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The Hunt Institute Archives contains hundreds of thousands of pages of historical content, writing and images, created by thousands of individuals connected to the botanical sciences.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of time and social context in which these materials were created, some of the collections contain material that reflect outdated, biased, offensive and possibly violent views, opinions and actions.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 Archival records are historical documents, and the Hunt Institute keeps such records unaltered to maintain their integrity and to foster accountability for the actions and views of the collections'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.

AGPE:MISC/4

THE DOCUMENTATION OF PLANT GENETIC RESOURCES

A Background Paper

by

David J. Rogers
Consultant on Documentation
Crop Ecology and Genetic Resources Unit

Food and Agriculture Organization of the United Nations
Rome, April 1974

WS/E8865

TABLE OF CONTENTS

1. INTRODUCTION

The purpose of this paper is to provide a brief, general background for the concepts of documentation of genetic resources. This is necessary because these concepts are more inclusive than past concepts of documentation within the overall activities of genetic resources centres (referred to frequently herein as "GRCs"). This paper is written for those who must make judgments about the usefulness and feasibility of any aspect of the work to be done, either in the individual GRCs, or in the coordinating role of FAO, or other organization.

Information is given :

(1) to indicate the relation of documentation to all other functions in GRCs ;

(2) to present an historical perspective to the work ;

(3) to describe the necessary components of the documentation function ;

(4) to indicate the requirements in personnel and equipment ;

(5) estimate costs for the work ;

(6) to present sufficient appended definitions of unusual terms to make the paper meaningful.

This document provides the basis upon which a detailed plan of action for documentation may be prepared, as well as the budgetary requirements. Clearly, any plan will have to be related to all the other requirements of genetic resources functions in FAO, and considerable effort will then be required to place the documentation function in the proper context. These necessary connections and their final integration are not defined in this paper as these must be a second stage in the documentation procedures for genetic resources.

In general, the documentation function must have the capability to supply answers to questions first, about the collections themselves, and second, to provide information which will be useful for the coordination, or management function which is the responsibility of the Crop Ecology and Genetic Resources Unit in FAO. Examples of the first category are as follows :

a) What types (species, varieties or cultivars) of wheat with resistance to stem rust are housed in which genetic resource centre ?

b) What are the dates of flowering of those almonds in Turkey with good fruit predictability and found at altitudes above 800 metres ?

c) What is the number of collections of cowpeas presently in the GRC at IITA, Ibadan ?

Examples of the second category are :

a) What are the regions where primitive cultivars of wheat may be found, but which of these regions are least well represented in present collections ?

b) What exploration experts are available, and when, to send to the Sahelian region of Africa ?

c) What are the storage conditions existent in any particular GRC or country ?

d) What are the rates of accumulation of all accessions in any country or GRC over the past five years ?

We cannot answer the above questions now - the data do not reside in data banks. But we can design the systems necessary to put present (and future) data into a documentation system which will allow these types of questions to be answered.

2. BRIEF HISTORY OF DOCUMENTATION

2.1 Early Plant Collection Practices

In the past, whenever wild or cultivated plant materials were collected, the data obtained with the material were generally sketchy or inaccurate, and to such extremes, for example, that the material was collected "in India", with no other designated location specified. No means could later make this information more definitive. From the earliest days of world-wide exploration some sort of naturalist accompanied the expeditions to bring back plants of interest to the sponsoring institution. Private individuals and organizations have, as well, continued the search for plants of economic significance, and frequently veiled their introductions with misinformation to prevent rivals from finding anything which might have had competitive advantages in production of some commodity of value. The results have been, until recently, a relatively chaotic condition. As more and more collections were made, there grew a realization that more precision was needed to identify the many collections, and as this realization grew, more emphasis was placed inevitably on accumulation of descriptive data with the raw plant material. Once the collections were made, they were stored in some manner, and the storage problem became acute when there was no means of selecting one or several previously collected accessions from a large mass of collections. Effectively, the larger the collection, the greater chance that the individuals would be lost.

2.2 Documentation for Utilization

Certainly, scientific plant breeding on a large scale cannot tolerate misleading or inaccurate data. This has led to the demand that all accessions of viable material be associated with the most efficient means of storing and retrieving the data associated with them. In addition, the many functions of management of the accessions such as provisions for further exploration, determination of personnel and facilities requirements, etc., demand that summaries of information be available to facilitate meaningful planning. Thus, documentation has come to play a more vital and central role than ever before.

