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The Hunt Institute Archives contains hundreds of thousands of pages of historical content, writing and images, created by thousands of individuals connected to the botanical sciences.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time and social context in which these materials were created, some of the collections contain material that reflect outdated, biased, offensive and possibly violent views, opinions and actions.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 Archival records are historical documents, and the Hunt Institute keeps such records unaltered to maintain their integrity and to foster accountability for the actions and views of the collections'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

Centre for
Agricultural Publications
and Documentation

Your ref.
Our ref. 10400/MD/AH

Date November 26. 1969

Professor D.J. Rogers,
Taximetric Laboratory,
University of Colorado,
BOULDER, Colorado
U.S.A.

Dear Mr. Rogers,

Many thanks for the manuscript on Numerical Taxonomy you sent me. I translated it - which took me some trouble because you treat a rather philosophical subject and I had to keep as closely as possible to your wording.

I have only a few remarks to make:

1. You introduce the term taximetrics, but (through being no latinist) I asked myself if it should not have been taxometrics: measuring a taxon, and not measuring a taxi. But, of course, I do not propose to change it. It is just a remark.
2. On page 6, line 2, you mention an imaginary "centroid". Is such a centroid not represented by the type specimens used by many plant taxonomists and should that term not be added?
3. In your literature list I found under the third Rogers: BioScience. It is not in the world List, but is it Biol.Sci.Bull.Fla St. Mus. 14? Under the same number you mention H. Fleming, whereas the next cites H.S. Fleming.

We prefere not to mention the editor of a book, but the place it has been edited (or both). Could you as yet supply this information for Lamarck, and for Sokal and Sneath?

I think that with these additional data your paper will give no more difficulties. If you want a reading proof, please let me know.

Yours sincerely,

E. Meijer Drees.

6a Duivendaal, P.O. Box 4, Wageningen (Netherlands)

Centre for
Agricultural Publications
and Documentation

Your ref.
Our ref. 10351/MD/AH
Date October 1, 1969

Professor D.J. Rogers,
Taxonometric Laboratory
University of Colorado
BOULDER, Colorado
U.S.A.

Dear Mr Rogers,

In April of this year you made a contribution to the Symposium on 'Biosystematics' in the meeting of the "Biologische Raad" at Amsterdam. As you know, the lectures delivered on that occasion will be published in a small book. Pudoc, the institute I am working for, has the task to make the manuscripts ready for the press, and in your case to take care of the translation into Dutch.

But up till now neither Professor Stafleu, nor Professor Voous, who are the members of the editorial committee, received a copy of your paper. As we are all very much value your contribution, I should like to ask you if you are able to send it on short notice.

We understand that you did not put on paper its details. But even then, a kind of summary of, say, five to ten pages would be very welcome to complete the "report".

Please let me know if we still can expect something, or not. We are waiting for it, as all other papers are ready for the press.

Yours sincerely,

Dr. E. Meijer Drees

6a Duivendaal, P.O. Box 4, Wageningen (Netherlands)

Taximetrics Lab., Armory 101 November 5, 1969

Dr. E. Meijer Drees
Centre for Agricultural Publications and Documentation
6a Duivendaal,
P.O. Box 4,
Wageningen (Netherlands)

Dear Dr. Dress:

I am sending under separate cover an paper for the symposium on Biosystematics in the meeting of the "Biologische Raad". Sorry that I have delayed so long.

Sincerely yours,

David J. Rogers
Director.

THE AIMS, SUCCESSES AND SHORTCOMINGS OF NUMERICAL TAXONOMY

David J. Rogers
Director, Taximetrics Laboratory
University of Colorado
Boulder, Colorado

Numerical taxonomy is a title proposed by Sokal in 1963, denoting an interest in, and development of models for, taxonomic methodology. A school of numerical taxonomy has grown to some proportions in the United States and the United Kingdom particularly, and to a lesser extent in other countries, embracing an operational philosophy. It was early stated by this group (Sokal and Sneath, 1963) that numerical taxonomy "deliberately set out to revise taxonomic theory and practise." That I do not espouse either the philosophy nor the objective, should be quite clear from my own publications (1960, 1963, 1964, ~~1965~~, 1967, 1969), and for this reason, I cannot speak to you as a representative of numerical taxonomy, as will become more evident later. In this paper I will try to indicate some of the objectives, developments, and areas of great interest and concern to all taxonomists.

The major purposes or objectives, in the areas which numerical taxonomy works, are to provide satisfactory mathematical models, and computer programs, to aid the taxonomist in the very heavy load of classification, evolutionary study, and in an area which taxonomist serve not only other biologists, but the larger community of science, as the organizers of information. To arrive at the objectives, several necessary steps are involved:

1. Formulation of taxonomic processes into logical, consistent, rules.
2. Restatement of the rules in mathematical terms.
3. Development of computer programs which follow from steps 1 and 2.
4. Evaluation of the programs on some problems in taxonomy.

