COMPUTER DATA ENTRY TECHNIQUES
USED IN SCIENTIFIC APPLICATIONS

K.C. Zinnel and M.F. Marozas

ABSTRACT

Prolific data production by the Long-Term Ecological Research (LTER) program has demonstrated that the critical path for analysis and interpretation of biological phenomena involves the initial step of converting observers' measurements of the environment to a machine-readable format. A historical perspective describes the evolution of current data entry procedures used by LTER sites and emphasizes where new techniques allowed the assimilation of additional LTER data sets. Recent electronic technology has made it cost effective to minimize, and potentially eliminate, human involvement in data entry procedures. Critical factors to be considered when making decisions regarding technological advances in data entry are discussed within the framework of a scientific research organization.

INTRODUCTION

Research data management (RDM) encompasses a broad continuum of functions, which are outlined in Figure 1. Scientific research organizations involved in ecological research tend to become oriented toward long-term projects with multiple investigators from different disciplines (Culliton, 1984). This approach produces a need for a "database" system to ensure accessibility of project information. One reason why some large studies have difficulty producing breakthroughs may result from keeping data on paper or in filing cabinets rather than in a RDM system. The role of data entry within RDM is important because measurements made during the course of an experiment must be encoded into a
Fig. 1. Capsule of research data management.

computer-compatible media before analysis can be performed. Analyses are necessary to answer the questions posed by the research design, thus producing publications on which applications for future funding are judged (Kennedy, 1985).

Scientists must be aware of the costs and benefits associated with data entry technology. Data entry encompasses several discrete steps, which are generally employed by most scientists (Fig. 2) and which are essentially the same steps used by corporate data-processing departments for
control over information input (Cerullo, 1983). Data entry may consume more resources than any other RDM step. Documentation may take almost as much time and analysis can cost more, but both areas are dependent upon successful data entry as an initial step. Among the 11 LTER sites, the actual mechanisms of data entry procedures have converged over time.

Fig. 2. Steps within the data entry process.

Data entry cannot begin until the experimental design phase has been completed. Time spent on designing appropriate field data recording forms will assure consistent and compatible measurements of biological phenomena. According to a pre-established and documented experimental design, data are then recorded either by human observer or by instrumentation. After encoding the data into a machine-readable format, the raw data are tabulated and statistically summarized. Data must then be verified to assure correct transcription from source form to machine-readable file. Finally, each corrected data set must be associated with documentation concerning origin and purpose before entry into the project database is considered complete. More than one set of documentation can exist, as projects often change emphasis between inception and completion. Tools for capturing documentation during data entry, which can be automated, greatly enhance future usability of data (Kellogg Workshop Report, 1983).

EVOLUTION OF DATA ENTRY TECHNIQUES

Precise definition of the steps for data entry first requires the choice of the machine that will be used to
physically encode the data. Four phases of historical
development can be recognized from the LTER group experi-
ence. Prior to 1978, most research projects keypunched 80-
column cards from coding forms and submitted decks in a
batch computer environment. Keypunch facilities were expen-
sive to maintain and operate, and the turn around time was
slow. For large data sets, researchers had to contend with
a backlog of data and had no access to any part of the data
while it was being entered. Another feature of large data
sets was that errors introduced during the transcription
process stood a greater chance of not being detected.

Because a high error rate was characteristic of the
process, data tabulation was performed prior to the encoding
process. Due to the high cost of double entry, verification
was accomplished by visually checking printed listings of
card decks against the raw data coding forms. Data were
corrected by repunching cards, and summary programs were run
using corrected data providing tabulations for comparison
with pre-entry calculations to further reduce errors. Ulti-
mate data bank disposition was generally a metal card cata-
log file of card decks with limited shelf life.

By the beginning of the LTER program in 1981, data
entry at most sites had progressed to an interactive editing
program and access to a mainframe computer through a time-
sharing terminal for entering and correcting data (Table
1). Again, a printed listing of data was produced and visu-
ally verified against raw data coding forms. Computer cen-
ters subsidized this type of data entry by promoting inter-
active usage through low costs for connect time and perma-
nent file storage. With many users sharing the cost, com-
puter centers could afford to buy statistical packages and
other software supported by vendors, thus sparing users the
cost of developing their own software to tabulate data and
flag outliers. A library of data files on magnetic tape,
sometimes with documentation, was the usual data bank.

In many cases, methods of data entry employed prior to
the start of LTER proved inadequate for handling the flood
of information generated by LTER scientists. As university
computer facilities increased their rates, a need for more
efficient forms of data entry was apparent; resulting in the
development of a distributed approach using off-line micro-
computers for entry and verification of data (Table 2).
Development of data entry procedures using microcomputer
technology has been adopted by most LTER projects (Fig.
3). The use of microcomputers helped alleviate raw data
backlogs, reduced computer usage charges, and increased
error detection through full-screen data entry software featuring validity checking. Scientific research organizations, which employ multiple investigators, technicians, and students working toward common goals, also began to employ communications packages in much the same manner as corporate

Table 1. Data entry techniques utilized by LTER sites when the program was initiated.