2.3 Computers and Information Retrieval

In the 1950's, when the first commercially-available computers appeared, the use of these complex devices was suggested to cope with the ever-enlarging load of data and literature. Most of the endeavours to employ the machines were in the realm of literature retrieval - finding rapidly and exhaustively the places where any (or all) articles on a specific subject had been published - but some investigations were financed to establish systems for handling data primarily. The enthusiasm of those who either sold computing machines or those who had access to them frequently exceeded the actual capacity of computers to "deliver" what had been so extravagantly proclaimed. We have learned to view with much scepticism those sweeping claims that the computer would be the saviour of us all. Computing machines were initially designed to accomplish tasks of calculation, and the memory units were needed only to hold necessary data as input to some mathematical calculation. As the memory units have grown in size, so has the

desire to use the machines as enormous storage and retrieval devices has grown. However, most computing machines still have their greater function as calculating devices. The processes of storage and retrieval do not require particularly complex mathematical functions, but they do require sophistication in the proper types of mathematical manipulations. These guarantee that the machines' capacities are used to their maximum for information/data storage and retrieval purposes. We refer frequently to these systems as IR.

The major efforts to develop information retrieval systems using computers began when the third generation of computing machines became available, in the middle of the 1960s. By that time, much more was known about IR requirements, the limitations of the machines, etc., so that some sophistication in the design of IR systems could be expected. However, the major concepts for IR computerized systems are less than a decade old, and under these circumstances, it must be understood that there is still much room for great improvement, both in the software 1/ packages and in machine design.

Today, there are a number of IR systems which have nearly the same capacities to serve the needs of the genetic resources community. It therefore becomes necessary to make good cost/effective comparisons to discover the one preferred system which will most satisfactorily provide the back-up required for GRC documentation. There are, for example, "proprietary" IR systems such as GIS (General Information System) and STAIRS (Storage and Information Retrieval System) owned and contracted by IBM. The General Electric Corporation also has a proprietary package, not only for IR, but also for networking. MARK-IV is another, owned and contracted by Informatics, Inc., in California. There are several other IR systems developed by governmental organizations that might be suitable for genetic resources, but the most outstanding of these is "SELGEM", a system designed and developed at the Smithsonian Institution, Museum of Natural History.

FAO itself is now working on two complementary types of systems, CARIS and AGRIS, but the functions of these two are complementary to, not duplication of, the type of IR systems needed for genetic resources.

3. DOCUMENTATION

This term has become the accepted one for all the functions of data management and information exchange in the milieu of genetic resources functions. Documentation refers to all the activities with respect to data and information from the initial point at which the information or data are derived during collection and through every stage of activity in a genetic resources centre to its ultimate application to the propagating material contained in the centre or group of centres, and the coordinating function of FAO. A full explanation of the activities is beyond the scope of this discussion, but a full exposition is given in the paper by Hersh, Rogers and Appan 2/.

The main emphasis in the above-mentioned paper is that documentation requires, inter alia, a new set of skills, referred to as management science, to fully coordinate the many functions of documentation. The management science function monitors continuously all aspects of documentation to insure that the objectives are met in the most cost/effective means.

1/ See appendix for definitions.

2/ Hersh, C.N., Rogers, D.J., and Appan, S.G. 1974 (in press). Documentation and Information Requirements for Genetic Resources Application. in Frankel and Hawkes (eds). IBP/FAO Report on FAO Technical Conference on Crop Genetic Resources, Rome, 1973. Blackwell, London, Publ.

Users of documentation fall into two major categories :

1) primary user,

2) secondary user.

The primary user is any individual who both generates data and/or information for genetic resources, and uses his data (and that of others) to satisfy some requirement. The primary user will use data either to give some answer to a question in plant breeding, ecology, agronomy, physiology (and many other disciplinary functions), or to provide lists of the included data and information for some scientific purpose.

The secondary user is a person who normally does not generate the data, but who may "call" information to satisfy some requirement from that which is included in the data bank 1/. We normally think of the scientist as the primary user of documentation and the administrator as the secondary user. (This is a small distinction, but nevertheless one of importance when decisions are made concerning the rights of individuals or organizations to query the system). Administrators (directors or chiefs) clearly need to know what material the genetic resources centre under their direction contains, where collections have or have not been made of the various crops, what areas require further collection or examination, how much space is still available for storage of the accessions, etc. Access to these data enables the administrator to meaningfully plan and budget for future operations.

We have used the terms "data" and "information" above. We should distinguish between the two because they are most likely structured differently, and have to be treated differently inside the computing machine. Data, in our sense, are the primary observations made on some accession. They are structured into descriptors 1/ and descriptor states 1/, and are usually discreet, non-overlapping types. In contrast, information is used to refer to published sources which may be overlapping or non-discreet. Clearly, this is a technical differentiation, but a very significant one for IR systems.

4. GENERAL TASKS INVOLVED IN DOCUMENTATION

4.1 Precomputer Tasks

Before one is ready to "feed the data to the computer" there are a number of considerations to guarantee that what is "fed to the computer" will accomplish the tasks required of this data/information. The first requirement is that those preparing the data about the accessions have an overall concept of the purposes which the data must serve, and a framework for structuring the data. The overall concept is one which guides the collection of data. Below we introduce a new set of ideas : a classification of types or classes of data. This has not, to date, been brought into focus by any of the earlier stages in the development of documentation.

