DEVELOPMENT OF A SAS DATA MANAGEMENT SYSTEM FOR THE GERMPLASM COLLECTION AT THE ASIAN VEGETABLE RESEARCH AND DEVELOPMENT CENTER (AVRDC)

John N. Hubbell, Jr., AVRDC

Research Environment

The principle objective of AVRDC is to increase yield and quality of vegetable crops for the humid tropics and subtropics in order to improve the quality of life for the large number of people at the low end of the income spectra. To reach this objective scientists from different disciplines have combined efforts. Six crops have been selected for improvement -- soybean, *Glycine max*, mungbean, *Vgna radiata*, tomato, *Lycopersicon esculentum*, Chinese cabbage, *Brassica pekinensis*, sweet potato, *Ipomoea batatas*, and white potato, *Solanum tuberosum*. Plant breeders have gathered germplasm from sources throughout the world. Pathologists and entomologists screen the collections for resistance to diseases and insects. Chemists identify accessions that are high in desirable nutrients and low in undesirable compounds. Horticulturists, soil scientists, and economists evaluate accessions within the physical and socio-economic environment.

Such a worldwide interdisciplinary approach is logically appropriate for an international center, particularly for one involved in vegetable crops. National funds for agricultural research are limited, and research involving grain and cash crops receive priority over vegetable research. Collection and maintenance of a large germplasm bank is expensive and need not be duplicated in each country as long as promising material and information relating to this material can be disseminated. Concentrating a critical mass of scientists from separate disciplines to produce improved material and valuable data is generally outside the scope and design of traditional national agricultural institutes.

Managing the inventory of the AVRDC germplasm collections presents problems because of the large number of items or accessions -- from 389 accessions for sweet potato to 9980 for soybean (Table 1). In addition, often an accession is obtained with only a limited amount of information. For some accessions only the country of origin is known. For others certain pedigree information is available. Later data are generated from trials and laboratory analyses at the research center and from trials throughout the world (Fig. 1).

Field and laboratory experiments as well as surveys which produce information (re socio-economic aspects) relating to our six crops produce a substantial quantity of data. Analyses of these data may involve analyses of variance, linear regression, calculation of correlation coefficients, creation of frequency tables, Duncan's multiple range tests, etc.

Requirements of a Data Management System

From this examination of the AVRDC research environment, the requirements for a support system for data management and analyses may be formulated. In early 1975, when I first formulated a system there were some very severe constraints. At AVRDC headquarters in Taiwan, there was no large computer in house. Only a small Munroe 1860 programmable printing calculator was available at AVRDC. If a large computer was to be used, debugging would have to be done via the mail to a large computer 330 km away. The system had to be initiated in the three months before I was to leave for a two-year post in the Philippines, thus leaving one BS-level staff member to assist the scientists in the development of the system.

With these constraints in our environment, whatever system was selected would preferably have these following characteristics: (1) it should be easy to generate and maintain. Adequate documentation and support should be available to facilitate generation and maintenance. (2) it should be user oriented so that the scientists who are not computer experts, can easily understand not only the output but also the input requirements. Also it is desirable that the user be intimately involved in the development of any application. This would be difficult in a system that is not user oriented. (3) it should be modular and flexible. The support staff could initiate the system immediately without a large investment in training. The users could begin by using only a few of the capabilities and later expand their use of the systems capabilities. (4) it should be a developing system, so that AVRDC would not be committed to a system with built-in obsolescence. (5) it should be a system made of components that have a large number of users, thus allowing easy exchange of ideas and allowing AVRDC to benefit from the developments made by other users. (6) it should be a system that would permit the number of components or software package to be kept to a minimum. Each additional package requires generation, maintenance and education of support staff and users.

System Software Selection

Lutz (6) classified software into three groups -- compilers, data management systems, and application packages. He considered SAS as belonging to the last group. However, Browne (3), Strand (9), Amara (2), and others have used SAS successfully for developing data management systems.

