The Botanist and the Computer

Modern taxonomic botany has changed little from its descriptive origins in the 18th century. With the modern understanding of the plant as a living system and its place in the larger system around it, our descriptive efforts have changed to place the plant in context and not simply consider it as a dried specimen on an isolated herbarium sheet thousands of miles from its origin. In spite of this, however, our science has remained largely a descriptive one, and perhaps it is the pressure of our own needs, plus the pressure from our academic companions in other fields, which has caused us to look to a more mathematical approach to botany. The advent of the computer was a natural beginning to numerical processing of taxonomic data. From initial enthusiasm about ten years ago for numerical taxonomy [1] as the cure-all for taxonomic problems, we have now settled down into a more reasonable and wider use of computers in the field of botany. For many botanists, this early enthusiasm was entered into without a thorough understanding of the tool which made it possible. Today, with 30 years of computer experience behind us and perhaps some ten years of botanical enthusiasm in the same area, we are in a position to look at the equipment available and to evaluate the potential and the pitfalls of the computer as a new tool in the botanist's laboratory. We must, however, recognize that the computer is only a tool and not in itself a solution to the problems of botany. The computer can be used by the botanist only to the extent that he is logical or mathematical in his approach to problem solving. The past literature of botany, which is the data bank of the science, has already grown to such an extent that it is a nearly impossible task to extract from it all the pertinent facts relating to any given taxon. With the number of botanists working today, the situation will continue to worsen unless we start now to logically orient our facts and to place this logical assemblage into a machine-readable form which will allow us to search it and process it as the needs of the individual scientist require.

In recent years there has been an increasing quantity of literature built up relative to the use of computers in the broad field of biology. It is not the purpose of this paper to review the applications of computers to specific problems. This has been done

in part by Crovello and MacDonald in their index of EDP-IR projects in systematics,[2] and also has been covered to a great extent by the papers reported in the *Biological Journal of the Linnean Society* in September 1971.[3] It is the purpose of this paper to present some of the basics of computer operations and the effects they must have upon the approach to botanical problems if proper use is to be made of this new tool. Those persons who have not taken up the new world of computers often continue to shy away from, and in some cases even to fear, the unknown. Many others who have become absorbed in this new technology have often gone to the other extreme and completely masked the products of their efforts in a foreign jargon which has served only to increase the gap between the users and the non-users. This paper is an attempt to both slay the dragons and to bring the prophets back to earth for the greater number of botanists who have remained until now on neutral ground. The central theme of this discussion is that the computer is simply a tool, and in no way is it an end unto itself. The problems to which this modern tool may be applied are not unique to modern thinking, nor does the use of this equipment mean, of necessity, the abandonment of familiar procedures and techniques. On the other hand, a proper comprehension of the potential use of the new tool can lead to entirely new approaches to old problems.

Perhaps it is best to start by looking at what a computer is and what it can do in the abstract. In its most fundamental form, a computer is a machine that can execute simple logical commands. It can be made to compare two sets of data and, based upon the result, can embark upon one or another of a set of previously determined courses of action. In its simplest form, the computer is not a mathematical machine, but it is as a consequence of specific series of logical operations within the machine that various mathematical operations are carried out. Computers in themselves exhibit no intelligence, and it is unfortunate that the anthropomorphic view of these machines has become the popular one. Problems cannot be solved nor questions answered except to the extent that a previously anticipated answer or course of action has been selected as the correct solution based upon a logical and previously determined analysis of the data in hand. In a perhaps over-simplified sense, what computers are best suited to do is to carry out a relatively simple manipulation in a repetitive fashion upon a large body of data. Alternatively, they can also be utilized to process a fairly complex series of operations on a relatively small quantity of data,

but since the operations must all be previously carried out by hand (*i.e.*, programmed), a decision must be made as to whether it is worthwhile to spend the programming effort necessary for a relatively low degree of utilization of the computer.

An early and continuing area of misunderstanding has been the supposed numerical nature of computers. Whereas most computers have been designed for the solution of numerical problems, there is nothing inherent in the nature of the machine which prevents representation within it of alphabetical characters. The concept of numeriClature as opposed to nomenclature is both unnecessary and undesirable. The early mechanical limits of fixed field from a punch card origin, and of numeriClature as a means to conserve storage space within the computer by codifying otherwise lengthier conventional language statements, have been two facets of the early use of computer techniques which have imposed unnecessary and unfortunate procrustean limits on data processing in botany. These initial difficulties, augmented by an unnecessary jargon and a number of unsuccessful projects initiated primarily in the name of fashion, have led to a slow start in the use of computers within the field of botany.

