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The Hunt Institute Archives contains hundreds of thousands of pages of historical content, writing and images, created by thousands of individuals connected to the botanical sciences.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time and social context in which these materials were created, some of the collections contain material that reflect outdated, biased, offensive and possibly violent views, opinions and actions.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 Archival records are historical documents, and the Hunt Institute keeps such records unaltered to maintain their integrity and to foster accountability for the actions and views of the collections'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

8 November 1967

J. F. Danielli
Theoretical Biology Center
4248 Ridge Lea Road
Amherst, New York 14226

Dear Dr. Danielli,

I would be pleased to review:

R. H. Flake and B. L. Turner: "Numerical Classification for Taxonomic Problems"

which has been submitted for possible publication in the Journal of Theoretical Biology.

Very truly yours,

George F. Estabrook

GFE:gm

24 October 1967

Dr. Murray F. Buell, Editor
Bulletin of the Torrey Botanical Club
Rutgers - The State University
New Brunswick, New Jersey 08903

Dear Murray,

Since I have very little complaint with Lems' paper, it will not be necessary to remain anonymous. I frankly think he has written a good paper. There are one or two suggestions concerning procedure that I will make but that is all. As far as his discussion on Echium is concerned I have no value judgement. I think he has defended his design making process for the new taxa eloquently. I wish that 90% of taxonomists would be as evident.

On the procedure I suggest that on page 3, characteristics employed, also list the states of each character employed. This is important for knowledge of establishment of the similarity measure. On page 7, it seems to me that Lems has not been sufficiently detailed in describing the development of his similarity index. If he has used his own methodology then it is necessary for him to describe it. If he has used others then he should give reference to the similarity measure adopted. Still on the same page, about midway down, is the sentence "It is clear that there are at least three distinct groups . . . . etc." I find the clarity not so clear. Please ask Lems to state the meaning of the distinction which is the basis of his clarity. Aside from these comments nothing strikes me as being "wrong" with the paper. I find it mostly "right." If there be anything wrong with Echium then there will have to be someone else who makes that noise.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm

RUTGERS · THE STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

NEW BRUNSWICK, NEW JERSEY 08903

October 13, 1967

Dr. David J. Rogers
Department of Botany
Colorado State University
Fort Collins, Colorado

Dear Dave:

Thank you for agreeing to review the manuscript by Lems. I am enclosing it herewith.

Sincerely,

Murray F. Buell, Editor
Bulletin of the Torrey Botanical Club

MFB/mk
Enclosure - manuscript

12 October 1967

Dr. David M. Prescott
Scientific Editor, BioScience
Institute of Developmental Biology
PSR Bldg. 1,
University of Colorado

Dear Dr. Prescott:

Returned herewith is Davidson's and Dunn's paper.

First, let me say that Davidson has prepared a number of papers which I have seen and most of which I have turned down for one journal or another. For that reason I would appreciate it if you would get an independent decision from mine before any final decision is made on the present paper.

Somewhere, I think, Davidson has a message but somehow he never seems to "cut the mustard." In a way he tries to be a philosopher, a psychologist and a biologist all rolled into one. I feel that he misses the boat on all three. Mathematically we find him belaboring the obvious with a set of mathematical symbols which looks like the score for an opera. We cannot even be sure we can follow all of his statements, but as far as we can determine it looks as though he mostly demands some a priori classification ((see page 15, comments in margin). Furthermore we are more sophisticated, even laymen, than to make a modern-day classification which includes whales, perches and rabbits. Our problems do not involve that sort of division and modern biologists are aware that their classifications do not necessarily reflect the "truth." The problem of weighting also is an unnecessary lot of discussion.

I would recommend to Davidson to go and talk over his problem with a non-statistical mathematician, listen very carefully to his words and try to stick to his business of development of some satisfactory algorithm and program. I don't know why he does not use ours - they work for the purpose of classification.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm

AUG 22 1967

August 21, 1967

103
PSREBJS'

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

Dear Dr. Rogers:

BioScience has received the enclosed manuscript by Drs. Davidson and Dunn entitled "A Probability Model of Biologic Classification" for publication. I would be grateful for your opinion concerning its merits. Do the authors have something worthwhile to say and do they say it well? If you recommend publication, are any revisions needed?

Many thanks for your help.

Sincerely,

David Prescott

D. M. Prescott
Scientific Editor
BioScience

DMP:mjs

Enclosure

EVOLUTION

Ralph G. Johnson, Editor
Department of the Geophysical Sciences
The University of Chicago
Chicago, Illinois 60637

August 18, 1967

Dr. David J. Rogers
Taximetrics Laboratory
Armory 101
University of Colorado
Boulder, Colorado 80302

Dear Dr. Rogers:

Thank you very much for agreeing to review the enclosed manuscript entitled "Maximum parsimony and maximum likelihood in numerical cladistics." I enclose a critic's checklist and a stamped return envelope for your convenience.

I hope that you will be able to return it within the next several weeks.

Sincerely,

Ralph G. Johnson

Ralph G. Johnson
Editor

2 October 1967

Dr. Ralph G. Johnson, Editor
Editor, Evolution
Department of the Geophysical Sciences
The University of Chicago
Chicago, Illinois 60637

Dear Dr. Johnson:

Returned herewith are our collective comments on Farris' paper "Maximum Parsimony and Maximum Likelihood in Numerical Cladistics." In addition to my own comments, part of which are editorial from my experience as editor of Economic Botany, there are included those of two of the Taximetrics Laboratory who are professional mathematicians. Where one stops and the other begins is probably evident, for we haven't made too much effort to "smooth out" the critique.

From this review and critique I think you can see that we consider both the author and the paper well worth encouraging, but also that we have many misgivings about the work. I hope that you can diplomatically transmit these sentiments to the author, a young man who has much promise as a contributor to this area. We do not wish to "blast" him with our comments, but rather to give him a nudge in a direction we think will be useful if he wishes to make himself heard.

Some of our comments are dictated by our own experience in reporting w work of this type to audiences primarily in systematics. We feel that, if biologists are to be the major audience, we should present our works in terms most meaningful to them. When we cross disciplinary lines to include mathematics, at least to this generation of biologists, we'd better be very careful to get it out with the least amount of confusion. In its present form it satisfies neither a mathematician nor a biologist.

I hope the time lapse hasn't been excessive, and that we have done the job to your satisfaction.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm

COMMENTS

1. Editorial

The most striking point is that the paper has no headings. This is critical, for there are no means to see the structure of the work, nor indeed to indetify the type of paper, whether a new method, a critical review of other work, or a combination of these. There should be a number of headings, major and minor, and a number of identified paragraphs (i.e. letter and number headings).

In this work, a precise abstract, placed at the head of the paper, would serve a useful purpose for those who want to know what the title really means. There are at least three ideas in the title which I am certain are not well defined in evolutionary biology.

There are some places where a different placement of the paragraph, sentence or phrase could be helpful. Some of the problems are mentioned below.

On p. 15, eleven lines from the top, Olson, 1964 is cited, but there is no citation of Olson in "Lit. Cit."

2. Contents

The introduction to the paper should clearly indicate what the paper is to cover, whether this be a review, or what (see above). Not until some time after one has begun reading does it become obvious that the author proposes some means to evaluate methodologies and papers written about parsimony as a study of evolution.

Shortly after the introduction, Farris should state his own initial assumptions about parsimonical evolution as such. Then he should define all his terms that he is going to be using in the rest of the paper. It is not correct to refer the reader to another paper for a definition when the

concept is going to be a major part of the present work. Furthermore, there are no agreed-upon definitions for terms that he borrows from others, such as cladistics, particularly when he adds "numerical" to it, nor unit character, likelihood function, distribution of the $X_{(i)}$, distribution vs. density, etc. If he wants, he can decide on his own definitions for these and as long as he sticks to them, I'll be happy. On one page (p. 6) Farris uses "patristic lengths" and "patristic difference." Are these the same, or are they different.

In the course of the mathematical development, he should insist on rigor only insofar as it is necessary for the argument. Introducing notions of "rational valued R.V.'s", "Markovian Processes," "discrete or countable distributions vs. densities," "notions of convergence" and "limits" are perhaps unnecessary to develop the conclusion desired, viz. that the probability of a "tree" is measurable by its "length." If a biologist or anyone else is forced to this conclusion as a consequence of the rationals being countable (e.g.) he will not be happy and possibly be suspicious.

With respect to accuracy and rigor, the following statements, quoted from the text, are mathematically false:

p. 4. "the difference between x(i) and x(i - 1) tends to zero* as n becomes large." *this, or course, should be defined.

p. 5. "as n becomes large [exp] L converges to the probability of the path considered as a continuous-time process, and D converges to the length of the path."

p. 5. "It is clear from the diversity of life* that d(i) is not zero with probability one." * the diversity of life has nothing to do with it.

p. 6. "The negative exponential law relating probabilities of paths to the patristic lengths* holds for any amount of elapsed time."

* see comments in first paragraph above.

Because of the shortcomings mentioned, and others not described, it is difficult to unambiguously interpret the assumptions and definitions. A number of interpretations of the assumptions have been tried, but from none of these do the conclusions follow. It would seem that w is not a constant but rather a function of the time interval initially chosen. Perhaps with some of the suggestions made, and some reorganization, we could be more certain of the appropriate interpretations.

Following is a suggestion. It is felt that the "proved" result is a reasonable way of describing the biological phenomena under study. On pages 8, 9, 10 of the manuscript good biological arguments are made supporting the result but diluting the mathematical precision of the argument on pages 5 and 6. Why not start with the assumptions defended from biology ~~and~~ On pages 8 - 10, viz, that rate of change in probability of D is proportional to the probability of D. This gets $\frac{dP(D)}{d(D)} = wP(D)$. In an easy argument we get $P(D) = e^{(wD)}$ as desired.

UNIVERSITY OF CALIFORNIA PRESS
BERKELEY · LOS ANGELES · NEW YORK
2223 Fulton Street · Berkeley, California 94720

July 31, 1967 AUG 2 1967

Professor David J. Rogers
Department of Biology, Taximetrics Laboratory, Armory 101
University of Colorado
Boulder, Colorado 80302

Dear Professor Rogers:

The Crovello manuscript and the art work arrived this morning. Thank you for your most thorough review which came on July 28th.

A form,"request for issuance of check", has gone to the University Accounting Officer. I have asked that the check be sent to the above address. I trust it will be written and mailed promptly.

Again, many thanks.

Sincerely yours,

Hazel Niehaus

(Miss) Hazel Niehaus
Administrative Assistant

AUG 10 1967

AUG 1 [illegible]

EVOLUTION

Ralph G. Johnson, Editor
Department of the Geophysical Sciences
The University of Chicago
Chicago, Illinois 60637

August 1, 1967

Dr. David J. Rogers
Curator of Economic Biology
The New York Botanical Garden
Bronx Park
New York, N.Y. 10458

Dear Dr. Rogers:

Farris

A manuscript entitled, "Maximum parsimony and maximum likelihood in numerical cladistics" has been submitted to the journal EVOLUTION for publication. Your name has been suggested to me as a possible reviewer. I would like your permission to forward it to you for critical review.

If you find that you cannot read the manuscript within the next several weeks, please refuse this request. I would appreciate receiving your recommendations for other possible reviewers.

Sincerely yours,

Ralph G. Johnson

Ralph G. Johnson
Editor, EVOLUTION

rl

Ddb gal wa + ppr

Card returned in affirmative. 11 Aug 67

Review of proposal "a Proposal for the Development of a Center for Information Services Phase II: Detailed System Design and Programming."

filed in Information Retrieval folder.

1R

- Taximetrics Laboratory

June 20, 1967

Refer to: Proposal N-2898

Dr. Jerry S. Kidd
Program Director
Special Projects Program
Office of Science Information Service
National Science Foundation
Washington, D.C. 20550

Dear Jerry:

I am sorry to have delayed so long in making comments concerning "A Proposal for the Development of a Center for Information Services Phase II: Detailed System Design and Programming." My comments are put onto a separate paper for your benefit.

The procedures listed in this proposal will be complementary to our own activities and we will certainly want to work closely with people who seem to have a good grasp of the situation.

Sincerely,

David J. Rogers
Professor of Botany

DJR/ch

Enc.

"A Proposal for the Development of a Center for Information Services Phase II: Detailed System Design and Programming"

In general we find the proposal well balanced, conceptually sound and useful. The generalized software philosophy seems good. We approve of generality rather than standardization in the approach such as this one. Clearly the principal investigators are on their toes.

It is unfortunate that there are few or no details for hardware or software. Had there been some given in the proposal, we would have had a clearer picture of the modus operandi. Perhaps the investigators have not found it possible to give a clearer picture than they have. I would like to see beforehand a little more precise description of their procedures.

The most distressing feature to me is the setup for their budget. I should think for the amount requested considerably more detail could have been given as to how the money will be spent. It would be more of a revelation about their operation than I now have. Also it seems to me the investigators have been somewhat chary in letting us in on the personnel who will make the system fly. Such detail would help us in the evaluation of such a proposal.

My overall evaluation of this proposal is good. The usual scale of research proposals of 1 to 5, where the lower numbers are the higher ranks, I would put this one between 1.5 and 2--excellent to good.

NATIONAL SCIENCE FOUNDATION
WASHINGTON, D.C. 20550

Office of Science Information Service

May 23, 1967

In reply refer to:
Proposal N-2898

Professor David Rogers
Department of Botany
Colorado State University
Fort Collins, Colorado 80521

Dear Professor Rogers:

We should be grateful for your assistance in evaluating the proposal from the University of California, Institute of Library Research entitled "Development of a Center for Information Service, Phase II: Detailed System Design and Programming." In making decisions on such proposals it is our practice to obtain and consider the reactions of a number of individual experts in the appropriate field, and in this instance we should like very much to have your views.

Any comments you can send us on the soundness and potential value of the study as outlined, the competence of the investigators, and the reasonableness of the budget will be most helpful to us. Your remarks will, of course, be held in confidence. In completing your comments please indicate clearly whether you favor support, have mixed feelings and are undecided, or favor declining the proposal.

Sincerely,

Jerry S. Kidd

Jerry S. Kidd
Program Director
Special Projects Program

Enclosure: N-2898

BIOGRAPHY -- Robert M. Hayes

Professor Hayes received his Ph.D. in Mathematics in 1952 from the University of California, Los Angeles. Since that time he has taught, done research, and solved problems in each of the successive areas of application of modern data processing technology and methodology--numerical analysis, real-time control, business data processing, information handling. He is presently Professor in the School of Library Service at UCLA and Director of the University of California's Institute of Library Research. Prior to joining the University, he established and headed a private consulting and research company, Advanced Information Systems. He has been employed by the National Bureau of Standards, Hughes Aircraft Company, the National Cash Register Company, and the Magnavox Company. He has taught special courses in information science for the University of Washington, American University, Georgia Institute of Technology, and the US Air Force. He is a member of the American Mathematical Society, the Special Libraries Association, Phi Beta Kappa, Sigma Xi and other societies. He is the author of several publications in the information sciences, and co-author of the book Information Storage and Retrieval: Tools, Elements, Theories. He holds several patents on equipment developments in the field of information storage.

THE REGENTS OF THE UNIVERSITY OF CALIFORNIA

INSTITUTE OF LIBRARY RESEARCH

LOS ANGELES AND BERKELEY, CALIFORNIA

A PROPOSAL FOR THE DEVELOPMENT OF A CENTER FOR INFORMATION SERVICES PHASE II: DETAILED SYSTEM DESIGN AND PROGRAMMING

PRINCIPAL INVESTIGATOR:	R.M. Hayes, Director Institute of Library Research
CO-PRINCIPAL INVESTIGATOR:	Robert Vosper, University Librarian University Research Library
TO:	The National Science Foundation
SUPPORT REQUESTED:	$ ~~309,330.00~~ 310,593
PERIOD:	1 July 1967 - 31 December 1968

R.M. Hayes, Director
Institute of Library Research
UCLA

Research & Extramural Support

Robert Vosper, University Librarian
University Research Library
UCLA

ABSTRACT

This proposal is for continuation of the existing effort in specification and development of a "Center for Information Services", designed to acquire and utilize a large variety of nationally produced, mechanized data forms. The work to date has resulted in specification of the required software and hardware for use in such a center. The proposed effort will produce a detailed design of the computer programs and operational procedures.

TABLE OF CONTENTS

I. Background

About a year ago the Institute of Library Research of the University of California, under funding from the National Science Foundation, started a program to study the problems in introducing into a single university campus library the means, the technology, and the administrative procedures for handling media of the newer kinds, including magnetic tape and microforms. These have been developed for a variety of purposes outside those that have normally been considered within the scope of the library. The concern is with the problems in acquiring such media, in cataloging them, and in providing service based on them.

The interest in mechanized information services in the library arises from the so-called "information explosion" and the consequent need felt to exploit new tools in coping with the increasing size of libraries, the development of data bases in machine-processible form, and the production of generalized computer programs that can process a variety of machine-stored data. The result is that research supported by automated processing of machine processible data bases will pervade virtually all academic departments of the University. Since needs are expected to cut across departmental lines, scholars and students in all disciplines should have access to the entire campus data base, as well as to the data bases in existence elsewhere.

Of immediate significance in this respect, are the large number of national programs which are now generating cataloging and indexing data in mechanized form. The Library of Congress is now embarked on an experimental project and will certainly be producing its catalog data on magnetic tapes

within two to three years. The National Library of Medicine and the Defense Documentation Center are already doing so. Chemical Abstracts, Historical Abstracts, and similar abstracting services are experimenting right now with similar mechanization. Innumerable socio-economic data banks are being established in mechanized form. There are a number of commercial organizations in book distribution which are installing mechanized catalog services. In summary, mechanized data on which "information retrieval" services can be provided will become readily available to every large library throughout the country. In addition, there are an equally large number of national programs to record books, documents, and other material in microform -- microfilm, microfiche, microstrips, microcards.