The first group, compilers, fails to be adequately user oriented for the AVRDC environment. Using one or more compilers, one staff member could not possibly develop or maintain the necessary programming support for data management and for statistical analyses. The large computer oriented data management system software packages that were available were designed to coordinate large files that are related. There actually is no present need for generally relating tomatoes to soybeans. The packages also require a good deal of man hours for education.

Of the second group, data management systems, in early 1975 none were available to AVRDC that were designed for germplasm collections, adequately documented, adequately supported, and could run on the computers available -- which ranged from the Monroe 1830 through an IBM 1130 to a 360 model 65. Certainly the Information Sciences/ Genetic Resources Program has produced some interesting work. However, as reported by the International Board for Plant Genetic Resources (5) as late as 1976 their computer packages required further documentation and development.

Of the third group, application packages, SAS was my first choice. SAS required establishing access to the IBM 360 at the IBM Data Center in Taipei which was the only computer in Taiwan with OS, a requirement of running SAS.

Rather than maintain more than one package to handle data management and statistical analyses, we chose to use SAS for both applications. Use of a high-level package such as SAS would allow rapid program development under conditions with limited programmer expertise, with high user involvement and with debugging by mail.

System Development

After 6 hours of introductory instruction in SAS in early 1975, the AVRDC scientists were able to evaluate how they could best use SAS. In planning for computerization, the entire system of germplasm evaluation, of which the computer system is only a part, has had its an interdisciplinary review. AVRDC scientists were surveyed to identify (a) which variables (e.g., yield, resistance to downy mildew) were of first or second priority and (b) which discipline is responsible for determining the values of the variables (e.g., pathologists are responsible for entering the value of resistance to downy mildew into the system). Using SAS, we considered each accession or entry of a collection as an observation (to use SAS terminology) with several variables called "descriptors" in gene bank terminology (8). Each variable assumes appropriate data values or "descriptor states". The scientists responded well to the review of the system and found the exercise enlightening.

Initially data management of the larger germplasm collections -- mungbean and soybean -- were computerized. One immediate benefit was improved documentation, an important factor when there is a change in personnel. For example, the new mungbean breeder on arriving at AVRDC in late 1974 found the documentation of the mungbean collection less than satisfactory. The initial mungbean data set had among its variables the accession number and the United States Department of Agriculture plant introduction number (USDA PI) if known. From a list of the data set in USDA PI sequence, when examining old accessions or when adding new accessions, the breeder can identity possible duplicates. The collection had quite a few accessions with the same USDA PI number. There are now 72 mungbean accessions having the same USDA PI number. These duplicates are being grown in the field side by side for confirmation of duplication. It is a useless expense to maintain true duplicates in a germplasm collection (7).

Having the SAS procedures, such as PROC FREQ, facilitates analyses. PROC FREQ can produce a quick picture of the distribution of the mungbean collection by country. A comparison of the distribution of the mungbean collection by country with the distribution of area planted to mungbean by country from data collected at the AVRDC First International Mungbean Symposium (1) indicates that it would be worthwhile to collect more material from Burma, Indonesia, Pakistan, Sri Lanka, etc. (Table 2).

As new procedures are being developed in the SAS package, many of these are useful to the analyses of the collections. For example PROC CHART, newly available in 1978, allows a bar graph to be produced in the same run as a list of the accessions. This is particularly worthwhile when distributing lists to scientists who are interested in selected variables on the data set. This continuing development is available to AVRDC without any investment in program development and only a minimum investment in keeping the programming support current.

In August of 1978, as part of an IBM/AVRDC partnership program, an IBM 2780 terminal -- a remote job entry terminal -- was installed at AVRDC and linked to the IBM/370 Model 158 of 1.5 megabytes of core storage located at the IBM Data Center in Taipei 330 km away. This terminal and in-house keypunch facilities permit greatly reduced turnaround time for analyses of data and for further development of the germplasm evaluation systems. As the result of increased use of the support services, the support staff has been increased. In the spring of 1979, an IBM 3270 terminal with display station and printer will be added to the configuration.

