	0.26	0.5	1.86	0.07	26.82	5.75	
Mak20	Thobola	2	s	3.28	0.27		0.51	0.055	33.79	6	
Mak20	Thobola	2	c	4.68	0.62	0.39	1.02	0.07	28.07	5.95	7.06
Mak20	Adisani	3	g		0.6	0.86	1.45	0.075	5.67	6.05	6.59
Mak20	Adisani	3	c	2.63	1.61	0.28	0.8	0.08	2.38	5.95	5.11
Mak20	Majoni	4	s			0.24	4.35	0.06	7.21	6.25	4.05

Table 6.8 - Plot level data

In order to link the two tables we include the key identifier from the experiment level table **(ExptID)** in the plot level table.

There still remains information in Table 6.6 that we have not captured. These are the units of measurement and the names for the species codes. In the database there is generally space for you to include a description of each field. Here you should include the units of measurement used. The species names are captured in the third table shown below in Table 6.9.

SpeciesID	SpecName
g	gliricidia
s	sesbania
c	control

Table 6.9 - Species codes

This table is linked to the plot table by the **SpeciesID** field, which is the key identifier for the species table. Thus we end up with the following data model.
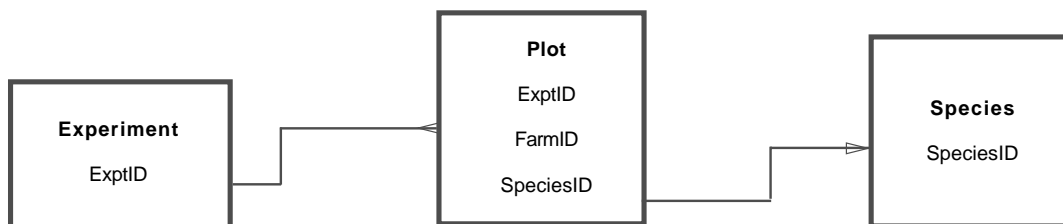

Figure 6.4 - Relationships between Experiment, Plot and Species

In Exercise 4 of Session 5 we created a query to select *Sesbania* plots on steep slopes. We set the criteria for the species as 's' because that was how the species values were stored in the plot table. Using the above data model (Figure 6.4) we can link these species codes to corresponding value in the species table and thus pick up the full name of the species. The full names are only stored once but through the link can be displayed many times.

Repetition and consistency of data values

You'll notice in Table 6.9 that the farmer name is directly related to the farmID. Including the farmer name in this table as well as the farmID is superfluous. At the data entry phase, repeatedly entering the same name can lead to errors (spelling mistakes for instance). Also if a correction needed to be made to a particular farmer name, you would have to change every occurrence of the name. Again this is error-prone as it is too easy to miss one or more of the occurrences thus leading to data inconsistency.

The solution is to remove the farmer field at the plot level and create a separate table for the farmers as shown in Table 6.10.

FarmID	Farmer
1	Chakame
2	Thobola
3	Adisani
4	Majoni

Table 6.10 – Farmer level data

FarmID is then the link from the plot level data to the farmer level data and is the key identifier at the farmer level.

Logical and physical design

We have briefly discussed 4 indicators to look for in your data, which would indicate the need to split your data into separate, linked tables:

1. Numeric suffixes on fieldnames.

2. Several answers to a single question.

3. Order of rows and/or columns not independent.

4. Repetition.

As you work out how to split your data, you build a logical design of your data structure. We strongly recommend you always draw a diagram of your data structure before creating your database. If you know what tables you need and how they are related, then building the database is much easier.

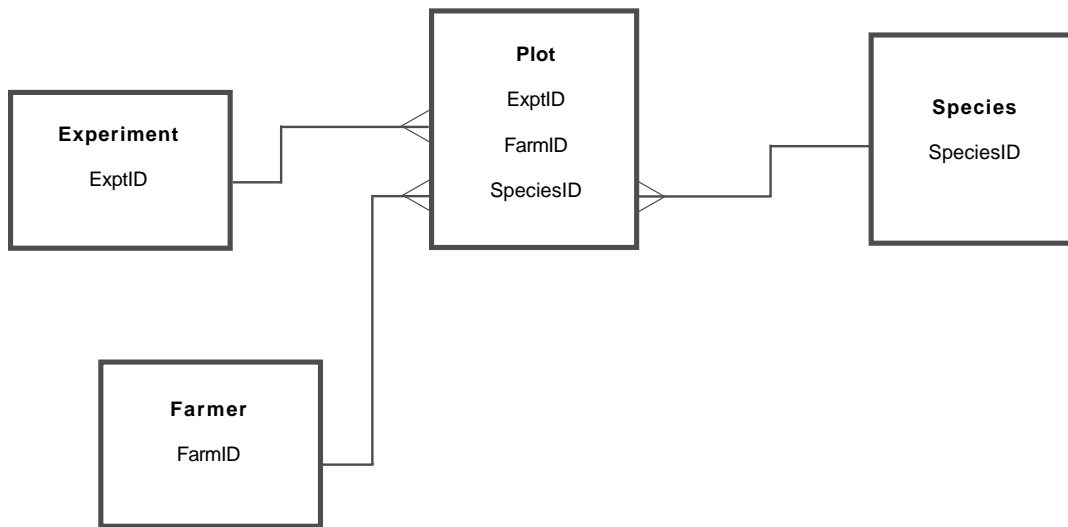The logical design of the data for the 'On-farm cropping trial' is shown below in Figure 6.5.