The large number of existing national projects amply demonstrate the extent to which mechanized information retrieval is a reality today, and the number of these national programs generating data in magnetic tape form is continually increasing. Although most of them have been oriented toward meeting the special requirements of the agencies involved, as a whole they now constitute a national resource of great magnitude and importance. The efforts to create information networks which would make the stored data readily available to users everywhere have been many and varied -- the Stafford Warren proposal,[1] the studies by COSATI,[2] the program of the National Library of Medicine,[3] the efforts of EDUCOM.[4] The planning for them has focussed on administrative

---

1. Kent, Allen, Library Planning for Automation, Spartan Books, Washington, D.C., 1965, pp. 3-33.
2. Carter, Launor F., et al, National Document-Handling Systems for Science and Technology, Information Science Series, John Wiley & Sons, Inc., 1967.
3. "The National Library of Medicine Index Mechanization Project," Bulletin of the Medical Library Association 49, 1 (1961), Part 2.
4. In preparation.

issues in such networks, on the physical problems of creating communication systems which will link computers together, and on the kinds of usage which such networks might satisfy. But in all of this planning, there has been the explicit view that the nation's major research libraries are important components in an adequately functioning system. An important issue, therefore, is the role which library automation can play in these developments of mechanized information retrieval services and national information networks.

The progress toward mechanization in the library can be seen already, as libraries throughout the country install mechanized circulation systems,[5] use data processing equipment in the production of catalogs,[6] and process acquisitions data for serials, documents, and books with computers. The extension to include mechanized data files is perhaps not so immediate, but it may result in more fundamental changes in the nature of the library and its services.

2. The State of Developments

To provide background, it is worthwhile to review a number of these national programs, with the intent of demonstrating the great variety of substantive areas and forms of data now available in magnetic tape form. In general, they can be roughly classified into the three classes: reference files, data files, and text files. However, the bulk of those of immediate interest and the ones which raise the largest number of pragmatic issues are the reference tapes.

---

5. Cox, James R., "Automated Circulation Control in the University Research Library at UCLA," UCLA, University Research Library, 1965.

6. Cartwright, Kelley L., and Ralph M. Shoffner, Catalogs in Book Form, Institute of Library Research, University of California, January, 1967.

American Bibliographical Center
800 Micheltorena Avenue
Santa Barbara, California 93108
Dr. Eric Boehm, Director (805) 962-6582
Mechanized production of the five year index for Hist. Abstracts, containing all bibliographic elements exclusive of the abstracts themselves.

American Chemical Society
1155 16th Street, N.W.
Washington, D.C.
(202) RE 7-3337
Joseph Kuney, Director of Publications Research
Steve Walcavich, Programmer in Charge
Began in 1966 to produce Journal of Chemical Documentation by computer-driven Photon 200. About 100 articles for 1966; 500-600 expected by end of 1968. Also putting some of the articles in the Journal of Chemical Engineering Data for 1966 in machineable form. Goal is to produce all of ACS journals in this way, possibly by 1968.

American Institute of Physics
335 East 45th Street
New York, New York 10017
(212) MU 5-1940
They are working on a thesaurus in magnetic tape format as the first step toward mechanization.

American Petroleum Institute - Division of Refining
Central Abstracting and Indexing Service
555 Madison Avenue
New York, New York 10022
Mr. Everett H. Brenner, Manager
Currently abstracting 1,000 doc's. and journal articles and 1,000 patents per month and storing in magnetic tape form.

American Society for Metals
ASM Documentation Service
Metals Park, Ohio 44073
Mrs. Marjorie Hyslop, Assoc. Director
(216) 338-5151
Literature in the field of Metallurgy, stored in magnetic tape form.

Applied Mechanics Review
Southwest Research Institute
8500 Culebra Road
San Antonio, Texas
Mr. Stephen Juhasz (572) 684-2000
Complete data in magnetic tape form for the WADEX (word and author index) system -- actually a series of programs and data to create the index to AMR.

Atomic Energy Commission
Germantown, Pennsylvania
John Sherrod, Asst. Director; Director of Information
(202) 973-4371
Two sets of magnetic tapes currently produced: 1) Used in production of Nuclear Sciences Abstracts; 2) A general type used in SDI programs

BioScience Information Service
3815 Walnut Street
Philadelphia, Pennsylvania 19104
(215) 386-0414
Phyllis V. Parkins, Director
Miss Louise Shultz, Asst. Director for Systems Development
Mechanized system is used for producing Biological Abstracts and its four indexes:

1. Author
2. Permuted fragments from an augmented title listing
3. Coordinate Posted Index - a precoordinate index, using only the terms which are used as headings and subheadings of BA sections. These then are large terms rather than detailed specific uniterms.
4. Biosystematic index - (Taxonomic)

R.R. Bowker Company
1180 Avenue of the Americas
New York, New York 10036
(212) LT 1-8800
Mr. John Berry III
Has automated the production of Pub Weekly, BPR etc. and wishes to encourage general use of tapes (Wish to market own tapes eventually). They have more than one data bank available to any interested subscriber.

Chemical Abstracts Service
2540 Olentangy River Road
Columbus, Ohio
Mr. Elden G. Johnson, Manager; Subscriber Information Service
(614) 293-7423
293-5022

1. Chem Abstracts [CA] are not yet automated in any way, except for indexes to some issues being on tape. Hope to have CA produced by computer-driven Photon by 1969.
2. Chem Titles [CT] began in 1962. This is a KWIC-type index with about 500,000 titles.
3. Chemical Biological Activities [CBAC] began in 1964 and now has about 21,000 titles with abstracts.

CAS has a registry system which takes two-dimensional drawings representing compounds and searches for those containing the same structure, no matter how displayed. Now have 500,000 compounds, with a capacity of 5 million.

Clearinghouse for Federal Scientific & Technical Information
5825 Port Royal Road
Springfield, Virginia 22151
Bernard Fry, Director
Peter F. Urbach; Asst. Director, Systems
(703) 321-8500
Embarked on the production of consolidated indexes to Federal STI (NASA, AEC and DOD) using computers.

Institute for Scientific Information
325 Chestnut Street
Philadelphia, Pennsylvania
(215) WA 3-3300
Dr. Eugene Garfield, Director
Dr. Irving H. Sher, Director of Research
Tapes available on lease or buy arrangement. File of Science Citation Index, ASCA, and ISI Search Service.

Library of Congress
First and Capitol Streets
Washington, D.C.
(202) 783-0450
Mrs. Barbara Markuson
Project MARC is an experimental, but continuing, effort to provide primary cataloging data in magnetic tape form. Presently limited in distribution to 16 participating libraries.

NASA (by contract with Documentation, Inc.)
4833 Rugby Avenue
Bethesda, Maryland
Mr. Herbert White (in charge of NASA-STAR program).
(202) 696-9500
NASA-STAR is produced using magnetic tape. Now developing an on-line service. Establishing sub-centers at several universities.

National Standard Reference Data Systems
Gaithersburg, Maryland
Dr. Steven Brady, Director
(202) 921-1000
Indexed bibliographies and data in magnetic tape form. Bibliographies run to 30K references with notes on data content.

National Library of Medicine
8600 Rockville Pike
Bethesda, Maryland
(202) 656-4084
Bethesda Office (301) 654-9190
Has been producing Index Medicus for several years using mechanized methods and is servicing requests for searching of the existing files. Establishing sub-centers throughout the world.

Office of Education
400 Maryland Avenue, S.W.
Washington, D.C.
Project ERIC (Educational Research Information Center) is proceeding to place citations and abstracts on magnetic tape.

Scientific Information Exchange
Smithsonian Institution
Madison National Bank Building, Suite 300
1730 M Street, N.W.
Washington, D.C. 20036
Dr. Vincent Maturi, Chief, Physical Sciences Division
Monroe E. Freeman, Director
Magnetic tape files include descriptions of grant supported research projects and an inverted subject index to them.

3. The Role of the Library

With this ever-growing store of nationally available mechanized data bases, the role of the library becomes a real and very important issue. For a variety of reasons, it seems clear that the library has a three-fold responsibility: 1) as the agency for acquisition of data of high utility; 2) as the point of entry into the national networks, and 3) as the point for assistance in the use of this form of data.

A small, but growing number of libraries are already acquiring and using magnetic tape data from a variety of sources. At UCLA, for example, the library is already receiving on a continuing basis, magnetic tapes from the Bureau of the Census, the National Library of Medicine, and the Library of Congress. As the above listing demonstrates, there are at least a half-dozen others which are immediate candidates for acquisition, covering every academic field from biology to engineering to business administration to art history to education. The fact is that magnetic tape is becoming as important a form for support of research as the book and the journal -- and for certain purposes, an even more useful one. And the library is a useful administrative organization for acquiring it.

But whether or not the individual library itself acquires a given data base, availability of it from wherever it may be stored will require points of access -- consoles, procedures and programs, methods of communication and control. Although one can visualize the individual investigator sitting in his own office and using his own console for communication to remote data bases (at other campuses and at national information centers), the realities -- of economics, of the operating schedules and available services at the remote points, of the sheer volume of data processing and transmission involved -- all suggest that the vision will apply only to a limited number of investigators and a limited set of data bases. And yet, the library can provide a convenient administrative mechanism for access to national networks by other investigators. It already serves this function for the traditional printed material; it can maintain the union catalogs and directories by which to locate the mechanized data as well; it has the procedural mechanisms to ask for it; it has its own needs to obtain such access; and it is now experimenting with the operational reality of rapid communication so necessary for use of it.

Finally, much though one can also vizualize the "machine-aided search," the fact remains that the use of mechanized data bases is extremely complex, requiring training not only in the methods of using machinery, but in the ways of formulating needs and of interpreting results. There may be a limited number of investigators sufficiently informed and willing to work with new data bases, but for the bulk of researchers, technical assistance will be a necessity. Libraries and library schools are now in the process of creating a new generation of librarians who will have the combination of capabilities required to provide such service.

The point of course is that the combination of requirements -- for acquisition of data on a continuing basis, for cataloging of it, for providing economic access to it, for aiding in the use of it -- all serve to

make the library a key agent in mechanized information retrieval services. Furthermore, a crucial issue is whether mechanized information retrieval services are called for in a specific situation (their value in general seems self-evident and demonstrated by the ever-growing number of mechanized data bases). There are always alternative resources, most of which are a traditional part of library service, and the librarian is in an ideal position to evaluate their relative utility.

II. A "Center for Information Services"

But having said all of this, we are faced with some enormous pragmatic and immediate problems -- economic, administrative, technical. To provide a framework within which to discuss them, consider a "Center for Information Services" in the library providing the mechanized services involved in using such data bases.

The Center, as an extension to normal library activities, would supplement both the media handled and the methods of operation. The usual library functions are those of acquisition, storage, cataloging, and circulation. The media presently include books, serials, microforms, and such special collections as manuscripts, incunabula, and archives. The Center for Information Services would emphasize the acquisition, storage, indexing, and dissemination of computer processible media, such as magnetic tapes. The emphasis on indexing rather than cataloging, and dissemination rather than circulation, brings out the special attributes of a computer based system. Thus, while circulation suggests a reader taking a book out of a library, dissemination includes both this and active transmission of the contents to a reader at a remote console. Cataloging suggests the standard author, title, and subject guides. Indexing includes not only these, but greater depth of detail as well, through the ability of computer programs quickly to process vast amounts of data according to detailed and complex instructions. In this respect, the Center must be able to process a great variety of machine processible records. Many of these are already in existence but others are not, and therefore the emphasis must be placed not just on the utilization of existing information retrieval programs. Rather, the need

is the development of a software capacity for processing a wide variety of material.

Achieving such a Center as a completely operational service by, say, five years from now, with recognition of the problems in effective transition from today's services, implies a series of scheduled tasks in the intervening period. We must define the underlying assumptions of the Center, its physical makeup, its organization within the University, and the procedures which it is to follow. The result of this development should be a complete, operational system, tested and proven in an operating environment, and suitable for replication at many universities. It should include not only computer programs but also recommended operating procedures, recommended relationships among the Center, the Library, and other departments and functions of the University, recommended policies for funding and administration, methods for interchange of data bases and associated computer programs, and recommendations concerning other issues involved in procuring an effective, viable information service.

The development of the Center therefore must have the following design goals: First, it should be operational: It should be designed to meet the daily needs of the University community, and not simply be a research system or an experiment. Second, the Center should be a general purpose system: It should be able to accept a wide variety of both existing and future data and should be able to satisfy a wide variety of requests. Third, the system should be adaptable, so as to meet needs not initially anticipated: It should incorporate a capability for monitoring the history of its own operation and for continually studying that histroy to introduce improvements. Fourth, the system should be replicative: It should be designed so that it can be installed at many places and serve many needs as they arise; this demands

a more polished design than if only a single center were to be developed, and requires careful attention to documentation and to procedures for installation. Fifth, the system should encourage increased receptivity and use: Education of prospective users, design for easy use, and responsiveness to changes in use must be stressed. Sixth, the Center as an administrative part of the Library should be designed so that library personnel could operate it and provide the ongoing services of acquisition, cataloging and indexing, storage, and dissemination of the material peculiar to the Center.

Physically the Center is envisioned as a storage facility with a small, library-based computer, including tape drives and disc packs. It would be linked to a campus computer network, including connection to on-line consoles. Requests to the Center would be handled by library personnel who could decide whether to use traditional resources, to use the library-based computer, to use it in conjunction with the more powerful network computer, or to allow the user to conduct searches himself at his console connected to the large-scale computer.

Because data bases from elsewhere will also be important, the Center must be designed as a potential node in an inter-university network. Before the end of the next ten-year period, computer networks and time sharing will have become a normal educational tool and an operational aid to research. Since networks connecting various universities will require well-defined points of access, suitable for those not familiar with them as well as for those sophisticated with their use, the Center will occupy a particularly important role. The library has been the traditional operational informational resource on campus and it should be natural to turn to it for the service we are considering here.

In summary, the Center for Information Services is engendered by the

developments of modern information technology. Its development will provide a supplement to the media and method of operation of the usual library and will include development of policies and procedures, relations to other organizations, and cooperation with other centers. Emphasis must be placed on integrating a wide variety of data bases and programs. The system must be operational, general purpose, adaptable, replicative, and designed to encourage easy use. Organizationally the Center is viewed as an administrative part of the library. Physically it is viewed as a storage and processing facility embedded in a large complex network of computers.

A. The Problems

The intent of the present, NSF - sponsored study then is to study the problems involved in such a development and particularly some specific issues. Some of the issues relate to the content: What kind of material should the university library acquire? Some of them concern library processes: How do we catalog magnetic tape material? Some of the problems are technological: How do we provide man-machine communication? Some of them are administrative: How do we finance it? How do we fit it within the traditional library structure? How do we relate the library to the computing facility? The study then is directed toward a detailed specification of what this library-based Center for Information Services (as we have called it) should look like, and what programs should exist in the computer facility to provide the kinds of services that are wanted.

As one step in accomplishing these aims, symposia are being held, bringing people together who are informed in each of several areas of the physical, biological, and social sciences. Essentially, these symposia will try to answer two questions, or at least pose two questions and then

see what answers arise: First, given such a Center as described above, what material, of magnetic tape form in particular, should it acquire from nationally available sources to meet the needs of the scientist on campus? Second, what kinds of services should the Center provide? These symposia then are directed at the first set of issues related to the Center, (viz., the content issues concerning what kinds of material such a Center should have). Another set of symposia are concerned with specific technical problem areas: the problems in cataloging this kind of material, the problems in acquisition of it, the problems in the man-machine dialogue with it, the relation to library clerical processes, and the relation to national networks.

With respect to technological issues, capabilities are now under development completely adequate to meet the needs for storage and processing of this mechanized data and microform records. The computers are fast enough; the programs, or "software," are sufficiently developed; the mass memories are large enough and provide rapid enough access. Perhaps more significant is the appearance of "on-line" capabilities which will make the computer virtually a utility with direct access to it through typewriters, input devices, and consoles of varying levels of sophistication. The technology for communication is so advanced that it places virtually no limits on the transmission of reference data, text, or images. The technology of microforms has undergone a minor revolution over the past ten years, particularly because of the needs in large government-supported programs such as NASA, AEC, etc. Finally, the abilities to reproduce copies of text material -- cheaply, rapidly, and with good quality -- have already introduced major advances in library service.

As a result, the dominating technical constraint appears to be the requirement for ability to handle data from a variety of existing files. The processing and output preparation of the data once it has been selected and extracted is a relatively straightforward (although by no means trivial) task. The heart of the matter, therefore, is the capability to respond to users' requests to read, select, and extract data from files prepared by other organizations. Specifically, how do we add data bases without proliferating programs to the point of virtual strangulation? As it now is, each data base has its own format, its own thesaurus, and its own package of "file management" programs which provide capability for maintenance and search. Each data base now requires a separate set of forms and procedures for utilization.

The answer might lie in standardization, but that seems hardly likely, in view of the enormous variety of purposes served by the data bases to those who originate them. It might lie in conversion of the data bases to some standard format and structure by the library using them, but this also seems unlikely, in view of the sheer bulk of data involved. It might lie in the use of generalized file management programs which can handle the variety of data bases and provide standardized services based on them.

The conclusions, based on the work done to date, are that custom programming for each data base is too lengthy, too costly, and too unresponsive to the needs of the Center and its users and that translation or transliteration of files for use in some standard system is impractical because of the possible loss of meaningful information, the costs, the continual changing of formats, and the difficulty in processing. These points, when combined with the uncertainty of future data base formats and the changing nature of user requirements, all suggest the development of a

generalized system appropriate to Center operations as the only solution.

The design of the generalized system will be special purpose insofar as it reflects the special requirements of the Center for Information Services. Many recently developed generalized file management techniques, however, will form the basis for system design. In the next section, these characteristics are described in terms of the needs of the Center for Information Services.

To use such generalized programs requires a careful description of each data base both so the generalized programs can operate on it and so the user can know what level of service he can call on. Usually, these programs provide a clearly distinguishable set of stages of processing, from fixed field processing (the simplest and most efficient) to open field processing to text processing. Their relative efficiencies differ so radically that the prospective user must be well aware of precisely what data from a given data base can be effectively processed by a given level of program.

This leads to questions relating to the cataloging of magnetic tape material. They combine issues of traditional library descriptive and subject cataloging with issues of inventory control and file format definition. If the kind of utilization of this form of material, which is being talked of in all the experiments and planning, is to succeed, cataloging standards will need to be established, and a national union catalog of available tape data maintained.

