Hunt Institute for Botanical Documentation
5th Floor, Hunt Library
Carnegie Mellon University
4909 Frew Street
Pittsburgh, PA 15213-3890
Telephone: 412-268-2434
Email: huntinst@andrew.cmu.edu
Web site: www.huntbotanical.org

The Hunt Institute is committed to making its collections accessible for research. We are pleased to offer this digitized item.

Usage guidelines
We have provided this low-resolution, digitized version for research purposes. To inquire about publishing any images from this item, please contact the Institute.

Statement on harmful and offensive content
The Hunt Institute Archives contains hundreds of thousands of pages of historical content, writing and images, created by thousands of individuals connected to the botanical sciences. Due to the wide range of time and social context in which these materials were created, some of the collections contain material that reflect outdated, biased, offensive and possibly violent views, opinions and actions. The Hunt Institute for Botanical Documentation does not endorse the views expressed in these materials, which are inconsistent with our dedication to creating an inclusive, accessible and anti-discriminatory research environment. Archival records are historical documents, and the Hunt Institute keeps such records unaltered to maintain their integrity and to foster accountability for the actions and views of the collections' creators.

Many of the historical collections in the Hunt Institute Archives contain personal correspondence, notes, recollections and opinions, which may contain language, ideas or stereotypes that are offensive or harmful to others. These collections are maintained as records of the individuals involved and do not reflect the views or values of the Hunt Institute for Botanical Documentation or those of Carnegie Mellon University.

About the Institute
The Hunt Institute for Botanical Documentation, a research division of Carnegie Mellon University, specializes in the history of botany and all aspects of plant science and serves the international scientific community through research and documentation. To this end, the Institute acquires and maintains authoritative collections of books, plant images, manuscripts, portraits and data files, and provides publications and other modes of information service. The Institute meets the reference needs of botanists, biologists, historians, conservationists, librarians, bibliographers and the public at large, especially those concerned with any aspect of the North American flora.

Hunt Institute was dedicated in 1961 as the Rachel McMasters Miller Hunt Botanical Library, an international center for bibliographical research and service in the interests of botany and horticulture, as well as a center for the study of all aspects of the history of the plant sciences. By 1971 the Library's activities had so diversified that the name was changed to Hunt Institute for Botanical Documentation. Growth in collections and research projects led to the establishment of four programmatic departments: Archives, Art, Bibliography and the Library.

Reading copy See attached paper

DATA PROCESSING FOR GENETIC RESOURCES*

David J. Rogers
Professor of Biology
Taximetrics Lab
University of Colorado
Boulder, Co. 80309

Japan 76
10th Computer Science Symposium
Theme: Food Supply & Agricultural Problems

Introduction

In this paper, I wish to stress three activities that must be considered together when discussing the management of data describing genetic resources. These three are: activities that must precede (or preliminary to) data gathering, activities associated with the computerized (or computer-aided) manipulation of the data, and activities concerning education and training for genetic resources data.

Historically, data management in genetic resources has been a matter of private (or individual) concern, and only with the advent of national and international cooperation

---

*Genetic resources are those propagating or regenerating structures that carry the inherited substances of plants and animals. "Germplasm" is an older term that is more or less synonymous with genetic resources. Frequently, we speak of "gene banks" as repositories of genetic resources.

has the matter of management begun to assume proportions greater than the individual. Genetic resources share the same historical perspectives - that is, no concerted effort was made in connection with the scientific functions incorporated today. It was the responsibility of the individual scientist or technician to gather, maintain and use any data describing his experiments, and no standardized procedures were required, as long as the end results served the individual's needs. In these activities, the individual should have received sufficient training so that he could carry out his functions adequately--managing his own files of data, developing his own means of data analysis, and reporting the results of his activities. Unfortunately, as I have examined the work and results of many biologists, agriculturists, and others interested in the life sciences, I have become painfully aware of the inadequacy of the training of our scientists, either individually, or together.