1/ See Appendix for definitions.

(i) Classification of data types in GRC work

1) "Finding" information

Accession information :

- accession numbers,
- accession dates,
- accession types,

Collection information :

- names of collectors,
- collector's numbers,
- collection dates

Nomenclature information :

- scientific names,
- common names,
- numeric designations of individual variants.

Origin :

- country of origin,
- state or province of origin,
- precise locality (may be latitude and longitude, or other designation)

Storage information :

- physical location of the accession in storage,
- conditions of storage - bag, envelope, box, can, etc.

2) Organismic information

- biochemical,
- physiological,

Genetic information :

- number of chromosomes,
- crosses

Morphological :

- plant dimensions,
- plant habit,
- root characteristics,
- stem (or tuber) characteristics,
- leaf characteristics,
- flower characteristics,
- fruit characteristics,
- seed characteristics.

3) Pest and disease information (resistance to, types of organisms, etc.)

- bacterial,
- fungal,
- viral,
- insect,
- nematode,
- other pest or disease

4) Environmental information

- ecological
- environmental damage :

  wind,
  water,
  drought.

5) Rejuvenation or regeneration information

- germinability tests
- history (dates, where work done, when returned to storage, etc.)

6) Use information

- breeder's data,
- specific application (food quality, advantageous or detrimental quality, etc.)

7) Other categories, as required.

Once the above type of classification is understood, it then becomes more reasonable and rational to consider the purposes and values of some sort of minimal standardization. Furthermore, it provides a basis for placing data from extant collections into the proper format for use. We know, of course, that in the past, the data on accessions is likely to be incomplete, but we cannot judge how incomplete, nor where and when new information on the accessions will be needed. With the above type of classification, a more rational basis for proceeding is permitted, which at the same time, gives proper direction to the establishment of standards.

The above categories were established from a pilot test 1/ of the applicability of the TAXIR 1/ system to genetic resources. We did not generate the classification a priori, but after some experience with the types of data to be found in genetic resources centres, the classification was developed a posteriori. (This represents a philosophy which we will adopt in the documentation section of the Crop Ecology and Genetic Resources Unit - we do not presume to establish by fiat some arbitrary set of standards which have not been actually tested on a sufficiently good sample of genetic resources material. We will allow individuals who have their own standards to employ them. We will compare the standards of various workers, and share with each of the workers (given their permission) the methods of several different researchers working in the same situation. If the workers then choose to accept some standard, their basis for doing so is much more real than from the results of proceeding in the a priori fashion which has typified so much previous work in this area. N.B. This does not exclude any of the previously agreed-upon standards, but puts them into proper context.

In the classification given above, we designate as "finding" information all those types which aid the user in precisely designating any one accession out of the total collection. One may be able to find some accession by several different routes, by a number, a collector's name and his number, by the scientific or common name, or by several other designating systems, but all serve the general purpose of "finding" or locating one accession. The many different types of "finding" information have developed over years of usage. None of them is any more or less valuable than any other. By applying all of them, the chance of loss is less than if only one were employed. Besides, each "finding" piece of information may give clues to the other, equally valuable types of use. For example, using the accession and dates, one may trace the history of a single accession. **Or, the total number of accessions, associated with their scientific and common names, and their geographic locations, in a centre or group of centres, gives the administrator the opportunity to judge what additional collections are necessary, and from what regions.**

Having established the above classification as a basic guide to the types of important information in the documentation of genetic resources, one may then proceed to discover which types of information will be important as a minimum set of data to be gathered. These minimum sets have been and continue to be a major area of concern to a small group of genetic resources specialists. We trust we will soon have completed the preliminary set of standards and proceed to use those available in actual work in genetic resources centres.

Furthermore, the same basic classification of types of information can be used in all the steps necessary to guide the proper movement of an individual accession from the moment when the plant material is first collected through its various steps of quarantine, assignment of accession number at the genetic resource centre where it is

---

1/ See Appendix for description.

deposited, and through multiplication, storage in appropriate situations, evaluation, and application. Within the unifying functions of this classification, there will be an orderly, understandable, continuing addition of information to the accession as more is learned about its properties. Above all, adoption of this system should eliminate the wasted duplication of records from office to office, from centre to centre.

One further word : the above classification is given only in outline form in this paper to illustrate the type of classification used. It is far from complete, and there will be additional major headings in the classification, and many additional sub-categories.

