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The Hunt Institute Archives contains hundreds of thousands of pages of historical content, writing and images, created by thousands of individuals connected to the botanical sciences.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time and social context in which these materials were created, some of the collections contain material that reflect outdated, biased, offensive and possibly violent views, opinions and actions.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 Archival records are historical documents, and the Hunt Institute keeps such records unaltered to maintain their integrity and to foster accountability for the actions and views of the collections'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

50 ± 12:00 – Rm 227. Geology Seminar
Thurs. Mar. 21.

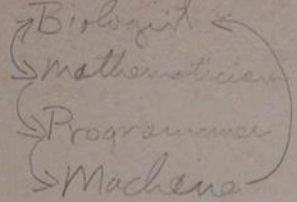
1. Generalities –
Needs

Useful descriptive references
1. Wirth, Crkn. + R-
etc.

Programs now running and available.
1. graph theory clustering.
2. character analysis
3. p-tree for cladistic studies.
Programs now approaching completion.
Computerized Inf. Retrieval.

Geol Seminar

General approach used in making computer machine work -



Important for biologist (or geologist) to establish the dialogue w/ mathematician

1. Isolate problem from other considerations
2. Define the questions
3. Discover how he (geologist) attacks problems.
4. Get statement from mathematician which puts some math behind the thinking to (a) be certain of necessary + sufficient condition to attack the problem (b) ~~to~~ allow an objective analysis of the test data.

Then, bring in programmer to write the program required to carry out the manipulations. Then, discover which parts of the computer will be required.

Description of graph clustering.

1. Similarity
   basis - characters
   some level classification of objects

   ratio ~~##~~ for each pair of objects

   $$S(a,b) = \frac{\text{all characters alike}}{\text{number of characters compared}}$$

2. table of similarities for all pairs of objects
   serves as input to clustering

3. Clustering - a series of partitions,
   each partition brings together objects
   at that similarity

4. Disjoint partition: all objects separate
   strictest measure of similar

5. "Relaxing" similarity measure causes
   ~~re~~ new partitions

6. "Conjoint" partition = all objects joined -

7. Other measures - nearest - distance from
   nearest object in a cluster to next object out -
   interconnectedness of objects in any cluster.

Flow Chart of taxonomic work -

1. Importance of some such structure for scientific use of computers.
2. ~~Cluster analysis~~ Classification -

A. Development of ideas -

Requires some assumptions -

1. Similar objects grouped together
2. Hierarchy is useful - a series of nested boxes.
3. A classification is a series of partitions, in which groups of objects at any one partition form into groups.

Similar objects defined by their "interesting" description (characters in biology) - ~~same~~ similarity is fixed by some figure between 0 + 1 -

Discover which objects defined by their similarities can be clustered

B Some other requirements -

Discover shapes of clusters after

Seminar

- Taxiametrics Laboratory

December 13, 1966

Dr. Paul Winston
Department of Biology
University of Colorado
Boulder, Colorado

Dear Dr. Winston:

Enclosed are two sketchy biographies for myself and Mr. George F. Estabrook. These you may use as you like or any part thereof for the seminar to be given January 13. I suggest as a title for the seminar "Towards a Biological Information Retrieval System." If this is not satisfactory, please do not hesitate to say so. We expect to divide the seminar into two parts (1) the biological methodologies and (2) the needed software for computer development. I will speak to the first part and introduce the topic. Mr. Estabrook will handle the second portion.

Any other required information will be gladly supplied.

Sincerely yours,

David J. Rogers
Professor of Botany

DJR:ch

Enc.

Biographical sketch for David J. Rogers

1. Native of Florida, BS, University of Florida (Botany) 1941. Graduate work at Washington University, St. Louis, under Edgar Anderson and Robert Woodson (taxonomy—monographic studies of Euphorbiaceae). Ph.D. received 1951.
2. Taught botany and general biology at Allegheny College, Meadville, Pa. from 1951-1957.
3. Curator of Economic Botany, and editor of ECONOMIC BOTANY, the New York Botanical Garden, 1957-1965.
4. Professor of Botany, CSU, 1965—date.
5. Interest in the systematics of cultivated plants, in particular, of Manihot esculenta (cassava, manioc, yuca, and tapioca), family Euphorbiaceae. Intricacies of the relationship amongst the cultivars, and the reticulate nature of the relationship of these plants, led to interest in new methodologies to untangle the relationships, and this led into use of computers as aids in classification.

The methodological studies became as absorbing as the actual classification, and we broadened our work to the development of an interdisciplinary group, (which includes taxonomists, mathematician and programmer) now devoted to general application of taxonomic methodologies.

Biographical sketch for George F. Estabrook

Native of New York, but really cosmopolitan. Received his training at Dartmouth College, where he graduated from the honors program with a double major--one in structural mathematics, one in plant physiology. Now continuing his advanced work in the math department of the University of Colorado.

Came to work with the present group while it was still in New York, and in this period has continued his interdisciplinary work of development of a variety of mathematical models for systematic biology.

Boulder, Jan. 13 Seminar - Bio. Inf. Retriev.

