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The Hunt Institute Archives contains hundreds of thousands of pages of historical content, writing and images, created by thousands of individuals connected to the botanical sciences.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time and social context in which these materials were created, some of the collections contain material that reflect outdated, biased, offensive and possibly violent views, opinions and actions.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 Archival records are historical documents, and the Hunt Institute keeps such records unaltered to maintain their integrity and to foster accountability for the actions and views of the collections'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

Royal Botanic Gardens
Kew Richmond Surrey

Telegrams Kewgar Richmond Surrey
Telephone 01-940 1171

Please reply to The Director
Your reference

Our reference

Date 3/5/74

Dear Dr. Rogers

EDP MEETING AT KEW, OCTOBER 1973

Arrangements are now in hand to publish the full account of this meeting. I am enclosing a draft account of the discussions, in which you took part. I would be grateful if you would kindly look at what you said, make any alterations you think necessary to your contribution and return it to us as soon as possible, and anyhow not later than [28th May 1974] If you do not return it by then it will be assumed that you have no alteration to make.

Yours sincerely

J P M Brenan
Keeper of the Herbarium
Deputy Director

Dr. Rogers

Discussion on Afternoon of Wednesday, 3 October

Dr P M Cowan (Introduction)

I am here acting as the representative of IAPT and convey greetings from Dr F Stafleu to this Conference. IAPT is willing to give what help it can to the establishment of an information management process to serve taxonomy. It is hoped that the computerised final part of ING will be ready for the Botanical Congress in Leningrad.

Mr R J Pankhurst

Standards are necessary for the identification of specimens especially the descriptive data. These standards also apply to non taxonomists.

[Methods (indirect) - prepared by computer but not necessary for use].

We must become accustomed to having our data as complete and consistent as possible. This is absolutely necessary for computerisation of such information. With this data it should be possible to evaluate automatic identification on a large scale. It is vital that a National Institution should be responsible for this task.

Professor J Heslop-Harrison

One of the most interesting details is the cost. How did Professor Gomez Pompa's Vera Cruz project come into being?

Professor A Gomez Pompa

It started in 1966. We were looking for a new approach in Herbarium technique and considered the use of computers. It was hoped that this would encourage interest from the University and indeed the Computer Centre of the University offered the use of its facilities and help in programming. A pilot scheme was carried out using a small part of the Herbarium, the ferns. At a Symposium on information Sciences, organised by the Smithsonian Institution, the idea of an EDP project for the Flora of Vera Cruz was put forward. It was necessary to have links with a large Herbarium and this was provided by Harvard University. They and the National University of Mexico provided funds for this scheme to be tried and as its usefulness was proved, interest in this project grew and further funds became available.

A training scheme was started for students who wished to work on the project. The computer time was about 20 hours per month.

Professor J Heslop-Harrison

What was the level of skill necessary for an operative to extract data from the labels?

Professor A Gomez Pompa

Technicians who were undergraduates were used in gathering the data from the Herbaria of Kew, Harvard and Mexico.

Dr A Hall

Technical assistants did this work in the Bolus Herbarium.

Dr J Raynal

How many specimens were dealt with and how were the names checked?

Professor A Gomez Pompa

About 30,000 specimens were examined. The identification of new material was done by myself at Mexico, by Navling at Harvard and other specialists who cooperated in the project. The older material was not checked but corrections can be made when monographers have worked on this material.

Dr D J Rogers

What kind of programmes were used and are these available?

Dr J A Toledo

The ALGOL System was used sequestial filing structure but the programme is not available for general study.

Dr R M Cowan

It is necessary that the data should be of value to other users in order that granting agencies will advance funds.

Professor A Gomez Pompa

The biological programme for F.V.C. included information which was of ecological and climatic importance.

Dr J F Mello

The cost of 50 files was approx. $ 1.57 per specimen. 23c. for the specimen and the rest for human time.

Dr D J Rogers

What was the level of technical skill required?

Dr J F Mello

A clerk/typist (Salary 7000 dollars a year)

Mr T W Davis

What machine was used in the Flora of Vera Cruz Project?

Professor A Gomez Pompa

A Burrough 6700

Mr J P M Brenan

Was there a significant amount of editing and how much of this EDP exercise could one assess as part of the normal curation operations?

Dr J F Mello

Yes, the editing was strictly a clerical aspect of the work. In the typing of the label data an average of between 50-100 specimens were dealt with per day (best 100 per hour). The use of computers to produce lists, labels etc. showed its advantages in the curatorial aspect of the work.

