Data Management Guidelines for Experimental Projects

March 1998

The University of Reading
Statistical Services Centre

Biometrics Advisory and
Support Service to DFID

Contents

1. Introduction

Research projects often involve the collection of a large volume of data. The data have then to be processed and analysed, with results and summaries being prepared for publication in some form. For this sequence to proceed smoothly, the project requires a well-defined system of data management. This booklet gives some guidelines on the components of such a system.

The main stages of the data management process in a research project are as follows:

- The raw data have to be entered into the computer, and checked;
- The data have then to be organised into an appropriate form for analysis (often in different ways, depending on the analysis);
- The data have to be archived, so that they remain available throughout subsequent phases of a project, and afterwards.

Most of the examples in this booklet refer to projects involving experimental data rather than survey data. Experimenters have sometimes been unaware of the value of careful data management until fairly late within their project, and the research has suffered as a consequence. We hope that these guidelines will help researchers to plan the data management aspects of their project from the outset.

2. What we mean by "data"

At the simplest level, "data" are the values recorded in the field books, record books or data-logging devices, that are to be entered into the computer and then analysed. An example of a simple dataset – a table of rows and columns – is shown below.

A simple dataset

Plot	Replicate	Treatment	Flower	Total weight	Head weight	Grain weight
101	1	4	26	25.2	6.6	1.7
102	1	2	28	32.7	8.8	2.4
…	…	…	…	…	…	…
416	4	8	26	19.7	4.9	5.3

The information in this table is certainly needed for the analysis, but it is incomplete. Additional information in the protocol which gives details of, for example, the treatments, the type of design, the field plan and the units used for measurements, is also needed, for both the analysis and the archive. Such information is sometimes called "metadata" – but whatever name is used, it should be considered as an integral part of, and equally as important as the data in the table.

We are now in a multimedia world, so photographs and maps can be considered as part of the "metadata", as can reports, talks and other presentational material. For most of this booklet we use the word data relatively narrowly, but we return to the broader meaning that encompasses such material in the section on archiving.

Roughly, one can regard the data management task in a project as *simple* if all the data to be computerised have been collected on a single type of unit, e.g. plots or animals. The task is *complex* where data have been collected from a number of different units or levels. For example, in an on-farm study there will often be interview data at the farm level and response measurements at the plot, animal or tree level.

Sometimes the complexity of the data management tasks differs for different parts of a project. An example is a regional project consisting of a variety trial at each site, where the data are to be keyed at each of the sites. In such a project, the set of varieties is often not identical at all sites. Then the data entry at each site is simple, i.e. it is a single rectangle, as in the example above. However, the regional co-ordinating office might need four additional sets of data, as follows:

- Data on each site, e.g. name, location, soil type

Site Number	Site Name	Country	Latitude	Longitude	Altitude (metres)	Soil Type	...
1	Dori	Benin	10.654	2.813	200	C	...
2	Gaya	Niger	12.109	4.171	175	D	...
...	...	...	...	...	...	...	...
46	Mara	Niger	12.576	2.543	140	D	...

- Data on each experiment at a site, e.g. date of planting, insect problems

Site Number	Experiment Number	Year	Planting Date	Stress	Pest Problem	...
1	1	1997	12 June	mild	minor	...
1	2	1997	16 June	none	none	...
1	3	1998	2 July	none	none	...
2	1	1997	19 June	severe	major	...
...	...	...	...	...	...	...

- Data on each variety used in the project, e.g. name, origin, type

Variety Code	Variety Name	Origin	Type	...
12	OFT1226	Mali	erect	...
14	PLO2279	Togo	spreading	...
...	...	...	...	...

- Yield data from each of the sites

Site Number	Experiment Number	Variety Code	Yield	...	...
1	1	6	4.1	...	...
1	1	14	2.9	...	...
...	...	...	...	...	...

In this example, the co-ordinating office would need to use techniques that are built in to relational database management systems (DBMS) to combine the information from the different data tables and so to provide an overall analysis across sites.