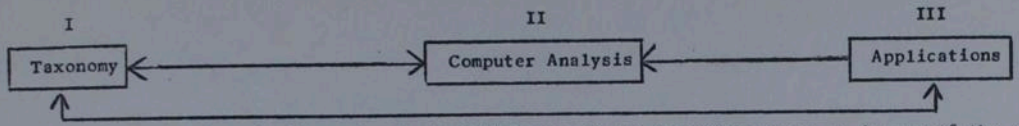
1. What is I.R.?
   Our data, in notebooks, card-file, or literature – pieces of information useful to us – may be in a multitude of forms.

2. The place of I R in our scheme of things.

3. In addition to the goal of a practical scheme, the investigations themselves provide a very stimulating experience, for it is possible to find some mathematical stimuli and models with the biological thinking process.

4. We will not talk about software for the computers, that being such a topic as to require several separate seminars.

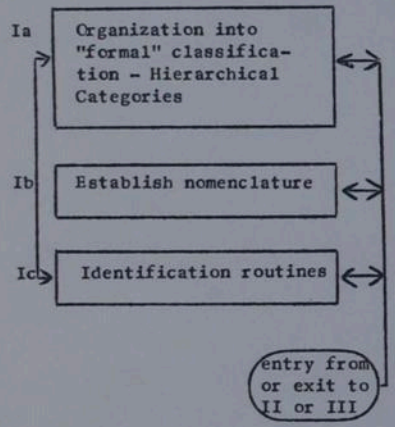
George:
Characters, logical arithmetic
etc.



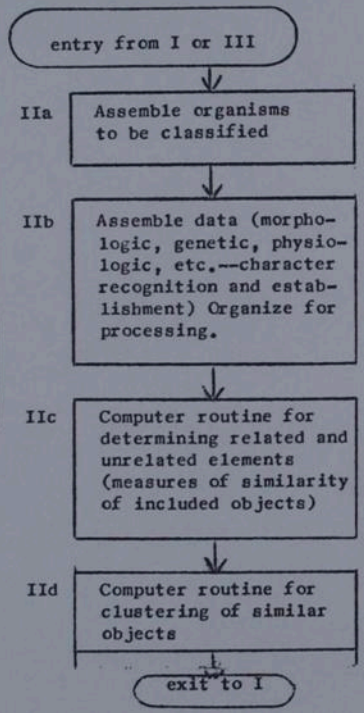
Note: The procedure begins and ends in either one of the taxonomic areas or in one of the applications.