Dr D J Rogers

How many items of data per specimen were recorded?

Dr J F Mello

About 20.

Dr J L Cutbill

On comparing costs it was 66.6p per item using the existing system and 65.6p per item by computer. Using a commercial quote for computer time, the actual computing cost was a very small part of the total cost - about 1p per item. The question we should ask ourselves is can we afford to document our collections imperfectly.

Dr S W Greene

A part-time typist was used for the Antarctic Survey Scheme. The running costs are small - £200 per year for machine maintenance and £150 per year for computing. We are able to put in as much information as we wish once per month and can question the computer once per month. About 5 days per month are taken in dealing with input, the rest of the time is used for standard curatorial duties.

Mr R Ross

Does the running cost take account of programme writing?

Mr J L Cutbill

No, it does not include such developmental costs. These were about £30,000 for programming costs. This was likely to increase to £100,000-150,000 in the future with an estimated ½ million items being processed each year.

Mr R Ross

To what extent could we use an already developed system?

Mr J L Cutbill

The system developed by the Museums Association (I.R.G.M.A.) should be available.

Dr S G Shetler

We should decide whether the goals we wish to achieve are valuable and if so to gather the necessary funds for carrying out an EDP programme. Editorial manpower is necessary for publication of the data in order to ensure consistency. The figures which have been quoted do not tell the whole story.

Mr R S Cowan

Many of these costs are add-on costs and must not be confused with the computing costs.

Dr D J Rogers

It is wise to hire expertise to uncover costs which are unknown or overlooked. These management scientists can gain details of costings and offer advice on ways of reducing them. An example of this was their suggestion of cooperation with other government agencies in the employment of disabled people who, with a short training programme, are able to carry out dull routine duties.

Professor A Gomez Pompa

Dealing with this problem of cost, we must ask is the worth worth it. If we embark on an expensive EDP programme we should have an aim in view, for example is one of the goals of computerising the collections of the major international Herbaria, a World Flora?

Professor C H Oppenheimer

We are required to go into a computerised system in order to handle the data efficiently. The urgent need is for compatibility to facilitate the interchange of information between the systems.

Mr R S Cowan

The environmentalists must be made aware of the value of computerising this enormous amount of data. When they do, funds should become available. The question is, do we want the data which EDP can provide for the use of the monographer?

Dr J Raynal

Compatible systems are not very important but it is essential that the information should be in a standardised format.

Mr R Ross

There is a curatorial need to have a method which would enable us to quickly find all the material from a particular area, at the moment this is lacking. An agreed geographical breakdown, by means of which a list of geographic areas and their species could be made is one of the priorities. The data capture for this is simple and the scheme worthwhile.

Professor J Heslop-Harrison

I should query whether this is true. Most of the information required is to hand on opening the cover.

Dr D Brummitt

We are talking about two levels of information. Processes at the level of a taxon need have no relevance to the specimens, while there are other processes actually dealing with the specimens. In dealing with Mr Brenan's list of priorities, nos. 1, 5 & 6 are specimen processes while 2, 3, 4, 7 & 8 are at the level of taxa. I agree that we should concentrate on the problems at taxa levels. In a few years (2-3/3-4) Kew's holdings could be dealt with in the way Mr Ross suggests. It is important to deal with problems which have a definite end point and which can be accomplished in a few years.

(Dealing with the type holdings would be a very long term project)

Mr P S Green

It might be possible to ask borrowers of material to extract the input data for EDP and this could be supplied to an Institution acting as a central European data bank.

Discussion in Morning of Thursday 4 October

Mr R J Pankhurst

The danger in putting too much reliance in the human computer is that the data/ expertise is lost with the death of the individual.

Mr J F M Cannon

Is it possible to update the microfiche record when new material accumulates or new type specimens are selected?

Dr F H Perring

Such corrections can be done in our own system and periodic renewals of other systems can be prepared.

Mr J F M Cannon

Is the value of the complete renewal worthwhile?

Dr F M Perring

This is a matter of detail for each user to decide.

Mr J P M Brenan

The photographs stored in the two files are in geographic and systematic order. How are these reorganised when additions are made?

Dr F M Perring

They are cut up and rearranged. This is easy if the labelling of the specimens is carefully done.

Mr L Ryvarden

There are great advantages in this microfiche system described by Dr Perring. With special coding on the labels might it not be possible for electronic sorting

of these to be done as is done by the Post Office using postal codes?

Dr F H Perring

This is certainly a possibility.

