Centro Internacional de Agricultura Tropical

SEMINARIOS INTERNOS

SE-8-84

September 7, 1984

A DATA MANAGEMENT SYSTEM FOR BREEDERS' MATERIALS AND GERMPLASM COLLECTIONS

N. R. Sackville Hamilton [1]

Attempts to manage large quantities of data on the computer invariably fail if they are based on conventional filing systems. There are two main reasons for the failure, namely redundancy of data and waste of time.

In an integrated data management system such as IDMS, all data are stored in one file, with a logical structure that closely reflects the natural logical structure of the data. The physical structure of a database is like that of a book; as with a book, searching for data in the database to answer a question involves looking up an item of data in the index, turning to the indicated page and them following up cross-references to related data until the question is answered. As with a book, questions are answered most quickly if all the relevant data are stored on one page.

Any breeding program is concerned primarily with the production of new, improved cultivars. A germplasm collection serves as a store of genetic variability for use in breeding programs. In both cases, what the scientists need to know about a line can be reduced to the two questions, 'what is it like?', and 'where is it from?'. Each season, lines are chosen, for use as parents of crosses, or for further selection or evaluation, or for distribution, on the basis of data collected on them in previous seasons.

The problem of chaaracterising lines is a problem of biology more than of database design. The phenotype of a line depends on its genotype, environment and the stage in its ontogeny. On a database, a score for a particular phenotypic can be stored as a three-way junction between the line, the phenotype character and scoring system used, and the time and location of recording. Such scores are of value only to the extent that they tell us how the line might perform in farmers' fields. It would be more useful to determine the genotype, and to combine the database with other databases, on the distribution of pathogen races, on agroecology etc, and hence establish patterns of stability and response to environment.

The origins of lines can be expressed in general as a many-to-many relationship, in which one line may have several parents and several offspring. One line can be derived from another in several different ways, such as by crossing, by selection from a heterogeneous or segregating population, or by donation from another institute.

The object is to develop a flexible package for interrogating the database, with which a user can search for any data for named line or list of lines, or search for a list of lines satisfying any given criteria. By combining the answers to a set of questions, a user could build up a set of lines for inclusion in a new experiment, or produce a report, catalogue or field book containing any data that he specifies for any group of lines.