DETAILED FLOWCHARTS

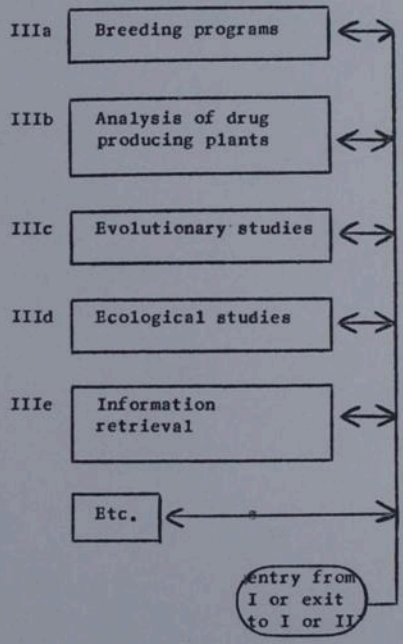
I - Taxonomy

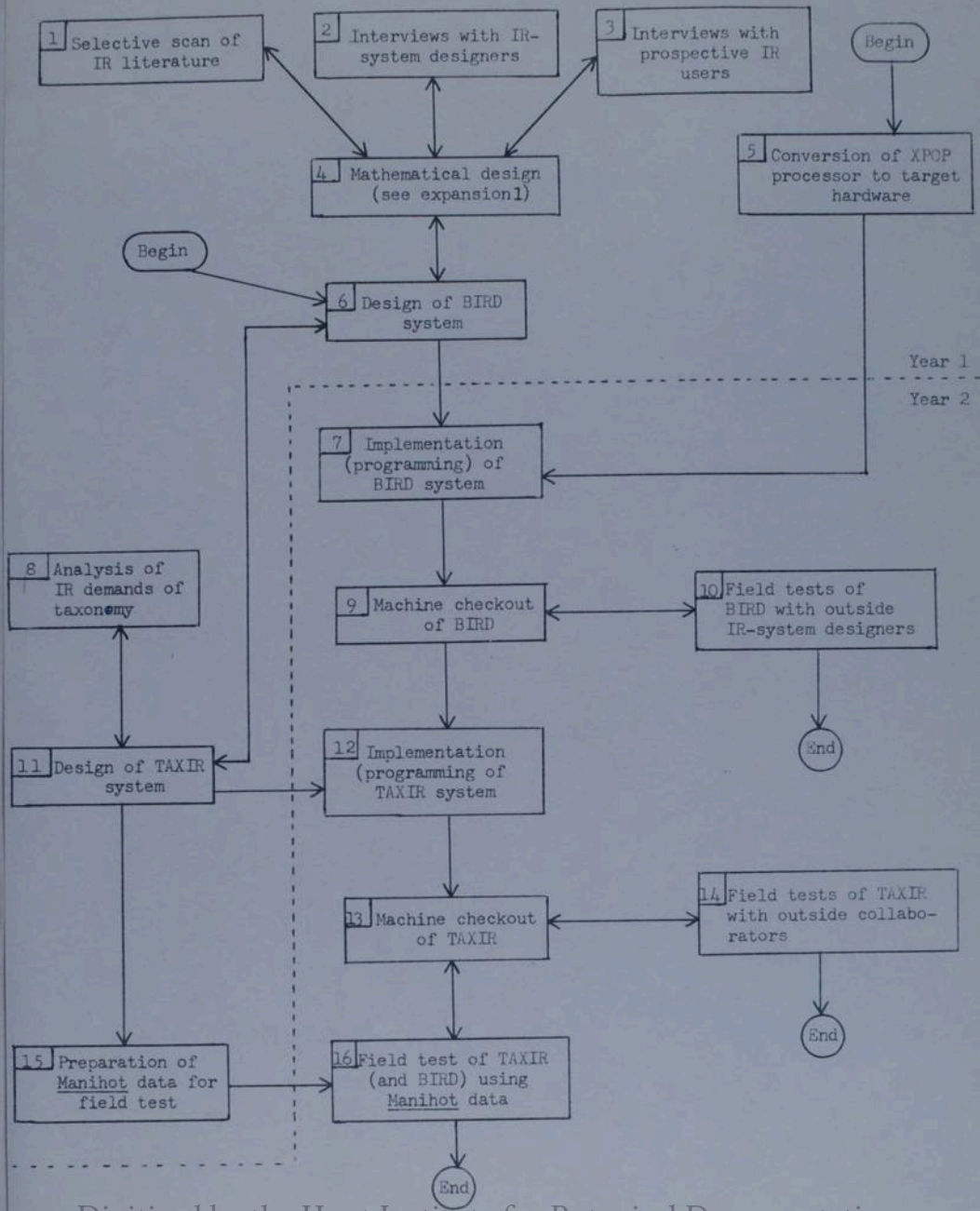


II - Computer Analysis

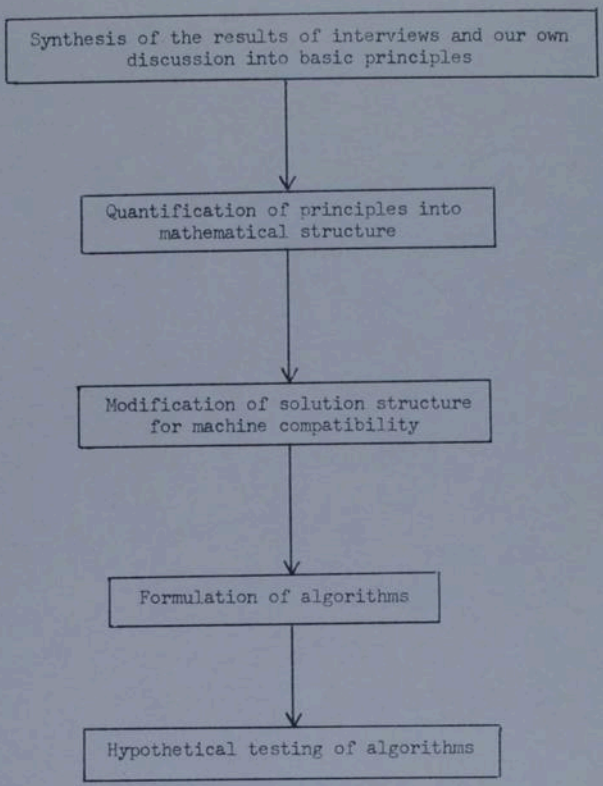


III - Applications





Expansion 1



MAY 8 1968

THE COLORADO COLLEGE
COLORADO SPRINGS, COLORADO 80903

Department of Zoology
6 May, 1968

Dr. David J. Rogers
Department of Biology
University of Colorado
Boulder, Colorado

Dear Dave:

For a number of years the University of Colorado Museum (in Boulder) has run a summer lecture series, this coming summer being the 12th series. Talks generally focus (though sometimes out of focus) on a particular theme, this summer's theme being various aspects of the Biosphere. There will be a lecture on antarctica, for example, on the far north, on New Zealand, the Indiana Dunes and succession, man's impact on the biosphere, the biosphere's oceans, and so on. I want to have one popular lecture on the earth's green mantle (you know, interesting tidbits on the vegetation, etc etc), and I'm hoping I can talk an old New York Bot Garden man into returning to the public arena to give such a talk. These lectures are scheduled on Sunday evenings during the summer, are ordinarily illustrated, and have a popular pitch. We are juggling dates and speakers now, but potential open dates include July 14, 21, 28, or August 4. If you have time and are interested, I'd be delighted to sign you up for this summer's series on a date of your choice. There is a small honorarium for the lecture (more details on that after we get our budget firmed up) and considerable (questionable) honor. Let me know how it strikes your fancy.