B. Generalized Programs for the Center

It may be said that any computer programming language is general purpose in the sense that it is not limited to particular files and functions. However, in order to relieve the programmer of some detail, the notion of higher

level languages was developed. The best known of these languages are COBOL, PL/1, FORTRAN, and ALGOL. The use of these languages is said to result in an average reduction of about 5 to 1 in the number of instructions which must be written by the programmer to perform a given application.

A generalized program introduces a still higher level of communication between the user and the computer. It provides a defined set of capabilities which can be applied to virtually any data base. By relieving the user of many of the requirements to communicate the tasks embodied in it to the computer, it permits use of the computer without formal training in computer programming. Through the concept of different levels of communication between the user and the computer, it may be used by faculty, students, library personnel, system analysts, or computer programming specialists -- at the appropriate level of detail. Thus, instead of employing assembly language or a higher level language, the user employs a small set of structured forms to describe his problem solution in the amount of detail required by the generalized program.

Thus, the most important advantage of a generalized program is its simplicity of use. The user merely answers a series of questions, describing the results he requires in standardized form. An ordinary search request, for example, can be described directly by the researcher or librarian in a few minutes. More complex and sophisticated problems and even the installation of a completely new data base can be described in a few hours.

The generalized program is then used for producing computer programs for normal day-to-day operations (as well as for specialized requirements). Typical functions which may be involved in such operations include the creation and maintenance of files from original input (e.g., punched card) data, the selection of records from files according to established or computed

criteria, computations involving data from selected records, extraction and sequencing of results dependent on these data, and the preparation of output in printed form or in the form of new files for use in other applications. For a given application, the file and the functions to be performed are independent of each other, thus providing great flexibility in use. In execution, however, they are tied together in order to minimize the information which must be provided by the user to the System.

The Center for Information Services as presently planned is based on such a general purpose file management system. It will accept descriptions of the files upon which it is to operate and of the operations which it is to perform. A great variety of file strucutres may be defined to it in a manner independent of the functions; similarly, functions to be performed with data from these files may be defined independently of the file structure declarations. Thus, the functions are not limited to any particular set of files or to any particular set of functions for a given task.

A range of capabilities is being planned at this time in order to accommodate a variety of user needs, trade-offs of capabilities, and specific data bases to be included. Some of these functions are:

1. Read existing files from punched cards, magnetic tapes, and other machine-readable input.
2. Maintain files by making additions and deletions.
3. Re-format files to reflect changing specifications and requirements.
4. Select, from files, records that contain data of interest in a problem.
5. Extract data items from the selected records, or use whole records.
6. Arrange output by sorting, sequencing, and grouping.
7. Format printed reports that contain such elements as Preface, Page, Title, Column Headings, Column Footings, Line Numbers, Detail Entries, Summaries, Statistics, Line Count, and other details that make a printed report or document informative and attractive.

8. Summarize data to as many levels of totals and sub-totals as required, with wide flexibility in format and content of printed output.

9. Compute new values based on values in the file, for use in selection, further computation, printed output, subfiles, or the updated file.

10. Produce printed reports or other printed documents such as 3 x 5 cards, labels, or output on preprinted forms.

11. Produce subfiles on cards, magnetic tape, disk, or other media for further processing by CISS or other systems.

12. Coordinate related files for simultaneous use.

The system will provide for the storage of source programs, in the language, in a "library" for subsequent compilation. By storing the source program, rather than the object program, the system enables the user to conserve space in his system library for other purposes. In operation the user has the option of re-running such programs by recalling them either in source or object language form and operating under the System. This capability supplements the ability to define new applications.

The generalized program is centered around the concept of "master files." In order to extract or retrieve data from files, the problem statement must refer to previously defined field names in one or more specific file. When processing requests are presented, the files with which they deal must therefore have been previously defined. The file definition specifies certain overall file parameters, such as record format and block size. Record structures can be fixed or variable in length and can contain:

1. Variable length fields and segments.

2. Repeated fields and segments of the same type.

3. More than one type of format of field or segment at any hierarchial level.

4. A number of hierarchical levels of segments within a record.

5. Various format types and sizes of records, segments and fields in a file.

The capability to maintain and query master files, once the user has defined the master file and the query specification, is then essentially automatic. This type of implicit specification is a basic design concept of the system. Simplicity has been emphasized in order to maximize the ease of use.

For example, a "standard" mode of operation will automatically be invoked unless the user specifically requests an alternative mode. These "De Facto" cases are applicable in many situations.

The retrieval capabilities of the system enable the user to select and extract data from the files, based on the logical selectivity capability of the system. These include appropriate comparators, Boolean connectors, and types of comparands. Conditional expressions may be combined and an adequate number of nesting levels is provided.

File access aids, such as 1) inverted files, 2) keyword indexing, and 3) dictionaries and thesauri, and hierarchically structured subject headings are treated as separate master files.

The system includes monitoring capabilities, with provision for:

1. Accumulating utilization statistics by user, file, type of request, etc.
2. Cost accounting and charging of accounts.
3. Protecting of proprietary files.

III. The Development Program

A. Phase I. Specification (Work to Date)

In the proposal which led to funding of the present work, the concept of the Center for Information Services was outlined and specific areas of work were described. These were:

1. Review of Available Programs. This has covered the relevant projects underway at UCLA and elsewhere. Specifically, it included the mathematics citation index project, the socio-economic data bank, the Medlars center, and the MARC project -- all as now being developed on the UCLA campus. The bulk of this work, however, has been devoted to survey of the characteristics of available programs and data bases from national sources. The institutions covered and a summary of their characteristics are tabulated in Table 1.

2. Programming. The work over the past year in the area of programming has been explicitely concerned with developing programs for console communication. The specific substantive vehicle used in the testing of these programs has been the on-line file of mathematical citations.

3. Planning for Services. As the first step toward involvement of the scientific faculty on Campus, a number of symposia have been held (one with the social sciences faculty, one with the physical sciences faculty, and one planned for the biological sciences faculty). These have had the purpose of alerting them to the potential represented by the availability of nationally produced data bases and to the kinds of services the Center would provide.

4. Specification and Planning for Implementation. The bulk of the effort under the present NSF grant has been concerned with the following

Table 1. Characteristics of Data Bases

Organization*	Type of Data Base	Available	Documentation Received	EDPE	Record Format	Maximum Block Size (Char.)
ABC	R	Yes	Some	Cards	F	
ACS	NT	Yes	Yes	1460	F	500
AEC	R	Yes	Yes	7090	V	2,004
AMR	R		Some	1401	F	63
API	R	Yes	Yes	1401/7090	F,V	12,000
Bowker	R,NT	Yes	Some			
BPA	R	Yes	Yes	1401	V	
BSIS	R	No	Yes	1440	F	10x90 = 900
CAS	R	Yes	Yes	1401/360	F	12x81 = 972
CENSUS	D	Yes	Yes	1105/1401		
CFSTI	R	Later	Yes	UNIVAC/IBM	V	1,024
EJC	WL	Yes	Yes	1401	F	10x60 = 600
F & S	D	Yes	Yes	1401	F	4x435 = 1,305
GE-FPD	R	Unknown	Some	7094	V	53,970
HLC	NT	Unknown	Yes	1410	V	3,001
ISI	R	Yes	Yes	1401/7074	F,V	900/1200
LC	R	Yes	Yes	1401/360	F,V	2,008
Lehigh	R,WL	Later	Yes	GE225		On discs
NASA	R	Yes	Yes	360/40	V	3,000
NBS	R	Unknown	No	7090		10x130= 1,300
NLM	R	Yes	Yes	800/7040-94	V	4,230
RAND	NT	Yes	Yes	7044	V	2,400
SIE	R	Yes	Yes	360/30		450
URBANDOC	R	Yes	Yes	1401	V	2,200

D = Data; NT = Natural Text; R = Reference; WL = Word List; F = Fixed; V = Variable
*See reference list of identifier codes and organizations on following page.

Cross-Reference List of Identifier Codes and Organizations

CODE	ORGANIZATION
ABC	American Bibliographical Center
ACS	American Chemical Society
AEC	Atomic Energy Commission
AMR	Applied Mechanics Review
API	American Petroleum Institute
Bowker	R. R. Bowker Company
BPA	Bonneville Power Authority
BSIS	BioScience Information Service
CAS	Chemical Abstract Service
CENSUS	Bureau of the Census
CFSTI	Clearinghouse for Federal Scientific and Technical Information
EJC	Engineers Joint Council
F & S	Frost and Sullivan
GE-FPD	General Electric Company, Flight Propulsion Division
HLC	Health Law Center
ISI	Institute for Scientific Information
LC	Library of Congress
Lehigh	Lehigh University
NASA	National Aeronautics and Space Administration
NBS	National Bureau of Standards
NLM	National Library of Medicine
RAND	The Rand Corporation
SIE	Science Information Exchange
URBANDOC	

specific tasks:

a) Specification of the requirements for generalized file management programs, applicable to a broad range of mechanized files.

b) Specification of an economic configuration of data processing equipment.

c) Specification of standards for library cataloging of magnetic tape material (including union catalogs)

d) Specification of the administrative organization for mechanized library services.

B. Phase II. Detailed System Designs and Programming

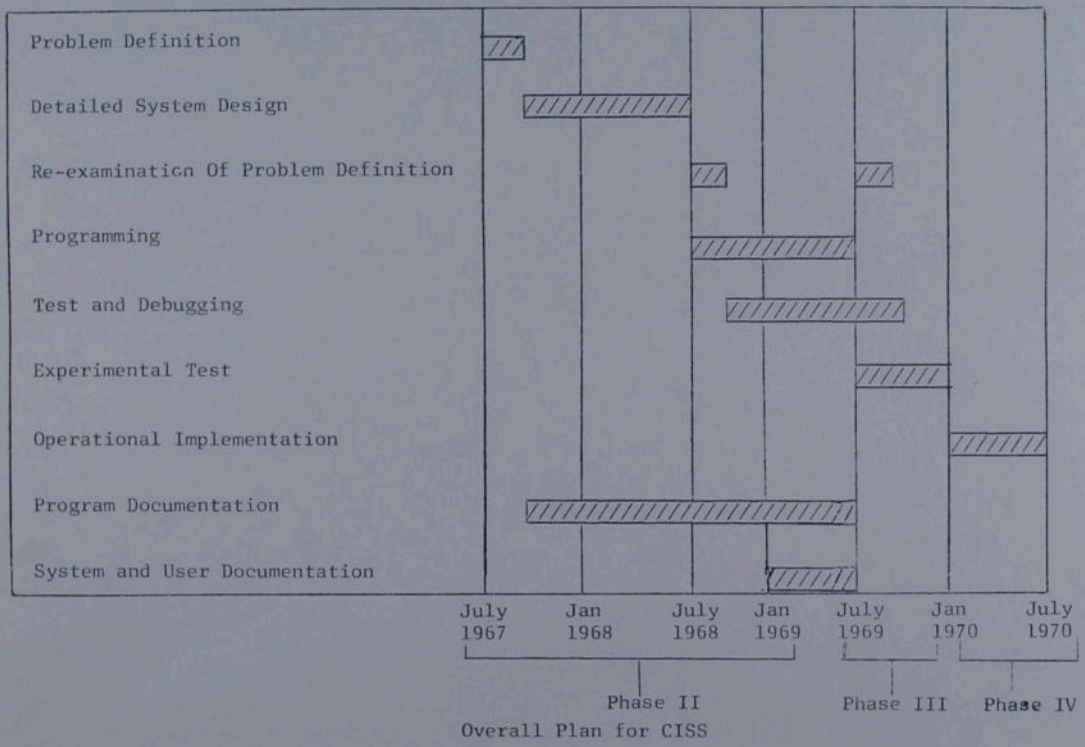
The purpose of the work to be accomplished at Phase II is to produce detailed designs for the computer programs required for the operation of the CISS. These specifications will develop the detailed designs for the functions to be performed by the system and the methods and procedures by which the system will be used by both researchers and computer oriented professionals. They will include a functional description of the system, language specifications, detailed specification of each system function in programming terms, interface requirements among the system modules and with the manufacturer's software, and other similar information. The product of Phase II will be a set of documents from which programming personnel can program, check out, test, and implement the system. These documents are designed to reduce to an absolute minimum the number and scope of decisions that such programmers will make regarding the form and substance of system functions. For this reason, continuing review of these functions is provided in Phase II.

Certain specific items will be available at the conclusion of Phase II. These are, basically, the:

External Design Specifications - a detailed description of the forms and procedures necessary for the use of the CISS in the performance of its functions (approximately 200 - 300 pages).

Internal Design Specifications - The detailed specifications of the CISS software and its interface with DOS. These specifications describe the means of implementing the functions specified in the Preliminary Users Manual (approximately 300 - 400 pages).

In Phase III, the system, as specified in Phase II, will be programmed, tested, and implemented. Phase IV represents operational testing of the system and provides the opportunity to evaluate the final system in the context of the environment in which it must operate.



Budget

1 July 1967 - 31 December 1968

CATEGORY	FTE	CENTER DEVEL.		UCLA CONTRIBUTION
ILR Director	.25	$ ---		$4,150.00
ILR Admin. Director	.25	---		6,000.00
Research Staff		15,000.00		
Library Staff		23,000.00		
Programming Staff		15,000.00		
Clerical Staff		10,000.00		
SUB TOTAL		$ 63,000.00		$10,150.00
Staff Benefits 10%		6,300.00	13%	1,319.50
Research Assistants 5 @ .50 FTE		22,770.00		
Consultants				
Sub-Contracting Software development		150,000.00		
Supplies & Expenses		4,500.00		
Equipment & Facilities		4,500.00		
Travel		4,500.00		
Publications		1,000.00		
Computer		20,000.00		
Overhead 42% Total Salaries		36,023.00		4,263.00
TOTAL		$312,593.00		$15,732.50

NSF Contribution 95%
UCLA Contribution 5%

NATIONAL SCIENCE FOUNDATION
Washington, D. C. 20550

MEMORANDUM

DATE: February 17, 1967

TO : Dr. David J. Rogers
Colorado State University

FROM : Systematic Biology Program
Biological and Medical Sciences Division

SUBJECT: Proposal, B7-1468R

The National Science Foundation awards grants on a competitive basis. We attempt to have before us as much professional advice as we can reasonably ask of the scientific community before making a final decision in regard to any proposal. Accordingly, we now seek the benefit of your experience and considered judgment in regard to the enclosed application(s) for funds. We are anxious to support the most worthy requests, and it is through the cooperation of such consultants as yourself that we are able to meet these responsibilities. Each evaluation is treated confidentially.

It would help us most if your comments would cover such points as the scientific merit of the proposal, whether it duplicates other research in progress, its relative importance, the scientific qualifications and productivity of the Principal Investigator, the adequacy of facilities both for research and student training, and the propriety of the budget. Any other comments which you believe will contribute toward a proper evaluation will be much appreciated. It is important that you numerically score the proposal. Intermediate scores using the first decimal place should be entered in lieu of plus or minus ratings (for example, 2.3 - which is regarded as being a lower score than 2.0).

For your convenience in recording and forwarding your comments, duplicate "Rating Sheets" are enclosed together with a self-addressed envelope. Please return one of these sheets to the Foundation as promptly as convenient; the other is for your records. It is not necessary to return the proposal, but it should be considered a privileged document.

Your cooperation is of importance to the field of systematic biology as a whole, and we will much appreciate it. Thus, we are thanking you in advance for your evaluation of this research.

Sincerely yours,

R. K. Godfrey

R. K. Godfrey
Program Director
for Systematic Biology

Enclosures

NATIONAL SCIENCE FOUNDATION
BIOLOGICAL AND MEDICAL SCIENCES

Prop. B 7-1468R Syst

PROPOSAL EVALUATION SHEET

Title Computer Programs for Comparisons of Natural Populations

Investigator Andrew P. Nelson Institution Dartmouth

Return by 1 MAR 1967

COMMENTS (If more space is required please use additional page).

I find the present request so incomplete that I cannot make an evaluation of it.

I strongly suggest that the investigator spend more time preparing this proposal, and that he make it more of a scientific approach than he has. It is not at all clear what reasons he has for writing the program, i.e. what thought processes are involved in the program, what model he is following in the program, what are the anticipated results of running the program, etc. Then too, I think many of the parts of the program thathe indicates are needed have been written many times for various computers, and are probably available at most computer centers. He should compare his project with others already available, and give some indication of the uniqueness of this approach.

The budget is modest.

Numerical Rating For Merit
(please check one)

- [ ] 1 Highly meritorious
- [ ] 2. Meritorious
- [ ] 3 Acceptable
- [ ] 4 Questionable
- [x] 5 Declined

Name David J. Rogers
(please print or type)

Institution Colorado State University

Date Feb. 20, 1967

NSF FORM 9A-2
OCT 1962

Dartmouth College, Hanover, New Hampshire

Summary of Proposal for Research Grant To

National Science Foundation

1. Andrew P. Nelson — Biological Sciences
   Principal Investigator — Department

2. Computer Programs for Comparisons of Natural Populations
   Project Title

3. 15 June 1967 (15 July 1967) — 1 year
   Desired and alternative starting date — Time period

4. $2600.00
   Amount of Support Requested

5. Endorsements — Telephone:

_____ (603)646-2394
Principal Investigator (signature)

_____ (603)646-2364
Raymond W. Barratt, Chairman
Department of Biological Sciences (signature)

_____ (603)646-2700
James Hornig
Associate Dean of Sciences (signature)

_____ (603)646-2872
Foster Blough
Assistant Comptroller (signature)

JAN 27 1967
Date of Submission

COMPUTER PROGRAMS FOR COMPARISONS OF NATURAL POPULATIONS

Attached is a paper presented by myself and Mr. Craig B. Ordway, Dartmouth class of 1967, before the Society for the Study of Evolution meeting with the AAAS in December 1966. Response to this paper included requests that the system of computer programs illustrated be made available to other workers in the field. With this system, a biologist with average statistical background could utilize the computer to reduce data for comparison of populations to standard statistical form without making premature assumptions concerning the weighting of characters or the choice of correlations to be attempted.