Figure 6.5 - Logical design for 'On-farm cropping trial'

The links or relationships are all **one-to-many**. This means that for each experiment there may be many plots; a particular species is potentially grown in many plots; a farmer may own/farm many plots. On the other hand a particular plot is related to just one experiment, has a single species growing on it, and is owned by just one farmer.

From this logical diagram of our data model we can create the necessary tables and relationships in the database. Figure 6.6 shows the corresponding relationship diagram from MS-Access.
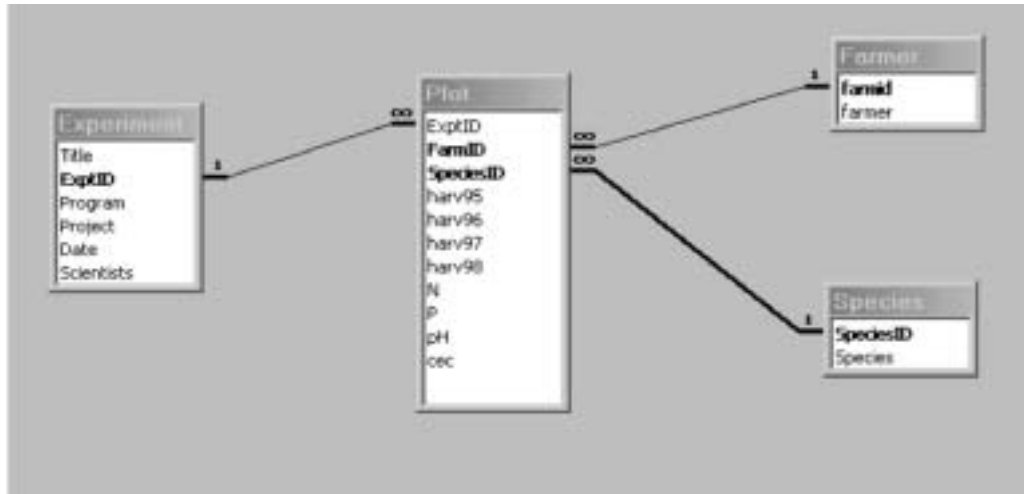
Figure 6.6 - Physical relationship diagram for 'On-farm Intercropping trial'

The process of data modelling that we have been discussing is known in database terminology as **normalization** and people often talk about various **normal forms**. Some software tools automate the process of data modelling when a dictionary of data items already exists. One such design tool is the Logbook Assistant, developed at ICRAF, which is designed to normalize data files created using MS-Excel worksheets. Although such tools exist it is still worthwhile knowing and understanding the steps that are involved in creating your data model.

Project Life Cycle Reference

In our project life cycle this entire session deals with the process of compiling the dataset.

You should by now be equipped to apply the data modelling process to set up your individual data tables. For example, if you have complex survey forms, you should be in a position to create appropriate tables to hold specific data items, rather than having to store all the data in one or two flat files.

9 6

Data modelling:
organizing data for easier querying

Exercise 1 – Survey data

Study the questionnaire for the timber business census with the file named *Timber business census questionnaire.doc*, and develop a logical design of the data structure. You should determine how many tables you need, what fields are needed in each table and how the tables are linked. Keep in mind that you need to store the study title and the name of the scientist involved, somewhere in your database. Also bear in mind that some of the questions could have many answers.

Exercise 2 – Split-plot experiment

In Sessions 2 and 3 you worked with data from the trial called 'Influence of improved fallows on soil phosphorous fractions' (*Improved fallows complete sheet.xls*). You designed in spreadsheet form these data in Session 2 and entered the data in Session 3. Your task now is to work in pairs to develop a logical model of these data. You should ensure your model contains space to store all the header information from the data collection sheet. Your diagram should include all the tables and their fields and the links between the tables. For each link or relationship indicate whether this is a one-to-many relationship, one-to-one, or many-to-many. For one-to-many relationships it should be clear from your diagram which table is at the 'one' side of the link and which is at the 'many' side. If you did not complete the exercises in Sessions 2 and 3 then use the data in the MS-Excel file *Improved Fallows complete sheet.xls*.

Exercise 3 – Your own data

Again working in pairs use your own dataset (or your partner's) and the spreadsheet you designed in Session 2 and develop a logical model for these data. Again ensure that you specify all the tables you need and all the fields in the tables. Also specify the relationships between the tables. If you have time you can repeat the exercise for your partner's dataset.

9 8

Building a data management strategy

Objectives

Œ To help participants develop a data management strategy for their own research activities.