(ii) Different functions within a genetic resources centre, and the documentation associated with each function

In any genetic resources centre, either presently functioning, or yet to be established, there will be a fairly common flow of functions, and information on the accessions will be gathered at each stage. In exploration, for example, certain data will be accumulated with the collected material. There are several requirements in the processing of the collected material before it becomes an accession in storage. The management of the movement and data accumulation requires a well-ordered and managed documentation system. There are several techniques by which the data, as they are generated, can be placed into some machine-readable format, either on mark-sensing cards, or on standard 80 column IBM punch cards. The sophistication of the process is dependent upon a number of factors which include training of the personnel, the size of the daily, weekly, or monthly accumulation of accessions, and the budget which is available. The particular configuration to be used in any GRC will be dictated by an analysis of all these factors, and the most cost/effective techniques will be employed. However, there should be a system which can be modified without interrupting its fundamental functions, should there be a valid need to change. The objective in all the documentation function is to provide the most efficient and least costly means which will insure that the data and information serve the purposes for which they have been accumulated.

4.2 Computer-related Tasks

(i) Background information - The difficulties of installation of IR systems

Almost every paper written recently about the establishment of genetic resources centres on a global scale have indicated that computers should be involved in the task of documentation. It is certainly realistic to speak of a computer-based documentation system, considering the quantity and quality of data to be managed. However, there has been little understanding about what is involved, and it is the purpose of this discussion to provide an introduction to the understanding of the procedures to be followed.

A further discussion is needed if we are to discover the real meaning of a network of genetic resources centres. We have a vague idea of the meaning, but no specifics have been provided. Clearly there are many connotations of the term network, and one of those deals with information and data. I will address this problem further on, after some discussion of the more basic problems.

Over the past decade there has been an intensive effort in many computing centres and other institutions to develop information storage and retrieval systems. One very successful endeavour for the retrieval of data is that which provides airlines with reservation information. These are very efficient, but they do not serve as a good model for the storage and retrieval of genetic resources data for the simple reason that the types of information into and retrieval from store by airline are much more restricted, and the data are not needed for long-term storage. As soon as one airline flight is completed, the computer's memory is purged of all information about that flight, and thus the capacity of the memory units does not have to be large. Considering the fact that there are a relatively large number of crops (estimated to be about 200 different crops by the specialists at the National Seed Storage Laboratory, U.S.D.A.,

Ft. Collins), that each of these crops needs a set of descriptors that do not exactly coincide with the need for another crop, and that most frequently the data are permanent records, we have a much larger and more complex problem. (This is one reason why it is very difficult to establish a set of minimum standards, except in the area of the classification given earlier under "finding" information).

Another set of difficulties with computerized IR systems has been the relatively small size of the memory units. Fortunately this picture is changing rapidly, but many of the computing centres available to GRCs are rather small. Therefore, they may have to be provided with computer service at units (sometimes provided by a computer manufacturer, sometimes by other governmental, or private organizations) away from the Centre itself. This procedure is quite common in the computing milieu since most individual organizations do not have a sufficient work volume to justify having their own computer. But returning to the size of memory required for IR systems, it is a fact of life that such systems will require large computing memory. This is true both because of the necessary complexity of the software packages for information retrieval and because the data load will be increasing continuously. Frequently, one is misled about the size of the memory by the (usually) very large indications of the size of the computing machines. The larger machines may measure their memories in the millions of bits (or sometimes "bytes" or "words") but these are designations which refer to single on-off electronic surges, several of which may represent only a single letter or numeric symbol. Given this information, one is not so impressed with the memory sizes, and further, anyone who builds systems must be extremely aware of the need for extraordinarily efficient and well-designed software packages. Of course, the memory units may be extended by the use of "peripheral", attached memory units, (magnetic tapes, drums or discs) but these operate at much slower speeds than the central unit. Therefore this limitation cuts down on the efficiency of the computing machine and increases costs. The most effective means to prevent high cost computing is to work carefully initially on the design of the data banks 1/, and to carefully analyze, on a continuing basis, the needs of the network or individual centre.

(ii) Types of functions required of the computer software package for IR

Given the above background, it is further necessary to describe the functions which must be included in any information storage and retrieval system which will be useful in genetic resources function. We get some useful indications of the types of functions from the previous statements written about the needs for information and documentation in genetic resources centres. (A useful set of documents for this purpose is the Plant Genetic Resources Newsletter. Here, in editorials and in the various articles one finds published statements on the requirements for computerized banks of particular elements of genetic resources centre functions, both in FAO and in various parts of the world).