As our concerns for an adequate food supply have grown, individuals, disciplines (such as plant breeding), organizations (either private or publicly supported) on a national or international level, have found a need to act in a more coordinated manner. We have learned that plant and animal genetic resources are not restricted by political boundaries, and that our most important varieties of cultivated plants come from different parts of the earth. Perhaps one of the best indicators of international interdependence on genetic resources is that concerning the

short-strawed wheats that have figured so prominently in the "green revolution". The source of the genes for short straw were first noted growing in northern Japan, were collected there and put into storage in the United States, in Pullman, Washington, where the variety was described, and "put on the shelf" until Norman Borlaug used these dwarfed varieties in his breeding program in Mexico. (see Harlan, 1976 and Anonymous, 1972). The source of the dwarf genes in Japan is unclear, perhaps the mutant was developed there, or it may have come from another part of the world. Whatever the circumstances, this illustration indicates the great interdependence we all have with respect to our pool of genetic resources, and further, the significance of data describing the genetic resources.

A further historical comment that is important for our concerns in data management is that the individual scientist usually did not realize that the data he collected would be important to other researchers either in his discipline or other disciplines. Today, we have come to recognize that data can be useful in many different settings, that the data gathered by the pathologist, the breeder, the agronomist, etc. would be useful not only among these disciplines, but that these data also could have very great importance in decision making by social and economic scientists. Slowly, our perceptions of the great importance and the value of data are improving and we now find that more individuals and their organizations are realizing the

necessity of emphasis on proper data management. A number of factors have been responsible for this new, and welcome, change in attitudes.

Preliminary Activities

Given the growing concern for coordinated data gathering, there must be careful design of the activities to be coordinated. The work done in Japan is exemplary (Matsuo, 1975). Perhaps the first function that one should be concerned about is a determination of the general types of work that must be carried on to produce the desired end result, which must be defined. I will use as an example the description of genetic resources work that has been developed in my laboratory. First, a definition of genetic resources work was required. It was important to learn about the requirements of the participants at each of the levels of involvement, so that there would be adequate consideration of each when designing an overall data management system. Only by contacting many different groups of individuals, all of whom had some function with genetic resources, could an adequate description by forged. Figure 1 identifies the major functions or activities carried out in genetic resources, and indicates the relationships of data (or documentation) and communications to these functions. The first activity that one can ideally expect to occur in genetic resources is collection, from farmer's fields, from markets, and/or from the wild. One must obviously do

Fig. 1

explorations in order to make these collections. Each of these activities, exploration and collection, produce data associated with the collected materials and these clearly must be documented. Next, evaluation follows collection and exploration through organizing what has already been collected, and determines what still remains to be collected in a particular area. Thirdly, the conservation process of genetic resources functions is complex, involving many disciplines and generating many data. Overlapping the function of conservation is the function referred to in Figure 1 as improvement, from simple selection through complex breeding programs. Both conservation and improvement generate large data banks. Finally, following improvement, production is a necessary step as a part of the overall function of genetic resources.

Lights

If the above description of genetic resources functions is a valid description of the overall functions, then there is a better opportunity to set out the objectives one may have with respect to a data management system. One objective, clearly, is to serve the persons who initially collected the resource material and data so that their goals may be more nearly accomplished. In addition to the primary producer of the data, there are other users in the chain of genetic resources functions who will benefit from it, and therefore, there are other objectives. At each level, some of the data from an earlier level can be employed. For example, data gathered in exploration or collection can be used by plant

breeders. If a plant breeder is aware that a particularly collection of Zea mays was collected above 3000 meters, he will be able to use that information in selecting his breeding stock since maize is particularly altitude-sensitive. Or habitat data from collections frequently will give some indication about the possible types of pathogens that might be encountered in a particular region, and breeders of pathogen resistance profit by knowing as much as possible about those collections that can be useful in their resistance breeding. Further objectives can be thought of--inventory control, needs for further collections in certain regions, evaluation data, etc.,etc.

Fig.2

Figure 2 identifies further the functions that must be integrated in an appropriate genetic resources data management system. The triangle contains the functions of documentation analysis and shows its central role throughout the activities in genetic resources, both in conservation and improvement. In both cases, data bases are formed. The purpose of maintenance, obviously, is to keep in good condition all of those genetic resources that have been collected. Each activity has a need for some data management function. Similarly, improvement has several functions that demand a data system of some sort, either statistical or multi-variate analysis techniques, to understand the data that have been collected in association with improvement.