Œ To identify training and other needs to complete it.

Œ To determine the action plan needed to implement the strategy.

Summary

A data management strategy for researchers may be prepared for individual studies, for individual scientists, for projects or whole institutes.

The strategy includes a description of all the steps in data collection, entry and processing and storage. It should indicate why, where, how and by whom each component is executed. It must be relevant to the objectives and constraints of the scientists, project and institute. At the same time it must meet the requirements of ensuring data quality, maintaining its long-term value and allowing efficient processing.

The strategy also needs an implementation plan. It will normally make sense to start at the lowest level, improving the management of data from individual studies.

A strategy needs commitment from staff at all levels of the organization. Managers as well as technicians need to understand the importance of, and the benefit gained, from good data management. Resources in terms of time and money must be allocated to these tasks and it is often the managers who have control over financial budgets and staff workloads.

Strategy

1. Brief presentation on (a) elements of a data management strategy, (b) tools to help define it, (c) steps needed to implement it, and (d) roles and responsibilities of project members from managers through to data organizers.

2. Group work involving participants thinking about their own data management problems and possible solutions. Participants will create an Action Plan for themselves.

Required skills

This session could be run independently of other sessions in the course and is suitable for managers as well as technicians.

Files needed for practical exercises

Œ *Metadata.xls* - A complete dataset.

Support documentation

Œ *Improving Research Data Management (RDM) in ICRAF* - Muraya P & Coe R, July 2001.

Œ *A review on Current Scientific Journal Requirements for Access to Data: What does it mean for a Data Management Strategy ?*

Œ *Issues in Data Ownership and Access* - Coe R, June 2001.

Œ *The 'Bromley Principles' Regarding Full and Open Access to 'Global Change' Data* - Data Management for Global Change Research Policy Statements, U.S. Global Change Research Program, July 2001.

Œ *Good practice in data management* - Statistical Services Centre - Case Study No. 6 .

Building a data management strategy

Introduction

During this course you have learnt why data management is important together with techniques for managing your data in an efficient way. Now it is time to put all you have learnt together and build a strategy that can easily be implemented and maintained.

This session will look at examples of good and bad data management. We get you to think about your current strategy, to think about problems and potential solutions. Often data are left disorganised and undocumented. The task of data documentation often falls into the category of 'I'll do that later when I have more time'. Of course you never do have more time so the task never gets done.

Let's now imagine a typical data management scenario. A particular organisation runs a research project involving several data analysts. Each analyst has taken copies of some, but not all, of the data, which are stored in several spreadsheet files. When they have noticed errors in the data during analysis they have tended to edit their own copies of the data with the intention of editing the master copy later on. Of course the master copy was never updated. At the end of the project everyone involved has other work to move onto, so the report is written and disseminated but nothing is done with the data. Over coffee and in general meetings project team members are often talking about 'archiving' the data 'when they have time'. Unfortunately no one is given the task of archiving the project materials and no time or money is available for this activity, thus it never gets done. A couple of years later when a protocol for a follow-up project is put together, staff spend many days (or weeks) trying to pull together all the information and trying to work out which is the most up to date of the data files. How much easier would this have been had they managed the data properly in the first place? Five or ten minutes now and then throughout the project with about half a day at the end to archive everything would have saved many days or weeks of work - **a little time now saves a lot of time later**.

Why is a research data management strategy important?

Good data management is not something that will 'look after itself', or evolve if left long enough. It will only come about when those responsible have a clear data management

strategy. The document 'Improving Research Data Management at ICRAF' has many examples of research data management problems. They are all true stories and could happen to anyone at any time if you do not have a well-defined strategy.

Issues include:

Œ   Managers unaware of data management requirements

Œ   Lack of skills

Œ   Intellectual Property Rights (IPR) - what authority do you have?

Œ   Data sharing policy

Œ   Confidentiality

Œ   Joint access rights in a collaborative project

Œ   Effectiveness of data management - ability to meet requests

Œ   Incompatible datasets

Œ   Standardization

Œ   Data archiving

Œ   Lack of documentation (metadata)

Œ   Illegible and disorganized log books

Œ   Requirements for paper storage

Œ   Files stored on old media (e.g. 5¼ inch diskettes) and no machine available to read them