In general, where the data management tasks are complex, a database management package should be used by the project. This enables all the information to be stored in a structured way. Whether the same software is used for all tasks, i.e. for the data entry, checking, management and analysis, is for the project team to decide.

3. Software for handling data

The different types of software used for data management include the following:

- database (DBMS) packages, e.g. Access, dBase, EpiInfo, Paradox, DataEase;
- statistics packages, e.g. Genstat, MSTAT, SAS, SPSS, Statgraphics, Systat;
- spreadsheet packages, e.g. Excel, Lotus-123;
- word processors, e.g. Word, WordPerfect; or text editors, e.g. Edit.

Database, statistics and spreadsheet packages have overlapping facilities for data management. All handle “rectangles” of data, as shown in the previous section. In these rectangles, each row refers to a case or record, such as an animal or a plot, and each column refers to a measurement or *variable*, such as the treatment code or the yield. Broadly, database packages are very good at manipulating (e.g. sorting, selecting, counting) many records or rows. They are also able to handle hierarchical data structures, such as observational data collected at both a farm and a field (crop) level, where farmers have more than one field. Statistics packages are designed primarily to process the measurements, i.e. they have powerful tools for operating on the values within the variables or columns of data. Spreadsheets do something of everything – though with limitations.

Our general views on software for data management are presented next.

- Transfer of data between packages is now simple enough that the same package need not be used for the different stages of the work.
- The data entry task should be conceptually separated from the task of analysis. This will help when thinking about what software is needed for data keying, for checking purposes, for managing the “data archive” and for analysis.
- Database management software (DBMS) should be used far more than at present. Many research projects involve data management tasks that are sufficiently complex to warrant the use of a relational database package such as Access.
- Spreadsheet packages are ostensibly the simplest type of package to use. They are often automatically chosen for data entry because they are familiar, widespread and flexible – but their very flexibility means that they can result in poor data entry and management. They should thus be used with great care. Users should apply the same rigour and discipline that is obligatory with more structured data entry software.
- More consideration should be given to alternative software for data entry. Until recently the alternatives have been harder to learn than spreadsheets, but this is

changing. Some statistics packages, for example SPSS, have special modules for data entry and are therefore candidates for use at the entry and checking stages.

- If a package with no special facilities for data checking is used for the data entry, a clear specification should be made of how the data checking will be done.
- A statistics package – not a spreadsheet – should normally be used for the analysis.

4. Database structure

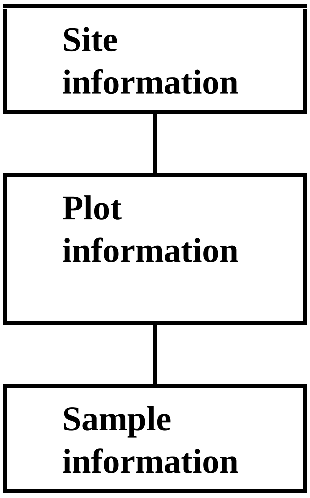
As noted in Section 2, the task of data management may be simple or complex. In database terms this distinction corresponds to whether the data tables are flat or structured – i.e. linked together in various ways. The database structure is *flat* if all the data exist at a single level and can be held in one database table. Familiar examples are: an address list, a card index of library book titles, and a list of references.

Experimental projects usually require several, linked tables to hold all the data. For instance, an experiment conducted regionally at several sites may produce a flat file for storing information about each site, such as the average rainfall at the site, maximum temperature, location of the site e.g. nearest village, or distance from village. Here the rows in the data file would be the sites, while the columns would provide different pieces of information about each site (as in the example in Section 2).

A second flat file is used to store the information on each plot. Rows of this file will include a code to identify the particular plot, while columns would correspond to plot information such as the time at which flowering occurred in more than 50% of the plot, plot yields, or the number of live plants in the plot.

Yet another flat file would be needed to store the information collected over time at each plot. Here the rows would correspond to the samples, with several rows on each date to allow for the different plots being sampled on that date. The first two columns in the data file would give the sampling date and an identification code for the plot, while other columns would hold the measurements, e.g. pod weight, or plant stand.