One of the major concerns in genetic resources is conservation, and the left side of Figure 2 is an indication

of the types of data management that are needed for these activities. There are, for example, needs for computer-aided methods merely as an inventory. As in any large scale endeavor, there is a continuing need to know what materials are on hand, which materials need replenishing, what space is available for new materials, etc. These common requirements are more complex with stored genetic resources because viability of the stored seeds must be maintained. As yet, most storage facilities are not sufficient to keep seeds viable indefinitely, and as a result there is a need to test viability of the stored genetic resources at regular intervals. When viability of seeds drops to a given percentage, the stored seeds must be taken out, grown in a suitable environment, harvested, and returned to storage. These are complex biological tasks, and require, therefore, exceptionally efficient data systems to keep up with the records of the various activities involved. Likewise, the contents of the seed storage facility must be catalogued, so the users of the seed banks may know what is contained. Since users may have different types of requests of the bank, the catalog should be responsive to these various inquiries. For example, one individual may not be certain just what seeds he needs for his breeding program, and may ask: "Please send me samples of all those seeds that have exhibited some type of resistance to stem rust"; this requires that a special search be made for those seeds, and no others; other individuals may request very specific seed

stored in the bank, such as the cultivar "turkey red". Finding such a seed may require many man-hours if the catalog is not arranged according to cultivar names, and a flexible, computer-aided cataloging system will be of immense aid.

That function of genetic resources management named "improvement" also will have several types of data management requirements, some in common with other functions, and some unique to that function. Frequently in improvement the data required for statistical analysis, to determine the best performance (such as yield, maturation time, freedom from diseases, etc.) once analyzed, are no longer needed. These data will then be discarded, and no storage of the data is required. Various types of statistics, from simple analysis of variance through various multi-variate techniques will be employed. Other efforts are, or could be well assisted by use of computer methods. For example, printing field books of crosses to be made would be an immense time-saver, and much more accurate than the laborious hand-written books now largely employed around the world. Although I have indicated here that much of the data used for improvement is not kept beyond its immediate use, it might be valuable if a record of all crosses made of all breeding lines could be stored, to prevent much duplication of effort in later years.

Lights

The last major function we have spoken about in the genetic resource work is that of production. Under this

heading are included all those functions that deal with actual crop production by farmers and the use of the products by consumers. There are, therefore, needs for data processing to record the activities and results in these areas--not only the data from the crops themselves, but also data on the economic values, and eventually, sociological data. The analyses of these data pose problems that are different from those of the agronomic, or plant-breeding type, and need well-defined, and well developed, computer-aided programs. The sociological data eventually will provide the information that determines all the other functions, since these deal with peoples' tastes, physical health needs, and their preferences. And yet this area of endeavor is perhaps the least well understood, and the least developed.

The above summary of events that should be considered before actual data management in genetic resources can proceed is merely a small indicator of the large number of opportunities for application of data management techniques. Since the agricultural community is just beginning the use computer-aided methods, it is safe to predict that the next few years, many more applications will be demanded, developed, employed. It has been true in other fields, and there is no reason to doubt that it will happen in agriculture. Therefore, it would be wise to consider development of a system that can be expanded, modified, and updated. This is necessary not only because of the needs of the various endeavors in agriculture, but because the field of computer

development and software programs is continuously changing--new equipment is developed, making older equipment obsolete. New methods of memory storage are increasing the capabilities of the machines to store incredibly large quantities of data; and the resulting continuous up-dating of software is a fact of life. I have personally experienced these changes over the last 20 years, and while the changes are not always improvements, frequently the new hardware or software does accomplish more. Learning to live with these changes is not easy, but it can be done.

ACTIVITIES ASSOCIATED WITH COMPUTER DATA MANIPULATION.

Some indications are given in the preceeding section about needs for a computer. In this section, I wish to speak about development that has taken place over the last few years in the Taximetrics Laboratory. The development consists of a Communications, Information and Documentation System (CIDS) for Genetic Resources (GR), and thus the acronym GR/CIDS. This development has been sponsored by the International Board of Plant Genetic Resources (IBPGR), which is an organization formed by, and supported by, the Consultative Group for International Agricultural Research (CGIAR). The IBPGR has as its charge the collection, conservation, and evaluation of wild and primitive cultivars of the major food crops. While it is primarily concerned with food crops, it also takes into account certain other crops that have global economic significance, such as

Gossypium species (cotton).