Starting with the types of data input, we recognize a variety of designations of data-alphabetic letters, numeric notations, and alpha-numeric combinations of differing lengths and complexities. The software package must, therefore, be designed to receive for computer manipulation any combination, in any length, which will be useful in the GRC. Therefore, free-field 1/, rather than fixed-field 1/ input is a primary requirement. The programmes should be written so that the user can apply his own particular descriptive terminology in the most effective manner. We have a very rich vocabulary associated with plant materials, and this richness should not be reduced by unnecessary restrictions placed upon the user by any limitations of the software package. Since GRC scientists are already using their own vocabulary, we must accommodate them by allowing them to use their own terms, in their own way. This is preferable to imposing a strict "thesaurus" of terms which may, or may not be sufficient for the inclusion of new, or different terminology. Therefore, the terms to be used should be accepted by the IR system and without any restriction. Since we have this set of requirements, it follows that the most efficient technique should be used for storing the terms in the memory units.

1/ See appendix for definitions.

To accomplish this, mathematical techniques should produce means in the machine of **internal coding, which guarantees the most storage in the least possible memory space.** Indeed the success or failure of an IR system probably depends more on this one point than any other single feature of the software. There are, needless to say, some more and some less efficient procedures of storage. The most efficient one known to date is that designated as "compressed Chi functions" 1/, which not only reduce the data to extremely efficient storage units, but at the same time, provide their own addresses of the stored data. To date, no better means has been described or developed than this procedure.

From the user's standpoint, the IR system must be sufficiently well-designed to permit the user to manipulate any single datum, or sets of data, at will, to permit an answer to any type of question which it is meaningful to ask of a set of data. There are two general types of questions possible : (1) those which request a list of all the items 2/ within the data bank, ordered by some specified set of criteria. Each item would be accompanied by an associated set of descriptive information, according to some specific demand, and (2) those which precisely pinpoint one or more items within the total data bank that have a common, unique, set of attributes. To accomplish the tasks imposed, a mathematical methodology, derived from the discipline, boolean algebra, is applied to manipulate the included data in the bank to guarantee complete and unequivocal responses.

Other necessary features to be included in a good IR system are correction, synonymizing, deletion, or additions to data banks, the merging of data banks, and re-defining terminology, etc. Also, a good system should permit editing of the input.

Several desirable features in addition to the above are : "report generators" (presentation of the data in either columnar or tabular form, with the decimals properly aligned, and each column with appropriate headings, sums of columns of figures, percentage figures, etc.) ; "generate" capacity, where the output from the IR system is "formatted" for processing by additional software packages, such as plotting routines, where the data must be prepared in a certain configuration for use by an automatic plotting device. In genetic resources work, such a feature is almost essential, because we want to have accurately plotted maps of distributions of certain crops. "Generate" should also prepare data for statistical analysis, so that with one computer command, one may extract subsets of data from the data bank, and manipulate them for such requirements as means and modes, Chi square, regression analysis, factor analysis, etc.

Considering all of these functions, one begins to see that the software programmes which direct them become necessarily rather large and sophisticated. The "state of the art" in computerized information retrieval systems is, as explained in the section on history, less than a decade old. It is therefore understandable that most scientists and administrators have no real concept of the problems involved in accomplishing the tasks with a computing machine which they consider simple. There are obvious needs for continuing investigation and development of IR packages, even though the presently available system - TAXIR, provides the basic needs and the framework for adding additional functions. One of the clear needs is for FAO to be involved in these developments through its documentation function, and to maintain close contact with all the leading developers of systems drawing on the expertise which exists in both commercial and non-commercial organizations.

1/ See paper by Estabrook and Brill, The Theory of the TAXIR Accessioner, 1969. Mathematical Biosciences 5 : 327-340

2/ See appendix for definition

5. NETWORKS OF GENETIC RESOURCES CENTRES

It is accepted by an increasingly wide range of individuals that, to accomplish the goals of gathering into secure places the primitive races of crops and nearly - related wild species, there is a need for some coordinated, global organization for information gathering and processing. We have come to accept the ideas or concepts of networks, a term which is very popular today, for nearly any function or group of functions. But before any real work may be accomplished, we must make an effort to define exactly what we mean by the term "network". Several definitions are available, and different disciplines employ different connotations. From the documentation for genetic resources point of view, a network is some (or several) type(s) of formal communications system. The ultimate communications network, given today's technology, is a world-wide system of satellites, with transmission and receiving equipment in each GRC. This would be an ideal, if unrealistic, approach to our communications network. Ranging down from the ultimate, there are hierarchical series of communications systems which can be effective although these may function at slower speeds than we might desire. The telephone provides a powerful network which serves well over certain distances, but becomes prohibitively costly over inter-continental distances. This brings us to the level of postal services which, although very time-consuming and, in many cases, unreliable, are still the most likely communication links between world-wide GRCs. Data transmission between the GRCs may be expedited by preparing magnetic tapes at one location, and then sending these to some appropriate collecting point where coordination between the GRCs may be accomplished. With proper design, we can, hopefully, secure a communications link, or network, for GRCs. Part of our endeavour will be to forge these links, working again through the management science concepts, on the most cost/effective basis.