<table>
<thead>
<tr>
<th>Site</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDREWS</td>
<td>Keypunch main encoding device, corrections made with interactive editor on mainframe, raw data stored on magnetic tape in mainframe library.</td>
</tr>
<tr>
<td>CEDAR CREEK</td>
<td>Microcomputer and commercially available software used for data entry and correction. Raw data archived on floppy disks and mainframe mass storage after upload.</td>
</tr>
<tr>
<td>COWETTA</td>
<td>Keypunch used to encode data, raw data maintained as card decks and archived on IBM 370 magnetic tapes.</td>
</tr>
<tr>
<td>JORNADA</td>
<td>Interactive entry/editing of data with mainframe system, raw data maintained on mainframe mass storage and archived to tape. Meteorological data logger read directly to mainframe disk file.</td>
</tr>
<tr>
<td>KONZA</td>
<td>Microcomputers and in-house generated, user friendly software used to enter/correct raw data. Meteorological data logger entered as microcomputer disk file.</td>
</tr>
<tr>
<td>LARGE RIVERS</td>
<td>Interactive entry/editing or keypunch of raw data.</td>
</tr>
<tr>
<td>NIMOT RIDGE</td>
<td>Keypunch, interactive terminal used for data entry/editing. Raw data archived as published reports and/or mass storage files on mainframe. Meteorological data logged onto cassette tape.</td>
</tr>
<tr>
<td>NORTHERN LAKES</td>
<td>Microcomputers and commercially available spread sheet software used to enter/edit data. Raw data uploaded and reformatted for archival as SIR mass storage files on mainframe.</td>
</tr>
<tr>
<td>NORTH INLET</td>
<td>Interactive software for full-screen management on mainframe system used for data entry/correction. Raw data archived as mass storage files on mainframe.</td>
</tr>
<tr>
<td>OKEFENOKEE</td>
<td>Keypunch used to encode data, raw data maintained as card decks, archived to IBM 370 magnetic tapes.</td>
</tr>
<tr>
<td>PAWNEE</td>
<td>Keypunches used for data encoding. Meteorological data logging on cassettes. Raw data archived on mainframe as magnetic tape files.</td>
</tr>
</tbody>
</table>
Table 2. Procedures evolved during the course of the LTER program.

<table>
<thead>
<tr>
<th>Location</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDREWS</td>
<td>Microcomputers with commercially available software used for data entry, correction, and report generation. Documentation on microcomputer mass storage as well as archived to mainframe magnetic tape along with data.</td>
</tr>
<tr>
<td>CEDAR CREEK</td>
<td>Microcomputers used for automatic data acquisition in laboratory chemical analyses, weighing plant samples, and animal location by telemetry.</td>
</tr>
<tr>
<td>COWETTA</td>
<td>Individual PIs enter data with microcomputers and are responsible for maintenance. Database administrator uses microcomputer to create and maintain data set documentation.</td>
</tr>
<tr>
<td>JORNADA</td>
<td>Microcomputers used for data entry/editing. Raw data - maintained on microcomputer mass storage, but archived with mainframe magnetic tape system. Minicomputer workstation also used for data entry/editing/access.</td>
</tr>
<tr>
<td>KONZA</td>
<td>Incorporated validity checking into in-house data entry software and established procedures for sending data files to mainframe.</td>
</tr>
<tr>
<td>LARGE RIVERS</td>
<td>Microcomputers with commercially available data entry and management software networked to database on in-house minicomputer.</td>
</tr>
<tr>
<td>NIWOT RIDGE</td>
<td>Microcomputers with commercially available software used for data entry, correction, and report generation. Multiple levels of data maintenance with data not necessarily in machine-readable form.</td>
</tr>
<tr>
<td>NORTHERN LAKES</td>
<td>Increased number of microcomputers and communication capabilities while maintaining procedural stability.</td>
</tr>
<tr>
<td>NORTH INLET</td>
<td>Meteorological data station and laboratory microcomputer used for automatic data acquisition. Microcomputers used for data entry and editing of data sets to supplement full-screen management software.</td>
</tr>
<tr>
<td>ONEFENOEKEE</td>
<td>Multi-user microcomputer system used for data entry, correction, archiving, and documentation of raw data.</td>
</tr>
<tr>
<td>PAWNEE</td>
<td>In-house minicomputer, interactive entry/editing of raw data.</td>
</tr>
</tbody>
</table>