Best regards,

Dick

Richard G. Beidleman
Professor of Zoology

Agreed, for July 21

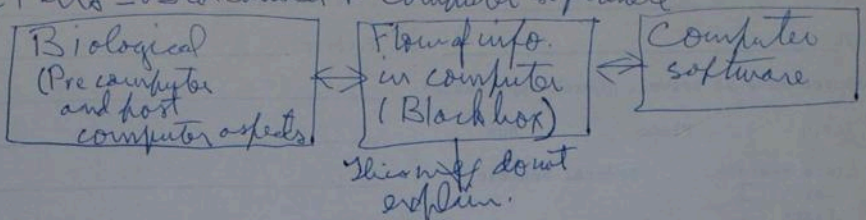
Econ. Bot. Class - Brunz Klinge, Cl. Jan. 20, 11 AM. 1967

1. Classification - Family, Genus, Species -
   Common names -
   One of 12 food crops between man and starvation.
   Root crop, largely carbohydrate.
2. Distribution 2 slides -
3. Plants - 2 "
4. Leaves - Veget + inflor. 4 slides
5. Roots - 2 slides -
6. Flowers - 2 "
7. Cultivation - 3 slides
8. Methods of preparation - 5 slides.

Botany Seminar, Oct. 17, 1966

IR Seminar

1. 2 Parts – Biol concept + Computer software



2. Problem – all included – not just literature that interests – digested (lit.) and non digested (unpubl.)

3. Needs in IR
   1. Get data described in unequivocal (unique) terms – an index
   2. Stored in most efficient manner
   3. Provide natural language entry and exit.
   4. Provide system common to many users.

4. Techniques – the pre computer aspects –
   1 Use input system similar to that in classification.
   ~~Characters and~~ a
   Objects = specimens, or other items which one uses – could be a compound, or a phenomenon of interest
   Characters + attributes
   " of morphology, of physiology or, in IR could be a reference –
   attributes, – unique descriptors under characters –
   example morph – leaf shape –
   Set up in binary, 0, 1
   See example –
   Give length of words of code –

THE TEACHING OF TAXONOMY

AAS MEETINGS MAY 2, 1966

LAS CRUCES, NEW MEXICO

1. Investigations of computer as a taxonomic tool eventually responsible for new look at all taxonomy.
   A. Requires examination of theoretical structures of the field.
   B. Requires that some basic structure or flow of process be established.
   C. Leads to the idea that many processes are at work.
   D. Gives the basis for a structured approach to teaching students in a logical manner.
2. Flow chart--not the only way to break up taxonomy, but useful.
   A. Indicates a clear separation of the parts, and can be used as a curriculum guide for majors in taxonomy.
      1. Taxonomy in column I.
      2. The methods of establishing the classification in II.
      3. The applications of the classification in III.
   B. Many courses--called local flores--given corresponding to part Ic, but may include Ia and Ib.
   C. Seldom do taxonomy courses get into column II, where the processes of forming a classification occur.
   D. A problem in taxonomy may be found by working in columns I or III--give examples.
3. The new course at CSU concentrates on column II.
   A. The flow chart gives the student an orderly sequence to follow in learning methodologies for classificatory work.
   B. The greatest effort is made to enlighten area IIb. Professor Fleming will speak about these problems--characters, in the next presentation.

C. Though we use the word computer in IIc and d, it is not necessary that a computer be used. Some problems will not require the hardware.

D. The clustering methods mentioned in IId are still in development. This is the point where decisions are made about the membership of a specimen in a particular group.

E. It is explained that when this column has been completed, still more work is needed in column I to complete a classification.

Abstract for the Las Cruces Meeting, SW AAAS

Title: The Teaching of Taxonomy

David J. Rogers

Most of the teaching in taxonomy focusses on presentation of attitudes, and customs, not on methodologies. Methods for the aspiring taxonomist must be discovered by the inefficient process of imitation of older taxonomists, some of whose methodologies leave much to be desired. A new course at CSU focusses on the methods involved in the process of classification. The discipline of systematics is divided into component parts, namely: (1) discovery of a "taxonomic problem"; (2) decisions as to methods of attacking the problem; (3) developments of measures of similarities and differences; (4) methods of clustering the objects; (5) the development of diagnostic keys; (6) the presentation of formal classifications; (7) and, methods of presentation of results. The major emphasis in the course is placed on areas 3, 4, and 5, for these areas leave most to be desired in "orthodox" taxonomic work. With the aid of computers, and with the development of certain types of mathematics, much can be done to dispell the aura of mystery in classification.

