DATA ENTRY

John M. Briggs
Kansas State University, Division of Biology/Ackert Hall, Manhattan, KS 665026-4901

Barbara J. Benson
University of Wisconsin-Madison, Center for Limnology, 680 N. Park Street,
Madison, WI 53706

Mike Hartman
University of Colorado, INSTAAR, Campus Box 450, Boulder, CO 80309-0450

Rick Ingersoll
Cornell University, Biometrics Unit, 441 Warren Hall, Ithaca, NY 14853

Abstract. This chapter summarizes the conversion of field- or laboratory-collected data into an electronic form. Techniques to make this conversion as quick and error-free as possible are illustrated, including descriptions of data entry software. Finally, the use of field data recorders and newer technological advances such as optical character recognition for data entry are discussed.

INTRODUCTION

One of the many issues that an information management specialist must consider is the need to convert data into a useable electronic format. This frequently means converting data collected in the field, usually on paper, into an electronic form that can then be used in a statistical or graphical package by the researcher. The purpose of this paper is to present guidelines that we have found useful in making this conversion as quick and error-free as possible.

PLANNING

The first and most critical aspect of data entry is planning. If at all possible, the information manager should be involved in the development of data entry strategies. Much like a statistician is consulted prior to a proper experimental design, an information manager can assist scientists in designing the means of data capture (recorded on paper in the field or laboratory). Guidelines for designing the forms on which data are recorded include:

- Field and laboratory data collection forms should facilitate data collection.
- On-screen forms should facilitate data entry in the computer.
- Field and laboratory forms should be as similar as possible to on-screen forms.
- Include fields for initials of data collection personnel and date.
- Allow room for qualifying comments and metadata.

On-screen data entry forms should incorporate the following features:

- Forms should be easy-to-read and arranged to facilitate data entry.

- Use color only to improve readability. Use of color can be effective, but should not be overused. Excessive use of color can generate eyestrain. In addition, color schemes may not "translate" if other systems don't support the same palette of color choices.
- Use automatic duplication, manual duplication, default values, and other keystroke-saving methods to speed entry and reduce tedious aspects of data entry. Data entry is a tedious job; anything you can do to speed it up and to reduce keystrokes is useful.
- Use quality-control features such as range checks, internal and external table lookup, as well as re-key verification. If properly established and implemented at the time of entry, these steps can greatly reduce the number of errors introduced into the data.

SOFTWARE TOOLS

Historically, data entry tools associated with mainframe computers were limited to data punch or Teletype machines. However, numerous options were introduced with the advent of personal computers. The tools we describe by no means comprise a comprehensive list, but have been found to be useful.

Spreadsheets are probably the most common software tools used to enter data. Their generic interface of rows and columns is familiar to most scientists and, with a little modification, can be quite powerful. For example, the North Temperate Lake LTER site uses EXCEL™ spreadsheets which have been customized to make date entry easier and incorporate error checks. To protect the data entry template from modification, the menu bar has been simplified to permit only a limited set of spreadsheet operations. *In addition, cells that define the form (and should not be changed) are locked*. There is a considerable amount of error checking built into the data entry sheet through formulas and look-up tables. After data entry, the data entry staff scrolls down the spreadsheet to an area that shows a duplicate of the entry area but with errors marked. For example, categories of error checks for the North Temperate Lake LTER fish data include: range check, checks of spelling of character-valued parameters, and comparison of length and weight.

For relatively small data sets, or for sites that can afford to hire people to do double entry of data sets and/or for research sites that use SAS™ for their data analysis, there is an excellent application called SAS™ DUALDATA. It is available at *www.npsc.nbs.gov/resource/tools/software/dualdata/dualdata.htm*. If you are going to use SAS™ as your analysis tool, this application will automatically create a SAS™ dataset for you. It uses double-entry techniques (i.e., you enter the data twice) for validating data entry. The user defines the variables, enters the values comprising the data set, and then reenters to validate. During validation, each value entered is compared against the corresponding value from the initial data entry. If discrepancies occur, the field is flagged and the user is prompted to enter the correct value. The application keeps track of validated observations so that the validation process can span multiple data entry sessions. The application also allows for the possibility that observations may be omitted or duplicated.

A powerful commercial package designed solely for data entry is EasyEntry™ (P.O. Box 2464, Chapel Hill, NC 27515-2464; Phone 919-933-3113; Fax 919-968-1350; Toll free 1-800-532-7573; Email: info@easyentry.com; Web: *http://www.easyentry.com*).
EasyEntry™ is easy to learn, thereby reducing training time and allowing for rapid data entry (minimizing of keystrokes). Numerous quality control features are included:

- Full screen design and modification
- Data field specifications
- Field validation
- Entry and modification

- Keypunch emulation

Under the data field validation, this package allows for:

- range tests
- validity checks
- internal and file table lookup
- selective and full re-key verification
- error messages-standard as well as user-defined messages

EasyEntry™ interfaces with SAS™, Oracle™, Rdb™, Informix™, and other software packages, thus allowing the data to be ported into almost any data analysis package. EasyEntry™ operates on a variety of platforms including: AS/400™, Unix™ [IBM™, DEC™, HP™, SUN™, SGI™], Windows™ (XVT), OS/2™, MS-Windows™, X-Windows™ and Mac™. Thus, it is truly hardware-independent. Future plans for EasyEntry™ include interfaces to new data input devices such as scanners, optical character recognition (OCR), and barcode devices.

Another option is for users to write custom programs. These can range from customizing spreadsheets (as described above) to computer-language specific (or proprietary) data entry. Custom programs are particularly well suited to long-term research for which data collection and entry protocols undergo little change. Nonetheless, the "hidden costs" associated with development and maintenance of custom data entry programs should not be overlooked. For example, evolution of modifications to data collection, programmer turnover and inadequate documentation, as well as the rapid evolution of computer technology can result in high maintenance costs. For these reasons, Konza Prairie LTER is relying less on custom data entry approaches than they have in the past and more on commercially available data entry packages.

Field data recorders, commonly used to make meteorological and hydrological measurements often capture data electronically and do not have to be manually entered. These are very common and useful tools. However, the fact that data are collected electronically does not imply that those data are accurate. Where appropriate, tools used to ensure reliability of manually entered data, such as field range checks, should be employed for electronic data collection as well. In addition, electronically collected data are excellent candidates for many of the approaches Edwards (this volume) has advocated.

FUTURE DIRECTIONS

Future technological advances such as optical character recognition (OCR), voice recognition, electronic "notebooks" and electronic writing devices will reduce the need for manual data entry. The decrease in hardware size and cost associated with an increase in computational power should augment this trend. However, as long as ecologists must collect their data under "field" conditions, there will always be a need for at least some manual data entry.