When these three flat files are considered together, they form a hierarchical structure, illustrated in the following figure.



One record exists for each site.
SITE is the unique identifier.

One record exists for each plot within the site
and includes a grid reference for the plot.
SITE + PLOT makes up the unique identifier.

One record exists for each sample.
SITE + PLOT + DATE makes up the unique identifier.

Site information resides at the top level of this structure, plot information resides at the next level, while information collected for each sample resides at the lowest level of the hierarchy. The information at different levels of the hierarchy is linked via *key variables* (or *key fields*). The key variable is a unique field or combination of fields that can be used to identify a particular record. One – and only one – record would hold a particular key value. Many database packages do not allow you to enter a new record where the key value is the same as the key value in an existing record. In the case of natural resources experimental data, the key field is typically one that combines the code of the plot with the date (assuming that there is a maximum of one measurement for each variable per day).

The key field values at one level of the structure link a record to the record (or records) at another level with the same values in the corresponding key fields. These links or relationships between the database tables define the database structure. The facility to store database structures is what makes a DBMS important for experimenters.

5. Designing a data entry system

In designing a suitable system for data entry, consideration must be given to several aspects of the data. These are discussed in turn.

5.1 Understand the structure of the data

Few projects generate simple data; most have a complex structure with more than one flat file which must be linked in a clearly defined way, as described in the previous section. It is essential that both the flat file components and the links are fully specified, to ensure that the information meets the database requirements of completeness, integrity and minimum redundancy (or duplication) of information. Modern, relational database software makes this task fairly easy. Spreadsheet software does not – in fact it can make the task more difficult.

5.2 Identify the types of information being collected

Try to foresee the full range of different types of data that will be collected, e.g. plot data may consist of crop yield from all plants in the plot, number of plants with pods for harvest, total pod weight and number of dead plants. Build facilities in the data collection sheet for recording all such information. Often data will be collected from the same plot on a number of sampling occasions. Dates of such records must be kept, with space available on the recording sheet for notes about the plot or farm at that specific time. Such secondary information will be valuable at the data analysis stage to explain any curious behaviour of the data.

Codes are needed to distinguish between information collected on different types of plots. Some plots for example may be primarily for recording disease incidence, while others are primarily for yield. Abbreviations may be used as suitable codes.

5.3 Specify the measurement units and precision

Ensure that the database system clearly specifies the units of measurement used for all quantitative variables. Changes in measuring instruments, or in field and research staff, or in methods of data collection, may bring about changes in measurement units. Consideration must be given at an early stage of the database design to allow for such changes to be incorporated into the data recording system.

Specify clearly the precision (number of decimal places) to which all measurements are to be recorded. The number of significant digits should match the real precision of the measuring instruments or recording devices.

6. Data entry and checking

We consider primarily the data that are collected in field books or survey sheets. First we discuss the overall strategies that can be adopted for data keying and for checking, and then give separate guidelines on the two aspects.

6.1 Strategy for data entry and checking

When planning a strategy for data entry, clearly distinguish between the data entry / data checking / data management activities and that of data analysis. The ultimate aim should be a fully-documented archive of checked, correct, reliable data that can be subjected to scientific scrutiny without raising any doubts in the minds of subsequent researchers. Unfortunately, many worthwhile research projects do not achieve this.

The process of data entry will normally involve a skilled person who designs the system, while more junior staff, e.g. trained data entry operators or field staff, carry out the actual keying. Checking is done both at the time of keying and afterwards. If the project is small, then the same person may plan the system, key the data and do the checking, but it is still useful to have a clear strategy for the activities.

When planning the system, aim to make the data entry stage as simple as possible. For example, in a replicated experiment it should never be necessary to type variety names or long treatment codes for each plot. A single letter or number is usually sufficient. Then, either the data entry system can insert the full code, or the full names may be available in a separate, “look-up” file, as outlined in Section 2. Simplifying the keying process will speed the task, make it less tedious and hence also less error-prone.

The logical checking phase should be done by trained staff who understand the nature of the data. Usually this phase involves preliminary analyses, plotting etc.