Within the CIDS, there is the concept of an overall data management system, with a series of computer programs that are capable of either separate use, or combining so that one program works with others to meet some data processing need. This concept is called EXIS, Figure 3. The generalized diagram merely indicates that a number of different functions can be made to operate together to assist in the processing of genetic resources data. While this concept has been employed in many other areas of human endeavor, it is the first time that such a concept has been put into action for genetic resources, as defined earlier. EXIS consists of a data storage and retrieval system, EXIR (EXecutive Information Retrieval, executive here meaning to carry out, rather than to direct); with a set of programs to prepare the data for printing (report generation), means to merge or concatenate different files (file manipulation) and various types of data manipulation or processing under the heading of statistical analysis.

Fig. 3

Figure 4 is a more expanded version of the EXIS concept, using different titles for some of the same functions shown in Figure 3. The basic "core" module is the data storage and retrieval system EXIR. From this module, subsets of data may be called out, in proper format, placed into other modules (such as "synthesis/analysis modules"), or, if no more processing is needed, directly into "display programs", which was indicated in Figure 3 as "report generator". The

Fig. 4

synthesis or analysis programs are sets of clustering or multivariate analysis methods, such as principal component analysis, to indicate multi-dimensional relationships between objects. The functions at the bottom of Figure 4 are more or less self-explanatory. One very powerful use of the graphic display capability of computer-associated machines is the ability to prepare maps of distributions on different scales. Collections, if provided with latitude and longitude data, may be plotted to any scale desired. Such information would be very useful in planning expeditions to make further collections. Other graphic display devices can produce density-gradient displays, which might indicate for a geographic region the frequency of occurrence of certain genetic traits (provided again, that these data had been associated with specimens in the computerized data storage).

Fig 5

Figure 5 is a further display of the types of programs we call "synthesis/analysis programs". Here, the programs are broken into three classes. In Class 1 are the synthesis or analysis programs most commonly used (at least in the United States). Program or package SAS is a very powerful, integrated set of statistical analysis packages developed at North Carolina State University, by a group who intended that the programs would be especially useful for agricultural scientists. SPSS, or Statistical Programs for Social Science, is another very widely used package not only in the social sciences but also in agriculture. In Class 2 programs are those referred to as single purpose, or self-contained

programs. Two of these listed (GRAPH, CHARANAL) were developed in the Taximetrics Lab. and are very powerful methods for classification. One may prepare data for these programs directly from the data storage and retrieval system, EXIR, and by proper commands to that system, have the data arranged in a formation for direct input into either GRAPH or CHARANAL. Class 3 type programs are, as indicated, useful sets of subroutines that perform various types of analysis, as needed. IMSL is a mathematical and statistical library of subroutines that is maintained by a special committee of the American Mathematical Society. There are many more such programs, or program subroutines, or program packages that one might conceive of as a system.

Lights

The point that I wish to make is that in a system such as EXIS, there are many opportunities for adding special data processing programs, and the only restriction is that the programs be compatible with the remainder of the system, and provide some general capability for genetic resources data processing. The attendant costs, both in financial and man-power expenditures, must be considered before one adds a particular set of programs to the system. Additional factors to consider in adoption is that of the ownership of the packages. Many software companies sell their packages with restrictions on the user, so that the user must conform to certain agreements. Other software programs are maintained by the computing companies that sell the hardware, and therefore, consideration must be made in developments such

as CIDS, that in general the software packages are available without restriction. This imposes a very serious problem when financial resources are restrictive, as they are in most agricultural endeavors.

