MICROCOMPUTER STORAGE AND RETRIEVAL OF TREE IMPROVEMENT RECORDS

by Kim C. Steiner, associate professor,
James J. Zaczek, research assistant, and
Henry D. Gerhold, professor,
School of Forest Resources,
The Pennsylvania State University,
University Park, PA 16802

ABSTRACT.--This paper examines the potential for using microcomputers and proprietary software to manage information in tree improvement programs. Important criteria for selecting software are described, and Penn State's system is presented as an example of how performance data and other information can be logically organized to facilitate storage and retrieval.

Record-keeping is one of the most difficult and frequently neglected of tasks in tree improvement. Even with good intentions and organized habits, it is surprising how rapidly information builds to the point that data may become annoyingly hard to locate. This is especially true if the data are on paper, because the arrangement of file drawers cannot accommodate the complex relationships that develop between projects, experiments, breeding activities, and data. Of course, simply having the information "on the computer" doesn't guarantee its ready access either; a program is needed to handle the housekeeping chores of data management.

The availability of data management software for microcomputers has opened new opportunities in information storage and retrieval. Although this software has been developed primarily for the business market, most of it has been designed and written by profession-

[1] Journal Article No. 6999 of the Pennsylvania Agricultural Experiment Station. Partial support of U.S. Department of Agriculture Regional Research Project NE-27 and the Pennsylvania Christmas Tree Growers' Association is acknowledged.

als using sound principles of information management. These principles have broad applicability, and some of the programs available now are perfectly suitable for the sizes and kinds of data sets common to tree improvement programs.

Although more powerful software packages are available for mainframe computers, microcomputer-based systems offer advantages in cost, accessibility, ease of use, and portability. Processing speed seems to be the only real limitation of the best micros compared to mainframes. However, with good software, micros can handle most information retrieval tasks within a few seconds or minutes, which is fast enough for the occasional use that is normal in a tree improvement program.

The record-keeping system we are developing at Penn State is designed around the program K-MAN ("Knowledge Manager", by Micro Data Base Systems, Lafayette, Indiana) running on an IBM-PC/XT. The program can be purchased from mail-order supply houses for about $300, and it runs on a machine that retails at about $3500 for the minimum necessary configuration (two disk drives and 192 K of memory). In choosing among the available programs, we considered the following criteria to be important and most likely to cause limitations in applicability to tree improvement records:

1. Maximum number of records per file (e.g., accessions per species or trees per plantation).

2. Maximum number of fields per record (e.g., information categories per accession or measured traits per tree or other experimental unit).

3. Maximum number of characters per field and per record (especially critical for text as opposed to numeric data).

4. Maximum number of files accessible simultaneously (e.g., ability to extract information from accession files and one or more plantation files at the same time).

5. Ease of adding, renaming, deleting, and changing fields and records.

6. Ability to export and import data in different coding formats for use in other programs and computers (e.g., for statistical analysis).

7. Miscellaneous editing, indexing, sorting, retrieval, and programming capabilities.

8. Protection against unauthorized changes to the data (password capability).

The K-MAN program is more than adequate in all these respects. Other programs for the IBM-PC that appear comparable in price and capabilities are "Savvy PC", "DataFlex", and "dBase III."

The advantage of information storage on the computer lies in the ability to retrieve data rapidly in whatever combination or format is suited to the task at hand. In order to fully exploit this capability of data management programs, one must carefully consider the relationships among data elements and data sets. Data irregularities that are normally handled with explanatory notes in paper filing systems must be eliminated or minimized. Thoughtful consideration should be given to how the data will ultimately be used. The discipline required to manage data with a computer program tends to result in cleaner, clearer data sets.

We identified four basic kinds of information that must be dealt with in our tree improvement programs, and they are probably common to most others:

1) accession information
2) breeding records
3) plantation maps
4) performance data

The logical relationships of information within and among these categories, and how that information will be used, determines how files, records, and fields must be defined.