Before going on to some of the possible advantages of computer utilization in the botanical area of activity, it is perhaps best to digress momentarily in order to discuss a matter which is basic not only to the utilization of computers but in many ways to science itself. Consideration of the broad spectrum of biological problems in light of possible solutions by computer means is, in effect, a consideration of the design of experiments. Perhaps one of the most critical areas in botany in this regard is the general field of taxonomy. Classical taxonomy has, since the time of Linnaeus, been rather in the nature of an art, in spite of all protestations as to its scientific status. The questioning of such status comes not from the end result which has stood the test of time well, but from the manner in which the decision is arrived at. In the physical sciences, an experiment can theoretically be conducted by anyone if the conditions and the equipment involved in the experiment are stated, and if this set of conditions and equipment can be duplicated. Given the same data and the same procedures, the same results should be obtained. What has been lacking in botanical taxonomy and, in fact, in taxonomy in general, has been the specific designation of the parameters involved in any given "experiment". It was precisely the rigidity of such re-

quirements which led to the early misunderstanding, apprehension, and derogatory comment upon the use of computers in botanical taxonomy. It is not, however, the computer which is in question in such circumstances, but the statement of the experiment itself. If the taxonomist is not able to precisely define his terms and with equal precision to define the steps by which he arrives at a given conclusion, then he is unable to describe the experiment. If such parameters cannot be precisely determined and defined, then it is not possible for other individuals to consistently arrive at the same conclusion, given the same starting point. Recent work in the computer construction of identification keys and in random access identification queries in on-line computer systems have shown that the unforgiving taskmaster, the computer, can considerably simplify the task of identification if the logical processes by which a specific identification is reached are rigidly stated.[4]

The complexity and redundancy of the human mental process, while it is yet to be mechanically duplicated, can be more fully appreciated today than in the recent past. The capability of visualization without apparent quantitization was apparently unique to the human mind. It is now certain, however, that the neuron of the human brain is comparable to the single flip-flop of the computer, but both are meaningless except as they participate in the larger context of the total machine system. The process of learning in the human being is directly comparable, in part, to the building of a data bank in the computer. Furthermore, the learning process consists of considerable reprogramming which, as anyone who has worked with large data banks finds, is also a necessity as one develops the uses of the stored information.

In many fields, such as computer design and chemical engineering, considerable effort has been expended in utilizing the computer both as a design tool and as an automatic means of control. In many cases where the problem has been reduced to a machine-soluble form, it has been the repetitive nature of the task which has lent itself to successful computer application. In other instances, however, the rigid test of logical statement that must be met before the problem can be reduced to computer solution has often times been sufficient in itself. That is to say, that having reduced the problem to a logical statement, it was no longer a problem and could be solved without the use of expensive computer equipment. It is precisely this rigid evaluation of logical processes with which the botanist is now faced, and it is in this context that the problems of botany must

be stated. A careful distinction must be made between the problems and their potential solutions utilizing current technology. Too often the question is asked, "How can I use the computer to solve my problems?", in situations where the problems themselves have not been stated. While it cannot be denied that some of the solutions to botanical problems, particularly those involved in the curating of large collections or extensive literature search, are possible only because of the potential use of computer technology, it should not be presumed *a priori* that such problems will only have their ultimate solution through the application of computer systems. If the botanist can develop the ability to ask himself what he wants from his data without burdening himself with the seeming limitations of his present technology, the solution can ultimately result in considerably expanded horizons. This has been particularly significant in the statistical handling of biological data. Many of the techniques of such statistical evaluation, some of which have been unfortunately labeled as a sub-science of numerical taxonomy, are really only mathematical manipulations of data which could be done equally well with pencil and paper, but unfortunately could not be carried out in that way on a large data set within a reasonable period of time. The computer as a tool has allowed the statistical technique to expand, but statistics and computers are not synonymous.

Whereas the advent of computers has offered a wider horizon to the botanist for the potential solution of previously unsolvable problems, this potential, in itself, is not the only basis upon which a decision for computerization can be made. A realistic evaluation of the utility of any information retrieval system or statistical evaluation must be made on a basis of the value of the net result without consideration of the means of possible solution. If the end result is justifiable, only then can the means of reaching that end be evaluated. The distinction must be clearly made between the organization of knowledge and the mechanization of that knowledge. In this respect, a realistic evaluation of time and cost for the solution to any given problem must be considered. The high internal operating speeds of modern computers are impressive statistics. The fact that data can be manipulated within the computer in fractions of a second is not a realistic statistic when evaluating the true time from presentation of the data to the availability of the solution to a given problem. The true solution time includes all that time which is involved in the entry of the raw data, all of the delays and waiting periods in submitting the problem to the