Mr R Ross

ING provides a list of generic names but does not distinguish between accepted names and synonyms and is not, therefore, adequate for the purpose suggested although it does provide a good base. Technicians are quite capable of extracting geographical data from labels with the exception of some early 19th century handwritten labels which need more expert interpretation.

Dr F H Perring

This should not prove a major problem as we found that technicians soon became very skilled at this sort of work. Botanists must agree on a list of names to use.

Dr D Brummitt

Significance for large herbaria.

Two minor problems have already been mentioned. Recording the geographical distribution of a species - for which a generic thesaurus is unnecessary but it is necessary to have an acceptable geographic list of country names. This meeting could ensure that this was produced. There is no easy method of relating the label data to the geocode. The production of a type register by photographic means - microfiche - can be undertaken in a finite time as was done with the Wallich Herbarium at Kew. The photographic types from all Herbaria could be centrally pooled. It would only be necessary to indicate what the specimen is a type of and not what the accepted taxon is represented.

Dr F H Perring

These details can be worked out by a discussion of the interested parties. The geocode includes a book of maps and thus can be associated with label data. Such a mathematical code has advantages. A central agency could collect all the photographs and arrange them in some agreed taxonomic order.

Dr J Raynal

The Herbarium itself, like Dr. Perring's system, is an efficient data file not requiring EDP technique. The specimens, arranged in a combined taxonomic and geographical system form a two-dimensional matrix which is easy to consult. The idea of a central collection of photographic types is a good one but it will be a great task to define the types.

Professor J Hawkes

The photographic system is admittedly useful but we should not forget that EDP can answer questions of great use to taxonomists. For example what Herbaria have collections of a given collector? What Herbaria have duplicates of given specimens? What families, genera and species have been collected from a specific geographic area - countries or localities? What habitat references can be obtained from a specific locality from all the specimens collected there? Other questions which EDP could help to solve quickly are what specimens were collected before and after a specific collection and which Herbaria have isotypes etc. for whole groups. This kind of information is very useful for monographers and for future expeditions.

Dr F H Perring

A separate file could be made for collectors but I would question the importance of this information. Such information could often be read directly from the photographic record of the label. An EDP system could easily be set up from the photographic record such as I suggest.

Professor J Hawkes

Such a system would not provde information for other taxonomic groups.

Dr S W Greene

With our system it is only possible to handle 1000 specimens per year on the computer. The photographic system would provide a method of doing this more rapidly.

Professor Dr M Riedl

While Professor Hawkes's suggestion is desirable, an Institution with small staff numbers and limited finances has to be realistic about what it can do.

Dr D L Rogers

We should examine how long a time it takes for the physical handling of the specimens for making the microfiche.

Dr D Brummitt

The Wallich Herbarium containing tens of thousands of specimens was done in approximately 6 weeks.

Dr D L Rogers

It is vital to make an accurate and detailed comparison of the times of different operations together with a comparison of costs. One must remember that even after the microfiche is made, it is necessary to extract the information from this photographic record.

Dr J Raynal

There is an advantage in having photographs because it avoids the handling of specimens.

Dr S G Shetler

This photographic record only creates yet another herbarium. The problem we face is indexing the data in our collections.

Dr G Panigrahi

The collections of type photographs would be very useful to developing countries. These often had large collections but few authenticated or type species and the photographs would assist in the writing of Floras. Part of the cost could be paid by the purchasing country.

Dr J F Mello

Could Dr Cutbill define what he means by standards?

Mr J L Cutbill

It is a set of tests applied to data (to test the test)

Dr J F Mello

What about standards for colour?

Mr J L Cutbill

We (?IRGMA) have no responsibility for this.

Professor A Gomez Pompa

Have you any standards for localities?

Mr J L Cutbill

Political names, longitude and latitude and grid references are used but latitude and longitude are recommended.

Professor Dr M Riedl

Can you apply your standards to old collections which have only poor data? Is there a means of translating scant data.

Mr J L Cutbill

This presents no difficulty; a low information point is indicated in the appropriate places.

Mr J P M Brenan

With interdisciplinary standards might it be necessary to develop this further?

Mr J L Cutbill

At the level at which standards exist there is little difference between the disciplines as they all have descriptive problems in common. Each thesaurus of key words may certainly be different.

Professor C Kalkman

The standards you describe are common to all international museums. In drawing them up who was responsible, curators of local museums, staff from national institutions or scientists?

Mr J L Cutbill

The initiative came principally from staff dealing with the humanities, mainly academics within IRGMA. It arose because of their awareness of their inability to answer questions from existing data.