In practice, the data entry and checking steps are usually designed at the same time. The way the data checking is undertaken will, however, depend on who is entering the data. Non-skilled staff should be expected to key exactly what they see on the data sheets or field books, and the logical checks (e.g. checks to rule out pregnant males, or minimum greater than maximum temperature) should be done by scientifically-trained staff after the (double) entry is complete. In that way, reasoned decisions can be made about what to do. If scientists are keying the data themselves, then the entry and full data checking can proceed together.

6.2 Guidelines for data entry

These guidelines may be summarised as “Do the data entry promptly, simply and completely.”

- The data should be entered in their "raw" form – i.e. directly from the original recording sheets or fieldbooks – whenever possible. They are therefore entered in the same order that they were collected. The copying out or transcription of data prior to keying should be kept to an absolute minimum.
- All the data should be entered. Entering "just the important variables, so they can be analysed quickly," limits the possibilities for checking, which can make use of relationships between variables. Often when short-cuts are attempted, the full data entry has to re-start from the beginning, or (more usually) the remaining variables are never entered.
- No hand calculations should be done prior to data entry. Software can be used to transform data into the appropriate units for checking and analysis, e.g. grammes per plot to kilogrammes per hectare, or to take averages of replicated readings, etc.
- One of the variables entered should give a unique record number. In field experiments this would generally be the plot or sub-plot number.
- In field experiments, the position of each plot should be entered. This enables data (and residuals during analysis) to be tabulated, or plotted in their field positions – very useful for checking purposes. Where plots are regularly spaced, with no gaps, the position can be derived from the plot number. Otherwise, two extra columns are keyed giving the co-ordinates.
- The data should be entered promptly – i.e. as soon as possible after data collection. For example, where measurements are made through the season, they should normally be entered as they are made. This speeds the whole process, because the data entry task at the end of the trial or survey is then not so large and daunting. It also helps the checking, because some checks can indicate unusually large changes from the previous value, and odd values can then be verified immediately. Feedback of any problems that are noticed to field data collectors can help maintain the data quality.

The above advice applies even when there are complications in the data. Typical complications that will require careful thought are as follows:

- Mixed cropping experiments, where plots have different numbers of variables measured, depending on whether they are sole or mixed plots.
- Agroforestry experiments, where data are often recorded on different subunits of each plot.
- "Expensive" measurements, such as neutron probe data, that may be collected for just a few of the treatments, or on only some of the replicates.

- Repeated measurements, where data, e.g. on disease score, are collected through the season.
- Animal studies, where the order of collection of the data may be different on each occasion.

6.3 Guidelines for data checking

The objective is that the data to be analysed should be of as high a quality as possible. Therefore the process of data checking begins at the data collection stage and continues until, and during, the analysis.

Checks when the data are collected

- Data should be collected and recorded carefully. Consider what checks can be incorporated into the data collection routine. For example, the best and worst animals could have a one-line report to verify – and perhaps explain – their exceptional nature. This will confirm that they were not written in error.
- Consider collecting some additional variables specifically to help the checking process. For example, in a bean trial, the number of plants with pods that are harvested could serve as a check of the yield values. It may be relatively inexpensive to take aerial photos (using a kite or balloon) to record the status of each plot. Where this is not possible, recording the "state" of the plot, or even of each row of plants within the plot (e.g. on a scale from 1 to 9) can be worthwhile.

Checks while the data are being entered

- If possible, use software for data keying that has some facilities for data checking.
- Recognise that ignoring the data entry guidelines given above may be counter-productive for data checking. For example, changing the order of the data, transforming yields to kg/ha or calculating and entering only the means from duplicate readings can all lead to errors in copying or calculation. It also makes it more difficult to check the computerised records against the original records.
- Do not trust reading or visually comparing the computerised data with the original records. Though often used, it is not a reliable method of finding key-entry errors.
- Consider using double entry, where the second keying is done by a different person. This does not take much longer than visual comparison and is a far better form of validation. Modern data-entry software has facilities for a system of double-entry with immediate or subsequent comparison of values.
- Build in further checks if your software allows. The simplest are range checks, but other, logical checks can also be used. For example, for a particular crop, grain weight might always be less than half of head weight.