Mr. Ordway and I have explored the problem of making this system generally available and have found that the following procedure should be technically feasible:

1) Conversion of the existing prototype system to an operative system in BASIC and EDIT using Dartmouth's time sharing system with the GE 625 and Datanet 30 computers; testing the system with research data.

In this form the system would be available to users of GE time sharing systems.

2) Translation of the programs into FORTRAN and ALGOL; testing of the translated programs.

These versions should be adaptable to a number of different computer systems.

3) Preparation of a report for publication, possibly in Science or Evolution, outlining the nature and possible uses of the system and stating how the programs may be obtained.

Completion of this project would require Mr. Ordway's involvement through the coming summer, as he is now the only one who knows the details of these programs and how they work. This, then, is a proposal to fund such activity.

Principal Investigator

Andrew Phillips Nelson

Born: December 13, 1936, at Auburn, New York

Attended Leavenworth Central School, Wolcott, New York

B. S., State University College of Forestry at Syracuse Univ., 1958.

M. S., State University College of Forestry at Syracuse Univ., 1959.

Ph.D., (Systematic Botany), University of California, Berkeley, 1962.

Married, two children.

Member: American Society of Plant Taxonomists, International Association for Plant Taxonomy, International Organization of Biosystematists, Society for the Study of Evolution, California Botanical Society, New England Botanical Club, American Institute of Biological Sciences, Sigma Xi.

Appointments etc:

Laboratory Assistant, State University College of Forestry, 1955-57.

Forestry Aid - Research, Forest Insect and Disease Laboratory, Rocky Mountain Forest and Range Experiment Station, Albuquerque, New Mexico, 1957.

Teaching Assistant, State University College of Forestry, 1958.

Research Assistant, Syracuse University, 1958-59.

Woodrow Wilson Fellow in Botany, University of California, 1959-60.

Teaching Assistant, University of California, 1960-61.

National Science Foundation Fellow, University of California, 1961-62.

Instructor in Biology, Dartmouth College, 1962-64.

Assistant Professor of Biology, Dartmouth College, 1964 -

Papers Presented:

A case of phenocopies in Prunella vulgaris L. ssp. vulgaris (Labiatae). Before American Society of Plant Taxonomists with AIBS, Amherst, Mass., August 26, 1963.

Subspecies and ecological races in Prunella vulgaris L. (Labiatae). Before American Society of Plant Taxonomists with AIBS, College Park, Md., August 17, 1966.

Methods in genecology. (Craig B. Ordway, joint author) Before Society for the Study of Evolution with AAAS, Washington, D. C., December 30, 1966.

Publications:

Nelson, A. P. & C. J. Lyon. 1962. Griggs' Key to the Families of Flowering Plants, Wild or Cultivated, in the Northeastern United States (a revision). Department of Biological Sciences Publication No. 3, Dartmouth College, Hanover, N. H.

Nelson, A. P. 1963. The spelling and derivation of the generic name Prunella L. (Labiatae). Bull. Torrey Bot. Club. 90:29-32.

_____. 1964. Relationships between two subspecies in a population of Prunella vulgaris L. Evolution. 18:43-51.

_____. 1965. Taxonomic and evolutionary implications of lawn races in Prunella vulgaris (Labiatae). Brittonia. 17:160-174.

_____. 1966. IOPB Chromosome Number Reports VI (A. Löve, ed.) Taxon. 15:161 & 162.

_____. 1966. Flora of Turkey and the East Aegean Islands, Vol. I. (review). Quart. Rev. Biol. 41:323.

Manuscripts:

Racial diversity in California Prunella vulgaris.

Hybridization and ecological races.

Other Personnel

A major portion of the work proposed above, including all computer programming, will be done by Mr. Craig B. Ordway, a senior biology major who expects to receive his bachelor's degree from Dartmouth this June. Mr. Ordway began his association with me as an N. S. F. Undergraduate Research Participant (GE 3979) in the summer of 1964 and continuing through the subsequent academic year. He was employed in my research program under NSF grants GB 1564 and GB 3281 during the summers of 1965 and 1966. Throughout this period he has been encouraged to pursue an interest in computer programming developed in connection with his URP project. The present project represents the culmination of this activity. I am confident of his ability and willingness to complete this project if it can be funded. In the attached budget I have proposed compensation which seems appropriate for a graduate biologist making a major professional contribution to the research project. It is understood that the entire summer is available for completion of the project even though compensation is scheduled to cover a period of two months.

Other Support

The research program of the principal investigator is currently supported by NSF grant GB 3281 (Initial Research inComparative Genecology in New England, $19,500.00 for two years), which terminates in March 1967. Funds in that budget are committed to other purposes and thus are not available for support of the project proposed here. No other proposals are pending and the current proposal is not being submitted to other possible sponsors.

Period of Proposed Support

Support is requested for one year specifically to cover the period June 15 to August 15, 1967. Support for the period July 15 to September 15, 1967, would also be acceptable.

Proposed Budget

	N.S.F.	Dartmouth College
SALARIES:		
Principal Investigator	no cost	
Assistants (Mr. Ordway) 2 months at $600.00/month	$1,200.00	
STAFF BENEFITS DISTRIBUTION:		
6%, Social Security and other fringe benefits	72.00	
COMPUTER USAGE:		
4 hours at $150.00/hour	600.00	
SUPPLIES:	50.00	
PRINTING: (reprints, publication costs, etc.)	150.00	
TOTAL DIRECT COST	$2,072.00	
Indirect Costs	528.00	$241.00
TOTAL	$2,600.00	$241.00

Methods in Genecology

Presented before the Society for the Study of Evolution meeting with the AAAS, December 30, 1966, Washington, D.C.

by Andrew P. Nelson and Craig B. Ordway, Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire 03755

A primary purpose of studies in genecology is evaluation of the genetic relationships between populations of a species and the illucidation and explanation of whatever natural patterns there may be in these relationships.

The work which I shall summarize here is aimed at development of techniques for collection of genecological data and for its reduction to interpretable form with a minimum of bias, experimental error and human error. At the same time, we hope to retain opportunities for the investigator to make conscious use of intuitive evaluations based on his experience with the materials and with the experimental methods.

The units of sampling in genecological studies are populations. However, it is impossible to collect a population. The population is represented in our experiments by a collection of individuals. Thus problems of sample size and selection must be considered at two levels.

1) The number and location of populations to be sampled.
2) The number and nature of individuals to be selected from each population.

Published considerations of sampling for genecological studies are mostly inadequate because they treat one or the other but not both of these considerations. It usually takes a number of years to complete a genecological study and there is not yet a sufficient quantity of adequately collected data for meaningful application

of statistical formulae to the question of sample size. We are currently working with the pragmatic tests of predictability and repeatability for evaluation of sample size in our studies.

Data are obtained from cultivated specimens. The basic assumption is that any differences expressed by individuals growing in a common environment must have a genetic basis. I personally consider the question of whether a single garden or a series of different gardens is used to be of second order importance. A single cultural environment will yield useful information, a number of different cultural environments will yield more of the same.

In the past, reduction of quantitative data with the desk calculator has been one of our most time consuming tasks. We have recently developed the prototype of a computer system designed to take raw data and deliver reduced data in tabular and graphic form.

Our data is collected in blocks which include all the values conveniently determined at one sitting. For example, all measurements of plant size, all measurements of leaf size and shape, or all measurements of flower size and shape are recorded together.

Data are recorded in numerical form and indexed by a six digit code number for filing in the computer. The first two digits identify the population sample, the second two digits identify the individual plant, and the third two digits identify the data block. SLIDE 1 (Figure 1)

For example, the number 012510 indicates plant height and diameter measurements (10) for plant 25 of population sample 01.

These six digit numbers become line numbers in a data storage program a small portion of which is shown in this slide. Upon collection, the data are immediately ready for typing into the

computer via remote teletype station. Data are handwritten once at the time of collection and typed once into the computer. Opportunities for errors in copying are thus substantially reduced.

Data are typed into the data storage program in whatever sequence may be convenient. The computer stores the data blocks in order according to the six digit line numbers so that we then have all data arranged in order according to sample and according to the individuals within the samples.

SLIDE 2 (Figure 2)

Computer handling of the data is illustrated by this diagram. Our main data storage program is symbolized at top left. On command the computer will reproduce and extract from the storage program any designated combination of lines. It rearranges the extracted lines for use in subsequent programs. Then, on command, it will select specific data for analysis.

SLIDE 3 (Figure 1)

Going back to the data storage program, the line numbers ending in --0001 contain a standard code number, the field collection number for the sample, and the latitude, longitude, and elevation of the collection site. Latitude and longitude are in minutes from an arbitrary base parallel and an arbitrary base meridian, elevations are in feet above mean sea level.

Lines ending in ----10 contain measurements of maximum height and maximum diameter of cultivated plants.

On the right you see data selected by the computer for analysis of maximum height of cultivated plants as it relates to latitude of the collection site.

SLIDE 4 (Figure 2)

We also have stored in the computer various programs for statistical reduction of data and for presentation of the results in tabular and graphic form. These are symbolized at top right. On command the computer will select and weave together designated data reduction programs and print out programs. Another simple command tells the computer to merge this composite instructional program with our program of selected data. Then we tell the machine to go to work.

SLIDE 5 (Figure 3)

Here is an example of what we get:

1) A table showing collection numbers, number of individuals measured in each sample, mean, standard error, and standard deviation of the measurements, and the highest and lowest measurement for each sample

2) A scattergram in which either latitude, longitude, or elevation of the collection sites is plotted over sample means.

In a working run each data parameter would be plotted against each parameter of site position. All scattergrams would then be edited and compared in a search for consistent elements of pattern. These would be checked against whatever is known about the habitats presented by the various collection sites in a search for correlations between genetic and ecological variation which might contribute to explanation of patterns of genetic diversity within the species.

```
LIST

PDP8I     11:13   DEC. 2, 1966

010001 DATA 888,114,480,445,1000,
012510 DATA 27,55,
012610 DATA 31,50,
012710 DATA 31,59,
012810 DATA 28,61,
012910 DATA 27,53,
013010 DATA 34,50,
013110 DATA 28,50,
013999 DATA 0,
020001 DATA 888,566,300,481,100,
020210 DATA 29,50,
020310 DATA 25,70,
020510 DATA 18,77,
021110 DATA 15,41,
021210 DATA 19,47,
021310 DATA 11,59,
021510 DATA 14,50,
029999 DATA 0,
030001 DATA 888,602,515,515,500,
031610 DATA 31,58,
031710 DATA 37,65,
031910 DATA 30,68,
031910 DATA 34,63,
032010 DATA 25,47,
032110 DATA 30,67,
032210 DATA 36,69,
032310 DATA 30,69,
032410 DATA 38,75,
032510 DATA 37,53,
039999 DATA 0,
040001 DATA 888,632,441,563,100,
041510 DATA 28,48,
041710 DATA 23,40,
041810 DATA 15,37,
041910 DATA 24,65,
042110 DATA 16,38,
049999 DATA 0,
050001 DATA 888,653,495,552,200,
051610 DATA 27,61,
051810 DATA 23,53,
051910 DATA 26,77,
052010 DATA 26,57,
052110 DATA 21,50,
052210 DATA 20,47,
052310 DATA 18,57,
059999 DATA 0,
060001 DATA 888,645,330,474,100,
064310 DATA 24,34,
064410 DATA 34,50,
064510 DATA 34,57,
064710 DATA 36,76,
064810 DATA 44,68,
064910 DATA 36,60,
065010 DATA 43,71,
```

```
TEST      11:42   DEC. 2, 1966

SELECTED DATA FØR SITE 114
LATITUDE 480  MINUTES NØRTH ØF 33RD PARALLEL
MAXIMUM HEIGHT, CM.
 27   31   31   28   27   34   28

SELECTED DATA FØR SITE 566
LATITUDE 300  MINUTES NØRTH ØF 33RD PARALLEL
MAXIMUM HEIGHT, CM.
 29   25   18   15   19   11   14

SELECTED DATA FØR SITE 602
LATITUDE 515  MINUTES NØRTH ØF 33RD PARALLEL
MAXIMUM HEIGHT, CM.
 31   37   30   34   26   30   36   30   38   37

SELECTED DATA FØR SITE 632
LATITUDE 441  MINUTES NØRTH ØF 33RD PARALLEL
MAXIMUM HEIGHT, CM.
 29   23   15   24   16

SELECTED DATA FØR SITE 653
LATITUDE 495  MINUTES NØRTH ØF 33RD PARALLEL
MAXIMUM HEIGHT, CM.
 27   23   26   26   21   20   18

SELECTED DATA FØR SITE 646
LATITUDE 330  MINUTES NØRTH ØF 33RD PARALLEL
MAXIMUM HEIGHT, CM.
 24   34   34   36   44   35   42

SELECTED DATA FØR SITE 650
LATITUDE 356  MINUTES NØRTH ØF 33RD PARALLEL
MAXIMUM HEIGHT, CM.
 10   21   18   23   25   23   17   21   19

SELECTED DATA FØR SITE 681
LATITUDE 270  MINUTES NØRTH ØF 33RD PARALLEL
MAXIMUM HEIGHT, CM.
 19   23   20   34   14   22   25   28   25   21

SELECTED DATA FØR SITE 582
LATITUDE 240  MINUTES NØRTH ØF 33RD PARALLEL
MAXIMUM HEIGHT, CM.
 45   37   37   29   32   50   25   31   34   43
ØUT ØF DATA IN  100

TIME:   2 SECS.
```

FIGURE 1

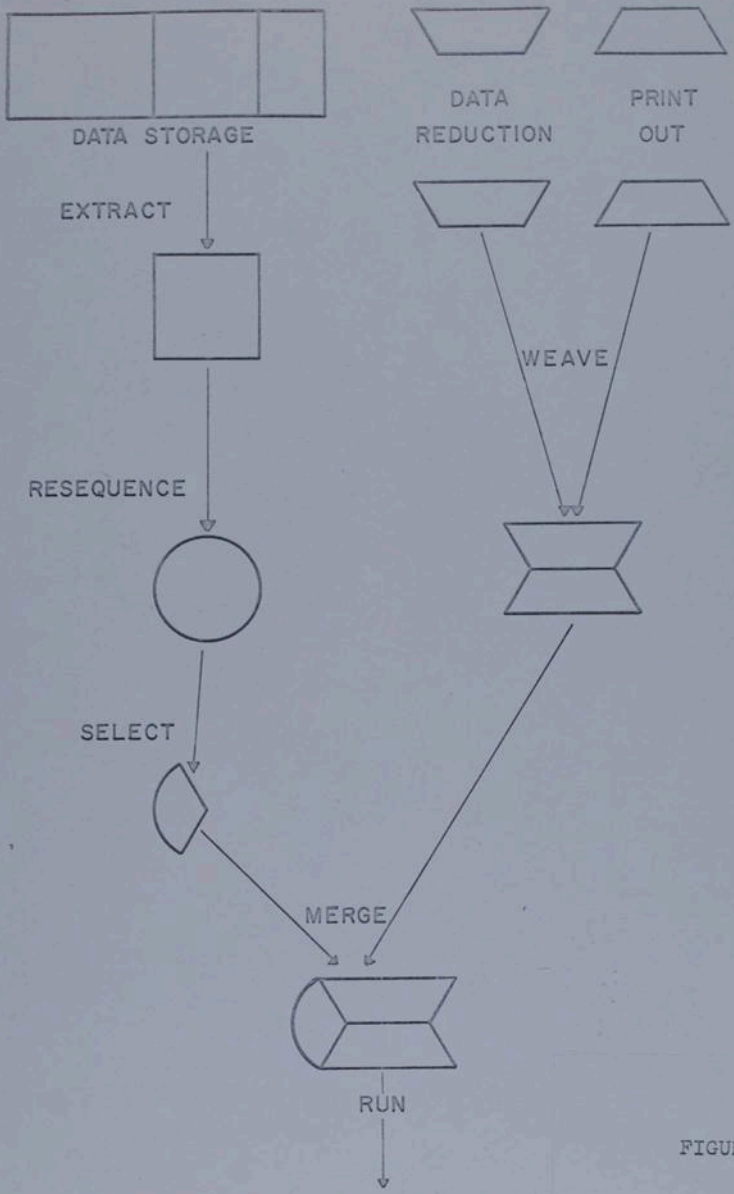


FIGURE 2

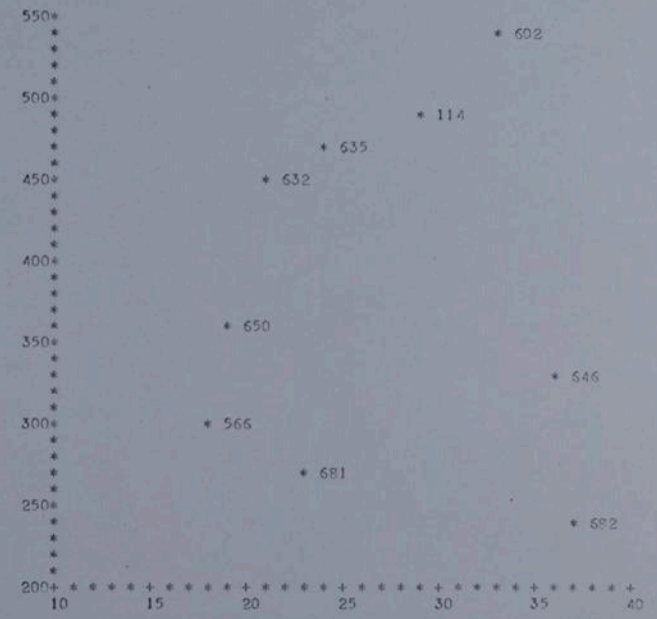
TREG 15:33 DEC. 2, 1966

IDNT	COUNT	MEAN	STD E	STD D	HI	LO
114	7	29.43	1	2.64	34	27
566	7	19.71	2.4	6.34	29	11
602	10	32.9	1.28	4.04	38	26
632	5	21.4	2.62	5.86	29	15
633	7	23	1.31	3.46	27	18
646	7	35.71	2.45	6.47	44	24
650	6	19.67	1.48	4.44	25	10
681	10	23.1	1.72	5.43	34	14
682	10	36.5	2.45	7.74	50	25

RESCAT 15:46 DEC. 2, 1966

DATA PRESENTED IN
SCATTERGRAM

VERTICAL AXIS
LATITUDE IN MINUTES NORTH
OF 33RD PARALLEL
HORIZONTAL AXIS
MAX HEIGHT IN CM



TIME: 2 SECS.