Professor C Kalkman

I find it difficult to see any value in setting up these standards.

Dr D L Rogers

We seem to be getting confused between standards for substantive data and a structure which makes this data into a generally useful system.

Mr J L Cutbill

The system was evolved because a general filing system was required for inter-disciplinary museums and existing systems were inadequate.

Dr D B or Dr T Williams

One problem is that we cannot predict in advance all the data which might be needed.

Professor Dr K Walther

There is a system available in which all colours are translated into figures; would not this provide a standard?

Mr J L Cutbill

Yes, if everyone accepts it.

Discussion on Afternoon of Thursday, 4 October

Mr B J Harwood

A system is wanted which will provide the taxonomists with the information they require. It is impossible to have a system which contains all the information for all requirements, therefore we must have one with some constraints. The standards will be the key words, descriptors and definition of the record. The system must be relatively simple in order that it may be understood by the systems analyst who will be required to put it right. It must be easy to maintain, modify and amend. The cost benefit must be in proportion to its usefulness. An international system will have both a language barrier in verbal communication and in the EDP programmes. The programmes must be compatible between terminals in order that there can be a full interchange of data, for example, by magnetic tapes. I wonder whether the system should be use oriented, where the user can state what the system should be capable of

or whether this should be the responsibility of the systems analyst?
How easy is it to maintain SELGEM and TAXIR? How long does it take to understand either system?

Dr D L Rogers

TAXIR is no more complex than any comparable system. It is impossible to answer the second question because it depends upon so many variables.

Dr J F Mello

ISD 70,000 dollars per year with maintenance and development of new programmes. It is possible to learn how to get SELGEM programmes operational without a great deal of expense.

Mr J L Cutbill

It takes us (?IRGMA) 3 months per year maintaining the programme (getting the bugs out).

Dr S G Shetler

We should look at the character of the available systems and assess them in how applicable they are to our requirements. These systems must be closely linked to the means of providing the data for them.

Professor C H Oppenheimer

With our Environmental data system (10,000 per year ENVIR, 64 system) it is possible to learn to use it in a few weeks. Students, with some knowledge of computers can use it easily and we have been able to teach visitors to use it very quickly because the system is easily understood - in the English language. It is very versatile in obtaining different types of information.

Mr J P M Brenan

Are there problems concerned with the interchange of information between TAXIR and SELGEM systems?

Dr D L Rogers

Each system has a different structure but there is no problem in writing a conversion programme which overcomes this difficulty.

Dr D M or Dr T Williams

How does the system handle data which contains groups of repeated entries for example a reference with author, data and title or identifications and identifier?

Dr J F Mello

There is an inability to handle this hierarchical data.

Dr D L Rogers

One way of dealing with this problem is to make as many data banks as needed. A merging command deals with this.

Information Management

and Use of TAXIR in Herbaria

David J. Rogers
Professor of Biology
Department of EPO Biology
University of Colorado

The purposes of this presentation are twofold: 1) to discuss the problems of management of information in herbaria and 2) to describe TAXIR (Taxonomic Information Retrieval) as a software package which may be used for the purposes of information management. EDP is a term which is so all-inclusive that it is necessary to restrict the definition for our purposes. The objectives of this meeting (as I understand them) are (1) to discover whether or not electronic data processing methods should be employed for storage and retrieval, and (2) to discover if there be any standards applied to (mostly) label information from herbarium specimens in various herbaria in Europe.

1. Information Management

It is useful to imbed the concepts of EDP in a framework, to know precisely the level of discussion at any particular time. The diagrammatic sketch and the concept presented below is by no means new, but one that is fairly common. We have labelled the concept as the "knowledge triangle" (Figure 1). The triangle has three essential sections. The base area of the triangle contains data and that part of EDP applicable to proper handling of these data is generally referred to as information storage and

retrieval. In this portion of the triangle are found all of the measurements or descriptions of the "real world" (in this case, specifically the collections in the herbaria). Data may describe anything about the organism or related to the organism represented by the specimen or collection in the herbarium. "Label data" frequently contain types of information that identifies the individual specimen. We are all aware that the herbarium, or label, name is not necessarily the most accurate name for the organism. However, it does indeed represent some means for identification of that individual specimen. There are other data on the herbarium label which have the same function, for example, the name of the collector and the collector's number further specify which particular specimen we refer to. In addition to these types of information on the label, frequently are found the dates of collection, the geographic locality of the specimen, and on more recent materials particularly, various habitat descriptions. We may also find notes about the use of the plants. Although the data on the specimens may indeed be erroneous in one part or another, we still use the data from the specimen as a means to locate and specify that particular specimen. We will also, in the normal work of a taxonomist, use these data as we process information to produce floras or monographs, and we frequently modify some of the information contained on the label, but we have adopted rather well-defined rules to make such modifications. The initial data on a label is

never changed - rather we annotate the specimen using separate labels, and many of you are familiar with specimens that have more annotations that there is plant material.