Œ   Backups stored next to the main computer

Œ   Files backed up using software no longer available

Œ   Software upgrades

These are mostly technical issues; there are people issues as well, exemplified by this illustration by Kim Finney of CSIRO, Australia.
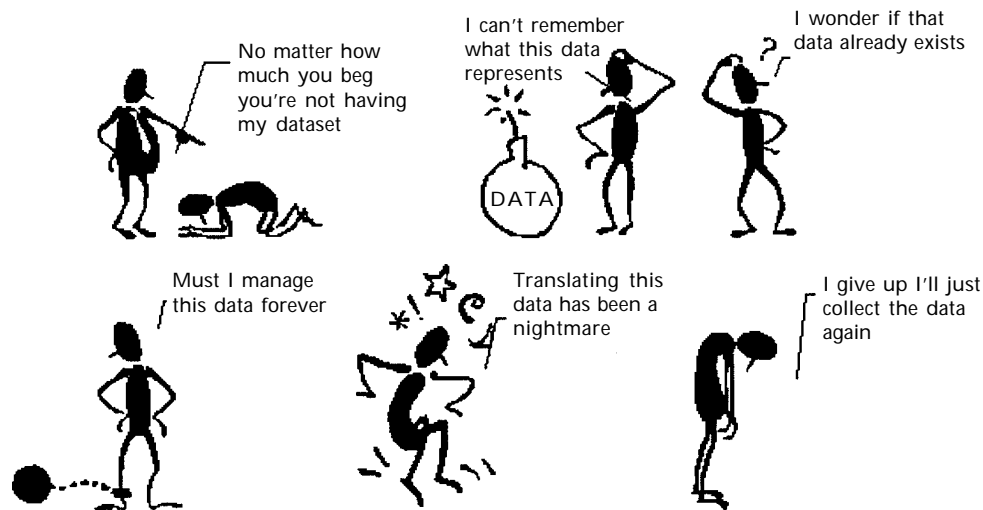
Research time wasted in seeking out, repeatedly reformatting or reprocessing old datasets, or re-collecting data is the opportunity cost associated with poor data management.

Requirements for building a data management strategy

1.   Commitment

We have put commitment as the first requirement as without it nothing will ever get done. You must not only want the end product of a good data management strategy, you must also be prepared to input the necessary resources to make it happen. Commitment must come from all members of the project team, particularly the senior managers who are often those in the position to allocate the necessary resources in terms of time and money; they are the ones with control over the financial budget and staff workloads. Each team member must play a part in data management and have a clearly defined role and responsibilities - the task should not fall to just one individual; organizing yourself is a difficult enough task, imagine how much more difficult it is to have to organize others too! Regular audits should be carried out on the data management strategy and data management should be part of staff performance reviews.

2.   Skills

This training course will have provided you with some of the skills needed for good data management. It is not an end-point though, but a beginning. You can and should continue to learn. You must be open to new ideas and new techniques and continue to review your strategy. To benefit fully from this course you will need to start using some of the skills you have learnt. In the practical session we will be asking you to develop an

action plan for improving your own data management. One suggestion we can give is to rerun this course for others at your institution or send them to the next run of the course.

It is important that junior staff have the necessary skills to undertake the data management tasks they are expected to carry out. Data entry, for instance, is often thought of as a trivial task for unskilled workers. However, if they do not understand the data or how to organize it, errors that could have been avoided will undoubtedly creep into the data involving more time and effort later on in the project cycle. A data entry clerk who understands the meaning of the data is more likely to spot errors before they get onto the computer. For example, imagine a survey involving taking anthropometric measurements of individuals in households. Weight was recorded in kilogrammes and height in centimeters. On one questionnaire the responses for height and weight have been put into the wrong boxes so it appears we have someone who is 66cm tall weighing 165kg. To an unskilled data entry clerk these are just numbers and will be entered exactly as given in the questionnaire - a skilled person is more likely to notice the error and enter the correct values.

3. Time

In any project you must allocate enough time for data management. You should consider your project team and each team member should be aware of the data management tasks they are expected to do. Data management must be part of their job description and their workload must be organized in such a way that they have the necessary time to carry out these tasks.

Often data management is not given the high profile it deserves. In the past funders and researchers alike considered the results of the analyses as being the key output from a project and so that area was given most of the resources in terms of time and money. Nowadays people are beginning to consider the data itself to be a key project output and thus are more willing to allocate the necessary resources to data management so that one of the outputs can be a well-documented, usable data archive.

4. Money

Along with time comes money. When developing project proposals it is essential to allocate a section of the budget for data management. All project proposals must demonstrate the existence of a comprehensive data management plan. The project budget must clearly indicate allocation of resources for data management and how these resources will be used. This includes allocating money for equipment. Computers must be powerful

enough to perform the necessary data management tasks we have been describing throughout this course; they must have facilities for making backups (zip drives and/or CD-writers); they must have network capability so that files can be shared between project team members; and they must have up-to-date software. This equipment must be budgeted for.

Key components of strategy

The four key components of the strategy are:

1. Transformations and their products: representing steps in research data management.

2. Metadata: documenting the ACTUAL data management steps and products.

3. Data management plan: a reference against which the ACTUAL steps can be evaluated to monitor performance.

4. Data management policy: generally agreed principles that guide structure and contents of 2 and 3 above.

Transformation products

In the introductory note for this course we talked about the project life cycle from problems to knowledge, through data transformations. Throughout the lecture notes we have periodically referred back to this life cycle.