computer (and the programmer), and in ultimately receiving the printout or other output from the computer. In a practical sense, what is often spoken of as microseconds of response can, in fact, represent weeks of waiting time for the individual botanist. In a similar manner, costs must also be evaluated. Can the expense of direct communication with a computer data bank be justified for information which is needed for a paper to be presented next month? The identification of an unknown plant by carrying out a question and answer session at a computer console is an impressive application of the computer, and, in fact, the random sequence in which the pertinent characteristics may in some cases be presented to a computer is a considerable improvement over the dichotomous key. The questions which must be realistically asked, however, are, "How much does it cost?" and "Is it worth it?" In small systems operating in small computers, the cost may seem reasonable, but when larger data banks are involved and consequently larger computer systems, the cost can quickly get out of hand. For direct communication with the computer, the data which are being referenced must be kept in a readily accessible storage medium. Even the least expensive of these can cost in the hundreds of dollars per day, while the most sophisticated high speed systems may run into the thousands of dollars per day for a data bank equivalent to the information content of a one-volume printed flora. As one who pays a rather healthy monthly computer bill for the operation of my own research projects, I have strong feelings that an unrealistic approach to botanical problems is presented in instances where major data banks are proposed on a large-scale basis without realistically evaluating the true costs of immediate availability. Realistic operating times and ultimate costs must be evaluated in the context of the true need of the user. This aspect of computer utilization has become most critical in recent times with the introduction of on-line terminals and time-shared computers. Can the individual scientist really justify the true costs of direct communication to a computer data bank as compared to looking up a similar piece of information in a more conventional reference work? This question must be asked each time one goes to a computer for the solution to a problem.

Oftentimes the on-line system is considered justification because of the supposedly random nature of the access need of the user. Critically evaluated, such access is too often not truly random but is limited to a finite number of rearrangements of the data. It must be remembered that the computer has not

only changed the mode of communication in terms of on-line access to stored data banks but has also greatly amplified the communication which is possible with the conventional printed word through the availability of permuted indices. By using computers for the preparation of multiple indices to a given reference work, the same ease of access that is available in an on-line operation can be made available in a printed report. In our own work, the preparation of a complete title index for a file of bibliographical references having more than 25,000 entries took only three minutes of computer running time. Actually, it took closer to a week in terms of response time from conception of the need to ultimate receipt of the print-out. Having once prepared this permutation of the data, however, it is now available on a direct access basis at a speed equal to that of an on-line computer inquiry, and, I might add, at a considerably lower cost of operation. When truly evaluated in terms of the total context of programming and equipment cost versus the time and effort required to manually look up a set of facts in a computer-prepared index to a data set, the justification for the use of on-line computer systems is difficult. The index which we prepared could not have practically been completed by other than computer means, but the off-line mode of operation was sufficient for our needs and continues to be so.

The preparation of multiple indices by computer means brings out still another aspect of the need for proper preparation and analysis of problems before entering into the computer operation. While the computer is particularly well adapted to searching large quantities of data for a particular set of facts, this search is made by comparing the information content of the body of data being searched against a standard which is the item being searched for. Comparison of two items to determine whether or not they are completely alike in every detail is a characteristic ability which is inherent in the computer hardware. Comprehension of the significance of two items which are functionally the same but which are stated in a different manner is not a built-in ability of computers. Such comprehension must be programmed and is one of the most difficult types of logic to program.

Thus, it becomes essential in preparing information for computer analysis that such information be presented in an absolutely consistent form in all cases. Such simple variations as an extra space or an incorrectly placed period are as different to the computer comparison as an entirely different word. If such variations are to be disregarded, the instructions to dis-

regard them must be specifically programmed into the computer. The solution to this problem is a simple one: consistency. This, however, is not an easy thing to achieve, particularly in an endeavor which extends over a long period of time. The preparation of indices requires the searching out and bringing together of all of a like item. Achieving this without programming comprehension into the computer can be done to a great extent by specifically identifying each particular type of information in each record. This "data language" is the responsibility of the individual botanist, and consequently data banks have the disadvantage, unlike museum specimens, that the information which is stored in the computer is already biased upon entry. If, however, the data are entered into the file in a consistent format and adequate recognition is given to the specific information units within the data and the potential extent of any piece of data, then for any given processing of the data the entire record need not be searched. Only that segment of the record which contains the specific information being sought must be processed. In more every-day terms, this simply means that the format in which data are presented to the computer must be rigidly adhered to, and that considerable attention must be given to this format before any data recording is begun. Only in this way will maximum utility of the data bank be possible for future needs.