The past ten years or so has been a time in which various parts of these processes have been examined, and in the development, some advances and some false starts have been made. But more important, the discipline of taxonomy

has been made more explicit, and it is possible to teach our students more in the spirit of science than as apprentices to the master.

Not many years ago, a taxonomist could be very fruitfull in the practise of his discipline without overtly expressed mathematical ability. Since computing machines were originally thought to work exclusively on (or with) numbers, we had to find some way to convert our works to numbers--ergo "numerical " taxonomy. Also, since that branch of applied mathematics called statistics had been applied successfully in genetic problems, in agricultural testing, etc., the idea developed that taxonomy could apply conventional statistical methods in the process of converting taxonomy to numbers to be manipulated by a computing machine. It is particularly in the area of applied mathematics for taxonomy that a number of false starts were made, and the most glaring of the shortcomings mentioned in the title.

Step 3 in the list above also was a problem for the developers of numerical taxonomy. The design and operation of computing machines was not understood by taxonomists at the beginning. If the taxonomist had any cause to manipulate numbers with a machine, it was with some sort of adding machine (or perhaps, if very sophisticated, a motor-driven desk calculator) where the taxonomist had in his own mind the steps necessary to add, subtract, divide or multiply. With the computing machine, it is necessary to write down in very exhaustive detail, step-by-step, each simple direction for number manipulation, in advance of turning on the machine. Today, the process of establishing the step-by-step processes--"programming"--is largely carried out by professional programmers, knowledgeable of the particular computer they are using, and of the sophisticated higher "language" packages provided by the computing machine manufacturers. Seldom are the programmers also knowledgeable of the substantive science which the programs are supposed to

support, and here again is one of the areas where there have been false starts or failures in numerical taxonomy: the naive assumption that the programmer "knows" about taxonomy, and can thereby write sophisticated procedures whereby the thinking process of the taxonomist is correctly carried out by the computing machine.

What are some of the objectives of numerical taxonomy? We can think of most of them, but certainly not all, and I will give you those that I personally espouse. A list seems to order priorities in descending order of importance, but this is not necessarily true.

Some of the objectives are:

1. To make an orderly synthesis of the facts about plants or animals.
2. To show the similarities and differences of organisms by placing them into hierarchical categories.
3. To provide a process for identification of unknown specimens.
4. To make available to other people the knowledgeon morphological, genetic, physiologic, and evolutionary information about the plants and animals.

This list of objectives certainly does not differentiate numerical taxonomists from those already pactising the discipline now, or in the past, nor should it. Only the methods of meeting these objectives are changing, and it is for this reason that I have coined the term "taximetrics," (Ro ers, 1963) and find this term more meaningful than "numerical" taxonomy, because many methods are needed to support taxonomic research.

Major Developments in Taximetrics

The most important developments in taximetrics have been in the areas listed as objectives 1. and 2. above. These two constitute the collective processes known as classification. If we are to design programs for the computer which reflect the thinking processes of taxonomists as they make classifications, we must obviously employ some orderly, step-wise "flow chart" of the p rocess. The following simplified chart does not intend to

be a computer-programmer's flow chart, but merely the major parts of the taxonomic classification p rocedure.

Flow Chart of Classification Procedures.

1. Select specimens representative of the problem in classification (Museum samples, population samples, etc.)
2. Derive classificatory information (characters) suitable to make the classification (morphologic, anatomic, ethologic, etc.)
3. Test the value of the characters for taxonomic purpose.
4. Develop some measure of similarities (differences) between individuals of the study.
5. Use similarity measure to place specimens into hierarchical clusters.
6. Designate taxon levels of the clusters.

Of the six steps in this listing, by far the most numerical taxonomic effort has been made on steps 4 and 5. In step 4, many statistical models have been made to consider various "distances," either in Euclidean or non-Euclidean space. That most statistical models have been unsuccessful for taxonomic work is well demonstrated by continued efforts to refine them. Eades (1965), for example, showed that the Pearson-Lee regression coefficient could easily put two organisms together as similar which, in fact, were not at all similar. Of the numer of possible statistical methods of comparing two objects with one another, the very simple method of counting the number of properties possessed in common by the two objects divided by the number of properties that have been used in the comparison has turned out to be the most successful.

The activities in step 5 generally are called "clustering." For the taxonomist, this activity is referred to as taxon designation (or "speciation" in some cases). As in step 4 above, several different types of clustering procedures have been designed. ~~The objective in the various clustering procedures have been designed.~~ The objective in the various clustering procedures may, or may not, be the objectives espoused by the biological taxonomist, and it is the obligation of the taxonomist to know the objectives

and methods employed in the clustering procedure. As an example, Rubin (1967) devised a clustering procedure, with well-defined objectives, which sought, and found, the most "efficient" divisions between the objects in a study, but without regard for the biological relatedness between the individuals within the divisions. The biological taxonomist is not at all satisfied with such a clustering procedure because the biologically disparate objects which may be placed together in the Rubin method are the ones which hold great interest to him, and are the most difficult to classify.