Checks after entry

- Transforming the data may help the checking process. It may be easier to see whether values are odd if they are transformed into familiar units, such as kg/ha.
- The initial analyses are a continuation of the checking process and should include a first look at summaries of the data. Useful things to produce at this stage are:
  - *extreme values*, in particular the minimum and maximum observations;
  - *boxplots*, to compare groups of data and highlight outliers;
  - *scatterplots*, especially if you use separate colours for each treatment;
  - *tables* of the data in treatment order.
- With experimental data, the initial ANOVA should also be considered as part of the checking process. Particularly with experimental data, it is difficult to do all the checking without taking into account the structure of the data – a value that is odd for one treatment may be acceptable for another. So make use of software for the analysis that allows you easily to display the residuals in a variety of ways.

7. Audit trail

An audit trail is a complete record of changes to the data and decisions made about the data and the analysis, rather like a log book. In fact, it is the equivalent for data management of the rather old-fashioned notion of a scientist's notebook, which is as relevant today as ever. A well-maintained audit trail, log book or notebook greatly eases the subsequent tasks of writing reports on the data and of answering data queries.

It is important to record everything you do at the time that you do it, as recollections are always poor at a later stage. For example, when errors are found during checking and changes are made to the master copy of the data, a note should be made in the audit trail. Keep notes also on the analyses that you do (including the preliminary ones done for checking purposes), writing down the names of all files created. Every entry in the log-book should be dated and initialled.

There is really nothing new here – we are simply re-stating a fundamental requirement of the scientific method, namely that you should ensure that your data management work is repeatable, by keeping good records of what you do.

8. Organising the data for analysis

We have recommended that the data be entered in their raw form. Hence the first step in the data organisation or management stage often involves calculations to re-structure the data into the appropriate form for analysis. This can either be performed in the software used for the data entry, or in the statistics package that will be used for the analysis. We recommend:

- A record must be kept of all changes to the data. This record becomes part of the database, and is kept in the audit trail. Many packages allow data to be transformed and re-arranged visually, but still generate a corresponding file that records the transformations.
- There should be a single “master copy” of the data. This is a standard principle of data management, to preserve data integrity.

The master copy will increase in size as data accrues. Even after basic data entry is complete, errors will be detected, and should of course be corrected in the master copy. It is therefore something which changes through the course of data entry, data management and analysis. Not only should this process be documented, but a consistent "version-numbering" system should be evolved and utilised by all data analysts and other users.

In our view the "master copy" should usually be stored using a DBMS. Only some of the data tables will be altered by data changes. For example, Mr A. the anthropologist may not immediately be concerned with changes to experimental records made by Ms B the biologist, but should be up-to-date with additions to the site list agreed by Dr C the chief. Keeping track of, and communicating changes to, the master copy of the data should be a project management activity like budgetary management.

Usually analyses and reports will be based on extracts from the master copy of the data. When final outputs for presentation or publication are being produced, it is important these are correct, consistent and complete in that they are all based on the final version of the master copy of the data. Interim analyses will have been based on interim data, and to avoid confusion and inconsistency, analysis data-sets, file-names and outputs should include a record of the master copy version number from which they were derived.

In the on-line version of this guide we show how problems can arise if multiple copies are kept of the same data in different forms, and also how to avoid them. We also illustrate some of the common transformations that are required before analysis.

9. Analysis

From the data management perspective, the analysis simply takes the raw data and produces summaries. The process can be viewed in two stages. The first is the production of results to enable the research team to understand their data. The second is the preparation of key summaries to be presented to others in reports and seminars. The statistics package used for the analysis should therefore satisfy the requirements for both stages.

- It should include tabular and graphical capabilities to facilitate exploratory investigations of the data. One use of these is to continue the checking process and hence ensure that the summaries presented are meaningful.
- The facilities for analysis should permit the presentation of the results in a form that assists the research team in their interpretation of the data.
- The package should permit the results to be displayed in a way that closely approximates to the tables, graphs and other summaries that will be included in any reports.