Abstract for the Las Cruces Meeting, SW AAAS

Title: The Teaching of Taxonomy

David J. Rogers

Most of the teaching in taxonomy focusses on presentation of attitudes, and customs, not on methodologies. Methods for the aspiring taxonomist must be discovered by the inefficient process of imitation of older taxonomists, some of whose methodologies leave much to be desired. A new course at CSU focusses on the methods involved in the process of classification. The discipline of systematics is divided into component parts, namely: (1) discovery of a "taxonomic problem"; (2) decisions as to methods of attacking the problem; (3) developments of measures of similarities and differences; (4) methods of clustering the objects; (5) the development of diagnostic keys; (6) the presentation of formal classifications; (7) and, methods of presentation of results. The major emphasis in the course is placed on areas 3, 4, and 5, for these areas leave most to be desired in "orthodox" taxonomic work. With the aid of computers, and with the development of certain types of mathematics, much can be done to dispell the aura of mystery in classification.

ABSTRACTS

PAPERS PRESENTED AT THE
FORTY-SECOND ANNUAL MEETING
(Fourth Las Cruces Meeting)

Southwestern and Rocky Mountain Division
American Association
for the Advancement of Science

And The

New Mexico Academy of Science

MAY 1 - 4, 1966

NEW MEXICO STATE UNIVERSITY

* * * *

TABLE OF CONTENTS

Program

FORTY-SECOND ANNUAL MEETING
(Fourth Las Cruces Meeting)

Southwestern and Rocky Mountain Division American Association for the Advancement of Science

And The

New Mexico Academy of Science

MAY 1 - 4, 1966

NEW MEXICO STATE UNIVERSITY

ABSTRACTS

PAPERS PRESENTED AT THE

FORTY-SECOND ANNUAL MEETING

(Fourth Las Cruces Meeting)

Southwestern and Rocky Mountain Division American Association for the Advancement of Science

And The

New Mexico Academy of Science

MAY 1 - 4, 1966

NEW MEXICO STATE UNIVERSITY

* * * *

TABLE OF CONTENTS

THE TEACHING OF TAXONOMY

AAS MEETINGS MAY 2, 1966

LAS CRUCES, NEW MEXICO

1. Investigations of computer as a taxonomic tool eventually responsible for new look at all taxonomy.
   A. Requires examination of theoretical structures of the field.
   B. Requires that some basic structure or flow of process be established.
   C. Leads to the idea that many processes are at work.
   D. Gives the basis for a structured approach to teaching students in a logical manner.
2. Flow chart--not the only way to subdivide taxonomy, but useful.
   A. Indicates a clear separation of the parts, and can be used as a curriculum guide for majors in taxonomy.
      1. Taxonomy in column I.
      2. The methods of establishing the classification in II.
      3. The applications of the classification in III.
   B. Many courses--called local flores--given corresponding to part Ic, but may include Ia and Ib.
   C. Seldom do taxonomy courses get into column II, where the processes of forming a classification occur.
   D. A problem in taxonomy may be found by working in columns I or III--give examples.
3. The new course at CSU concentrates on column II.
   A. The flow chart gives the student an orderly sequence to follow in learning methodologies for classificatory work.
   B. The greatest effort is made to enlighten area IIb. Professor Fleming will speak about these problems--characters, in the next presentation.

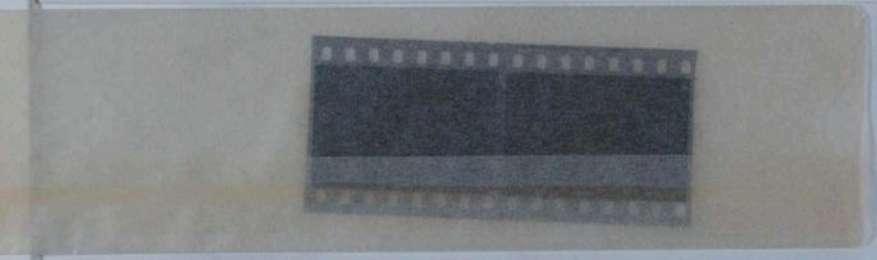
C. Though we use the word computer in IIc and d, it is not necessary that a computer be used. Some problems will not require the hardware.

D. The clustering methods mentioned in IId are still in development. This is the point where decisions are made about the membership of a specimen in a particular group.

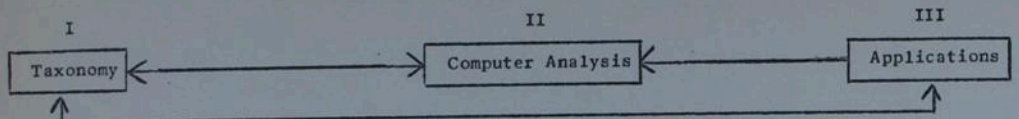
E. It is explained that when this column has been completed, still more work is needed in column I to complete a classification.

COLORADO STATE UNIVERSITY
DEPARTMENT OF BOTANY AND PLANT PATHOLOGY
FORT COLLINS, COLORADO 80521

Slide - Flowchart



GENERAL FLOWCHART OF TAXONOMIC PROCESSES



Note: The procedure begins and ends in ~~either~~ one of the taxonomic areas or in one of the applications.