FIGURE 3

THE NEW YORK BOTANICAL GARDEN
BRONX • NEW YORK 10458 LU 4-8500

January 3, 1967

Dr. Dave J. Rogers
Department of Botany
Taximetrics Laboratory
Colorado State University
Fort Collins, Colorado 80521

Dear Dave:

Thanks for the comments on the Antennaria paper. Your remarks obviously need to be passed on to the author, but I am not sure I can paraphrase them adequately, and I can't send a literal transcript without disclosing the identity of my consultant. I should have thought of this when I wrote you before. If you think it is appropriate, I will pass on your remarks verbatim, together with some of my own. If you would prefer to remain anonymous, I will do what I can on my own hook. Please advise me. Thanks again.

Yours,

Arthur Cronquist
Senior Curator

AC:dz

1/5/67

This paper suffers from the same problem that so many taxonomists fall into when they attempt to use some "quantitative" methodology. First, it is not clear what the author wants to do. Does he want to show that species of Antennaria are rather arbitrary? Then the paper does no more than illustrate the obvious. Second, there is no reason given, mathematically, for selecting one technique rather than another. Third, he has not really grasped the significance of one or another method which he mentions in the way of a summary of "numerical taxonomy." To take but one example, on p. 3 (bottom), "All of this work has assumed clustering of individuals or taxa in "phenetic space." That simply is not true. Nor has he summarized the later literature on the subject.

The real masterpiece of obfuscation occurs at the top of page 3--"Because of the relative simplicity of the confusion of this group, etc."

He has chosen for his work a similarity measure called the MCD. While this is permissible, he then garbles the whole by adding a discussion (bottom of p. 5) about correlated versus uncorrelated characters. Nowhere does he define correlated or uncorrelated, or show how to discover them. The limitation of the method chosen is indicated by the fact that he cannot find a way to use qualitative characters, such as those mentioned on p. 12, "shape of leaf margin, pubescence, glands, and bract shape."

Now to quit knocking the paper, and say something good about it. In his discussion, he recognizes that for taxonomic purposes, you probably can't recognize more than two taxa, but for those interested in the biological mechanisms to be found in the group, there are discernable differences caused by polyploidy, apomixis, and the like. This is a good way to do taxonomy, in my estimation, i.e., don't recognize the instability and clutter up the nomenclature with formal names for fleeting variation.

The ordination technique apparently works in discovering the variations amongst the plants studied. It is a heuristic technique, not one based on sound mathematical logic. This is not necessarily bad, but we never will develop ourselves until we quit using the approach that says "let's try this method--maybe it has something in it that we can use."

Paper reviewed:
A Taxonomic Continuum in the
Genus Antennaria in Wisconsin.
By Edward W. Beals - Dept.
of Botany, University of Wisconsin
Madison, Wis. - Taximetrics Laboratory

December 28, 1966

Dr. Arthur Cronquist
The New York Botanical Garden
Bronx Park
Bronx, New York 10458

Dear Art:

This paper suffers from the same problem that so many taxonomists fall into when they attempt to use some "quantitative" methodology. First, it is not clear what the author wants to do. Does he want to show that species of Antennaria are rather arbitrary? Then the paper does no more than illustrate the obvious. Second, there is no reason given, mathematically, for selecting one technique rather than another. Third, he has not really grasped the significance of one or another method which he mentions in the way of a summary of "numerical taxonomy." To take but one example, on p. 3 (bottom), in reference to my 1960 paper, "but these were more or less representative of known varieties." How the hell did he figure that anyone knows a variety in this complex? What definition is there for the taxon "variety", with relation to cultivated plants? He states that (last sentence, p. 3) "All of this work has assumed clustering of individuals or taxa in "phenetic space." That simply is not true. We, particularly, have not assumed a cluster. He has either not seen, or for some reason chooses to ignore, our more recent clustering method published in Systematic Zoology this year. While he does not want to use a computer and we do, he still could get through a test for his Antennaria had he followed that methodology. I'm not pushing him into this method, but he should have used a more recent citation for work done by our group. The 1960 paper is way out of date.

Nor has he announced the late literature on the subject.

The real masterpiece of obfuscation occurs at the top of page 3— "Because of the relative simplicity of the confusion of this group, etc."

He has chosen for his work a similarity measure called the MCD. While this is permissible, he then garbles the whole by adding a discussion (bottom of p. 5) about correlated versus uncorrelated characters. Nowhere does he define correlated or uncorrelated, or show how to discover them. The limitation of the method chosen is indicated by the fact that he cannot find a way to use qualitative characters, such as those mentioned on p. 12, "shape of leaf margin, pubescence, glands, and bract shape."

Now to quit knocking the paper, and say something good about it. In his discussion, he recognizes (as you did) that for taxonomic purposes, you probably can't recognize more than two taxa, but for those interested in the biological mechanisms to be found in the group, there are discernable differences caused by polyploidy, apomixis, and the like. This is a good way to do taxonomy, in my estimation, i.e., don't recognize the instability and clutter up the nomenclature with formal names for fleeting variation.

The ordination technique apparently works in discovering the variations amongst the plants studied. It is a heuristic technique, not one based on sound mathematical logic. This is not necessarily bad, but we never will develop ourselves until we quit using the approach that says "let's try this method--maybe it has something in it that we can use."

On another subject, we received your Christmas card, and were pleased to have it. Happy Hanukka to you!

Sincerely,

David J. Rogers
Professor of Botany

DJR:ch

Enc.

SCIENCE
1515 MASSACHUSETTS AVENUE, NW, WASHINGTON, D. C. 20005

Dear Professor Rogers:

Thank you for your book review manuscript, which has just reached us. If any questions arise in the editing, you will be consulted shortly. Otherwise galley proofs will be mailed to you for your approval.

Yours sincerely,

Mrs. Sylvia Eberhart
Book Reviews, Science

SCIENCE
1515 MASSACHUSETTS AVENUE, NW, WASHINGTON, D. C. 20005

JUL 14 1966

Thank you for your helpful comments about the paper we recently sent to you. Your cooperation is much appreciated.

JUL 18 1966

Sincerely,

John E. Ringle
Assistant Editor

- Taxonomy Laboratory

February 24, 1966

Dr. Walter Hodge, Program Director
Systematic Biology
National Science Foundation
Washington, D. C. 20025

Dear Walter:

Herewith is the expanded critique of the proposal by Rohlf. Hope this is satisfactory. We feel that these comments are very important to be considered by the panel.

Sincerely yours,

David J. Rogers
Professor of Botany

DJR/pam

Supplimentary Evaluation Sheet for Proposal B 6 1251 R (SYST)
Preliminary statement made on February 19, 1966.

At the outset we would like to clear up some confusion surrounding Dr. Rohlf's use of the term "on-line." On-line means "done by the computer" and off-line means "done by auxilliary equipment." As an example, data on cards may be entered into the computer in two ways: (1) the cards are placed in the computer's card reader and read directly into main storage (on-line), or (2) the cards are first put through a card-to-tape conversion via auxilliary equipment (off-line), then the tape thus prepared is read directly into the computer (on-line). From the foregoing it is clear that there is no such thing as off-line computation as mentioned on P. 3 of Dr. Rohlf's Research Proposal. All computation is done by the computer and is therefore on-line. The terms on-line and off-line have significance only when referring to the mode of input or output. However, from the context of Dr. Rohlf's proposal it is apparent that he means "on-line computing" to be synonymous with "computation under manual control of the operator" as opposed to the usual method of operating the machine at top speed under automatic program control.

The principal advantage of an electronic computer over a hand-operated desk calculator (and the sole economic justification for its use) is the extreme speed with which it can perform. Dr. Rohlf's so-called "on-line" approach to computation removes this advantage and effectively turns the computer into

an expensive desk calculator. It is true that the statistical routines of the Culler-Fried system run at top computer speed. However, between routines the machine idles while the investigator examines the results and decides which routine to try next. The time gained in using programmed routines is frittered away while figuring out what to do next.

We feel that the additional expense of Dr. Rohlf's manual approach brings with it no compensating advantages. On the contrary, we feel that the investigator who does his thinking on-line is under a disadvantageous pressure as he tries to work out his problems while an expensive machine is idling. In the usual approach to computation, an investigator runs his data with a program specially written by him or for him or with a regular package of routines such as the Bio-medical series mentioned by Dr. Rohlf on P. 3. He submits this program and data to the computation center and goes on to other tasks while the job is being processed. The job is run under the control of an automatic monitor which runs job after job in a highly efficient manner. When the results are returned, the investigator may give them a thorough examination without pressure, call in specialists or other colleagues to consult,[1/] and plan his next computer run after a well thought out decision, during all of which time he is not paying for the use of a computer.

The above should demonstrate that the same man-machine dialogue can be established under automatic control as under manual control, but with savings in computer time and benefits in additional thinking time. It may be argued by advocates

of the Culler-Fried system that overall real time is saved by their method, since the closed shop turn-around time [2] is avoided. This observation is true. However, real time is not a critical factor in biological research (as it is critical in, for example, a radar defense system). Therefore the justifying criterion must be economical and one should choose the system which minimizes the expense of operating the computer.

Dr. Rohlf tries to justify the use of the Culler-Fried system in the first paragraph of his proposal. He criticizes the standard packages of statistical programs (such as the Bio-medical series) by saying, "It is all too easy for investigators to simply run their data blindly through standard programs." This may or may not be a true observation, but in either case, it is not a criticism of the packaged programs but of the slipshod methods of certain hypothetical investigators. Anyone who runs data "blindly" through a program is in trouble to start with and can certainly not be rescued by the Culler-Fried system or any other system. If we can assume that an investigator knows what he is doing, then we can assume he can solve his problems by a judicious use of standard programming techniques.

Dr. Rohlf criticizes the standard statistical packages also by saying that these programs give either too little output or too much. This too may or may not be a true observation, but it is not a criticism directed at the automatic vs. manual question under debate. If these programs do provide inadequate or overabundant output they must be corrected by modifying the

programs. Better output cannot be provided by establishing manual control of the computer.

Now we would like to turn to the three specific research projects enumerated in Dr. Rohlf's proposal.

(1) Development of operations for general statistical computations. The task that Dr. Rohlf has set himself here belongs properly to the domain of mathematics and statistics, and although biologists can sometimes usefully apply statistical operations, the development of these operations is not a part of biology. Therefore, we feel that this section of Dr. Rohlf's proposal is improperly submitted to the Systematics Panel.

(2) Computer assisted identification of organisms in numerical taxonomy. It is difficult to evaluate this section of the proposal as no explicit description of Dr. Rohlf's methods is given. We suggest that he expand this portion of his proposal and resubmit it.

We can say this much, however. Inasmuch as the taxonomic identification problem is one of information retrieval and file maintenance it is not suited to Dr. Rohlf's "on-line" approach. Appendix A of Dr. Rohlf's proposal is the User Manual for the Culler-Fried system. On P. 3 is a check list for judging the appropriateness of using this system for any particular application. By means of this check list we discovered that the taxonomic identification problem fails to be appropriate on four out of five counts.

(3) Computer assisted instruction. In this final area

of the proposal we feel that the Culler-Fried system comes into its own, for here in the teaching of the uses of statistical analysis the instant feedback of the manually controlled programs, together with the graphic display feature, are perfect pedagogical tools. We feel that this equipment for this usage should not be under the administration of a life-science department, but should properly come under the jurisdiction of the mathematics department or statistics laboratory where it can be of service to all who are teaching and studying statistical methods. Moreover, the equipment should not be installed so that long-distance communication is required to a computer in Santa Barbara, California. It is probably the case that the existing data processing equipment at the University of Kansas can be modified to handle this procedure. Some standard version of a remote keyboard inquiry station, together with some standard cathode ray tube output device, together with whatever computer already exists at the University of Kansas could be made to serve very well for manually controlled teaching exercises by writing a special control program around any standard statistical package already running on the machine in question. This not only makes more sense from the point of view of centralizing control of the operation and also simplifying it, but it would certainly save considerable expenditure. Dr. Rohlf's proposal calls for $55,950 for long-distance transmission of data along and this solely for his own use of the equipment.

Moreover, Dr. Rohlf informs us (P. 7) that a new and faster computer is scheduled for Santa Barbara. This, in a

manually controlled system, confers no benefit whatever. On the contrary, if the new computer is not completely compatible with the existing model (and such incompatibility is typical of hardware changeovers), a nasty conversion problem could arise. On the other hand, we read in this proposal (footnote, p. 15) that the University of Kansas is firmly committed to move in the direction of on-line computation. If so, then let Kansas initiate, develop, and operate such a system of their own.

We suggest that if Dr. Rohlf wishes to use the techniques he describes for teaching statistical methods to his biology classes at the University of Kansas, he might try to interest the mathematics department in submitting a grant request (but not to the Systematics Panel) for the necessary equipment and programming. If this should bear fruit, Dr. Rohlf could then rewrite his own grant request to include funds for computer time on the Kansas machine.

In section XI of his proposal, Dr. Rohlf sets out his proposed budget. From this we learn that all three research projects proposed are to consume one day a week of the investigator's time during three nine-month periods out of the three years. In view of this light time commitment we feel that the budget proposed is high, particularly when it could be reduced by the suggestions made above.

We can summarize as follows: Project 1 - not for the Systematics Panel. Project 2 - not well defined. Project 3 - not under the right administration and not for the Systematics Panel. "On-line computation" not applicable to the first two research projects.

FOOTNOTES

1/ It is generally agreed that programming a computer is as much a full-time specialty as being a biological investigator, and most biological investigators whose work involves computer usage rely on the training and talent of professional computer programmers and/or mathematicians. The "free" time between runs permits the proper division of labor among these specialists before they put their heads together for a decision. Under the Culler-Fried system the team of specialists at the console must have a phenomenal quick-wittedness and rapport, or else, and this is rarer, the sole operator must combine in himself the talents of all those who should be helping him.

2/ Turn-around time: the time that elapses between submitting a job to a closed shop computation center and receiving the results. This varies from shop to shop and from day to day, but it is generally several hours minimum and can be as long as several days under poor operating conditions.

NATIONAL SCIENCE FOUNDATION
Washington, D. C. 20550

MEMORANDUM

TO : Reviewers DATE: 9 FEB 1966

FROM : Systematic Biology Program

SUBJECT: B 61251

The National Science Foundation awards grants on a competitive basis. We attempt to have before us as much professional advice as we can reasonably ask of the scientific community before making a final decision in regard to any proposal. Accordingly, we now seek the benefit of your experience and considered judgment in regard to the enclosed application(s) for funds. We are anxious to support the most worthy requests, and it is through the cooperation of such consultants as yourself that we are able to meet these responsibilities. Each evaluation is treated confidentially.

It would help us most if your comments would cover such points as the scientific merit of the proposal, whether it duplicates other research in progress, its relative importance, the scientific qualifications and productivity of the principal investigator, the adequacy of facilities both for research and student training, and the propriety of the budget. Any other comments which you believe will contribute toward a proper evaluation will be much appreciated. It is important that you numerically score the proposal. Intermediate scores using the first decimal place should be entered in lieu of plus or minus ratings (for example, 2.3--which is regarded as being a lower score than 2.0).

For your convenience in recording and forwarding your comments, duplicate "rating sheets" are enclosed together with a self-addressed envelope. Please return one of these sheets to the Foundation as promptly as convenient; the other is for your records. It is not necessary to return the proposal.

Your cooperation is of importance to the field of systematic biology as a whole, and we will much appreciate it. Thus, we are thanking you in advance for your evaluation of this research.

PLEASE RETURN BY 17 FEB 1966

SYSTEMATIC BIOLOGY

NATIONAL SCIENCE FOUNDATION
BIOLOGICAL AND MEDICAL SCIENCES

Prop. B 6 1251 R (SYST)

PROPOSAL EVALUATION SHEET

Title ON-LINE COMPUTATION FOR STATISTICAL ANALYSIS OF BIOLOGY

Investigator F. JAMES ROHLF Institution UNIVERSITY OF KANSAS

COMMENTS (If more space is required please use additional page).

Numerical Rating For Merit
(please check one)

- [ ] 1. Highly meritorious
- [ ] 2. Meritorious
- [ ] 3. Acceptable
- [ ] 4. Questionable
- [ ] 5. Declined

Name _____
(please print or type)

Institution _____

Date _____

NSF FORM 9A-2

N.B. This isa preliminary statement given to meet your deadline. A more detailed review follows on Feb. 25. I received this application one week before the deadline given and have attempted to finish the necessary work on it, but the nature of the application made it impossible to gather sufficient information to make an effective evaluation in the time allotted. The statements made below will be more fully explained in a supplementary review to be submitted on Feb. 25.

1. This application should not be considered by the Systematic Panel because it is preponderantly statistical, only one part of three being related to systematics.

2. The one part related to taxonomy is insufficiently explained to give an indication of the methodology intended. No theoretical foundation is mentioned, and there is no explanation of the proposed model for identification of organisms. Furthermore, from the sketchy description given, one must assume that a large computer is needed to store the information required to make identifications. Though no description of the RW-400 is provided, it is likely that the machine has a relatively small storage capacity.

3. The idea of "on-line computation" for research work is impractical, but the use of such techniques for teaching a wide variety of subjects in mathematics is good. The machinery should, therefore, batplaced in the hands of the math department at Kansas, and used by a wide variety of students.

4. There seems to be no real justification for maintaining a tie-line to a small computer in Santa Barbara when it is quite likely that the peripheral equipment requested could be attached to hardware in Kansas.

5. Almost $60,000.00 is requested to maintain the Western Union Broadband service. This seems inordinately high for the purposes of just communicating with a computer.

6. The principal investigator apparently will devote only 1/5 of his time to this project.

David J. Rogers

Colorado State University

Feb. 19, 1966

x

7. There are inconsistencies between statements made in the proposal. For example, the statement is made in one place that no programs will be made, but in another, that a library of programs will be made.