We find that most of the current statistics packages have good facilities for exploratory graphics that help the researchers understand their data, but their facilities for presentational graphics do not match those of the specialist graphics packages, at least for ease of use. If this is important in a particular study, the statistics package must be able to manage the summary data in a way that can easily be exported to a graphics package.

10. Backing up

It is essential to develop a system for regular "back-ups" (copies) of your data and command files. Omitting to do so may result in important parts of the research data being lost. Project managers should establish a documented routine for regularly making safe copies of the data, and should insist that all members of the research team follow the routine.

There are several types of hardware you can use for back-ups. The most common are diskettes, tapes and zip disks. Tapes and zip disks have the advantage of higher storage capacity than diskettes. Whatever hardware you use, it is advisable to have at least two sets of back-up media and to take alternate back-ups on each set. It is also important to ensure that the back-up media are stored in a safe environment.

11. Archiving

The data and programs from a research project must be archived in such a way that they are safe and can be accessed by a subsequent user. For an example, see the booklet *Project Data Archiving – Lessons from a Case Study*. The media used for the archive might be diskettes, tapes, or CDs – similar to that used for back-ups.

Although the copying of data to the archive comes at the end of the project, the way the information will be transferred to the archive should be planned from the outset. Careful planning will be helpful throughout the project, because it helps to promote a consistent directory structure and naming convention for computer files, and also encourages the recording of all steps in the project (see Section 7).

The archive is more than a permanent storage place for the files used for the analysis. It must give access to all the information from the experiment or project. During the operational phase of a project, the information about the research is partly in the computer, partly on paper and other media (such as photographs) and partly in the minds of the research team. The archive need not all be computerised, but it must include all the relevant, non-ephemeral information that is in the minds of the research team. Where data cannot be archived electronically, the sources of information should still be recorded in the archive.

In the absence of a proper archiving scheme, the usual outcome is that the researchers leave, carrying with them the only copy of their part of the data, and hoping that the analysis and write-up will be continued later. Eventually the hope dwindles and the datasets become effectively lost to further research. To avoid this outcome, we believe that (i) at least one full copy of the archive should be left locally, and (ii) the final report should detail the structure of the archive and the steps taken to ensure its safekeeping.

– – – – ⚐⚐⚐⚐ – – – –

The Statistical Services Centre is attached to the Department of Applied Statistics at The University of Reading, UK, and undertakes training and consultancy work on a non-profit-making basis for clients outside the University.

These statistical guides were originally written as part of a contract with DFID to give guidance to research and support staff working on DFID Natural Resources projects.

The available titles are listed below.

- *Statistical Guidelines for Natural Resources Projects*
- *On-Farm Trials – Some Biometric Guidelines*
- *Data Management Guidelines for Experimental Projects*
- *Guidelines for Planning Effective Surveys*
- *Project Data Archiving – Lessons from a Case Study*
- *Informative Presentation of Tables, Graphs and Statistics*
- *Concepts Underlying the Design of Experiments*
- *One Animal per Farm?*
- *Disciplined Use of Spreadsheets for Data Entry*
- *The Role of a Database Package for Research Projects*
- *Excel for Statistics: Tips and Warnings*
- *The Statistical Background to ANOVA*
- *Moving on from MSTAT (to Genstat)*
- *Some Basic Ideas of Sampling*
- *Modern Methods of Analysis*
- *Confidence & Significance: Key Concepts of Inferential Statistics*
- *Modern Approaches to the Analysis of Experimental Data*
- *Approaches to the Analysis of Survey Data*
- *Mixed Models and Multilevel Data Structures in Agriculture*

The guides are available in both printed and computer-readable form. For copies or for further information about the SSC, please use the contact details given below.

Statistical Services Centre, The University of Reading
P.O. Box 240, Reading, RG6 6FN United Kingdom

tel: SSC Administration	**+44 118 931 8025**
fax:	**+44 118 975 3169**
e-mail:	**statistics@reading.ac.uk**
web:	**http://www.reading.ac.uk/ssc/**