SAS has served as a basic tool for development of computer support services at rapid pace with a low initial investment under conditions of limited programming staff.

Table 1. AVRDC germplasm collection 1977 & 1978.

Crop name	1977	1978
Soybean (*Glycine max* and other *Glycine sp.*)	9,765	9,980
Mungbean (*Vigna radiata*)	4,151	4,927
Other *Vigna* species[a]	398	398
Tomato (*Lycopersicon esculentum* and other *Lycopersicon sp.*)	4,706	4,755
Chinese cabbage (*Brassica pekinensis* and other *Brassica sp.*)	661	699
Sweet potato (*Ipomoea batatas*)	389	389
White potato (*Solanum tuberosum*)	1,295	1,295
TOTAL	21,365	22,443

[a] Includes 179 Black gram (*V. mungo*), 85 Adzuki bean (*V. angularis*) and 134 Rice bean (*V. umbellata*).

Table 2. Comparison of distribution of mungbean accessions by country with distribution of area planted.

Country	Accessions collected		Area planted	
	Number	%	1000 ha	%
India	2,510	50.9	1,940.0	72.5
Thailand	235	4.7	222.8	8.3
Burma	10	0.2	184.0	6.9
Indonesia	63	1.2	147.4	5.5
Pakistan	54	1.1	68.4	2.6
Philippines	301	6.1	39.3	1.5
Bangladesh	1	0.0	15.2	0.6
Iran	249	5.1	30.0	1.1
Sri Lanka	4	0.1	8.3	0.3
Korea	140	2.8	8.0	0.3
Taiwan	34	0.6	0.6	0.2
Others	1,205	24.5	unknown	0.0
TOTAL	4,927	100.0	2,676.1	100.0

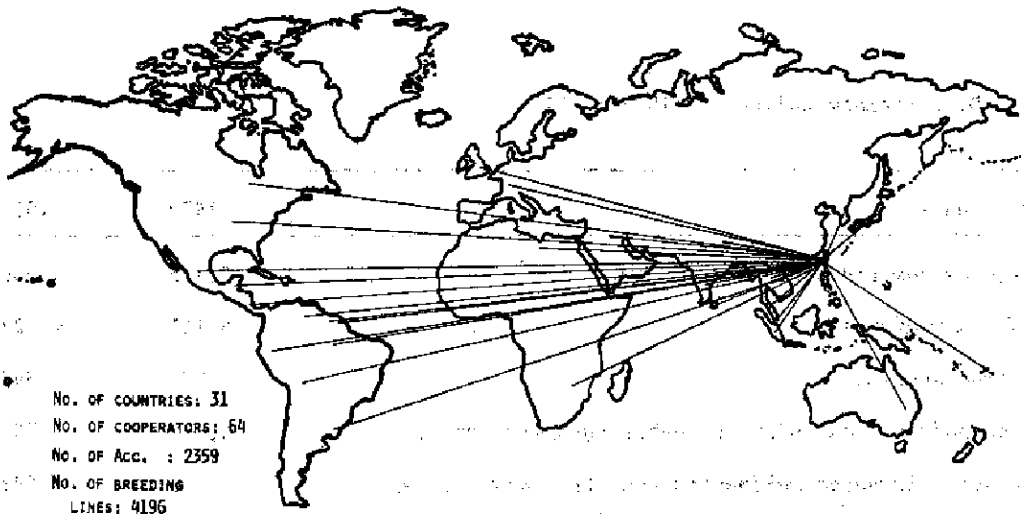

Fig. 1. WORLDWIDE DISTRIBUTION OF SOYBEAN GERMPLASM AND BREEDING LINES TO COOPERATORS

Fig. 1. Worldwide distribution of soybean germplasm and breeding lines to 64 cooperators in 31 countries.