8. If the principal investigator makes an appraisal of his projects in keeping with the questions posed on page 3 of the appendix to the proposal, only one of the three parts would be justified.

CONFIDENTIAL

B6 1251 R

SYSTEMATIC BIOLOGY

RESEARCH PROPOSAL

to the

National Science Foundation

Computer Sciences Program

January 1966

I. Principal Investigator: F. James Rohlf
Associate Professor of Statistical Biology*

II. Institution: Department of Entomology
The University of Kansas
Lawrence, Kansas 66045

Telephone: (913) 864-3706 (direct line)

III. Title of Proposed Research: On-line computation for statistical analysis in biology. MANUAL Control of the program by the operator

IV. Desired Starting Date: 1 February 1966

V. Time Period for Which Support is Requested: Three years

Signatures:

Department Chairman:

Robert E. Beer, Chairman
Department of Entomology
Telephone: (913) 864-3401 (direct line)

Principal Investigator:

F. James Rohlf

Official authorized to sign for
The University of Kansas

William J. Argersinger, Jr.
Associate Dean of Faculties
Director of Research Administration
Telephone: (913) 864-3126 (direct line)

---

* As of 1 July 1966. At present, Asst. Prof. of Biology at the University of California, Santa Barbara, California 93106. Phone: (805) 968-1511, ext. 713 or ext 4145. Will move to Kansas 1 February 1966.

VI. Description of Proposed Research

A. Introduction and Abstract

This proposal is to permit continued investigation and implementation of on-line statistical analysis into the Culler-Fried on-line computing system at the University of California, Santa Barbara (see Appendix A for a brief Users Manual giving a general description of the present state of the system). Three areas of application are of principal interest: 1. The development of operations for general statistical computations and data handling using simple list processing techniques; 2. Computer assisted identification of organisms in numerical taxonomy; and 3. The use of the system as a teaching aid in advanced courses in mathematical biology such as population genetics and population ecology where much time is usually spent considering the implications of various mathematical models.

In order for the above work to be carried out a Teleputer Control Unit (TCU) and associated operating console must be obtained as well as funds for transmission line costs so that the TCU may be connected to the computer at UCSB where the basic time sharing system has already been developed.

B. Detailed description of the proposed research.

Three types of activities are planned.

1. Development of programs for on-line statistical computations.

In recent years, more and more statistical analyses have been programmed for digital computers. Excellent packages of statistical programs (such as the Bio-medical series prepared at the Health Sciences Computing Facility, University of California, Los Angeles by Professor W. J. Dixon) have been prepared and have a rather wide distribution. A problem which arises with their use, however, is that it is all too easy for investigators to simply run their data blindly through standard programs. If the proper program is selected, the correct tests of significance will be given, but little opportunity is usually provided for the investigator to really study his data by making graphs, frequency distributions, tests of various assumptions, and to follow up a posteriori tests suggested by the data themselves. Programs which allow for some of these possibilities tend to produce voluminous output, much of which is not needed in a given application but is computed "just in case" and then discarded.

On-line computation promises to alleviate these problems if satisfactory software for specifying statistical computations is available. The user could then perform the desired analyses, display frequency distributions, scatter diagrams, etc. and, depending upon these results, try other analyses which the investigator now believes appropriate and of interest. The principal applications will be to those situations and investigations in which the investigator is personally interested in detailed analysis and exploration of his data. It is also assumed that the investigator has a sufficient knowledge of statistics and data analysis to know what he wants to do. No computing system can prevent a user from carrying out improper analyses.

During the tenure of the grant, I plan to study the implications of on-line computing for basic statistical data analysis to determine which types of analysis would be more practical and useful for on-line rather than off-line computing. Both univariate and multivariate analyses will be considered. From detailed studies of those analyses which would seem to benefit most from an on-line approach, macro-operations will be defined which best facilitate the specification of the computations. An attempt will be made to make the operations as general as possible rather than highly specific (although efficient for their intended purpose). For example, there would seem little point in an operation which computes a mean since it can be simply expressed in terms of more general operators.

In cooperation with the staff at the Computer Center at UCSB, I shall program these operations and add them to the basic system.

Some types of statistical computations can be easily specified using the operations already in the basic system. For example, console user programs (consisting of lists of basic operators) have been prepared, tested, and found to be very useful for carrying out a posteriori multiple comparison tests

for differences among a set of means using a technique (simultaneous test procedure, STP) recently proposed by Gabriel (1964). This method which represents a significant advance in the theory of multiple comparison tests, generally leads to very large amounts of computation since all possible combinations of means taken 2, 3, 4, ..., a at a time need to be considered. Conventional off-line approaches (using an IBM 7040) were found impractical when more than 16 means were involved. The amount of programming and computation time was found to be shortened considerably using the Culler-Fried system. The user program was written and checked out on-line in about an hour (including the "manual" on-line computation of sample data to test the program). The program was written so that the investigator could key-in the particular combination of means to be tested (actually only simple identification numbers not the actual means needed to be keyed-in each time) and the program would perform the necessary computations and display the results of the test. From these results, different means of interest are then tested. It was found that a user would soon realize that certain combinations were not worth testing. By proceeding in this manner, the investigator and the computer perform those portions of the task for which each is most suited. The investigator makes decisions about what is important and interesting to test and the computer performs the arithmetic computations. Many other applications of the STP approach are practical and of interest, e.g., partitioning means into maximally non-significant subsets, partitioning rows and columns of contingency tables, as well as various multiple comparison problems in multivariate analysis.

For other common types of analyses where the data structure is more complex, the present system is more awkward. For example, in a typical 4-level nested analysis of variance with unequal sample sizes, there is no provision for storing the data in a way which reflects the nature of the experiment. The lowest level groups must be represented simply as columns in a matrix (the computations can be carried out but this type of data storage can be considered convenient only for a single classification analysis of variance with equal replication). For this reason, it will be necessary to extend the present operations to handle data in terms of list structures. Since i) the lists will almost always contain many items, ii) there will be relatively little insertion and deletion of items within a list, and iii) for a given application, it is usually possible to specify the maximum length of a list at the time it is to be defined, it will, therefore, be possible to store successive items on a list in consecutive memory locations. This will make it possible to design a set of list processing operations which will easily fit into the present organization of the basic Culler-Fried system. This will be done in such a way that a user need not concern himself with this added complexity unless he wants to make use of it. The precise set of operations to be defined and included in the system has not yet been determined but will include operations such as defining a list, erasing a list, making lists of lists, tracing tree structures and retrieving lists from a tree structure of arbitrary complexity. Once a list of data has been retrieved the currently defined operators can be used to manipulate the data.

Unless one had many similar batches of data to be processed, the analysis of variance (probably the single most important set of statistical techniques) has always been a border line case between those analyses which should be done by hand on a desk calculator and those which are practical for computer processing. This is partly because the calculations are not very complex but

also because there is such a variety of data structures in the analysis of variance. It is difficult to prepare a general program without having the preparation of the program and its input unduly complex. Yet, there are relatively few basic operations which need to be performed and they follow very simple rules. In a given case, the sequence of operations is usually obvious to the investigator well versed in the analysis of variance. What he needs help with, is the arithmetic computation.

What I propose to do, therefore, is not to prepare a library of statistical programs, but, rather, to develop (within the framework of the Culler-Fried system) a convenient set of operations for specifying a large variety of statistical computations on-line.

2. On-line computation in numerical taxonomy.

Numerical taxonomy (Sokal and Sneath, 1963) is a field concerned with the numerical evaluation of the relative degrees of similarity among items (usually species) on the basis of their observed characteristics and the ordering of the items into clusters or classes. Important applications of these techniques have been made in such diverse fields as biology, geology, sociology, and linguistics (see Sokal and Sneath, 1963).

A topic which has not yet received much attention in taxonomy is that of developing practical systems for computer assisted identification of organisms. I plan to set up a pilot study using mosquitoes as research material. The TCU would be appropriate for this activity since it would allow graphic displays of organisms, the typing out of questions generated by the computer, and the monitoring of the users responses. I do not plan to just store and retrieve fixed taxonomic keys (one does not need a computer to do that), but rather to develop a system which can make use of a large file of descriptions of organisms and the results of cluster analyses so that there need not be a fixed pathway to the identification of a particular organism. The system will, therefore, be able to allow for the fact that a specimen may be incomplete. As new organisms are added to the file, the parameters of the system would be allowed to change so that the system can "learn by experience." It is important to note that taxonomic data in biology are highly structured so that the amount of computations and searching may be considerably reduced from what would seem to be implied by the above statements. For example, it will seldom, if ever, be necessary to match an unknown specimen against all of the specimens in the file.

3. Computer assisted instruction.

This past semester at Santa Barbara, I have used the on-line computing laboratory as a statistical laboratory in my biometry course. I have found it to be extremely useful for demonstrating various statistical theorems and principles (even though we are not yet able to permit students to use the different consoles independently for working problems). This is a "methods" course since most students do not have an extensive background in mathematics. The emphasis is on the understanding and correct application of statistical methods and not on formal proofs. By sampling experiments, it is easy, for example, to demonstrate the central limit theorem, the asymptotic approximation of the normal by the binominal and Poisson distributions, and the t, F and X^2 distributions. Since the demonstrations were on-line, I could

change the parameters of the experiment in response to student questions. This approach should be of even greater value in a course such as population genetics or population ecology where a large amount of time is typically spent considering various mathematical models. As an experiment, I have programmed most of the models which are covered in a graduate level course on population genetics and, as a result, I believe that the content of the course could almost be doubled. A program to solve (i.e., evaluate Δ_q and stationary frequency distributions) for the effects of mutation, migration, selection, and inbreeding in finite populations with two alleles at a single locus, required only two evenings' work. The effects of changes in the parameters could then be observed. In courses of this type, there is a great need for an instructor to be able to demonstrate such material and for the students to be able to experiment with and manipulate the parameters of a model in order for them to fully appreciate and understand the consequences of a given model. An on-line approach will permit a better interaction between the instructor and the student. I believe that in these advanced and specialized courses this type of computing ability is much more important than the use of the computer as a teaching machine (although a teaching machine approach could be developed which contained these computing abilities as a subset). For most of these applications, I should be able to use the operations in the basic system. But I may need to develop new operations for simulation experiments.

Should be under the auspicies of Math or Stat. dept, I think this use is GOOD

If the proposal is approved, I plan to use the TCU next fall when I teach Biology 343, Population Genetics.

Bibliography

Dixon, W. J. 1964. BMD, Biomedical computer programs. Health Sciences Computing Facility, University of California, Los Angeles. 585 pp.

Fried, B. D. 1964. STL on-line computer, volume I - general description. TRW Space Technology Laboratories, Redondo Beach, California. 31 pp. (This paper gives a general account of an earlier version of the Culler-Fried system.)

Gabriel, K. R. 1964. A procedure for testing the homogeneity of all sets of means in analysis of variance. Biometrics 20:459-477. (An important paper in the theory of multiple comparison tests. It contains references to other important papers in this field.)

Sokal, R. R. and Sneath, P.H. A. 1963. Principles of numerical taxonomy. Freeman: San Francisco. xvi + 359 pp. (Basic exposition on numerical taxonomy. Describes the principles, basic methodology, and gives many references.)

VII. Facilities Available

Space and ordinary laboratory facilities will be made available by the University of Kansas. Computer time and a Data Set Control Unit will be made available on the computer at the University of California, Santa Barbara, for three hours per day. speed limited by operator.

At present UCSB has a RW-400 computer but it is expected that it will be replaced by a larger and faster model by next fall to handle the increased load of servicing many remote stations (on the campus at UCSB and at several places across the country). The change of computers will have little effect upon the proposed research. NSF

VIII. Personnel

A. Principal Investigator: F. James Rohlf.

1. Academic biography: born Blythe, California, October 24, 1936.
   A.B. (Zoology) San Diego State College, 1958.
   Ph.D. (Entomology) The University of Kansas, 1962.

2. Positions:

   Research Assistant in Entomology, the University of Kansas, 1958-1959.

   United States Public Health Service pre-doctoral fellow, the University of Kansas, 1959-1962.

   Research Associate, The University of Kansas, summers of 1962 and 1965 and spring of 1966.

   Assistant Professor of Biology, University of California, Santa Barbara, 1962-1966.

   Visiting Assistant Professor of Entomology, The University of Kansas, spring of 1965.

   Associate Professor of Statistical Biology, The University of Kansas as of July, 1966.

3. Supplemental information:

Dr. Rohlf has been associated with the Culler-Fried system at Santa Barbara since its installation in November 1964. He has written many of the machine language basic system programs. Previous to this, he has had four years of experience in programming the IBM-650 computer in SOAP and then four years of experience in programming the IBM 7040, 7094, and 1620 computers (mostly in FORTRAN but subroutines were sometimes written in assembly language). Most of these programs were written for statistical data processing in numerical taxonomy and involved such things as basic statistics,

correlation analysis, cluster analyses of various types, factor analysis, and other types of multivariate analysis. Programs have also been written to simulate the genetics of populations and the behavior of simple organisms.

4. Publications:

Dr. Rohlf has published 17 papers on various applications of computers and statistics to biology. Most of them are concerned with the "taxonomy problem," a field called numerical taxonomy in biology. Some representative publications are listed below. A textbook of biometry and a set of statistical tables in co-authorship with Robert R. Sokal are nearing completion.

1962. Rohlf, F. J. and R. R. Sokal. The description of taxonomic relationships by factor analysis. Systematic Zool., 11:1-16.

1963. Rohlf, F. J. Congruence of larval and adult classifications in Aedes (Diptera: Culicidae). Systematic Zool., 12:97-117.

1964. Orias, E. and F. J. Rohlf. Population genetics of the mating type locus in Tetrahymena pyriformis, variety 8. Evolution, 18: 620-629.

1965. Rohlf, F. J. Multivariate methods in taxonomy. Proceedings of the IBM Scientific Computing Symposium on Statistics. (October 21-23, 1963). pp. 3-14. IBM: White Plains, New York.

1966. Rohlf, F. J. A randomization test of the hypothesis of non-specificity in numerical taxonomy. Taxon, 14:262-267.

B. Other personnel associated with the project.

In addition to Dr. Glen Culler and his staff at Santa Barbara, a number of people at the University of Kansas will be associated with the project. Several research workers who do statistical computations will be invited to use the system and to try out the operations which are developed in order for their usefulness to be determined. It is expected that their experiences on a variety of problems will be very useful in making decisions about particular ways of implementing changes in the system. Among the people who are expected to make use of the system (in addition to various graduate students in the Departments of Entomology, Zoology, and Botany) is:

Dr. Robert R. Sokal

a. Academic biography: Born Vienna, Austria, 1926 (U.S. citizen). B.S. (Biology) St. John's University, Shanghai, China, 1947. Ph.D. (Zoology) University of Chicago, 1952.

b. Positions:

Assistant in Zoology, University of Chicago, 1950-1951.

Instructor in Entomology, The University of Kansas, 1951-1953.

Assistant Professor of Entomology, The University of Kansas, 1953-1958.

Associate Professor of Entomology, The University of Kansas, 1958-1961.

Professor of Statistical Biology, The University of Kansas, 1961- .

Senior Postdoctoral Fellow, NSF, England, 1959-1960.

Visiting Professor, Fulbright Program, Israel, 1963-1964.

NIH Career Investigator, 1964- .

c. Supplemental Information:

Dr. Sokal has had considerable experience in statistical analysis in biology. He and his staff routinely carry out a variety of statistical analysis of data from several biological fields.

d. Publications:

Dr. Sokal has published 49 papers in several fields of biological research most of them covering some aspects of biometrical work. He and Dr. P. H. A. Sneath are co-authors of a book on the principles of numerical taxonomy. A textbook of biometry and a set of statistical tables in co-authorship with F. James Rohlf are nearing completion. A few representative papers are listed below:

1957. Michener, C. D. and R. R. Sokal. A quantitative approach to a problem in classification. Evolution, 11:130-160.

1958. Sokal, R. R. and C. D. Michener. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull., 38:1409-1438.

1958. Sokal, R. R. Thurstone's analytical method for simple structure and a mass modification thereof. Psychometrika, 23:237-257.

1958. Sokal, R. R. Probit analysis on a digital computer. Jour. Econ. Ent., 51:738-739.

1959. Sokal, R. R. A comparison of five tests for completeness of factor extraction. Trans. Kansas Acad. Sci., 62:141-152.

1962. Sneath, P. H. A. and R. R. Sokal. Numerical taxonomy. Nature, 193:855-860.

1962. Sokal, R. R. Typology and empiricism in taxonomy. J. Theor. Biol., 3:230-267.

1964. Sokal, R. R. and I. Karten. Competition among genotypes in Tribolium castaneum at varying densities and gene frequencies (the black locus). Genetics, 49:195-211.

1965. Sokal, R. R. Statistical methods in systematics. Biol. Rev. (Cambridge), 40:337-391.

1965. Camin, J. H. and R. R. Sokal. A method for deducing branching sequences in phylogeny. Evolution, 19:311-326.

When as is expected the University of Kansas will install one or more additional TCU's the Computation Center staff and researchers in the physical, biological, medical, and engineering sciences will also have access to the system.

first

XI Proposed Budget

First Year

	Total Cost	Requested from NSF	Contributed by Univ. of Kansas
Salaries			
Dr. F. J. Rohlf, Principal Investigator 20% of time for 9 months *	$ 2,500	$----	$ 2,500
Research Assistant, 1/2 time for 11 months	2,689	----	2,689
SALARY SUBTOTAL	$ 5,189	$----	$ 5,189
Employee Benefits	218	----	218
Permanent Equipment			
Teleputer Control Unit with single console	39,000**	39,000**	----
Polaroid Camera	450	450	----
Expendable Equipment and Supplies			
Film, Paper, Cards	50	50	----
Travel	300	300	----
Publication Costs	----	----	----
Other Direct Costs			
Western Union broadband service	15,750***	15,750***	----
Computer time at UCSB	5,800**	5,800**	----
TOTAL DIRECT COSTS	$ 66,757	$ 61,350	$ 5,407
Indirect Costs (computed at 50.59% of the Salary Subtotal)	2,625	2,625	----
GRAND TOTAL	$ 69,382	$ 63,975	$ 5,407

* Summer salary provided from other grants

** Approximate figure; may be somewhat less.

*** The budget is lower on these items for the first year since there is expected to be a three month delay after the project is approved and the equipment ordered before they can be delivered.