DETAILED FLOWCHARTS

I - Taxonomy

Ia Organization into "formal" classification - Hierarchical Categories

Ib Establish nomenclature

Ic Identification routines

entry from or exit to II or III

II - Computer Analysis

entry from I or III

IIa Assemble organisms to be classified

IIb Assemble data (morphologic, genetic, physiologic, etc.--character recognition and establishment) Organize for processing.

IIc Computer routine for determining related and unrelated elements (measures of similarity of included objects)

IId Computer routine for clustering of similar objects

exit to I

III - Applications

IIIa Breeding programs

IIIb Analysis of drug producing plants

IIIc Evolutionary studies

IIId Ecological studies

IIIe Information retrieval

Etc.

entry from I or exit to I or II

File under seminars CSU
10/20/65 63

Opening comments

1. Description of study of M. esculenta. Fam. Euphorb. genus M. 1-200 sp

2. Darwin's comments of use of cult. plants in study of evolution

3. Before use of M. esculenta as tool in evolutionary studies, required up to date classification. Classification needed because of lack of biological understanding of the group.

4. Before satisfactory classification of the variation in the cult. plants, had to relate to the wild species of the genus.

Slides

1. the distribution of the species of Manihot

2. the morphological types of plants in the genus
   1. trees--some rubber producers, "ceara" rubber, others not
   2. Shrubs--"weedy" types, some are bee plants.
   3. sub-shrubs
   4. some are "climbers"

3. Habitats of species--most are heliophyles, in open conditions, not forest-dwellers.
   Soils of many types, acid to base
   rainfall--from low to high, semi-desert to river margins.
   All in tropical lowlands, all frost sensitive.

4. Morphology of M. esculenta
   All shrubs, from low, many-branched to tall, nearly unbranched.
   Leaves, vegetative--deeply lobed, simple leaves, the lobe number varying from cultivar to cultivar.
   Leaves, inflorescence--most frequently recuded in no. of lobes, with some leaves unlobed, simple.
   Roots--enlarged, the edible portion. Have an outer peel (phelloderm) an enlarged cortex containing the carbohydrates, and a central vascular strand.
   HCN concentrations in the roots--from low to high, depending on various factors, not correlated with any known morphological characters--some non-poisonous vars. look poisonous, and vice versa.
   Flowers--inflorescence--monoecious. the pistillate with 5 sep. tepals. the staminate with 5 tepals half-united, 10 staments.
   Pistillate fls. open first, long before staminate flowers of same plant. Therefore, outcrossing made possible and probable. Source of variability.

   Hybridization, between cultivars and with wild species. Results.

5. Use of M. esculenta. Food producers, from very primitive to very (or relatively) advanced agricultural methods.
   Used in a few places as an ornamental--pigmentation similar to poinsettia.
   Well enough thought of in some places to be subject of a stamp.

   Preparation--slides

6. Origins--problems--Brazil? Meso-America? Question involves people, their orgins and movements.

October 15, 1965

Dr. Larry Leslie
Apartment M-26
Jardine Terrace
Manhattan, Kansas 66502

Dear Dr. Leslie,

We enjoyed having you stop in to see us during your stay here at CSU. I may have misled you. Dr. Barkley is not himself a "numerical taxonomist" but a good "classical taxonomist". I doubt that he is up-to-date on the developments in computer methodology. It is interesting that you are attempting to get the bacteriologists there going. How would it be to ask us to give a seminar there? We might be able to do it if an invitation is forthcoming. We may even like to encourage you to collaborate with us on your various problems in the Enterobacteriaceae.

Sincerely,

David J. Rogers
Professor of Botany

DJR/ec

Larry Leslie
Apartment M-26
Jardine Terrace
Manhattan, Kansas
66502

September 25, 1965

Gentlemen:

I wanted to write to you and express my appreciation for your hospitality (and the fascinating discussion on taxonomy) during the recent Genetics Society meetings.

Here at Kansas State University I've been "encouraging" the Bacteriology Seminar Coordinator to contact a botannical taxonomist, Dr. Theodore Barkley (whom I believe you suggested to me), about giving a seminar on numerical taxonomy. I believe that if most of our bacteriologists could see the many advantages and simplicity of this method for microbial classification they would be encouraged to investigate its possibilities further, and what a revolution might ensue!

I am looking forward to seeing copies of your work.

Notes for Mathematics Colloquium, Nov. 4, 1965 -- Dr. Rogers

I. Introductory remarks only --

Divisions in field of biology

2 types -- "analytical" and "synthetic"

Ours is the latter type.

II. Classification -- many operations, many types

1. Monographic

2. Floristic

Ours is the former

III. In monographic work, several objectives

1. To calssify, with all possible data available.

2. To apply the best possible nomenclature to the classified objects.

3. To give some rapid method of identification (keys).

Our operations mostly dealt with the 1st of these -- best possible clustering methods.

IV. In clustering, a knowledge of rules for the biological aspect are the most critical.

NOTES FOR MATH COLOQUIUM - Nov. 4, 1965 - Geo. F. Estabrook

I am going to discuss how some of the fundamental ideas of graph theory can be used to suggest tentative classifications for biology.