Perhaps one of the best ways to indicate the value of an EXIS type system is by example. Consider the problem of collection of genetic resources. An expedition to collect Japonica types of Oryza sativa is to be carried out. Several questions should be answered before the expedition is sent out, for example: What materials do we already have in our collections? Where were these collections made and when (precise locality, habitat, time of year)? What is the most likely geographic region in which to find that genetic resource of Oryza that is required? With these questions in mind, a good data management system can provide helpful direction, assuming the data of previous collections have already been entered into that system. First, the data storage and retrieval system should be able to provide answers to questions about geography, time of year, habitat, local names of cultivars in those localities, and the display functions should be able to produce maps of the source of specimens already known. These answers give the prospective members of an expedition information about the best localities from which to derive new collections. They may indicate that some particular region has not been previously visited for Oryza collections, or that it is a location already heavily explored. Density of certain

characteristics, for example disease resistance, if known can be plotted on geographic distributions as most likely sources for new genetic materials to introduce disease resistance. Eventually, after collections have been made, there should be several types of evaluations. These evaluations will provide additional data that should be recorded with the accessions. Morphological descriptions, etc. will, or should, be included with evaluation. Many different types of evaluations will be made--environmental responses, agronomic characteristics, fertilizer responses, etc., and these should all be fed into a data bank for most efficient record keeping. In addition to the evaluation of the collections, there should be other studies made. For example, taxonomists should base a reclassification of the plants on the new data provided from the collections and from the results of evaluation. Various computer-aided methods make it possible to reclassify plants on the basis of new evidence at the same time that collections and evaluations are being made. Often there will be cytological and genetic information, as well as analyses of various chemical constituents (i.e. amino acid profiles, carbohydrate, fat, and other constituents of importance) to be evaluated, and these can be valid characters for use in taxonomic work. The computer-aided programs that serve taxonomists are such packages as the NTSYS (Numerical Taxonomy System) or GRAPH and CHARANAL. These provide objective means to classify the plants, and classification of cultivated plants is a very

necessary activity for the theoretical aspect of genetic resources work. There are many reasons that are as important, but these are illustrative of a computer-aided data management system. Several of the international agricultural centers are asking for development of such systems for their own use.

Fig. 6

The above may be summarized and generalized, as illustrated in Figure 6. In this figure, one may also identify other important features that should be a part of the work of data management. These include the development of standards for measurements and standards for recording. There has been considerable discussion of these problems in the international community, but the process by which such standards are established has not been fully clarified. It is necessary, before such standards are accepted, that there be a clear understanding that there are the two different types of standards, as indicated in the figure. Another point made in Figure 6, is that there is a difference between "data" and "information". Data are the observations or facts derived by individuals and recorded, whereas information is derived from correlations from the data. We make this distinction in Figure 6, because it is important to recognize the hierarcy of data/information/knowledge.

Lights

Education and Training

The above discussion, while incomplete, gives some idea of what has been developed in the Taximetrics Lab.

We have discovered that the mere presence of a well-designed computer system is far from sufficient to ensure that there will be the desired result. I mentioned earlier that the contemporary requirements of national and international cooperation, in a systematic manner, is a considerable change from the past, where the concern with data for scientific purposes was an individual matter. The change to a new attitude will only occur through adequate educational processes. In addition to our standard disciplines in biology and agriculture, we must add one more facet--computer management of biological data. There are skills involved herewith that have not been emphasized in university programs, but need to be added.

It is necessary that people involved in the work of data management for genetic resources be well grounded in the subject matter for which they are responsible. It is unrealistic to expect that individuals trained as data managers completely divorced from the subject matter of their data will provide service of value to the endeavors involved in genetic resources. We expect most of our students today to have some understanding of computer programming, and in most cases, at least introductory statistics, but we do not focus any attention on the subject of data management, for which there is a considerable body of knowledge.

Today, I am not aware of any institution in the world where programs for educating data managers for genetic

resources is actually turning out students. It is my concept that an educational program equivalent to a master's degree would be necessary for most institutions, and that a few programs for Ph.D. level would be satisfactory. Since the field is new, some experimentation will be required to produce the most efficient curriculum. Part of this experimentation will involve the selection and training of qualified faculty for such a program and it will take some time to assemble a well experienced faculty. At the beginning of any new era, considerable experimentation both with means of educating and will the time requirements are to be anticipated. All that can definitely indicated today is that there is a crying need for such education to be carried on.