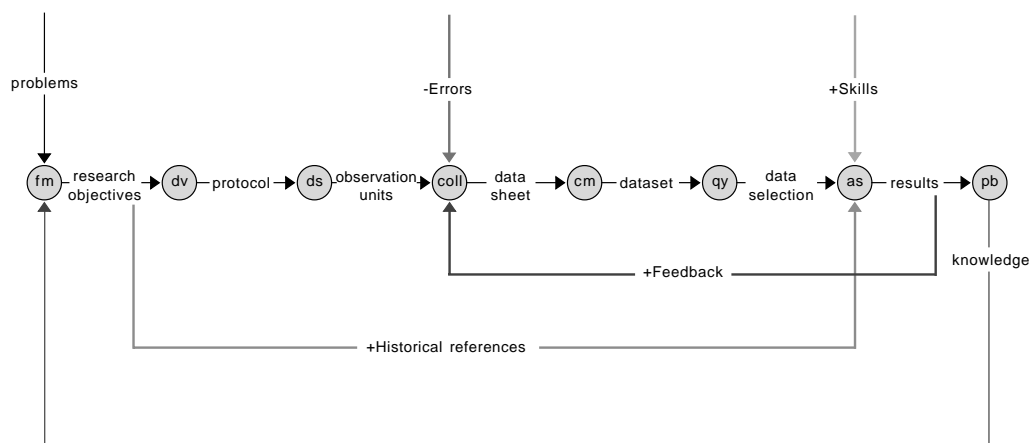In Figure 7.1 we show the project life cycle again.

Figure 7.1 - Research transformation-products life cycle

This time we have added extra inputs, which affect the performance at each data management step, each of which we will explain. First of all **errors** are a negative input. Unfortunately they can come in at any time, although in the diagram we have just illustrated this for the data collection stage. The aim is to minimize their occurrence and/ or to minimize the time (steps) between their occurrence and detection; this latter point underlines the importance of the feedback loop in data management. The earlier the feedbacks are available, the more likely we can detect and correct errors before they propagate down the chain. The worst case is when research objectives are formulated, data are collected etc., only to realize that the results are invalid because the objectives were not valid! From some of our exploratory data analysis we are likely to spot errors in the data so the **feedback** loop is there to correct these errors.

Skills are a required positive input to the process. Again these are needed at all stages although on the diagram we have just the one arrow indicating skills so as not to clutter the diagram. Sessions 2, 3, 4, 5 and 6 were designed to broaden your skills in compiling and querying your data. Without them your datasets remain as simply a collection of hard paper data sheets, which are in turn very inefficient to query. The lack of these skills is the main reason why data usually lie around in paper form for such a long time un-analysed.

Because the immediate outputs from one step are rarely sufficient to perform the next step, we need to be methodical about the way we organize ALL the products from each transformation for future reference. For instance, to effectively analyse a data selection, you need information generated earlier (such as the research objectives) because it is often not possible to deduce it from the data being analysed. In the project life cycle figure (Figure 7.1) this is indicated by the **historical references** loop. When this input is not well-managed it is difficult to complete the analysis when the data originators are no longer available, e.g. when there is staff turnover.

At the analysis stage we can use not only data from the current project but also data and knowledge produced from earlier projects (this is another example of an external input). This is sometimes known as secondary data analysis and can help establish a starting point for the analysis of the current project data. Of course we can only do this if the data from the earlier projects have been well-managed and documented and are in a usable form.

Finally the knowledge resulting from the project will feed into further projects, as inevitably the results will lead to further questions. Knowledge of course covers many aspects. Not only answers (hopefully) to the original research questions will be found, but also knowledge of the methods used to achieve the results; details of the experiment or

survey; what worked well, and what could have been done better. The entire package of knowledge can be used to generate more research questions and other methods of working.

Metadata

The most common definition of the term metadata is **data about data**. A fuller definition is **descriptive information about data, which allows a potential user to determine a dataset's fitness for use**. Metadata has many applications. It can be used to:

Œ	Concisely describe datasets and other resources using elements such as the name of the dataset, the quality, who is the custodian, how to access the data, what is its intended purpose, and whom to contact for more information about the data.

Œ	Enable effective management of data resources.

Œ	Enable accurate search and data resource discovery.

Œ	Accompany a dataset when it is transferred to another computer so that the dataset can be fully understood, and put to proper use and to duly acknowledge the custodian of the dataset.

Why should we use metadata?

Metadata help the users find the data they need and determine how best to use it. In addition metadata is also of benefit to the data-producing organization. For example as personnel change in an organization undocumented data may lose their value and sometimes data are lost. Also new employees may have little understanding of the contents and uses of a dataset and may find that they don't understand the results generated from these data. Lack of knowledge about other organizations' data can lead to duplication of effort and waste of time.

Information needed to create metadata is often readily available when the data are collected. A small amount of time invested at the beginning of a project will save money and time in the future. Data producers cannot afford to be without documented data. The initial expense of documenting data clearly outweighs the potential costs of duplicated or redundant data generation.

How much metadata should be stored?

There are now many different 'standards' for metadata and there are metadata-authoring tools available, which may or may not suit your needs. To help you decide how much metadata to store we suggest you think what, where, when, how, why and who.

Œ **What** do the data represent? **What** was the name of the project that generated them? Perhaps include an introduction or abstract to the project. **What** is the format and structure of the data? Include here any naming conventions used for the files.