Second Year

	Total Cost	Requested from NSF	Contributed by Univ. of Kansas
Salaries			
Dr. F. J. Rohlf, Principal Investigator 20% of time for 9 months*	$ 2,625	$----	$ 2,625
Research Assistant 1/2 time for 11 months	2,823	----	2,823
SALARY SUBTOTAL	$ 5,448	$----	$ 5,448
Employee Benefits	229	----	229
Permanent Equipment	----	----	----
Expendable Equipment and Supplies			
Film, Paper, Cards	50	50	----
Travel	300	300	----
Publication Costs	50	50	----
Other Direct Costs			
Western Union broadband service	20,100	20,100	----
Computer time at UCSB	7,800	7,800	----
Rental of card reader, punch and controller	2,544	2,544	----
TOTAL DIRECT COSTS	$ 36,521	$ 30,844	$ 5,677
Indirect Costs (computed at 50.59% of the Salary Subtotal)	2,756	2,756	----
GRAND TOTAL	$ 39,277	$ 33,600	$ 5,677

Third Year

	Total Cost	Requested from NSF	Contributed by Univ. of Kansas
Salaries			
Dr. F. J. Rohlf, Principal Investigator 20% of time for 9 months*	$ 2,756	$----	$ 2,756
Research Assistant 1/2 time for 11 months	2,823	----	2,823
SALARY SUBTOTAL	$ 5,579	$----	$ 5,579
Employee Benefits	242	----	242
Permanent Equipment	----	----	----
Expendable Equipment and Supplies			
Film, Paper, Cards	50	50	----
Travel	----	----	----
Publication Costs	100	100	----
Other Direct Costs			
Western Union broadband service	20,100	20,100	
Computer time at UCSB	7,800	7,800	
Rental of card reader, punch and controller	2,544	2,544	----
TOTAL DIRECT COSTS	$ 36,415	$ 30,594	$ 5,821
Indirect Costs (computed at 50.59% of the Salary Subtotal)	2,822	2,822	----
GRAND TOTAL	$ 39,237	$ 33,416	$ 5,821

Three Year Summary

	Total Cost	Requested from NSF	Contributed by Univ. of Kansas
Salaries			
Dr. F. J. Rohlf, Principal Investigator 20% of time for 9 months*	$ 7,881	$----	$ 7,881
Research Assistant 1/2 time for 11 months	8,335	----	8,335
SALARY SUBTOTAL	$ 16,216	$----	$ 16,216
Employee Benefits	689	----	689
Permanent Equipment			
Teleputer Control Unit with single console	39,000 **	39,000	----
Polaroid Camera	450	450	----
Expendable Equipment and Supplies			
Film, Paper, Cards	150	150	----
Travel	600	600	----
Publication Costs	150	150	----
Other Direct Costs			
Western Union Broadband service	55,950	55,950	----
Computer time at UCSB	21,400	21,400	----
Rental of card reader, punch and controller	5,088	5,088	----
TOTAL DIRECT COSTS	$139,693	$122,788	$ 16,905
Indirect Costs (computed at 50.59% of the Salary Subtotal)	8,203	8,203	----
GRAND TOTAL	$147,896	$130,991	$ 16,905

Justification of the Budget

A. Teleputer Control Unit

A teleputer system as manufactured by Bolt, Beranek, and Newman, Inc. (BBN), Van Nuys, California, for the Culler-Fried on-line computing system at the University of California, Santa Barbara is required. The minimum complete teleputer system consists of a Teleputer Control Unit (TCU) incorporating a single user console at the user end of a transmission circuit and a Data Set Control Unit (DSCU) at the other end. The equipment is designed to operate over a full duplex 2400 baud data transmission system as supplied by Western Union Broadband Switching System. Type 201B data sets are required at both the user and the computer site. The TCU is designed to be operated with the RW 400 computer at UCSB and will have modifications for the new computer at UCSB. BBN will be able to supply a TCU with associated operating console, F.O.B. Van Nuys, California, at a price of about $39,000 (allowing for the fact that we will make use of an existing DSCU at Santa Barbara).

The TCU provides the interface between the telephone data link and the Teleputer Console(s). It is specifically designed to receive data formated for the control of the display oscilloscopes. It also formats the keyboard information, from operator key depressions, for transmission to the DSCU. The TCU controls from one to 32 teleputer consoles. The University of Kansas may separately purchase one or more additional consoles (at $5,000 each) for the use of the computer center's staff in gaining experience in this type of computation and for research scientists in various fields.* When and if the system is implemented on the GE computer at the University of Kansas, these TCU's will continue to be useful. The card reader and punch (see below) will be one of the 32 stations and can therefore be used at the same time as the other consoles are being used.

why remote

This particular type of remote station is appropriate for the proposed applications since it provides convenient graphical displays (using points and vectors) of alphanumerical text, special characters, and functions. Since it uses a storage type display oscilloscope (and hence does not permit a dynamic display which is not needed in the proposed work) it is practical for use at a remote station over voice grade transmission lines.

this makes nonsense.

why be so emphatic with determined procedures?

---

* The University of Kansas is firmly committed to move in the direction of on-line computation. After the first TCU is installed in Lawrence we expect to install additional ones using the Culler-Fried system in order to see whether this system meets many of the needs for on-line computing at Kansas. If our experiences with the Culler-Fried system continue to be favorable, we hope to implement it on our GE-600 system with the assistance of Dr. Rohlf. We are, however, also planning to experiment with other systems.

At fantastic Expense.

isn't this backward and somewhat expensive if the C-F system proves to not meet the needs?

C. Polaroid camera and film will be used for making permanent records of the results of computations.

Type C-13 camera (#413, f/4.5 - 1:0.7 lens)	$360.00
Roll film pack (part # 1220603-00)	75.00
Tek part # 016-217 Mounting Bezel	15.00
	$450.00

D. Travel

This item is included in the budget to allow for trips to Santa Barbara to consult with Dr. Glen Culler and his staff if case problems arise and to permit more extended discussions on details of implementing system changes.

E. Western Union Broadband Switching Service

The budget is based on an estimated average use of 3 hours per day, 5 days per week or about 65 hours per month. On this basis the operating costs per month will be:

Broadband Service, fixed cost	$ 30.00
201-B Data Set	72.00
40 mile private line from Kansas City (@ $2.22 a mile)	88.00
65 hours of operating time (55 cents per min. 40% discount after first $500)	1,485.00

The estimated 65 hours of operating time per month is expected to be adequate.

F. The computation center at the University of California, Santa Barbara will make computer time and a DSCU available for $10.00 per hour.

G. Card reader and punch

A card reader and punch is desired in order to make more efficient use of transmission time. Data and programs may be punched off-line and then read in while other programs are being run from the console. After the data and programs have been transmitted, the programs can be run and checked out from the console and the data analyzed. The punch will be used to obtain copies of programs and data for later use. Since these activities can go on simultaneously and since relatively low volumes of information will be transmitted inexpensive low volume units are sufficient. These units can be rented from IBM at a cost of $212 per month.

IBM 1051 Controller for reader, punch, and printer	$100
IBM 1056 Card reader	70
IBM 1058 Card punch	95
Total	$265
Less educational discount	53
	$212

H. A half-time research assistant to Dr. Rohlf will be furnished by the University of Kansas from the Computation Center staff to help with the development of the system. soup5 for.

XII. Current support and pending applications

A. Dr. Rohlf is completing at present an NSF research grant (GB-1424) "Numerical Taxonomy-Theory and Application." Sept. 1963, for approximately two years. $14,500 (at the University of California, Santa Barbara).

B. A joint research proposal with Drs. J. H. Camin and R. R. Sokal "The development and testing of methods of numerical taxonomy" will be submitted to the National Science Foundation soon. Proposed starting date: June, 1966, 3 years, approximately $143,000 for three co-principal investigators.

C. This proposal is not being submitted to any other agency.

Acknowledgement:

This system has been made available for your use through an equipment donation by the Bunker-Ramo Corporation and through cooperation with Rome Air Development Center, USAF. The pilot program under which we developed software and showed feasibility of telephone driven curvilinear display consoles is supported by the Office of Naval Research. The program to develop a communication laboratory and associated classroom applications is supported by the Advanced Research Projects Agency. The dissemination of the system over national carriers is made possible via the support by the National Science Foundation.

The form and style of the mathematical communication within this system has been developed jointly with Dr. Burton D. Fried in research programs at Bunker-Ramo and Space Technology Laboratory. It emphasizes constructive techniques in problem solving which, over the years, we have found most effective in probing the structure of apparently complicated problems in applied analysis.

Finally, we owe an unpayable debt of thanks to a host of co-workers, colleagues, and friends who have contributed a considerable amount of personal effort, time, and influence and thereby made possible the early realization of our dreams for direct, user control of a computer as a problem solving tool.

G.J.C.

TABLE OF CONTENTS

INTRODUCTION

Through many recent hardware and software developments, it is now possible to provide a variety of computer users with direct and effective access to a computing system. This access is very well organized and carefully controlled; it therefore differs markedly from open-shop operation that was popular in the early 1950's. The drive to produce more efficient systems which earlier led to the closed shop and batch processing approaches has more recently led to on-line computation and time-sharing. These more recent efforts attempt to combine some of the valuable assets of earlier modes of operation to create systems more readily related to user needs both in terms of problem language and computer response. Our primary interest is in human control of a computer at a rather sophisticated mathematical language level. In this manual we will suppress all details related to hardware and as much detail related to software as is consistent with our desire to make this manual self-contained.

One can characterize the system described here as a problem oriented language which provides an on-line capability in the area of classical mathematical analysis. It is possible to do a variety of algebraic things with this system, but the basic constructs are those of analysis rather than algebra. There are many problems which are, so to speak, natural for this type of system, but there are also many which are utterly inappropriate and which violate the premises under which the system was designed.

For the sake of aiding you in judging such appropriateness, we suggest the following check list:

(1) Is your problem primarily analytic in nature?

(2) Is your problem of enough direct interest to you to make you want to solve it yourself?

(3) Is your problem difficult to do by conventional means?

(4) Do you need to work through the problem to see how it goes?

(5) Is the visual inspection of results of subproblems of benefit to you?

If most of your answers turn out to be "no", we recommend that you use some other approach than that presented here. You, as user of the system, must really provide the mode of problem solution yourself, and problem solving - even with the best advantages - is always difficult enough to merit full consideration of the approach.

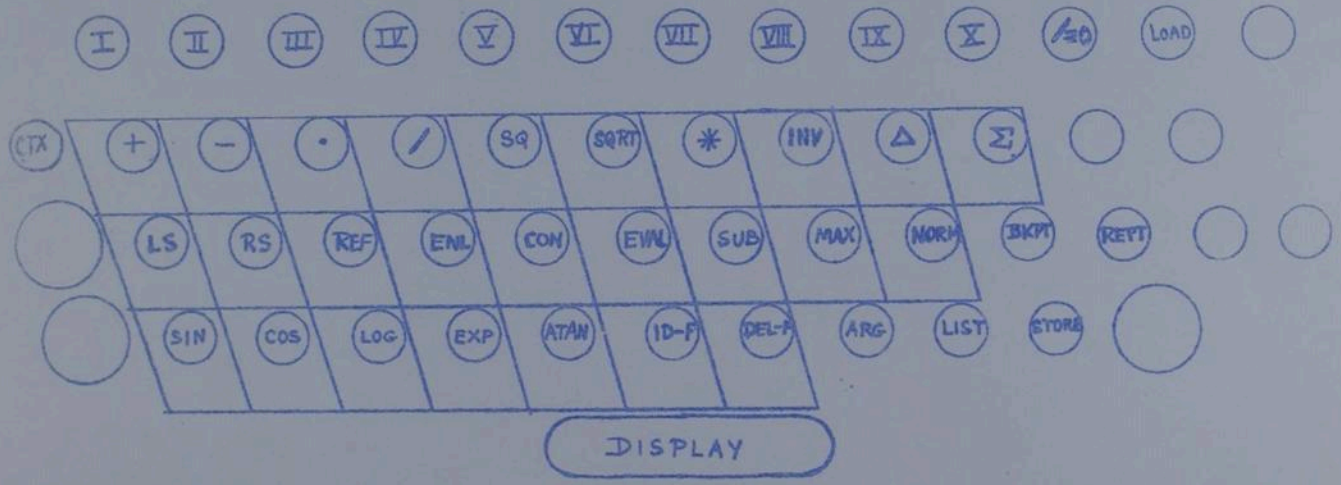
Taken altogether, the mathematical capabilities represented by the software lying behind this system are quite comprehensive; consequently, the user should be aware that to use the system one need not understand everything that can be done with the system. A good operating philosophy is to only pay detailed attention to that part of the system required for the solution of your problem.

Control of the computer is provided by means of the input keyboard as shown on the next page. This keyboard consists of two halves; the upper half permits access to operators and the lower half to operands. Thinking of each of these halves as separate keyboards we designate the types of keys as alphabetic, numeric, and punctuation. The labels on the upper keyboard do not reflect this nature but the color coding on the keys clarifies the correspondence.

DATE ___

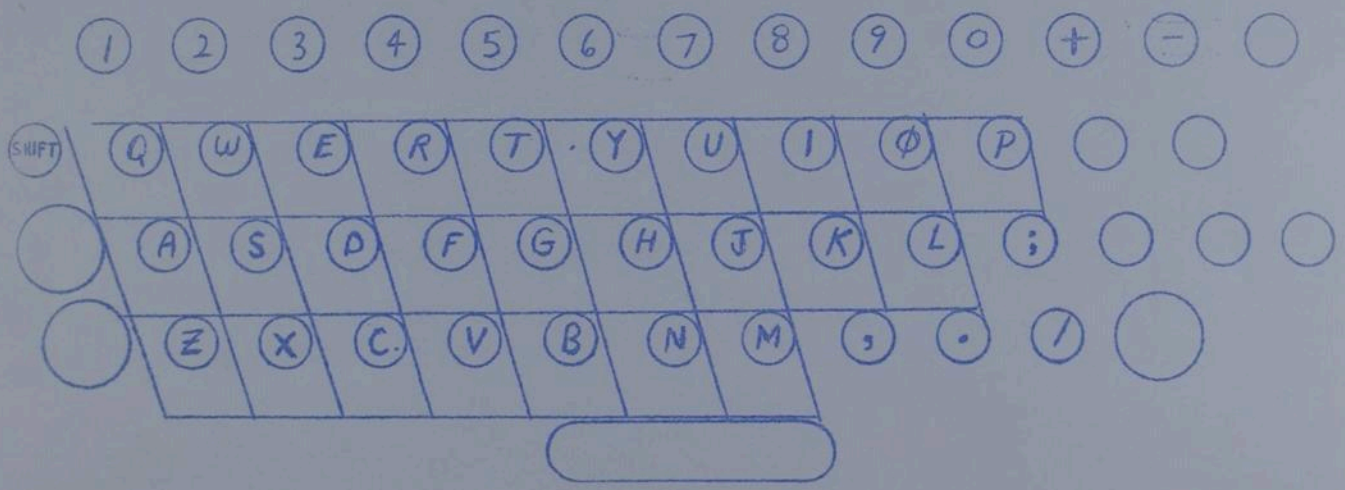
NAME ___

LEVEL ___



ARRAY ___
BANK ___

DATE ___
NAME ___



OBJECT KEYBOARD
(Right Hand)

The white keys are the alphabetic keys; they provide locations for storage. Programs and operations are stored under the white keys of the upper keyboard, and data may be stored under the white keys of the lower keyboard whether single numbers, one dimensional or two dimensional lists of numbers.

The blue keys, or numeric keys, provide a means for changing the reference to things stored under the white keys. That is to say, on the operator keyboard the roman numerals I, II, ... IX provide nine different sets of operations that can be accessed by depressing white keys, and the numbers 1,...,9 permit access to as many as nine different sets of data objects that can be stored under the white keys of the lower keyboard.

The green keys, as well as the black and the red, provide control of a non-mathematical and non-data type. Those on the upper keyboard provide what may be considered program control and those of the lower keyboard typewriter control.

The response of the computer to you as a user is provided by means of a display tube associated with each user station. This display tube is capable of presenting alpha-numeric display as well as curvilinear display and part of the system is designed to permit you rather easy means of using this tube for observing what the computer is doing with your problem.

The description of the system which follows is arranged in sections which for the most part correspond to the levels of

operations which are available in the basic system. Again, only those levels of direct significance to you as a user need to be studied. However, the section on user programs called Operators for Program Control needs to be read and understood by all users. After each chapter we will include a short list of exercises which will illustrate some elementary but typical uses of the operators on that level. We suggest that when learning a new level you do the exercises as a means of checking your own readiness. If you have difficulty, please call upon some other user to help you. Since you may be far away from the computer site and from those of us who have designed the system, we suggest that age old "each one teach one" philosophy. In case of trouble which requires our attention at the computer site, please call the UCSB Computing Center: Area code 805 968-1511, Extension 4145.

I

FLOATING POINT ARITHMETIC and SYMBOL GENERATION

On this level the data objects consist of single numbers represented with seven decimal digits and a decimal scale. You have storage for twenty-six such numbers under A through Z on the right hand keyboard on each of four different operand levels, 1, 2, 3, 4, and 21 numbers A through U on level 5. To gain access to data on a given operand level, just press SHIFT followed by the operand numerical key corresponding to the desired level number. Once on the level, the system will remain there until you specifically change levels. Thus at any time there is storage for twenty-six floating point numbers directly available for your use with 124 totally available on operator Level I. To enter a number into the system press LOAD and type in the desired number using the operand numerical keys. We have attempted to make this operation quite free of formating; to do this our program makes the following presumptions:

1. If no sign is used, the number is positive.
2. If no decimal point is used, the number is an integer.
3. If any key other than +, -, 0, 1,...,9 is pressed then you are through typing the number.
4. If a sequence of sign symbols occurs then the last one is correct. If a sign symbol occurs after a digit, then the following integer moves the decimal point and if no sign symbol comes after the digit then the decimal place is left unchanged.
5. If more than seven digits are used then all but the first seven are ignored. If less than seven are used the number is exactly represented.