I am sure that some of you are familiar with the basics of graph theory, but because I will be concerned with a rather narrow view of what is otherwise a rather broad field and because our terminology may not be the same, I would like to discuss briefly the aspect of graph theory that will be of concern to us here.

Many of the statements that I make will not always be true, but must be qualified by adding " This is the way I am going to use this concept to address this particular biological problem". Please assume this qualification for the rest of the discussion.

GRAPH A set of points called vertices some pairs of which are connected.

EDGE Every pair of distinct vertices gives rise to an edge. The graphs that will be of interest to us in this discussion will have each vertex connected to itself. The edges of our graphs will have no direction. In this way a graph can be thought of as a collection $\to V \to$ of vertices together with some symetric subset containing the main diagonal $\to$ of the cartesian product of V with itself.

Thus, any binary reflexive symetric not necessarily transitive relation defined for a set of vertices determines a graph.

ARC An arc is said to exist for two vertices V_1 V_2 if there exists a sequence $A_1A_2A_3$ and so forth up to A_i of distinct vertices in the graph with V_1 connected to A_1, A_1 connected to A_2, A_2 connected to A_3 and so forth up to A_i connected to V_2.

GRAPH A graph is said to be connected if for any pair of vertices in the graph there exists an arc.

A subgraph for a graph is a subset of the vertices together with the edges determined by members of this subset. A subgraph can be thought of as a graph in its own right → In this way it is meaningful to speak of connected subgraphs.

A subgraph is said to be maximal connected if it is strictly contained in no other connected subgraph.

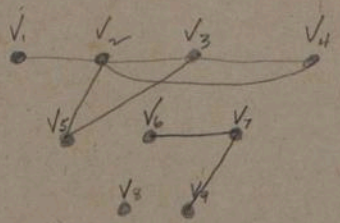
A maximal connected subgraph of a graph is frequently called a component of a graph (in case anyone is more familiar with that terminology).

Notice that if a graph is connected and contains M vertices, then it must contain also at least M-1 edges. This is easily shown to be true by induction. A graph with one vertex has no edges. Assume that a graph is connected and determines a minimum of edges. If one more vertex were added at least one more edge must be determined in order that the graph remain connected.

Of course, the maximum number of edges of non multiplicity the A graph can determine is simply the number of ways to choose distinct pairs from the set of vertices.

An articulation point for a connected graph is a vertex of that graph with the property that the subgraph of the graph consisting

Every vertex except the articulation point is not connected. Every definition and property of a graph that we have so far discussed can be illustrated by drawing the picture of a graph for its heuristic value.



The vertices are the points -- the non edges are the drawn connections. This graph is not connected because for V_1 and V_6 there does not exist an xxx. arc.

V_1 V_2 V_3 and V_4 determine a subgraph of this graph this subgraph is connected but is not maximal connected for it is contained in V_1 V_2 V_3 V_4 V_5 which is maximal connected subgraph. V_6 V_7 V_9 is an example of a connected subgraph which exhibits a minimum number of edges. V_7 is an articulation point for this subgraph as the subgraph V_6V_9 is not connected. We said earlier that a graph could be thought of as a collection of points (vertices) together with a symetric reflexive not necessarily transitive relation defined over them. Let that relation be denoted with the letter G. From our picture we see that V_1 G V_2 but not V_1 G V_5 an arc exists for V_6V_9 because V_6 G V_7 and V_7 G V_9.

By an extension for the relation G, I will mean any relation G' with the property that AGB implies AG'B. The smallest equivalence relation extension for G will be the equivalence relation extension of G of which every other equivalence relation extention of G is an extension. (Discuss lattice) The relation R where ARB if and only if the vertices A and B determine an arc is the smallest equivalence relation extension for G. R contains G because two points which are connected certainly determine an arc. Thus we need only show that R is transitive. ARB implies there is a sequence of connections from A to B BRC implies that there is a sequence of connections from B to C by combining these two sequences there must be sequence from A to C. A class under the relation R will be a maximal connected subgraph of the graph for which R is the arc relation. Thus the maximal connected subgraphs for a graph partition the vertices of the graph. R is the smallest equivalence relation extension for G for assume that R' is strictly contained in R→ R' an equivalence relation. Then different members of some maximal connected subgraph of the graph in question must be in different classes of R', hence R' cannot be an extension of G.

Let us turn our attention now specifically to the biological problem: the establishment of classifications for collections of biological objects. We h ve been able to establish from biological considerations three guiding principles. These principles are:

1. The definition of classification. A series of partitions for a collection of objects with the property that classes determined by later partitions consist wholly of classes determined by earlier partitions. This is known in biology as the hierarchical property of classification for example genera consist wholly of species. No species can be in two different genera.

2. Two specimens which are judged to be similar (the notion similar will be made more precise presently) should not be separated into distinct classes.
3. Good classes for a biological classification should exhibit some morphological genetic or some other measurable discontinuity from one class to another.

The method I am about to explain accepts these guiding principles as necessary conditions. They are not absolutely necessary conditions in the sense that biological classifications must embody them; but theyy are necessary conditions in the sense that a mathematical model for classification must be based on specific conditions set forth by the biologist in whose aid this method has been devised and in that sense we have accepted these as the bases for our classification.