Œ **Where** were the data collected? Give details about the site and perhaps the sampling frame used.

Œ **When** were they collected? State the time period covered by the data.

Œ **How** were the data collected? Describe the measuring instruments used.

Œ **Why** were they collected? What was the purpose of the research?

Œ **Who** collected the data, or **who** were the principle researchers involved? **Who** holds the data? **Who** has property rights to the data? Include names and contact details.

These are just ideas but if you can answer these questions then you are well on the way to having a well-documented dataset. Of course for this information to be useful for others it must be stored somewhere that others can access - in other words not just your memory!

Below we give an outline of what might be included in the metadata.

Title	The name of the dataset or project
Authors	Name of principal investigator and other major players in the research. Include mailing address, phone number, fax numbers, email, web address, etc.

Data set overview

Introduction or abstract. Time period covered by the data. Physical location of the data. Any references to the Internet.

Instrument description

Brief text describing the instrument, with references. Figures or links if applicable. Table of specifications.

Data collection and processing

Description of the data collection. Description of any derived parameters. Description of quality control procedures used.

Data format Data file structure and naming conventions. Data format and layout. Data version number and date. Description of codes in the data.

Data documentation should be sufficiently complete, so that persons unfamiliar with a given project could read the documentation and be able to use and interpret the data.

Part of your data management strategy might be to develop a database of metadata for all your research projects. Of course if your datasets are currently in disarray and undocumented, this will be a mammoth task, but you could start with current and future projects. We don't expect you to solve all your data management problems overnight.

Data management plan

A plan, which describes how data will be recorded, processed and managed during the life of the project and archived at the project's end. A data management plan will include:

Œ Clearly defined roles for all staff, including who keeps what, who is to organize the data entry, data-checking, etc.

Œ A regular backup procedure including details of when backups are to be made, where they are stored, etc.

Œ Details of data quality checks to be done on the data and the mechanism for recording and correcting errors.

Œ Putting data management on the agenda of project meetings to keep all team members abreast of the current situation.

Œ Procedure for upgrading software - there is a great deal of benefit in using unified systems for data from all projects. For instance if all databases have a similar structure, then updating to the next version becomes a routine task.

Œ Details of how an archive is to be produced at the end of the project.

Œ Details of how the archive is to be maintained. The archive needs to be readable in the future using systems and software not yet available. To keep your archive current you must allow for periodically rewriting the archive to ensure its continuing readability. This may involve changing the format of the files and/or the media on which they are stored in order to prevent them becoming obsolete. For example before you throw away your one remaining PC with a 5¼ floppy drive make sure you have copied everything of any importance from the stack of 5¼ floppy diskettes collecting dust on the shelf!

Data management policy

You may wonder what the difference is between a data management plan and data management policy. Data management policies are the directions, which guide the activities of the division to ensure there is a consistent approach to data management. In other words these could be thought of as the overall ideals for the institution, and the data management plans for each project detail how these ideals are to be implemented. A typical policy might include the following objectives:

Œ   To establish and distribute high-quality, long-term datasets.

Œ   To standardize quality control procedures for all datasets.

Œ   To ensure data and other project materials are archived in a timely fashion, and the archive periodically reviewed to ensure continued readability.

Œ   To reduce the times between data collection, entry and analysis.

Œ   To maintain adequate backup protocols for all datasets.

Œ   To facilitate data access and usability through improved metadata.

The following is an example data policy you may find useful to consider. We also refer you to the 'Bromley Principles' mentioned in the supporting document 'The "Bromley Principle" Regarding Full and Open Access to "Global Change" Data' -Session 7.

Example data policy

Introduction

The data policy of <Name of institute> is intended to augment long-term success of research, ensuring that valuable data collected by researchers are properly and effectively utilized and managed, thereby making best use of allocated project resources.

Principles

Data management issues will be addressed at the project-level. Project-level data management is a continuous process spanning the life of the project and beyond and is essential for the dissemination and utilisation of project results. This process should be captured within a data management plan, which must be formulated and provided as part of each research project proposal.

<Name of institute> recognises that project data management is a specialised activity requiring skills and resources that may not be available or accessible. However, developing a data management plan is a value adding activity that will:

1. Reinforce the principles of data management, thereby improving the longevity and effective use of information resources.
2. Provide a framework against which advice can be given.
3. Identify and document problem areas where there is a need for improved data management capacity.

<Name of Institute> will use primarily a decentralised data management and distribution system. The 'Centralised' component will be a comprehensive inventory of metadata with 'pointers' to the data location and key contact persons.

Implementation

Principle investigators should formulate a data management plan as part of the research proposal for a project. The data management plan should outline how the project will address:

1. Information and data flow.
2. Data documentation (e.g. metadata).
3. Data quality.
4. Technical issues (databases etc.).
5. Dissemination of final project data.
6. Longevity and final archiving.
7. Performance indicators.

Roles and responsibilities

Your data management policy is likely to include roles and responsibilities for individuals either within or outside the project team. Here we give examples of such responsibilities for different categories of individual.