In the knowledge triangle, the next third of the triangle is labelled "information." In other words, we move from data through the triangle to information - correlating the data by various techniques and in the process of correlation, use other electronic data processing systems to aid in these types of correlations. Many statistical analyses fall into the category of data correlation and the results of the statistical analyses provide the second part of the triangle, namely, information. Thirdly, knowledge is at the top of the triangle. By adroit use of both data and information, we eventually produce knowledge and in the particular case of the taxonomist, knowledge resides in the taxonomic reports, whether a note, a monograph, or some floristic study. In the production of knowledge from data and information, we employ another set of EDP methods, frequently referred to in the computing milieu as "clustering techniques." In these techniques (of which there are several) the various pieces of information (referred to generally as characters) form the basis for clustering the specimens into taxa, and from the technique, we derive conclusions about the organisms, the classification thereof, and the hierarchy of taxa contained in our study.

Each step in the study of a specimen requires a special technique for use of the data or the information and each of these steps requires some management technique. Not only in the actual scientific study of the specimens do we concern ourselves with management, but

we may also consider that the curator of the collection is also, in a sense, a manager. He is required to store his specimens in such a manner that he can readily retrieve them, for whatever purpose. As the manager, the curator must oversee all of the necessary functions to be certain that his institution serves the best purposes for which the organization was established and which it serves today. The number of different ways in which an herbarium and its contained specimens serves are essentially limitless. Each day brings some new application of the specimens in an herbarium collection. There is no predicting today what values any one specimen or any one collection of specimens may have for some future work. About all we can predict is that somebody will find new uses for our materials that we have been so cautious, so careful, to maintain.

In an ideal herbarium every specimen would be well identified, with no uncertainty. The realities, however, are so different that we all hesitate to expose our collections to anybody but our professional colleagues. However, if we consider that whatever information exists on the specimen, even though it be written in some very difficult script, is a means to identify the specimen, we are able to use the specimens for various purposes. The older herbarium specimens clearly are the least useful for ecological data; they also may have very poor geographic data. The many variations in kinds of data and means by which the data were recorded (or ommitted) on the specimens is so familiar to this audience that it

needs no repetition. Nevertheless, with whatever variations that exist, we do generally manage to find the necessary specimens. If we cannot do so immediately, we usually correct mistakes at some later date and continue to improve the knowledge and information content of our herbarium.

What could be done if there were sufficient funds, to assist in the process of transferring all the data from the herbarium labels to some device which would use the many different facets of information which reside on the label, which would aid, not only for the purposes of taxonomy, but for the many allied purposes which we know our specimens could provide, given the best system available? Nobody in this audience need be told how extremely difficult it is to get additional funds for any purpose in any herbarium. I believe, however, that by careful analysis of our problems and by careful management techniques, we can indeed provide the herbarium curator with some assistance from electronic data processing techniques, specifically, information management systems. First of all, in management, we must choose a set of priorities for our most important work: to which activities we will devote most of our time; which of our collections contain most valuable material; what parts of the world are most significant for our individual herbarium functions? Priorities are already established on these bases, and we spend most of our time in providing certain kinds of information concerning these priorities. The organizational

priorities between herbaria, although informal, are sufficiently well-defined for different herbaria to carry on major functions without overlapping the works of others. For example, the priorities between the collections of Kew and the Natural History Museum had been established, and a management decision made in which it was decided that certain different functions and certain different kinds of collections would be emphasized by each of the two institutions. Likewise, other herbaria know that Kew and the Natural History Museum provide these functions and therefore do not try to duplicate those specified. There were informal management decisions, and a set of priorities established, indicating that curators of collections are, indeed, managers.

As we move from the present methods of data recording into techniques in the electronic age, we must exercise great care, in order to achieve the most effective, least costly procedures. At this point, it would seem most efficient to employ persons whose specific training is in the field of management science. In management science students are trained to work in large organizations to achieve the most efficient function in the most effective and least costly manner, since the functions and data of herbaria are as complex as any organization in the world. It seems reasonable to me to employ a management scientist as a consultant at the beginning of our conversion to EDP methods.