executives. Using microcomputers as replacements for remote job entry terminals or data entry stations, files of database additions were uploaded to a host mainframe and reports and/or summary statistics were downloaded to personal computers for display (Ferris, 1983).
A fourth phase in the historical development of data entry procedures has been the use of machines to take measurements and encode data directly into a machine-readable media. For experimentation requiring continuous monitoring, sampling for long periods of time, or high frequency sampling intervals, an automated data acquisition system is cost effective. Automation of measurement processes has lead to vast increases in the amount of data generated in every scientific discipline (Lide, 1981). Because instrumentation generates large volumes of data, only an automated data entry system can minimize the lag time between raw data collection and machine-readable data encoding. For these reasons, a majority of LTER sites now use data acquisition systems for meterological monitoring and/or chemical analyses (Fig. 3).

![Data Entry Techniques Utilized](image)

Fig. 3. Evolution of data entry techniques employed within LTER.

ANALYSIS OF PRESENT DATA ENTRY TECHNIQUES

Selection of a microcomputer system for data entry involves weighing factors such as availability and cost of hardware and software, plus the time required to convert from an old procedure to a new system. The LTER sites selected different hardware, partially according to budgetary constraints. Most software packages selected, such as the STAR family by MicroPro, included characteristics to reduce transcription errors. Among these is full-screen data entry

245
software, which allows designing a "form" that appears on the screen as an imitation raw data sheet. With cursor movement, the data entry technician is guided to fill in the blanks with the values from the data sheets. These packages also allow data formatting controls and data validity checking which informs the operator that the data value does not meet specified criteria.

Other features include the ability to integrate repetitive information from other databases and generate reports and summaries. Raw data files can be listed for visual verification or another data entry technician can access each record and verify data entered. Previously, an expensive form of verification involved entering data twice, into two different files by two different people. Verifying the two files against each other with a "compare" program was an excellent, but expensive, way to locate transcription errors. Because of increases in speed of entry when utilizing full-screen software and off-line microcomputers, it currently is cost effective for sites to perform this type of double entry verification, which has the highest rate of error detection. A communications package capable of data file transmission between the microcomputer and a mainframe computer system is also needed for retrieval and archiving.

Data acquired by automatic instrumentation is cheaper on a per sample basis than data recorded by a human observer. If operating procedures are correctly followed, acquired data will be virtually error-free (Enke, 1982). Human involvement is required to ensure that sensors, signal conditioning, and recording equipment are performing accurately. Because automatic instrumentation is capable of generating far more data than can be manually checked by a human observer, elaborate software must be used to ensure data integrity. Currently, sophisticated validation software, developed in-house, is an essential requirement for any project utilizing data acquisition. Periodic checking of formats and data field ranges, plus testing for data accuracy, must be required as part of normal operating procedures. Final assessment of data integrity can be performed by an experienced technician through inspection of graphical data summaries.

Verification programs for data acquired by instrumentation should check for error conditions and extract usable data. Control totals such as record counts and average values per batch are used to evaluate machine performance during operation. Adequate operator training and
supervision are essential to producing quality data. A specific technician should be assigned responsibility for determining if errors have occurred during operation; logging the errors, and reentering corrections. Recommendations have been made to record and preserve all auxiliary information (such as calibration standards or ambient temperature and pressure when the analyses were run) in order to use the data with confidence (Lide, 1981). Future possibilities for data entry techniques include the use of optical character recognition (OCR) for entering field notes directly and the use of voice recognition system for field data recording.

DISCUSSION

Tables 1-3 illustrate an evolving trend for LTER projects in the use of technological advancement, such as microcomputers, to meet the challenge of entering an increasing amount of source measurements into RDM systems. While the costs of computing hardware has continued to plunge, computing power and ease of interfacing different devices have increased (Enke, 1982). Although it is technically easy to put together hardware to perform complex operations, the software required for varying applications, or combinations of complex operations, is not as easily implemented. Testing is an especially critical requirement as each application is inherently different. There always will be modifications before a system becomes fully operational. As expectations for software capabilities have increased and hardware costs declined, software has rapidly become the major consideration and limiting factor in applying microcomputer technology to scientific research needs (Enke, 1982). Most LTER sites use commercially available software for data entry as well as word processing, but in-house programs for data reformatting and archival are common.

Within the last two years, direct application of microcomputer techniques for control of scientific instrumentation has promoted increased use of data acquisition involving one or more sensors without direct human involvement. Initially, research organizations desiring this capability had to invest in expensive laboratory minicomputer systems with expensive support software such as LIMS/2000 or RS/1 (Robinson, 1983). Within a year, half a dozen firms were offering controllers and data stations that would allow scientists to interface APPLEs and IBM PCs to numerous instruments (Robinson, 1984). Despite the great power of
microcomputers to capture immense amounts of data, it is important to remember three caveats: (1) a machine is not a perfect observer, it is only as good as its electronic calibration, which is performed by humans; (2) larger amounts of data acquired by instruments take longer to analyze (software and human filters can reduce data to manageable levels); and (3) beware the salesman propaganda.