Other clustering techniques provide a print-out which indicate the clusters by using dendrograms, where only the ultimate ends of the branches indicate the relationship of the specimens (taxa), or "operational taxonomic units." While little criticism has been formally directed towards such representation of clusters, it is obvious that much information about the objects in the study is lost in the dendrogram, and it is almost impossible to follow the development of the clusters from the dendrograms.

In 1957, Sneath proposed a method now referred to as the single-linkage clustering method. Wirth, Estabrook and Rogers, in 1965, took up from this idea, developed the necessary mathematics and procedures for the single-linkage concept into a comprehensive clustering program which has been tested on many actual problems in various taxonomic groups with great success. See, for example, the classification of a section of the genus Cassia (Irwin and Rogers, 1967). It was particularly gratifying when Jardin, et al. (1967) gave mathematical demonstration that the single linkage clustering procedure proved to be the only satisfactory one with respect to the objectives of biological taxonomy. In the single-linkage process, each specimen is "connected" to the specimen most related to it (using the over-all similarity measure), and clusters are built up by discovering all other elements (specimens)

which share some common values. In contrast, the "average-linkage" procedure takes some arbitrary, imaginary "centroid", and measures the distance of each real specimen to it, and the defined clusters must be formed by setting an a priori limit to the distance from the "centroid" beyond which a specimen cannot be included in the cluster. Clustering procedures with the centroid feature assume that all clusters have similar "shapes," an assumption intolerable for biological reality.

After the initial efforts with steps 4 and 5 of the classification process had been studied for some time, numerical taxonomists began to focus more attention on problems arising in earlier parts of the overall classification procedure, i.e. steps 2 and 3 of the flow-chart. The early workers in numerical taxonomy claimed that the use of "unweighted" characters were significant in the numerical methods, and these ideas were generally attributed, though incorrectly, to an early French taxonomist, Adanson. Actually, Lamark, some years earlier had urged such a step in the introduction to his work on the Flora of France. Unfortunately, no distinction was made between the statistical concept of weighting and that of biological concepts of weighting, and this lack of distinction led to much fruitless argument. Clearly, the biologist must select those characters which are significant to reflect the genetic mechanisms which determine the relations between specimens, or between taxa, before putting the data into some computing machine. If some other process is used, then the chances of producing a good classification are poor, with or without the computing machine. Weighting, from the taxonomists's point of view, is the selection of meaningful biological information, or the rejection of useless data which cannot reflect the real relationship between the organisms. No significant computer methods have been devised to exclude the well-trained biologist in the classification process. nor will this happen.

These statements are not intended to intimate that nothing can be done to improve the taxonomic method by which information for classification purposes can be derived. Indeed, by a more objective understanding of the role of characters in classification, we can improve the taxonomist's results. A character, in terms of objectives of the taxonomist, is a level classification of a group of organisms. If another character gives the same or similar level classification, then the two characters are correlated, and in the biological sense, are "weighted." The definition given does clarify the ideas about what a character must be, but it does not indicate how many genes cause the character to exist. However, this definition clearly causes the taxonomist to exercise his knowledge of genetic mechanisms, such that the character does reflect the genetic mechanism as well as can be done in the absence of complete knowledge of the genetic constitution of the organisms. This character definition places a burden on the taxonomist to prevent the incorporation of information greatly influenced by environmental stress. Fortunately, the emphasis on methods, more forcefully impressed upon those who use computing machines, brings such definitions to light, which in turn, improved the scientific value of taxonomic work.

Returning to the objectives listed on page 3, emphasis in taximetrics is now turning to those areas listed as objectives 3 and 4. Once the information used to classify a group of organisms has been properly structured, and the characters used to classify the organisms well correlated with the different taxa, it is possible to construct an identifying process. Several attempts have been made recently to construct keygrams for computer-aided keys similar to the dichotomous keys of the regular taxonomic procedure. Such attempts, where the computer responds to each item of information by suggesting the next necessary piece of information to continue in the process of identification of an unknown, is inefficient of computer use and represents a lack of know-

ledge of computer potential. The process of computer identification of organisms follows very closely the work done by information retrieval experts for finding precise literature references, and not to look into work done in this area is clearly to deny that some other disciplines can have useful procedures for taxonomy. In information retrieval, the unknown is described by the properties it possesses, applying a vocabulary of characters contained in the memory of the computer. Using this vocabulary, the descriptive words (or properties) which are pertinent, and known features of the unknown are run through the computing program, and the known species possessing the same features as the unknown are automatically returned. Thus, rather than requiring the computer to make human-like, step-wise choices, the automated identification routine employs all information simultaneously, and directly discovers the unknown.