Continuing with the concept of networks, but coming to the individual GRC, we encounter different problems. First, an informal survey of the locations of GRCs indicates that these centres are seldom in locations where the organizations to which the GRCs are attached (either agricultural institute or university) themselves possess large-scale computer facilities. This may not be true for GRCs in the developed countries, particularly in the United States, but many of our functions are outside the boundaries of the developed countries. Second, a search will have to be made in the geographical area where the GRCs are located to determine if computing centres are available from large, unrelated organizations. Frequently, commercial establishments in the areas (such as oil companies) will have computer facilities which can serve as service units by contractual arrangements. We think that the GRC function of documentation can best be served by contracting with these types of computing centres, rather than requiring that the GRC either rent or purchase their own computing equipment. (Computing equipment in this sense is the main computer frame, and does not include peripheral equipment, such as key punches). Clearly, each GRC should have its own key punch (or two, if the load of work demands) so that data capture may proceed on a continuing basis internally within the GRC.

After confirming the presence of a nearby computing centre, then the size of such an installation must be checked, and the availability of competent computer programmers whose expertise is sufficient to understand and instal the IR system. It is my experience that, like computer-machine salesmen, programmers of any computer installation are likely to over-estimate their own skills, and will tell the unsuspecting that they can manage any size of software package that one may wish. Caution and skill, are required to determine just what capabilities exist. But, once a determination is made, one may proceed to design the software configuration which can be incorporated on a particular piece of hardware. This set of tasks should rest with the central documentation group in the Crop Ecology and Genetic Resources Unit in FAO.

The network for an individual GRC, then, may consist of internal data preparation (using standard IBM input cards), with a relation to a nearby computing installation, and finally, to the coordinating centre via magnetic tapes and the postal service. The FAO documentation team will aid in the design and installation of whatever appropriate software can be made to function efficiently on the local (or nearby) computing hardware. It will also determine what functions must be done at a larger, distant computing centre, and the most effective procedures to guarantee satisfaction and efficiency to each of the individual GRCs.

We cannot predict what the final configuration will be, not only because of the computer-related problems, but also due to the problems of budgeting for this function by the individual GRCs. Maximum communication links to the GRCs is necessary. A move to establish these links is now in progress through questionnaires to be sent out according to the recommendations from the FAO Panel of Experts in Plant Exploration and Introduction.

6. TASKS REQUIRED OF THE DOCUMENTATION FUNCTION IN FAO FOR GRCs

Given the above considerations, we can begin to see which tasks are essential to meet the as yet ill-defined objectives. The most immediate tasks are two-fold : (1) to outline, in detail, the documentation function acceptable to FAO and to the various users in genetic resources, and (2) establish the software package, TAXIR, either on FAO's computing machine 1/ or contracted to some other suitable computing centre, and determine what software configurations can be efficiently installed at certain GRCs.

Following on, we must develop two types of manuals to support the task of implementing these systems in the GRCs. These are (a) User's manuals, in FAO official languages and (b) Software manual, also in FAO's official languages.

After these developments (or perhaps simultaneously with them) we must arrange training sessions for GRC workers in the proper procedures that are to be followed. This explains the need for the manuals. Much of the basic work for development of the manuals has already been done. Preliminary manuals are already prepared, and now must be re-written to include the specific details for genetic resources documentation functions. Also, the expenses of the documentation of the software package for TAXIR have been met. (N.B. Documentation as used in this last sentence is quite different from documentation as used in GRCs - here, it means a line-by-line analysis of each of over 2 000 FORTRAN 2/ statements in TAXIR, telling what each line of FORTRAN instruction accomplishes, and the relation of that one line to any and all other FORTRAN instructions with which it could be associated. This type of documentation is absolutely essential to the proper installation of any complex programme on any computing machine, and the work was done in anticipation of the use of TAXIR by FAO).

Other significant tasks concern various aspects of documentation already initiated in this Unit. For example, there is a continuing survey of crops, and centres where accessions are to be found. If these data are incorporated into a data bank to be employed with TAXIR, they can serve many useful purposes. Also needed is a data bank

---

1/ See Appendix for explanation.

2/ See Appendix for definition.

of plant breeders and others interested in genetic resources, their organizations, addresses, specializations, etc. Furthermore, an important task is to use the documentation system for in-house functions, such as the seed exchange programme in FAO. In a real sense, the seed exchange programme is a rather specialized GRC, and should employ the same computerized systems as those outside FAO. Other units within this division are very closely related to genetic resources functions around the world, and there is no reason why these units should not use the same system for their activities.