Last in this discussion on management, it is necessary for those of you who have not devoted any effort to find out about electronic

data processing techniques, to have someone who can help sort out the many different processes and means by which EDP is accomplished. You have heard here today, and in previous discussions, many different claims for different EDP methodologies. It is necessary to discover which of these processes is most appropriate to the needs of the collections, of the data, and of the objectives and priorities of the herbarium curator. Again, the management scientist can be a great help in these functions.

2. Description of TAXIR

The Taximetrics Laboratory, under my direction, has had a long history of development of efficient computer-aided methods to assist in various aspects of the taxonomic process. In 1966 we became aware that one of the most impressive needs for the taxonomist was some means to handle most efficiently the enormous data loads, including the specimens in herbaria. In 1967 we began to develop an information retrieval system to incorporate data in any of the different forms in which the data are expressed - either alphabetically, numerically, or combinations of alphabet and numbers; to organize these data to retrieve any subset or combination; to add, delete, or change any data. We wanted the programs (software) to allow us to question the data in the most meaningful ways to the taxonomist. In addition to the requirement from the user's standpoint, which would include the ability to use his own language and his own terminology as freely and as openly as possible, we also required that the most efficient means of the use of the computing machine would be a part of the design of our system. We designed the means of

storage and the means of retrieval of the information to take advantage of the capacity of the computing machine to carry out calculations, as described in one of our publications (Estabrook and Brill, 1969). This is but one example of our efforts to increase efficiency of use of the hardware.

I have summarized some of the major attributes of the TAXIR system in an appendix to this paper. I have also included a very short glossary which provides definitions of selected terms that are not generally well known in the herbarium milieu.

Using TAXIR for herbarium applications. Having described the software package of TAXIR in a cursory manner, it is important to give an example of the TAXIR system application. The most important element to consider in the use of this, or any other computerized information retrieval system, is the division of the sets of data which we wish to capture into logical groups. We use a data bank to divide the data into logical subsets. Even with the most powerful machines with the largest memory units, it is necessary to limit the size of data banks to use the computing capacities at their greatest effectiveness and at the least cost. In large herbaria one could think of dividing data banks into two basic types: the one of most significance to this audience may be designated "curatorial" data banks, containing only the label information; the other type, containing data describing the plant material, does not concern us at the moment. Separate data banks may follow the logic already established in the herbariums: monocots separately from the

dicots, and within these two major divisions, further division into appropriate size of data banks according to some classificatory scheme. For example, all of the palms or all of the grasses might form the basis of a single data bank.

In the TAXIR system it is possible to merge different data banks if needed. For example, if certain types of information were common for several data banks (such as collector name, localities, etc.), the TAXIR system can merge data from separate banks and put together the required information in a single data bank, to be used in the questions on the merged system. By dividing our data banks into logical sets we are always able to keep items in some manageable size and by the same token, have available all the information upon call when necessary. Other kinds of data banks could be considered as required, either by the herbarium curator or by one of the taxonomists working on a specific problem in the herbarium. If one of the taxonomists, for example, specializes on one particular group of plants it would be possible for him to build his own data banks which could later be added to the general store of information for the whole institution. By using the TAXIR system in conjunction with all of the different functions of each of the curators (or visiting taxonomists) in the herbarium, we would be able to more rapidly capture the data from all of the herbarium. With careful management, data capture would proceed more rapidly than would be anticipated by a single calculation, (i.e., it would take one hundred man-years to capture all of the data from 4 million specimens

at Kew). Such a prediction has very little value because we have not thought about the means by which we presently capture data and recognize that many workers are involved in a collective effort.

After consideration of the structure of our data banks themselves, there are clearly additional types of requirements to be certain that we will proceed in the most orderly manner for the efficient use of whatever software package we may choose to operate in our data bank. One of the needed features is a careful standardization to be certain that all collections are uniformly treated. In this connection, the discussion by Dr. John Cutbill of Cambridge who has spent a considerable amount of time on the problems of standardization of museum materials, is extremely meaningful. Careful analysis of the means Dr. Cutbill has designed will be beneficial to the curator's of herbaria as they begin the process of deciding the types of data which would be included in the data banks.

Once again, I should emphasize that in the considerations that must be made here, we must clearly make a separation between the substantive content of the data that may be found on the labels of herbarium specimens and the structure of these data. By this I mean the following: we recognize that many of the labels on herbarium specimens are inadequate, that the names given on the specimens may be wrong, and that there is generally a great problem with interpretation of cryptic symbols, etc., but we still use whatever is presented to find things. Placing the information, no matter how faulty or sketchy, into a system using good management techniques

permits use of the specimen and at the same time, much more rapid correction and improvement. Again, considering the structure of the data, we use these types of information - incorrect though they may be - as means to find any specimen. Corrections or additions of data on herbarium specimens can be made much more efficiently when all of the data have been placed into some orderly data bank with some orderly use of an information management system.