Data owner

The <name of institute> is the data owner of all the research data and holds copyright to its policies, manuals and compilations of its information.

Data custodian

A manager of the institute who has been delegated responsibility for a portion of the information resource on behalf of the institute, in order to ensure its integrity and accuracy. Responsibilities:

Œ   Identify items of corporate data and distinguish primary source data.

Œ   Identify and document who is allowed access to the data and what level of access they should have.

Œ   Authorize downloads and uploads of corporate data.

Œ   Identify and document the process for authorizing and granting access to individuals.

Œ   Implement processes that maintain the integrity, accuracy, precision, timeliness, consistency, standardization and value of data.

Œ   Arrange appropriate training for staff and others to ensure data are captured and used accurately and completely.

Œ   Understand and promote the value of the data for institute-wide purposes.

Œ   Ensure compliance with the principles of the Privacy Act.

Data user

An individual, who has permission from the data custodian to access and use the data. It is the responsibility of all levels of management to ensure that all data users within their area of accountability are aware of their responsibilities as defined in this policy. A data user:

Œ Is responsible and accountable for all data access made through their user account and the subsequent use and distribution of the data.

Œ May not use the data for their own personal gain, or for the gain or profit of others.

Œ May not access data to satisfy their personal curiosity.

Œ May not disclose data to unauthorized persons without the consent of the data custodian.

Œ May not disclose their password to anyone.

Œ Must abide by the requirements of the Privacy Act and other relevant statutes.

Security administrator

The person responsible for security administration. Responsibilities:

Œ Provide access to users as specified by the data custodian.

Œ Ensure that the proper logical safeguards exist to protect the data, and that appropriate disaster recovery procedures are in place.

Œ Provide adequate procedural controls to protect the data from unauthorized access.

Œ Designate a custodian for all corporate data.

Information Systems Group

The Information Systems Group's responsibilities include a data management role. Responsibilities:

Œ Promote the management of research data as a corporate resource.

Œ Understand and promote the value of data for institute-wide purposes and facilitate data sharing and integration.

Œ Document and promote the logic and structure of institute data.

Œ Manage the use of common standard codes and data definitions throughout <name of institute>.

Other examples appear in the document 'Improving Research Data Management at ICRAF' which considers three groups of staff, namely:

1. Organizers - these people handle raw data on a daily basis. They set up data filing systems, enter and check data and maintain data banks.

2. Analysers - these people analyse and interpret data, reducing raw observations to useful information.

3. Managers - these people are responsible for providing an enabling environment for the first two groups and ensuring all commitments to stakeholders are met.

The document goes on to present indicators of research data management quality for each of these groups, plus a set of actions for improvement for each group.

Hints on building a strategy

Document current procedure

This gives a good starting point for your strategy. Determine what works well and what doesn't work.

Seek consensus

Unless all team members or at least the majority are in favor of the strategy it is unlikely to succeed. Seeking consensus makes all team members feel they have some control over the decisions that are made rather than having them imposed from above. They are then more likely to follow the procedures.

Establish a data management forum

This gives data management a higher profile and if discussed regularly the topic will stay in the minds of team members. Use meetings to update staff on progress and to set deadlines for particular tasks.

Standardize

Use similar data management plans for all projects. Not only does this save time but it also makes staff more familiar with the procedures.

Obtain funding

In any project proposal include a detailed data management plan and budget for it.

Remember a perfect data management strategy is not something that can be reached overnight (if ever - you should always be considering ways of improving your strategy). Most likely you have datasets that are totally disorganized from projects that are long finished. Documenting these and producing order from chaos is a daunting task. Consider as an example the data files on your own PC. Ideally you would want each file to be documented, but as a start to this procedure how about documenting just the first or second directory level. You could say for instance that files and folders below this level are all to do with a particular project. Later, when you have more time, you can go to the next directory level. In this way you turn a mammoth task that you are unlikely to find time for, into a series or more manageable tasks.

Table 7.1 below gives an example of a 3-phase approach to establishing a data management plan. The table summarizes the key objectives at each phase:

Phase I	Phase II	Phase III
Obtain consensus on strategy.	Identify core data elements for internal users.	Enhance the metadata repository.
Identify core data elements for external users.	Establish database certification process.	Continue to identify and retire obsolete data.
Establish a data management forum.	Establish data stewards.	Make shared data available in a timely manner.
Establish an inventory of databases.	Begin to identify and retire obsolete data.	
	Establish a metadata repository.	
	Identify data that needs to be shared.	

Table 7.1 - Key objectives at each stage of a data management plan

Building a data management strategy

Exercise 1 - Metadata

Work in pairs for this exercise.