REFERENCES

1. Anonymous, 1972. Genetic Vulnerability of Major Crops. National Academy of Sciences, Washington, D.C., U.S.A.

2. Harlan, J.R. 1976. Gene centers and gene utilization. Unpubl. paper presented at AAAS meeting, Boston, Mass., February, 1976.

3. Matsuo, T. (ed.) 1975. Gene Conservation. J.I.B.P. Synthesis Vol. 5. 229 p. Tokyo.

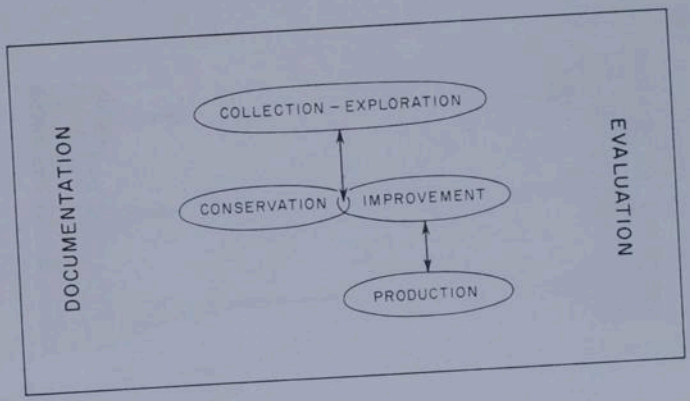

Figure 1.

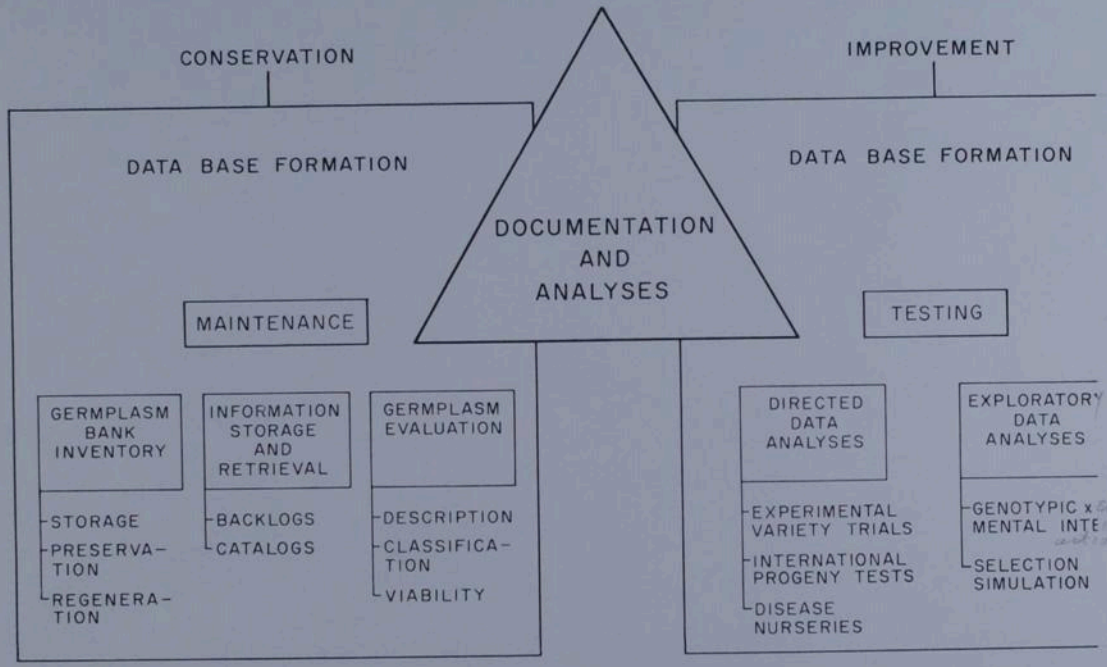

Figure 2.