In order to make precise the second principle, we define what we call a similarity measure for the collection of objects to be classified (this collection will henceforth be called the study) The similarity measure is a real valued function with range $[0,1]$ defined over all unordered pairs of objects in the study. $S(a,b)$ has the following properties:

1. $S(a,a) = 1$
2. $S(a,b) > S(c,d)$ implies that the pair of objects (a,b) is more mutually similar than is the pair (c,d)

The details of the definition of this function are quite involved and I will not go into them here. I will ask you to assume that such a function can be satisfactorily defined. The discussion of this function might make an interesting topic for some future meeting.

I will now define the relation G_c where$_c$ is some number between 0 and 1. G will be defined for the objects in the study as follows:

$$A\ G_c\ B \text{ if and only if } S(A,B) \geq c$$

Suppose that it is possible for us to agree on a value of c some number between 0 and 1 with the property that any pair (a b) of objects in the study are judged to be similar in the sense of Principle 2 whenever $S(a,b) \geq c$.

The relation G_c is symetric and reflexive but not necessarily transitive, thus the study together with this relation constitutes a a graph. We now ask ourselves what is the finest partition [that is to say classification containing the larges number of classes] which does not violate this interpretation of principle 2? Clearly this is going to be the partition established by the smallest equivalence relation extension for the relation G_c which was shown to be the arc relation for the graph whose vertices are the study and whose connections are given by G_c. Lets call this arc relation R_c. The classes under R_c are just the maximal connected subgraphs of the graph determined by our interpretation of Principle 2.

There is one question still remaining however. Namely, How do we choose this value C? There seems to be an arbitrarily large number of choices between 0 and 1. This question can be solved by eliminating it.

If c' is greater than c G_c, will be an extension of G_c. Thus $R_{c'}$ will be an extension of R_c. Another way of interpreting principle 1 is to say that a classification is a series of equivalence relations each an extension of the preceding relation in the series. It is meaningful to ask for a finite collection of points. What is the longest series of distinct equivalence relations with the extension property just described? Whenever there are N members of the study then this

sequence is at most N long.

Proof: Let L(i) be the number of classes in the Ith partition of the series.

$L(i) \geq L(i+1) - 1$ and $L(1) \leq N$ thus

$L(i) + i - 1 \leq N$ $L(T) = 1$ where T is the terminal value for i.

thus $1 + T - 1 \leq N$ as asserted.

This means that although there seem to be a large number of ways to choose a value for c there are only at most N ways to partition the study and these partitions satisfy Principle 1. We will thus choose our set of c values to interpret principle 2 as follows: c is in this set of values if for all $c' < c$ R_c, is not the same as R_c. This set will have at most N members and it is reasonable to ask a computing machine to discover all of at most N partitions.

It is interesting to note that because of the defined hierarchical property for biological classification in this set of at most N partitions for the study there can be at most 2N - 1 distinct classes

Proof: Assume initially that there were N distinct single point classes. When the ith partition is formed the number of new classes in this partitions (i.e. classes not determined by the i-1 partition) is bounded by L(i-1) - L(i). Since L(i) varies from 1 to N there can be at most N-1 new classes formed which are not included in the disjoint partition.

Let us turn our attention to Principle 3 from the considerations taken so far we know that our method can discover at most 2N-1 distinct classes. How can we measure the isolation for these classes? A class will be called a C- cluster if it is a class under the equivalence relation extension R_c but for no $c' < c$ is it a class under $R_{c'}$.

The moat of a c-cluster is defined to be the number c minus MIN [S(a,b)] where a in ℰ the c-cluster and bm is not. The moat of a c-cluster can be thought of as the amount of similarity by which the closest candidate for membership in the cluster fails to qualify. It can also be thought of as an indication of how likely we are to confuse some member of a cluster with some non-member. If the moat for a c-cluster is high the cluster is isolated from other clusters and we are unlikelyy to confuse its membership. Clusters with high moats are good candidates for named groups in biological classification. We can measure each of our at most 2N-1 clusters for moat and decide which ones are best suited in terms of principle 3.

It is also possible to determine a measure of connectedness for a cluster, by taking a normalized ratio of actual connections made within a cluster to the total possible number. We find this measure not particularly useful because the total possible number of connections increases as the square of the total number of objects in a maximal connected subgraph whereas the actual number of connections seems to vary directly.

In order to best interpret the results of the computing machine we have found it useful to actually draw the pictures of the resulting graphs. On these graphs similarity relations are shown with actual connections between objects. Articulation points in these graphs have proven especially interesting since by definition their absence would mean the c-cluster of which they are members would have to be considered two distinct cluster. In one study of orchids that was done in New York such an articulation point was found to be a hybrid with one parent in each of the two otherwise isolated articulation branches. As a function of the

way the objects have been described to the machine articulation points are always intermediate forms of some kind between their articulation branches.

In conclusion let me say that we are very willing to discuss this method, etc.