7. ESTIMATING THE COSTS FOR DOCUMENTATION OF GENETIC RESOURCES

The following discussion is, at best, an educated guess. But before we start, it is well to recognize that at present there are costs associated with documentation, although seldom are these costs exposed. If one estimated time as a cost, and together with those individuals who are now associated with whatever system is presently employed, it could be demonstrated that between 30 and 50 percent of the work now being done in GRCs is data-related. Therefore the itemizing of the total costs will not greatly exceed the present expenditures, but for the first time, such costs will be quantified.

To provide **a basis for the costs**, the following estimate is given of the numbers of accessions which will be incorporated on a global basis into GRCs. These figures are deduced from consultation with several groups or organizations. The staff of the National Seed Storage Laboratory, USDA, Ft. Collins, Colorado, estimated that 200 major crops are of sufficient importance to be considered in genetic resources centre work. Of these 200, about 30 presently have "large" collections, 50 have "medium" size collections, and 120 have "small" collections. A "large" collection is represented by Zea mays (Indian corn or maize), of which there are presently some 40 000 accessions extant from Mexico and South America. The maize experts estimate that this may be 40 percent of the total global collection comprising 100 000 accessions. A "medium" collection is represented by the tuber-bearing species of the genus Solanum, in which collections of three competent sources now total about 25 000 accessions 1/. Those may represent 50 percent of the necessary global collections, a total of 50 000 accessions. A "small" collection may be that of the cowpea (Vigna species) at the International Institute of Tropical Agriculture, Ibadan, Nigeria, where there are currently about 6 000 accessions, representing 50 percent of the world collection **of about 12 000 - 15 000 accessions.**

Using the above estimates for the distribution of collection size, we extrapolate the following total estimates :

1/ **I recognize that many of these are duplicates.**

Collection size	Number of such collections	Number of accessions/collections	Totals
large	30	100 000	3 000 000
medium	50	50 000	2 500 000
small	120	15 000	1 800 000
Total :	200	-	7 300 000

N.B. The range of accuracy is (±) 30 percent for a range of 5.1 to 9.5 million accessions.

If all the data of the above accessions were in "machine-readable" format, that is, ready to be placed directly into the computer, the costs would be very much smaller than if, as expected, most of the data are not in such format. At current prices, it may cost from $0.10 to $0.50 to transform the information on one accession to a machine-readable format.

The costs from the planning through to the implementation stages of the basic computerized system may run from $250 000 to $500 000 depending on network complexity. These costs could be greater if it were decided to proceed at a more rapid rate.

The running cost of use of the system, once it is established, is the most difficult to estimate, because of the lack of information about the computing system(s) to be used. Because of the complexity of a full-scale information storage and retrieval system, the costs may be inverse to the the size of the computers employed - the smaller the machine, the greater the time to install the system and the greater the length of computer running time after the system has been installed. The costs could be as low as $50 000 per year, but could be as much as $250 000 per year.

The costs of installation at any given centre is very dependent on the equipment available at or near the centre and the skill of the personnel who are available there. We will attempt to select centres large enough to have competent personnel and equipment. If a relatively standard set of equipment is available with which to work (IBM largely, but perhaps a Siemens, CDC, or UNIVAC), the conversion, installation and training costs for the software package might be as low as $2 000 per centre if the centre personnel do most of the work, to about $7 500 per centre. At this point, we emphasize another of our philosophies : that the computer-based documentation system should be made an integral part of the centre where it is installed, for any purpose the computing centre, and the administrators of the organization may determine. But this type of philosophy imposes another factor, and that is that there must be a competent team maintained on a continuous basis at the coordinating centre for the global system. Experience tells us that such a team is an absolute essential to the continued smooth-running of the system. Initial installation of the programme and training procedures will not make the individual centres completely independent. Computer-based IR systems are in continuing need of up-dating and of personnel to keep the "bugs" out of the system.

I must emphasize that the figures given above are based on current prices, and at best, are extremely tentative.

APPENDIX

Definition of some terms used

Accession. An individual collected sample of viable seed or other propagating material (a cutting, some leaves, stems or roots).

Data bank. A logically related set of items (q.v.) with their associated descriptor-descriptor states. For example, all the accessions and their descriptor-descriptor states made by a field collector ; or, all the accessions of wheat in a particular genetic resources centre, etc.

Descriptor. A single basis for comparison, defined over the objects (items) in a set. A descriptor may either be intrinsic to the item, or extrinsic, giving some information about the item. (see attached illustration).

Descriptor state. Under each descriptor, the precise data or information which applies to a specific item. (see attached illustration).

	Descriptor	Descriptor state
Intrinsic	flower colour	red blue white red-blue
Intrinsic	leaf length	10 cm 11 cm 12.5 cm variable
Extrinsic	collector	Smith, R. Smith, J. Rogers, D. Rogers, W.