THE EXIS SYSTEM

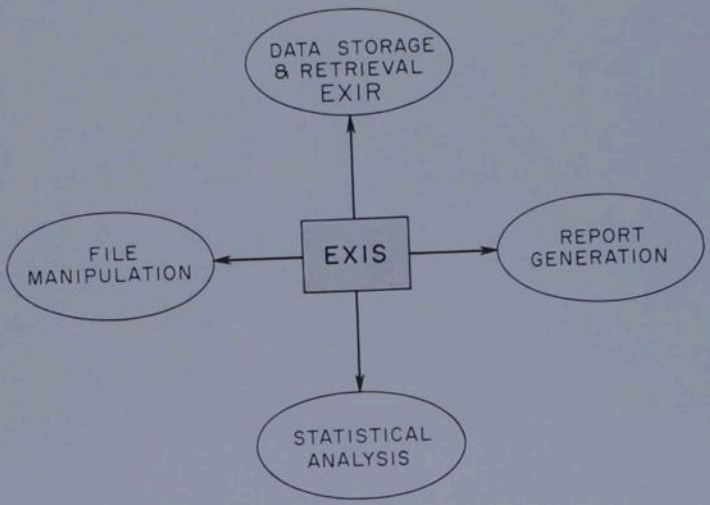

Figure 3.

EXIS DEVELOPMENT

SYSTEMS DIAGRAM

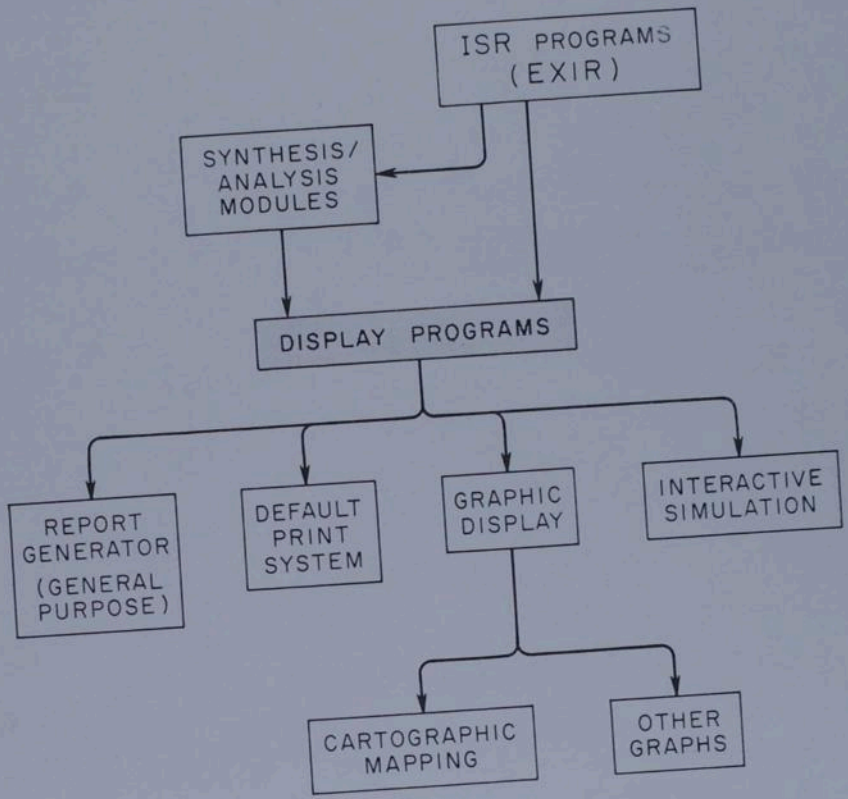

Figure 4.