Table 3. Decision basis for changes in data entry procedures and other pertinent factors.

<table>
<thead>
<tr>
<th>Location</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDREWS</td>
<td>Using micros for data entry more cost effective than maintaining keypunch machines and personnel. Data to be entered at centralized facility growing steadily at 10% per year. Faster turnaround time with micros.</td>
</tr>
<tr>
<td>CEDAR CREEK</td>
<td>Cost is primary factor. Initial hardware investment less than maintaining technicians' salaries. Number of data items entered increased 170% from Year 1 to Year 2, and 250% from Year 1 to Year 3.</td>
</tr>
<tr>
<td>COWEETA</td>
<td>Data administrator maintains data documentation and data directory, but does not supervise data entry or maintain data library.</td>
</tr>
<tr>
<td>JORNADA</td>
<td>Use of micros and workstation facilitates data screening and error detection. Reduced use of mainframe computing increases flexibility and project control of computing resources.</td>
</tr>
<tr>
<td>KONZA</td>
<td>A reassessment of project needs necessitated redesign of data entry software and validity checking. File reformatting needed because of procedural changes.</td>
</tr>
<tr>
<td>LARGE RIVERS</td>
<td>Specialized application package for geographical data required operating system stability not possible when using university mainframe system. Easier to predict and budget for computer costs of in-house system.</td>
</tr>
<tr>
<td>NIWOT RIDGE</td>
<td>Data dictionary maintained on microcomputer used to document location and status of data sets. All raw data published as CULTER data reports. No funds available for data entry and maintenance of machine-readable data library.</td>
</tr>
<tr>
<td>NORTHERN LAKES</td>
<td>Procedural stability most important factor. Data entry with micros adequate for task. Inefficient use of resources and time acquiring unnecessary technology.</td>
</tr>
<tr>
<td>PAVANEE</td>
<td>Backlog of preexisting data from International Biological Program. University mainframe costs out-of-control. Difficult to obtain funds for data management of pre-LTER data. In-house minicomputer stabilizes costs.</td>
</tr>
</tbody>
</table>
there always will be hidden costs. Troubleshooting problems, such as faulty grounds, are an inescapable part of normal maintenance. System performance criteria such as accuracy or linearity will be directly affected by the extremes of the operating environment (Chase, 1982). Unless care is taken when monitoring data acquisition system operation, there is the risk of producing large data sets (in nice neat formats) that are highly unreliable. This point concerning expert supervision of data acquisition systems is especially important in government-supported projects, which tend to be under-funded with regard to support (staff) personnel. Before any research project opts to use the latest technology, the question must be asked whether or not the benefits of using instrumentation to carry out experiments offset the costs of proper maintenance.

Intelligent tools require intelligent operators to supervise their operations effectively. Using microcomputers for data entry and data acquisition means that programming support is also required. Because in-house software support is affected by personnel changes, commercially available software packages are preferred by most sites. Gains in stability may be offset by losses in flexibility. Commercial software procedures tend to be less efficient than software custom designed for a specific application. Data managers involved with microcomputers also find that they have additional duties, including responsibility for maintenance and supplies, and consulting on the usage of various microcomputers and software packages.

CONCLUSION

Scientific and technical data are deemed necessary for the solution of societal problems such as energy supply, environmental quality, and industrial productivity (Lide, 1981). Since the late 1970's, ecologists have become increasingly aware that certain questions could be answered only by long-term data sets. Electronic tools have made mathematical analysis and management of such data feasible. One of the most labor-intensive R&D aspects resulting from this information explosion is the development of data entry procedures. A key consideration is the time lag between turning in data sheets and receiving access to a verified, corrected, and documented data file. Obviously, publications cannot be produced unless data are accessible for analysis or merging with another data set.
Many large-scale research projects, such as LTER, have found the critical path between measuring biological phenomena and interpreting experimental results to be the initial RDM step of data entry. In most cases, methods of data entry employed prior to LTER proved inadequate for handling the flood of information once the research was underway. As part of their strategy, many large research projects, including LTER, are automating as much of the data entry procedures as resources permit (Table 3). Reorganization of personnel and redefinition of procedures are usually necessary since data entry issues change as research data management in the ecological sciences evolves. Regardless of which combination of data entry techniques are chosen, any method will prove successful only if project leaders provide support for adequate software and hardware, as well as supervisory personnel.

LITERATURE CITED