A staff member is about to leave the organisation and as part of the hand-over process has given you a data file containing these data in Table 7.2 below.

block	plot	subplot	row	treat	wt1 (gms)	wt2 (gms)	MC%	15MC%	t/ha
1	1	1	1	Control	2291	528	14.4	2386	10.6
1	1	1	2	Control	1156	228	14.4	1204	5.4
1	1	1	3	Control	871	199	14.4	907	4.0
1	1	1	4	Control	970	219	14.4	1010	4.5
1	1	1	5	Control	505	88	14.4	526	2.3
1	1	1	6	Control	748	138	14.4	779	3.5
1	1	1	7	Control	541	134	14.4	564	2.5
1	1	1	8	Control	453	110	14.4	472	2.1
1	1	1	9	Control	648	143	14.4	675	3.0
1	1	1	10	Control	388	116	14.4	404	1.8
1	1	1	11	Control	490	148	14.4	510	2.3
1	1	1	12	Control	380	97	14.4	396	1.8
1	1	1	13	Control	583	153	14.4	607	2.7
1	1	1	14	Control	294	83	14.4	306	1.4
1	1	1	15	Control	689	170	14.4	718	3.2
1	1	1	16	Control	715	166	14.4	745	3.3
1	1	1	17	Control	516	131	14.4	738	2.4
1	1	1	18	Control	790	159	14.4	823	3.7

Table 7.2 - Data file

The complete dataset is in the file called *Metadata.xls* . What would you like to know about these data? Write down as many questions as you would like to ask the staff member while he/she is still around.

Exercise 2 - Problems & solutions

Working in groups of 4 or 5, discuss the data management problems of each member's institute or organization. Consider possible solutions to these problems. It is likely that some of you have similar problems. In Table 7.3 below write the problem, the number of group members with that problem and any possible solutions you come up with. Continue on a separate sheet if necessary.

Problem	Frequency	Possible solutions

Table 7.3 - Problems and solutions

In the plenary session we will look at the most frequent problems occurring within the groups and the solutions that have been suggested.

Exercise 3 - Tools to help define a strategy

Again working in groups of 4 or 5, consider each of the topics in Table 7.4 below, together with any others you might think of. Write down what your current situation is, and then state what changes are needed. The topics are not presented in any particular order.

Area of Concern	Current situation	Changes needed
Objectives of data management scheme Œ What sort of data is concerned? Œ Will this be a scheme for a few simple activities or many complex or long-term activities? Œ Will the data be used by a few individuals working in isolation or by teams with many people working together? Œ Will the data be used at the time and place it is generated or elsewhere?		

Area of Concern	Current situation	Changes needed
Responsibilities. Who is responsible for: Œ Designing data collection sheets? Œ Checking field data recording? Œ Designing spreadsheets and databases? Œ Entering data? Œ Checking and correcting the raw data entry? Œ Pre-processing the data for analysis? Œ Checking and correcting data after pre-processing? Œ Analysing data? Œ Archiving data? Œ Overseeing the whole process? Œ Do all these people have the necessary skills?		
Timing Œ When (relative to field data collection) is raw data entered? Œ When is it checked? Œ When is the first analysis done?		

Area of Concern	Current situation	Changes needed
Data integration Œ Can data from different phases of the study be **integrated**? Œ Does research planning ensure data are comparable across trials, locations and seasons? Œ Are data from different parts of the study sufficiently well identified to be able to link them?		
Location Œ Where are paper data records stored? Œ Where are computer files stored? Œ Are copies sent to other locations? Œ Which is the 'master' copy and how do you ensure all versions are consistent?		
Data sharing Œ Who owns the data? Œ Who has access to it? Œ What happens when staff responsible for the data, leave the institute? Œ What rights and responsibilities do partner institutions have?		

Area of Concern	Current situation	Changes needed
Institutional Œ Are there institute - wide data management guidelines or practices? Œ Does your data management strategy have to be congruent with any others? Œ What priority is given to data management in the institute? Œ Does it receive adequate resources? Œ Does it receive attention during evaluation?		
Other		

Table 7.4 - Defining a strategy

Exercise 4 - Action Plan

As a final output from this course we want you to consider the current data management strategy at your institute or organization. Think about what works well and what could be improved. What steps will you take to improve your strategy? Think about the short-term (next week), the mid-term (within the next 6 months), and the longer term (within the next 2 years). Think about what you would like to achieve within each of these time frames and how you might reach those goals.
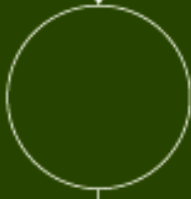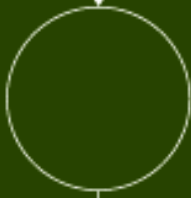
Table 7.5 shows an **Action Plan** template; fill this in stating your goals and a series of manageable steps you will take to achieve those goals. Keep your plan somewhere prominent in your office so you are constantly reminded of your goals, and tick each point as it is achieved. Continue on a separate sheet if necessary.

Data Management Action Plan			Today's Date: _____	
Deadline	**Goal**	**Actions**	**Progress**	**Date achieved**
(within a week) Date: _____				
(within 6 mths) Date: _____				
within 2 yrs) Date: _____				

Table 7.5 - Action Plan template

World Agroforestry Centre
TRANSFORMING LIVES AND LANDSCAPES