EXIS DEVELOPMENT

SYSTEMS DIAGRAM

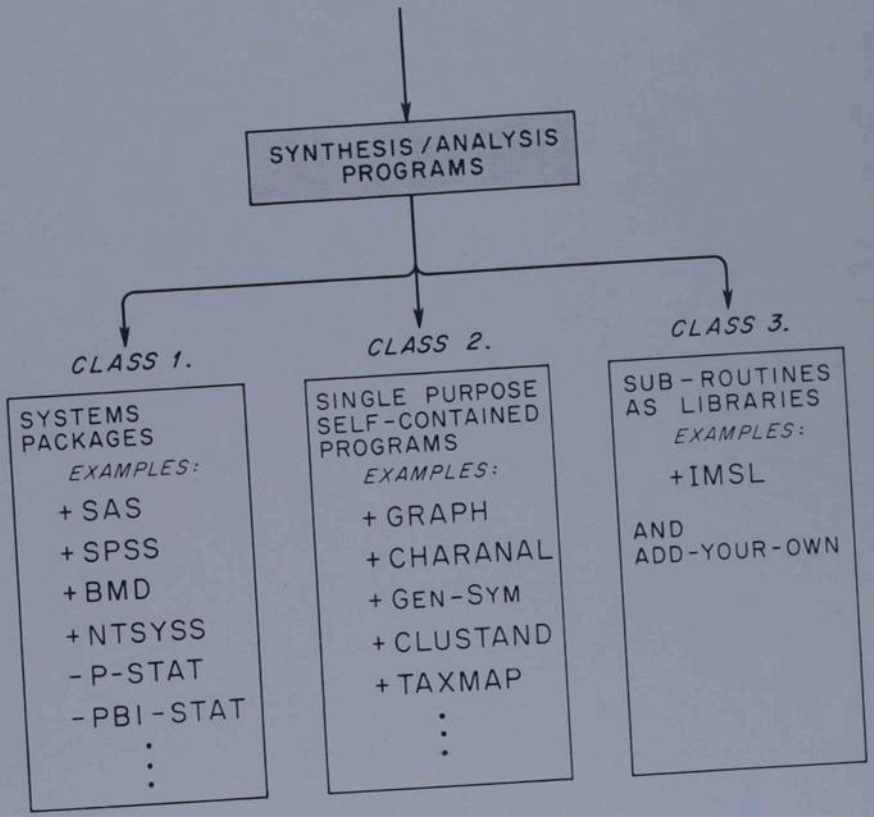

Figure 5.

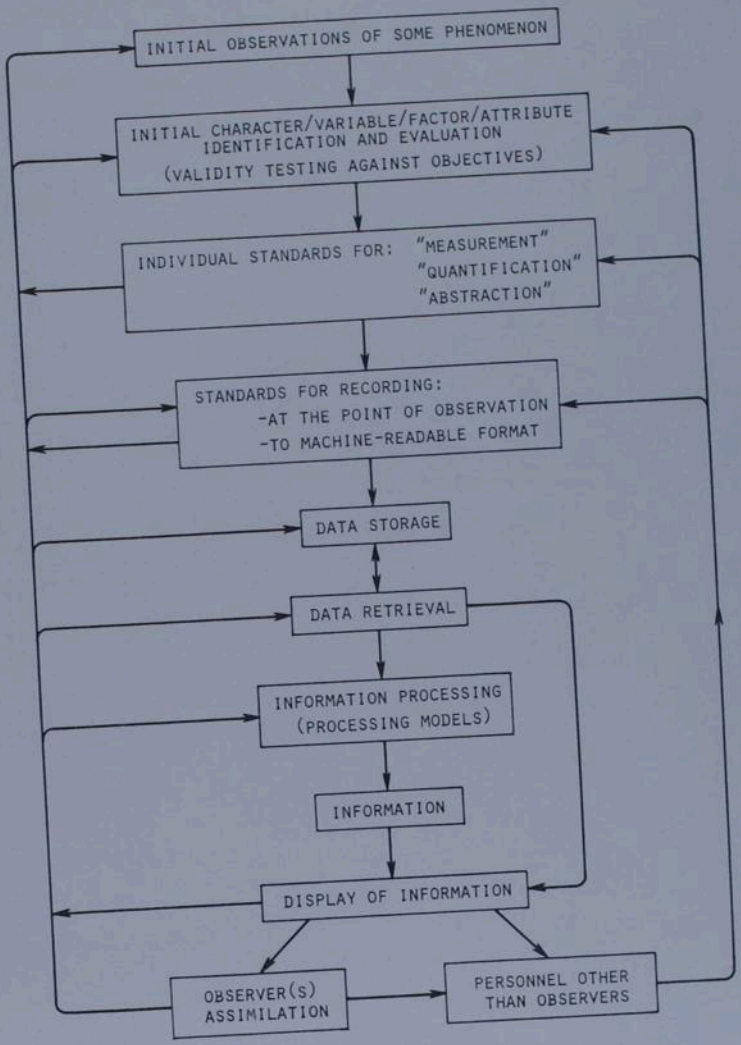

Figure 6.