Different solutions are possible, but we have organized our system around two principal file categories: accessions (a separate file for each species) and plantations (a separate file for each plantation). The primary, or unique, field for accession files is the accession number, and for plantation files it is the tree number (ordered serially in whatever pattern

is used in measuring the plantation). One of the fields for every record in a plantation file is the accession number associated with the tree. Since K-MAN allows access to many files simultaneously, it is possible to retrieve performance data by accession number from several plantation files and, at the same time, appropriate accession information from the accession file (e.g., provenance latitude and longitude of origin, progeny pedigree, alias accession numbers, etc.).

The minimum necessary accession information is of course the accession number, a label used simply for maintaining the identity of a genetic entity. In the process of organizing our records, we have rediscovered the need for absolute consistency and clarity of purpose in assigning accession numbers. The rule is this: Each genetic entity used for experimental or breeding purposes must have a unFie accession number. It may also 'nave aii ¯aTias, but at least one number should be unique to that entity. With three or more people creating accessions at Penn State, we avoid duplication by incorporating the researcher's initials as part of the number (see Table 1).

We have found, not surprisingly, that even minor departures from this rule can lead to later confusion. A tree in an experimental plantation may be considered a statistical sample of the population or family of which it is a member, and share that population's or family's accession number. But if that tree is selected for clonal testing, or used in breeding, it must be assigned a new number. Pollen or seed must be considered genetically distinct from the tree that bore it, and, for some purposes, distinct from collections made in other years. Roguing a population or family produces a genetically different group of individuals which, if used in further testing or breeding, must have a unique number.

In addition to labelling a genetic entity, accession records should also describe it well enough to enable its reconstruction or at least permit its genetic history to be traced. Our greatest difficulty in adapting accession records to the system was devising a record format that worked for a variety of accession types (open-pollinated provenance collections, clonal material, pollen, control-pollinated seed, etc.) with maximum opportunity for information recovery and minimum wasted disk space. The solution to

this could be different for each tree improvement program, but Table 1 shows a record format that is usable for a variety of accession types. This format is wasteful of space if such flexibility is not needed, but still it permits over 2000 accession records to be held on a single floppy disk.

Breeding records can be incorporated into accession files if each tree or pollen mix used in breeding is given an accession number. The origin of control-pollinated progenies can then be completely described with "Male Parent" and "Female Parent" fields in the accession file (Table 1), which may be left blank for other kinds of accessions. Query procedures in K-MAN and similar programs can be used to trace the pedigrees of accessions over several generations, if available, or to obtain the accession numbers of all progenies that share a common parent.

Thus, our accession files contain the first two kinds of info/mation, accession information and breeding records. Our plantation files contain the other two: plantation maps and performance data. Mapping is accomplished by including fields for accession number, row, column, and replicate designations in the record for each tree (Table 2). Although we use separate, typed maps in practice, maps could be reconstructed from this information if necessary. If a single tree is later given its own accession number for breeding purposes, entering plantation number/row/column (or plantation number/tree number) as part of the source information in the accession record will enable us to quickly recover other information about that tree from the plantation file.

Fields for tree identification and location compose the basic information in the plantation file. Additional fields for performance data can be added as characteristics are measured (Table 2). (This is where limitations on number of allowable fields and ability to add fields after initial record definition may be critical with some data management programs.) Thus, the plantation file becomes a single, structured repository for all the information about the plantation, with the exception of narrative descriptions of its location or establishment history. One can also include in the plantation file such information as cultural treatments (e.g., pruning or shearing) or soil characteristics if it is deemed necessary to keep such information on individual trees.

The ultimate advantage of using K-MAN and similar programs lies not in the capability of retrieving information quickly but in combining and summarizing data in different ways for different purposes. K-MAN commands enable this to be done with little effort and usually in a straightforward manner; and the data can then be printed in report format, exported to another program or computer for analysis, or analyzed (rather slowly) using K-MAN's own, elementary programming language. For illustration, Table 3 shows some of the K-MAN commands and their possible syntaxes.

Data management programs do not entirely eliminate the need for written records. However, they can do a superb job of handling the bulk of the information generated in tree improvement programs, and in fact the existence and location of additional information can itself be coded as a field variable. Until recently, the capabilities that data management programs offer were available only to those with the time and programming expertise to develop their own systems. Adapting this program to our needs has been relatively simple, and we anticipate many benefits.

Table 2. -- Sample record format for a plantation file.