Another problem facing the individual botanist when using computers is that of compatibility. Actually, this involves three separate and distinct problems in data processing. Machine compatibility is the ability to enter data automatically into the computer file. In its most acceptable form, the data should be able to be entered into the file without the need for human intervention. Recording the shape of the leaf scanned by a computer input device is, for all practicality, still a dream of the future. The more practical entering of numerical data from mark-sensed cards is a direct machine entry procedure. The capability of entering literature files, such as *Index Kewensis*, by means of direct optical character reading is a rapidly developing technique of the present. For the taxonomic botanist, however, most of the data which he will need in the computer file must be entered manually, and only after this manual data processing will a machine-compatible form of the file exist. The second type of compatibility is that of the data bank itself. While there are various formats for internal storage use in different computer systems, the transition from one format to another is almost always possible by direct computer processing,

and thus the data file compatibility is not a major concern, so long as the specific items of information which are stored in the file are analogous to any other file with which the data are to be utilized. This is not a new problem to botany but one which must be squarely faced if the botanist is to use a computer. Classification systems, based on the shape of leaf in one plant and the color of flower in another, do not lend themselves readily to the impartial comparison of a computer data file. Finally, the last level of compatibility is program compatibility, and this involves not only the operating programs which process the data but also is unfortunately tied closely to the individual hardware systems in use. The plea for compatible programs which can be exchanged between scientists is regrettably somewhat premature, considering the developmental stage of the computer industry. Computers began from a standing start at the end of World War II and now number in excess of 100,000 machines in use. The technology began with operations taking tens of milliseconds, and now has developed into operations described in nanoseconds. The hardware has developed from relatively small computers occupying the entire space of a large building to highly sophisticated computers occupying little more than the volume of an average television set. The development in this industry is not over, and with the rapidity of advance comes inevitable change. For the foreseeable future, true program compatibility is something to be sought but hardly to be achieved. What approach one uses to compatibility of programs depends in large part on which of two basic approaches to programming are utilized by the individual worker. If your computer needs are sufficient to maintain your own programming staff, then one can be relatively independent of changing hardware specifications, since new programs can be written to meet current needs and old programs can be modified as necessary. If, on the other hand, the computing needs are small and one uses the packaged programs which are available with most commercial computers, then one must be careful in this rapidly developing industry to choose a manufacturer who will provide a reasonable degree of stability. As a general rule, the more interchangeable the program, the longer the running time of that program in the machine, but here again, the true costs of using a computer must be evaluated, not in the simple running time of the computer, but in the total concept of the staff and equipment necessary to process the data.

If the botanist is to use the computer well, he must understand the processing of his problem in the computer. He must,

in essence, be able to program his own problems. Significant uses of computer data sets will come only from the individual who himself fully comprehends the abilities of the processing equipment. If the botanist is completely dependent upon another individual to provide programming, then he will be able to derive advantage from the computer only to the extent that he is able to state his problem in terms of the computer's ability. Programming has unfortunately become synonymous with the operation of translating the logical solution to the problem into the specific commands to be given to the computer. In reality, the most important part of programming is the ability to state the process of problem solution in simple logical terms. This is analogous to considering a skilled translator as equivalent to a talented novelist. The knowledge of the words of a language is not the same as the ability to use that language well. The botanist who does his own programming will soon learn to make the distinction between these two facets of the job. Whereas most university programming courses are exactly equivalent to university language courses, the course, Computers in Biological Systematics,[5] which is now offered at Michigan State University, is much more comparable to a course in composition and writing of a foreign language. Both knowledge of the language and fluency are necessary, but ultimately the latter is essential to the successful use of this new tool. A recent paper by Cutbill,[6] on new methods for handling biological information, should become required reading for any botanist involved in computer data processing.

In closing, I should like to say that the computer offers the botanist the means of coming from the 18th century into the present, but in order to do so, a retraining program is required so that the individual botanist may become completely familiar with the abilities of the computer and fluent in its language.

GILBERT S. DANIELS
Director, Hunt Institute
for Botanical Documentation
Carnegie-Mellon University

References

1. Sokal, Robert R., and Sneath, Peter H. 1963. Principles of Numerical Taxonomy. San Francisco & London: W. H. Freeman and Co.
2. Crovello, Theodore J., and MacDonald, Robert D. 1970. Index of EDP-IR projects in systematics. Taxon 19(1): 63–76.
3. Biological Journal of the Linnean Society 1971. 3(3).
4. Morse, Larry E. 1971. Specimen identification and key construction with time-sharing computers. Taxon 20(2/3): 269–282.
5. Furlow, J. J., Morse, L. E., and Beaman, J. H. 1971. Computers in biological systematics, a new university course. Taxon 20(2/3): 283–290.
6. Cutbill, J. L. 1971. New methods for handling biological information. Biological Journal of the Linnean Society 3(3): 253–260.