In identification routines, the same skill and biological knowledge necessary to describe the characters for classification is required to give good results in identification. The same tests (or criteria) are applied to key characters as are applied to characters for classification. If the states of a character do not clearly divided the specimens into non-overlapping sets, identification cannot be made. In a well-designed computer identification routine, one of the strictures of the dichotomous keys is overcome: characters do not have to be stated as one of two alternatives, a necessity which frequently requires elimination of many useful characters from standard keys. The biological information contained in the character may be stated in as many ways as the organisms demonstrate variations in the character. The searching and memory capacity of the computer can keep track of the variables with much more speed and certainty than can an individual taxonomist.

In objective four, which is to make available to others the knowledge of morphological, genetic, physiologic, and evolutionary information about

plants and animals, the taxonomist has spent much effort already, and the normal techniques used to write up the results of his studies, either in floristic (faunistic) or monographic style represent one of the most precise ways of information retrieval ever devised. Most taxonomists do not fully appreciate the power of their techniques of correlating such a vast amount of information about the natural world. In taximetrics, new aids to these well-established techniques are being developed, in many areas of the world. In herbaria and museums, much effort is now being devoted to information retrieval systems for the vast wealth of knowledge existing in our cabinets, shelves, bottles, and on herbarium sheets. Much information now locked up in inaccessible places in the collections can be made more rapidly available using well-designed computer-aided information retrieval systems. The system designed and programmed in the Taximetrics Laboratory was precisely structured to aid the taxonomist in his work. The system, called TAXIR, makes it possible to include in the computer memory, as much information as we desire about each and every specimen in the collections. With such a system in operation, the main objectives of the taxonomist listed earlier come closer to reality. If taxonomists recall their main role in the biological disciplines, and consider as one of their highest callings, the ideas of objective number four, then it becomes absolutely imperative that we employ computing machines to aid in the immense task of correlating biological information for ready access, no matter whether derived by molecular biologists, by ecologists, or any of the other biological disciplines.

In conclusion, we can state that taximetrics is much more concerned with aiding the taxonomist in his regular tasks and objectives than it is with the development of a new science. This emphasis is now more clearly being employed by the numerical taxonomists, much to their credit, and some of the earlier drastic comments, such as those by Ehrlich (1964) who said that we

should soon be able to completely do away with our specimens, and close up all the natural history museums, can be discounted as a radical fring not representing the main thrust of taximetrics. We can also begin to fulfill the functions and objectives of taxonomists, bring them to a central role as the correlators of biology, more fully than they already are.

References

Bades, D. C. 1965. The inappropriateness of the correlation coefficient as a measure of taxonomic resemblance. Syst. Zoo. 14: 98-100.

Ehrlich, P. R. 1964. Some axioms of taxonomy. Syst. Zoo. 13: 109-123.

Irwin, H. S. and Rogers, D.J. 1967. Monographic studies in Cassia (Leguminosae-Caesalpiniodeae). II. A taximetric study of section Apoucouita. Mem. N. Y. Bot. Gard. 16: 71-118.

Jardin, C.J., Jardine, N., and Sibson, R. 1967. The structure and construction of Taxonomic hierarchies. Mathematical Biosciences 1(2): 173-179.

Lamarck, J. B. A. P. 1778. Fl. Francais, 1st ed., 3 Vol.

Rogers, D. J. and Tanimoto, T. T. 1960. A computer program for classifying plants. Science 132 (3434): 1115-1118.

_____, 1963. Taximetrics--new name, old concept. Brittonia 15: 285-290.

_____, and Fleming, H. 1964. A computer program for classifying plants. II. A numerical handling of non-numerical data. BioScience 14: 15-28.

_____, Fleming, H. S. and Estabrook, G. 1967. Use of computers in studies of taxonomy and evolution. In Evolutionary Biology, Vol. 1, 169-196. Eds. Dobzhanzky, T., M. K. Hecht, and W. C. Steere. Appleton-Century-Crofts .

_____, and Appan, S. G. 1969. Taximetric methods for delimiting biological species. Taxon 18: December.

Rubin, J. 1967. Optimal classification into groups: an approach for solving the the taxonomy problem. IBM Jour. of Res., New York.

Sneath, P. H. A. 1957. The application of computers to taxonomy. Jour. Gen. Microbiol. 17: 201-226.

Sokal, R. R. and Sneath, P. H. A. 1963. Principles of Numerical Taxonomy. W. H. Freeman, Publ.

Wirth, M., Estabrook, and Rogers, D. J. 1966. A graph theory model for systematic biology, with an example for the Oncidiinae (Orchidaceae). Syst. Zool. 15: 59-69.