Free-field or fixed-field. These two terms refer to the means of structuring data for input on the standard IBM, 80-column punch card. The fixed field format indicates that any and all data must be placed in a certain number of predetermined columns (field) on the punched card, whereas the free-field format indicates that any number of columns, as required by the data, can be assigned to any and all data. The former is very restrictive, and frequently requires very limiting procedures on those who prepare data, whereas the free-field format permits the entry of any desirable data without restriction no matter what their length may be. TAXIR is designed to accept either free-field, or fixed-field format because of the different ways that data have already been prepared for more restrictive, less powerful systems. Thus, any presently machine-readable data may be input to the TAXIR package.

FORTRAN. Most computing machines today are designed to use one of several so-called "higher" computer languages. FORTRAN (which stands for formula translation) has become one of the most universally recognized languages for scientific purposes. It is much easier to convert FORTRAN packages from one machine to another, even though (as mentioned under the Appendix description of TAXIR), there are still

problems which must be faced because information retrieval packages are all (by the nature of the work to be done) very machine-dependent. There are several versions of FORTRAN, each more or less designed for more or less powerful computing equipment. There is today no application of FORTRAN I because it has been superseded by later developments, however, there are FORTRAN II, FORTRAN IV and V, with FORTRAN II used on smaller machines (and thus having less flexibility) with FORTRAN IV and V used for the larger machines. FORTRAN II can be used on larger machines, but FORTRAN IV or V cannot be used on machines designed for FORTRAN II.

Item. A single object about which data and/or information is collected. (Predominantly, in genetic resources centres, the item is an individual accession).

Pilot test. The pilot test referred to on page 6, 2nd paragraph, was carried out by the Taximetrics Laboratory, University of Colorado, under my direction, as a result of a recommendation made at an informal workshop on documentation systems for genetic resources which was held at the Department of Botany, University of Birmingham, July 1972. The recommendation was made to test the facility of the system, TAXIR (q.v.) for applicability in genetic resources centres. Two separate pilot tests were made, each specific to a single crop. One crop was potatoes, and the test was made in cooperation with three separate potato centres, each with slightly different types of data. The centres cooperating were the International Potato Centre, Lima, Peru ; The Commonwealth Potato Center in Scotland, and the U.S. Potato Collection at Sturgeon Bay, Wisconsin. Data on potatoes from the National Seed Storage Laboratory, USDA, Ft. Collins, Colorado were also included. This test has just been completed, and the data sent to each of the above-named centres for their review and recommendations. The system TAXIR met all requirements.

The second pilot test, on Zea mays was done in cooperation with CIMMYT, in Mexico City, The National Agricultural Center, Chapingo, Mexico, and The Colombian Institute of Agricultural Sciences, near Bogota. Again, the results indicated that TAXIR was satisfactory for the purposes.

Programme. This term refers to a set of directions to accomplish some task with some body of data in the computer. Frequently, "software package" is used to denote a series of programmes, each with some related set of functions. One who makes (writes, or designs) programmes is a programmer, and there are various levels of skills, such that a programmer has lower skills than a systems programmer. The more complex the tasks, the greater is the skill in programming which is required. A frequent misconception is that good programmers are mathematicians. The better programmers are those more competent in linguistics, and a good indication of a competent programmer is one who enjoys the details of grammatical construction of languages. Mathematicians are important in the programme design milieu, to design the functions in the machine and to test the necessary and sufficient conditions of a programme to guarantee that the programme will accomplish the tasks required.

Software. This is a collective term used when speaking of any (or all) computer programmes. Software contrasts with Hardware, which is any part of the machinery found in a computing centre.

TAXIR System. This acronym stands for TAXonomic Information Retrieval, and the acronym was chosen because we used as a model that discipline within biology, taxonomy, which has long been the major information system for all biological functions. The software package of TAXIR was developed by a team of biologists, mathematicians, and programmers under my direction. The software design began in 1967, and was completed in 1969. The development was supported by a grant from the National Science Foundation, in the amount of $300 000. Since 1969, various modifications and improvements have been made, with funds from a variety of sources. Today, the system runs on two different machines. The most up-to-date version is that

which is operational on an expanded CDC 6400 at the University of Colorado. A second, less-well developed TAXIR package runs on an IBM 360/65. Several different applications have been made of the system, including one at the USDA Regional Center for genetic resources at Pullman, Washington. Various other centres are testing (or using) the system, but their present status is unknown. TAXIR provides most of the features required of an IR system, and has demonstrated capability for genetic resources.

We must convert the most up-to-date version of the system to run on local hardware either in FAO or elsewhere. However, from initial indications, the machine at FAO, though of the latest IBM series, is at the lower end of the scale in terms of the needed memory size. Conversion to the IBM machine will be a major task of primary importance, because we wish to start with the most up-to-date version of TAXIR. Since all IR systems are very machine-dependent, it will take some time for a competent systems programmer to make the system operational on the FAO hardware.