Interchange Between Different Software Systems. We have heard presentations describing several different software packages that are used in various parts of the world. I do not wish to say that my system is better or worse than any of these other systems. Each of the systems is designed with specific purposes in mind and with specific attempts to solve problems. It is necessary to discover which of these systems is most appropriate in any particular setting. However, we must be aware that it is not impossible to employ different software systems with different computing machines and still share the collective data in the various systems by means of careful conversion programs. Each system discussed has a special format design by which the data are put into the computing machine. If we know the format design of any one system, we can convert from that particular format to the format of another system such that the data then becomes "compatible." Conversion routines can be generalized to the extent that they run extremely rapidly and very efficiently and with very little cost. Eventually, after some application of differing systems, it will be possible to compare the systems on an objective basis and after the comparisons have been made,

then, perhaps, will be the time for the adoption of one system over another. But my thesis is that we should not at the moment restrict ourselves to any one system, but test several general systems, each institution using that software package most easily available to him. Eventually we will discover the common needs by this means rather than by trying to demand that all people conform to the same software system at the beginning. It is much more efficient and much less costly to progress as we are at the moment with many different systems running. We will be in a more secure situation to decide which system accomplishes our needs after a period of experimentation. Information retrieval systems which you have heard described here are not over six or seven years old and most of them are much younger than that. At this stage in the development of the powers and capacities of the computing machines for the uses in information management we are still in a development phase. We have efficient systems, to be sure, but none of the systems can be called perfect. In this light, this is a scientific endeavor. We expect some exciting developments in terms of both computing capacity and in terms of software efficiencies. We believe that the herbarium contains the most valuable sets of information describing our environment that exist in the world. It is time to put the full information contained in the herbarium to work and this can only be accomplished using good management practices and efficient EDP methods.

Reference

Estabrook, G. and Brill, R., 1969. The Theory of the TAXIR Accessioner. J. Math. Biosci., 5:327-340.

Appendix I

Some attributes of TAXIR

-- TAXIR is an information retrieval compiler. That is, with the sets of instructions included in the program, differing data banks may be input to the computing system and from these, a specified information retrieval system designed by simple, common language instructions, with associated vocabulary, are achieved.

-- TAXIR is designed in modular fashion, with several "add-on" subroutines for various requirements. Examples of add-on routines are: plotter programs to make maps and graphs; report generators to print columnar data with headings; editing routines; and various statistical packages (at the moment, the statistical packages attached were designed by the Institute of Behavioral Sciences at the University of Colorado).

-- TAXIR is written in FORTRAN IV, the most common compiler language found in the world. Because of the exclusive use of FORTRAN IV, it is relatively easy to convert from computing machines of one manufacturer to computing machines of another manufacturer (the only requirement being that the machine possess a FORTRAN IV compiler). At present, TAXIR is running on Control Data Corporation machines (6400 and 6600), on International Business Machines (several 360 and 370 series) and on UNIVAC 1106 and 1108. Because of the modular design,

it is possible to use smaller computing systems than those mentioned above, such as the IBM 1130, or the IBM 360/20 models. The only requirement for the smaller machines is that they have associated random-access peripheral storage devices.

-- The TAXIR version on the University of Colorado CDC 6400 is adapted to run on the inter-active, time-share system, from remote terminals. It also runs on "batch" mode for long runs with extensive print-out to take advantage of lower cost operations.

-- The TAXIR system achieves extraordinarily efficient storage and retrieval capability by employing as a base of design set-theoretic functions which guarantee the most efficient use of storage capacity of the computer, and at the same time provide extremely rapid access to the stored information.

-- The TAXIR system alone occupies about 15 K 32 bit words of memory, and sizeable data banks can be contained within 32 K memory. Large data banks take advantage of drum, disc and/or tape peripheral memory to extend the size of the banks almost indefinitely. The largest data bank ever tested on TAXIR had 106,000 items, with 50 descriptors per item. Time to retrieve a single item in answer to a complex question from the bank with 106,000 items in storage required ca. 2 seconds of central processor time.

-- Any modern language using the Latin alphabet may be used as input to TAXIR with rapid, accurate translation to other languages possible.