Field	Number of Characters	Description and Comments
TREENUM	3	Primary field. Ordered serially in pattern used to measure plantation.
PLOTNUM	3	Ordered serially in pattern used to measure plantation. Used to facilitate extraction of plot means. May be used as primary field instead of TREENUM if data are to be kept on a plot basis only.
ROWW	2	Letters or numbers used to designate plantation row on map. "Row" is a reserved word in K-MAN.
COLUM	2	Letters or numbers used to designate plantation column on map.
BLOCK	1	Number of block or replicate.
ACCESSNO	7	Accession number of genetic entity to which the tree belongs.
COMMENTS	25	Miscellaneous comments not pertaining to codable performance characteristics or cultural treatments.
HEIGHT84	3	Unit of measure and date of measurement can be automatically added to value in all records with no cost of storage space.
SURV84	1	"1" if alive and "0" if dead. Date of measurement can be automatically added to value in all records with no cost of storage space.

Table 1. -- Sample record format for an accession file.

Field	Number of Characters	Description and Comments
ACCESSNO	7	Primary field. Year-Collector-Number (e.g., 84KS059)
ALIAS1	8	Optional, variable format
ALIAS2	8	Optional, variable format
SPECIES	0	"Logical" field (virtually no disk storage space required) if all records in a file are for the same species.
CATEGORY	2	Mutually exclusive codes as follows: SC = seed from a stand collection OP = seed from open-pollinated tree CP = seed from control-pollinated tree TR = select tree (for use in breeding) VP = vegetative propagule PO = pollen
NATIVE	1	(Y)es, (N)o
STATE	2	Zip Code abbreviations
COUNTY	20	
NEARTOWN	20	Nearest town or municipality
ELEV	5	Filled in only if NATIVE = Y. "Feet" or "meters" can be automatically added to value in all records with no cost of storage space.
LONG	6	Longitude. Filled in only if NATIVE = Y.
LAT	5	Latitude. Filled in only if NATIVE = Y.
PLANTAT	13	Plantation. Filled in only if NATIVE = N. For mapped plantations, use plantation number and, if necessary, tree row and column designations (for CATEGORY = OP, TR, VP, or PO). Since it is desirable to give accession numbers to all individual trees and pollen mixes used in breeding, this field can normally be left blank when CATEGORY = CP.
FPARENT	7	Filled with ACCESSNO if CATEGORY = OP or CP.
MPARENT	7	Filled with ACCESSNO if CATEGORY = CP.
COMMENTS	40	Miscellaneous comments.

Table 3. -- Selected commands and abbreviated syntax for file and record manipulations using K-MAN.

Command syntax	Comments
DEFINE file	Program responds with prompts for interactive definition of field names, sizes, and types (string, numerical, logical). Allows specification of read and write access codes for each field.
REDEFINE file	Program responds with prompts for interactive changes to field names and sizes, field read and write access codes, and field additions or deletions.
SHOW file	Displays current field names, types, and sizes and current number of records.
BROWSE file FOR conditions	Used to browse through the records in a file, editing data as desired.
CHANGE field IN file TO expression FOR conditions	Changes values of an indicated field that satisfy stated conditions.
CREATE RECORD FOR file	Program responds with prompts for interactively entering values for additional records in the specified file.
ATTACH FROM file1 TO file2	Used to incorporate records from a standard text (ASCII) file to a defined K-MAN file.
SORT file BY direction fields	Creates a new, sorted version of the specified file.
INDEX file2 FOR file1 BY fields	Creates an index file (#2) for file #1. Indexing permits faster retrieval of data by key fields.
OBTAIN position RECORD FROM file	Retrieves records one-at-a-time for display. A PLUCK command can be used instead for indexed files.
SELECT fields FROM files FOR conditions	An extremely flexible command that allows the retrieval of any combination of fields from any number of files.
STAT expressions FROM file FOR conditions	Calculates values of specified arithmetical operations on individual fields and field combinations, as well as summary statistics.
CONVERT fields FROM file1 to file2	Writes field values (or results of arithmetical operations on field values) from file #1 to file #2 in ASCII, DIF, or BASIC-compatible format. For exporting data to other programs.