-- TAXIR does not require a pre-determined thesaurus of terms. Because of the design of data input as descriptor/descriptor state, the descriptions of each item becomes the language by which the data bank may be questioned.

-- Items, descriptors and/or descriptor states may be added, deleted, or corrected with a single instruction that is contained in the TAXIR compiler.

-- Data banks may be merged or reformatted under control of the TAXIR compiler.

-- TAXIR will (within the next 3 months) be completely flow-charted and documented, and a user's manual available. Complete flow-charts and documentation, along with listings, and the user's manual will be available at cost. Since the system was built with public funds, the system is in the public domain. No request from a serious user will be turned down. It is requested that any user who adopts the system become a member of a "user's group," and that any further development of the system made by any user be shared with the developers of the system. (This is, of course, not enforceable by any means, but it is anticipated that the spirit of free scientific exchange will prevail.)

Appendix II

DEFINITION OF SOME COMMONLY ENCOUNTERED TERMS IN COMPUTERIZED INFORMATION RETRIEVAL. (For a much more complete glossary, see: Vocabulary for Information Processing, published by American National Standards Institute, and/or CDP Review Manual. A Data Processing Handbook,

Eds.: R. A. MacGowan and R. Henderson. Auerbach Publishers, Princeton, N. J., 632 pp.).

batch mode. (contrast inter-active). When using the computer, program and data banks are submitted to computer to be run at the convenience of the computer center. There are no possibilities to alter the program, nor to ask further questions during the computer processing of the data. Batch mode is much less expensive to use, and is most often employed when large or long printouts are expected.

character. In the computing milieu, a single letter or numeric symbol.

compiler. A software package which converts a set of instructions (program) to machine language.

data bank. A collection of items (a set) with associated descriptors and descriptor states (q.v.). The proper design of a data bank is very critical in cost-effective use of any computer-aided information retrieval system.

descriptor. In TAXIR, a single basis for description of an item. Example: collector, collector number, generic name, species name, country where collected, flower color, date of collection. etc., etc.

descriptor state. In TAXIR, a series of non-overlapping (mutually exclusive) descriptive statements (values) under each descriptor for each item (q.v.). Examples:

Descriptor	Descriptor State
Flower Color	red white blue red-blue
Leaf Length	10 cm. 11 cm. 15 cm.
Collector	Smith, R. Smith, J. Rogers, D. Rogers, W.

Note that the combination descriptor, descriptor-state conforms to the same construction as genus (noun), species (adjective).

Note also that descriptor states may be alphabetic, numeric, or combinations of these two. In TAXIR, there is no limit to the number of either descriptors per item, nor descriptor states per descriptor. The length (number of letters or numbers) per descriptor state is presently set at 90, but may be lengthened if need be.

In TAXIR, descriptors may be names, coded, or ordered (from-to), to give complete flexibility in description of the items.

field. With reference to the number of letters or numbers used in any descriptor state, the place on the punch card (paper or magnetic tape) where one places data.

a. fixed-field. A pre-determined number of spaces allotted on the punched card (frequently used in coded data).

b. free-field. Within limits, any number of spaces allotted on the punched card. This is a feature of the TAXIR system.

hardware. (contrast software). All the physical components of a

computing machine, including input and output devices, central processing unit, storage units, cathode ray tubes, etc.

inter-active. Any set of software and hardware computer configurations which permits the user to ask questions and receive answers sequentially, without resubmitting his program and data each time a new question or direction is submitted.

input. Any method of data preparation for computer manipulation and the machines that accept data in the computer.

item. In TAXIR, anything or concept which may be defined as a basis for description, i.e., a specimen or a taxon.

"K" (as in 15 K). Shorthand, or jargon, in the computer milieu, standing for 1,000. Thus, 15 K = 15,000. Refers generally to the size of memory in any computing machine.

on-line. When a user employs a computer with a remote terminal or time-share capabilities, he is said to be "on-line."

output. Any means of presentation of the results of computer manipulation of some input under control of the computer program--may be "hard copy" (typical computer print-out), a display on a cathode ray tube, or microfilm.

program. (see also software). A set of instructions that direct the function of the computing machine to accomplish some task or set of related tasks.

remote terminal. A piece of hardware connected at some distance from the computer via telephone line or microwave transmission. A means by which a computer user may communicate with the computer without having to visit the computing center. There are many levels of complexity of remote terminals.

software. (see also hardware). A generic term that includes all types of programs that direct the functions of a computing system.

time-share. The more sophisticated computing systems provide means by which several users may have programs running nearly simultaneously in a single computer.