Now suppose we press LOAD and type in a number. That number is held in temporary storage much in the same way the number twelve is held in temporary storage during the mental arithmetic operation 3 x 4 + 7 = 19. It is possible to combine a number in temporary storage with any of the other numbers previously stored under the alphabetic keys by using any of the elementary operations on the Level I. It is also possible to save the number which is in temporary storage by pressing STORE and then any other alphabetic key that you wish to store it under. If you wish to store it on some other operand level, then before storing merely change operand levels as described above. It is frequently useful to be able to read a number which has been stored in a particular alphabetic location. To do this merely press DISPLAY followed by the appropriate alphabetic key.

EXAMPLE.
On the operator keyboard press I, then press LOAD. Now type 30.7 and press STORE A DISPLAY A LOAD 4.72 + 1 STORE B DISPLAY B LOAD A + B STORE C DISPLAY C.

Your answer should be 36.42. Notice that we did not type L O A D but rather depressed the button with the symbol LOAD on it. For general information concerning the typing capability within the system refer to Section IX.

The operations +, -, ·, /, LOAD, STORE, and DISPLAY, all require the further specification of an operand before they take action. In contrast to this the operations SQ, SQRT, *, INV, SUM,

LS, RS, REFL, ENL, CON, MAX, MOD, SIN, COS, LOG, EXP, ATAN, ARG, DEL all take direct action; that is to say, they do not need the further specification of operand before the program runs to completion. The other operators DIFF, EVAL, SUB are non-operational on this level; if they are depressed the computer simply ignores them. For information concerning the operator punctuation keys other than DISPLAY refer to the section on Operators for Program Control and for information concerning the operand punctuation keys see the section on typing controls in Chapter IX.

In the following definitions let α represent an arbitrary alphabetic operand and n represent a sequence of numeric keys interpreted as a number in the sense described above; let τ stand for temporary storage.

OPERATOR DEFINITIONS FOR LEVEL I

One Place Predicates

+
-
....
.
/

if followed by α or n these compute the indicated combination with the number in τ and leaves the answer there.

LOAD.....if followed by α or n, then the resulting number is placed in τ .

STORE....if followed by α , then the number is placed in the indicated permanent storage position. If followed by n, it is ignored.

DISPLAY..if followed by α , then a decimal display is generated on the display tube. If followed by two operand digits, it initiates the display process for symbol generation by displaying a point. After this initiation, the operand keys marked with the small vector symbols will respond by displaying that vector on the tube attached to the end of the last vector. This symbol construction can be halted at any time, but if the symbol is to be stored for further use, then press SHIFT n STORE . For each new user, the symbol table on Shift 1 is initially provided with upper case letters..25 per line and 20 lines per tube. The remaining cases, 2,...,9 are open for your own definition. The number m of characters per line and the number n of lines per tube can be defined by using SHIFT on Level IX as detailed in Section IX.

PLACE FREE PREDICATES

SQ..... squares the number in τ.

SQRT.... square roots the number in τ.

* ..... negates the number in τ.

INV..... takes the reciprocal of the number in τ.

MOD..... takes the modulus of the number in τ.

SIN, COS, LOG, EXP, ATAN,.... perform the indicated operations on the number in τ.

SUM..... totals all numbers presently in permanent storage and leaves the answer in τ.

LS..... moves all numbers in permanent storage one position to left; that is, B goes to A, C goes to B, etc. through each of the operand levels with the first A moved to absolute last and with the first number on the next operand level becoming the last number on the prior level.

RS..... reverses the process described for LS.

REFL.... reverses the order of the numbers in permanent storage.

ENL..... enlarges the mantissa of τ by the factor of 2 and corrects the scale.

CON..... contracts the mantissa of τ by the factor of 2 and corrects the scale.

MAX..... places the largest of the numbers in permanent storage in τ.

ARG..... if the number in τ is positive, it is replaced by zero. If it is negative, it is replaced by $3.141593 \approx \pi$.

DEL..... if the number in τ is not zero, it is replaced by zero. if it is zero, it is replaced by 1.

ID..... erases the tube.

CTX..... if followed by CON, contracts the definition of the vector data in the Y-register to a set of floating point numbers appropriate for Level I and places these in permanent storage. If followed by ENL, enlarges the definition of the set of floating point numbers in permanent storage to vector data and places the resulting vector in the Y-register.

II

REAL FUNCTION OPERATIONS and CURVILINEAR DISPLAY

The data objects used on Level II can be thought of as vectors or as lists of real numbers. We denote such vectors in a typical vector-coordinate manner such as:

$$Y = (Y_1, Y_2, \ldots, Y_n)$$

The number N of coordinates allowable in our present system is restricted to be less than or equal to 124; the number M of such vectors allowable in an array is likewise restricted to be less than or equal to 124. The number of arrays is restricted by the total amount of storage used, but from the point of view of names, we are specifically restricted to arrays A through Z. These arrays may in turn be either real or complex and for a description of how they are defined and what is implied concerning subscripts and indices and component vectors, refer to Section X on data structures. Throughout our description of Level II, we will presume that the arrays needed have previously been defined and are available for our use. To understand how to change reference from one array to another refer to Section XII on operations for program control.

The mathematical capability represented by Level II is a form of discrete calculus, but for the sake of convenience one can frequently ignore the discreteness and think of data vectors as values of functions defined on some real interval. The selection of

operations has been made in such a way to make a happy balance between ease of formula construction in mathematical expressions and simplicity of operator definition. As a means of conveniently picturing what takes place during application of these operations, we like to think of temporary storage as consisting of two list registers, an X register and a Y register. At the time a function or curve has been displayed the picture shown on the display tube is precisely the picture of the function lying in the X and Y registers. Now we also must treat combinations of functions and for this purpose it is convenient to think of auxiliary function registers U and V. The alphabetic keys on the operand keyboard provide a means for referencing the set of functions presently available to the user. Each set A, B,...,Z is one level of some data array, and depressing the SHIFT key followed by a number will change levels within this data array and thereby make some other 26 functions available. The real operations are defined on the Y register for place free predicates. If the array presently referenced happens to be complex, then the X register will contain the X coordinates of the last complex vector that has been loaded either by LOAD or DISPLAY. The X coordinates are left unchanged by most of the operations on this level, the exceptions being SUB and EVAL, and for complex arrays LOAD.

In the following definitions let α represent an arbitrary alphabetic operand and n represent a sequence of numerical keys interpreted as a number, as discussed in section I.

OPERATOR DEFINITIONS FOR LEVEL II

ONE PLACE PREDICATES

LOAD.....If followed by α then the function under α is placed in Y (or in the case of a complex array in X and Y). If followed by n, then a constant function each of whose coordinate values is the number n is placed in the Y register.

STORE....If followed by α then the function in the Y register is stored under α for real arrays and for complex arrays the arc in the X and Y registers is stored under α. If followed by n it is ignored.

DISPLAY..If followed by α then a curvilinear display of the function contents of α is shown on the display tube and the X and Y registers are filled (as in LOAD α) by the data shown on the display. If followed by n then the numerical value of the n th coordinate of the Y register is displayed. If n is zero then the binary scale of the Y register is displayed.

+
-
.....
.
/ If followed by α the function stored under α is placed in V (or for complex arrays in the pair U V) and combined with Y as indicated by the operation. The resulting function is left in Y. If the denominator in a division has some of its coordinate values zero, then these same coordinate values will be zero after the division has been completed. If followed by n then a constant function whose values are n is loaded into V and the indicated combination with the function in the Y register is formed and left in the Y register.

SUB.....If followed by α then the Y coordinates of α are placed in the X register. If followed by n then the constant function determined by n is placed in the X register.

EVAL.....If followed by α then the function under α is placed in V, (for a complex array in U V). For each value v_k in V, the least upper bound and the greatest lower bound of the X register is obtained relative to v_k and a linearly interpolated value is then computed from the function in the Y register. If the array is complex then the U coordinates replace the X coordinates and if the array is real the X coordinates are left unchanged. If followed by n then the interpolated value corresponding to n is placed in the Y register.

PLACE FREE PREDICATES

SQ.....Squares the function in the Y register.

SQRT.....Square roots the function in the Y register giving a zero result wherever the square root in undefined.

*.....Negates the function in the Y register.

INV.....Divides 1 by the function in the Y register.

DIFF.....Forms the forward difference of the function values in the Y register and supplies the right hand value by a second order extrapolation.

SUM.....Is a running summation of the values in the Y register, ie...the corresponding sub-total is stored in each Y coordinate.

LS.....Places the k + 1 st coordinate of the Y register into the k th position, and places the first coordinate in the last coordinate position.

RS.....Places the k th coordinate of the Y register in the k + 1 st coordinate position and the last coordinate in the first coordinate position.

REFL.....Interchanges the order of the coordinates in the Y register.

ENL.....Doubles the values of each Y coordinate and diminishes the binary scale by one.

CON.....Halves the value of each Y coordinate and adds one to the binary scale.

MAX.....Makes a constant function equal to the maximum value in the Y register.

MOD.....Takes the absolute value of the function in the Y register.

SIN, COS, LOG, EXP, ATAN,....Perform the indicated operations on the Y register leaving a zero wherever the operation is ill-defined.

ARG.....Replaces negative function values by π and positive function values by zero.

DEL.....If the function in the Y register has a zero or if it can be interpolated to have such a zero then the nearest coordinate value is replaced by one and all others are replaced by zero.

ID.....The X and Y registers are loaded with a representation of the internal (-1,1) consisting of N equally spaced points.

III

OPERATIONS ON COMPLEX ARCS

There are two strong reasons for having easily available operations on complex arcs. One of these is the mathematical need for dealing with complex functions which is so necessary to go very far with classical mathematical analysis. The other, which is trivial beside this in importance but of considerable value to the person solving problems, is that operations on complex arcs provide the easiest means of mathematically specifying how to control display information. Putting these two requirements together we are led to a simple generalization of Level II, which is a proper base for applications involving analytic functions and conformal mappings and also provides an easy means for geometric motions. The data objects on which Level III operations are defined consist of vectors whose coordinates are complex numbers. If these complex numbers are visualized as points in the complex plane which are joined by line segments, then our data objects are representations of polygonal arcs. It is equally possible to think of these data objects as consisting of two real vectors each of which would be appropriate for use on Level II; that is, one of these providing the real part and one the imaginary part of the so-called complex arcs. In order to fix our discussion let us emphasize this relationship between Levels II and III. Suppose that a complex array has been defined and is available for our use (refer to Section X on data structures). Let Z be a column of that array then ...

$$Z = (Z_1, Z_2, \ldots, Z_n)$$

$$Z = (X_1 + iY_1, X_2 + iY_2, \ldots, X_n + iY_n)$$

$$Z = (X_1, X_2, \ldots, X_n) + i\,(Y_1, Y_2, \ldots, Y_n)$$

$$Z = X + iY$$

A comparison of the first and last of these expressions illustrate the different views described above. With the last of these in mind, we think of a complex arc as having its real part in the X register and its imaginary part in the Y register. Whereas the Level II operations primarily transform the information in the Y register, the Level III operations transform the data in both X and Y. Again, it is useful to have an auxiliary register and because of the complexness we have two of these, U and V, which provide temporary data storage required for performing the combinations for one-place predicate operations.

Most of the operations on Level III can be defined as extensions of the Level II operations based entirely on the relationship between the arithmetic of real numbers and the arithmetic of complex numbers; however, since we must deal with analytic functions defined on complex arcs, the mathematical definitions of such functions must be carefully provided in a manner consistant with the usual applications of this level. This means we must provide for branch cuts, multiplicities etc. in such a way as to minimize the work that must be done by the user in adapting his problem to the operations we have selected. Our approach is based on the assumption that the data objects really are discrete samples of continuous arcs in the plane and therefore the image of such an arc under transformation by elementary functions should itself be continuous. To do this in a consistent fashion, we first define the argument function by specifying that the first point of the function shall have an argument in the interval $[0, 2\pi)$ and that the argument of each subsequent point is determined by summing the changes in argument as one passes along the points of the curve. This choice makes it meaningful to use the argument operation as

a means for computing the winding number of the curve about the origin. The functions LOG, ATAN, and SQRT are then defined in terms of this argument function and these are the only ones on Level III that have branches or multiplicities. In order to make an appropriate definition of evaluate we need to provide facilities to define substitute on the complex level in such a way as to extend the relationship between SUB and EVAL on Level II. To do this we introduce a new register called the Zeta register, ($\mathcal{Z}$), which can hold a complex arc. Substitute will then store data in the Zeta register analagous to Level II SUB storing data in the X-register. With $\mathcal{Z}$ as the combination of the X and Y registers, we can then have a direct correspondence between $\mathcal{Z}$ and $\mathcal{Z}$ representing $\mathcal{Z} = f(\mathcal{Z})$ analagous to $Y = f(X)$ for the real function level.

It was possible to perform significant real operations on the Y-coordinates of even a complex array on Level II. On Level III we consider a real array just to be the real part of a complex array, the imaginary part of which is zero. Consequently, for the one place predicates operating with CTX set to a real array, we begin by filling the V register with zeros. In the usual case, with CTX set to a complex array, then the incoming data is placed in U and V. In both instances the results of the operation are left in the X and Y registers.

In the following definitions let α represent an arbitrary alphabetic operand and n represent a sequence of numerical keys followed by a comma and this followed by a second sequence of numerical keys. That is, we enter a complex constants $n_1 + i\, n_2$ by typing n_1, n_2.

OPERATOR DEFINITIONS FOR LEVEL III

LOAD.....If followed by α , then for a complex array the arc under α is placed in the X and Y registers (for a real array, the real vector under α is placed in the Y register). If LOAD is followed by a constant n_1, then n_1 is placed in the X-register; if n_1 is followed by comma and a constant n_2, then n_2 is placed in the Y-register. Thus the sequence for loading a complex constant $n_1 + i\, n_2$ into Z is: LOAD n_1 , n_2

STORE....If followed by α , then for a complex array, the arc lying in the X and Y registers is stored under α , (the vector lying in the Y register is stored under α for a real array). If followed by n, it is ignored.

DISPLAY..If followed by α , then the curvilinear display of the arc under α is shown on the display tube and left in the X and Y register just as in LOAD. If followed by n, then the numerical values of the n th X and Y coordinates are shown on the display tube in decimal form.

\+)
\- (---
. !
//)

If followed by α , then the arc stored under α is placed in U and V for complex arrays (the vector stored under α is stored in U and zero is placed in V for real arrays) and the indicated combination is taken with the arc in the X and Y registers; the result being left in the X and Y registers. If followed by n, say in the same complex formating as in LOAD, then U and V are filled with constants and the combinations with X and Y are taken as above.

SUB.....If followed by α , with CTX set to a complex array, then the arc under α is placed in a correspondence storage register. We will call this the $\mathcal{F}$ -register and denote a typical correspondence by:

$$Z = f(\mathcal{F})$$

Regarding the Z register as the pair of X and Y registers, then the ($\mathcal{F}$, Z) correspondence on Level III is analogous to the (X, Y) correspondence on Level II.

If followed by α , with CTX set to a real array, then the vector under α is placed in the X register. If followed by n then the constant n is placed in the $\mathcal{F}$ -register.

EVAL.....(a) If followed by $\mathcal{K}$ and CTX is set to a complex array, then an interpolated image of $\mathcal{S}$ is placed in Z and the arc under $\mathcal{K}$ is placed in $\mathcal{S}$. This interpolation is obtained from a local extension of the mapping correspondence.

$$Z = f(\mathcal{S})$$

to a region containing the arc $\mathcal{S}$.

(b) If followed by $\mathcal{K}$ and CTX is set to a real array, then $\mathcal{S}$ is used in place of X and Y-registers for a Level II EVAL. The resulting pair is then treated as an incoming $\mathcal{K}$ and (a) is carried out.

(c) If followed by n then after n is formed, either (a) or (b) is carried out according to whether n is real or complex.

PLACE FREE PREDICATES

SQ....Squares the complex arc in the Z-register.

SQRT..Square roots the complex arc in the Z-register such that the ARG of the answer is one-half the ARG of the original function.

*.....Conjugates the complex arc in the Z-register.

INV...Computes the complex reciprocal of the complex arc in the Z-register.

DIFF..Forms the forward difference of the complex points in the Z-register and supplies the right hand end point by a second order extrapolation.

SUM...Computes the running sum of the complex values in the Z-register, ie, the corresponding sub-total of the X and Y coordinates is stored in each X and Y-coordinate position.

LS....Places the k + 1 st coordinate of the Z-register into the k th coordinate position, and places the first coordinate in the last coordinate position.

RS....Places the k th coordinate of the Z-register in the k + 1 st coordinate position and the last coordinate in the first coordinate position.

REFL..Exchanges the data in the X and Y registers which is equivalent to reflecting the Z-register about the 45 degree line.

ENL...Doubles the value of each Z-coordinate and diminishes the binary scale by one.

CON...Halves the value of each Z-coordinate and adds one to the binary scale.

MAX...Makes a complex constant equal to the maximum value of the X-register for the real part and maximum value of the Y-register for the imaginary part.

MOD...Takes the modulus of each of the points of the complex arc, and stores these in the X-register while placing zeros in the Y register, thus MOD of a complex arc is real.

SIN, COS, LOG, EXP, ATAN,...Perform the indicated operations on the Z-register, and when required, we use ARG as a sub-operation to make the complex operation well defined.

ARG...Computes the argument of the first point in the Z-register in the half-open interval, $[0, 2\pi)$, and computes the argument of the remaining points by summing the difference in arguments for adjacent points.

DEL...Consists of the product of the real DEL operation applied to the X-register and the real DEL operation applied to the Y-register. Thus, if the complex arc passes through a small rectangle about the origin, then the nearest point will have the value 1 and other points outside this small rectangle will have the values zero.

ID....Loads both the X and Y-registers with a representation of the interval (-1, 1) consisting of n equally spaced points.

37

SCIENCE AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE
1515 MASSACHUSETTS AVENUE, NW, WASHINGTON, D.C. 20005 • 387-7171

8 October 1965

Dr. David J. Rogers
The New York Botanical Garden
Bronx Park
Bronx 58, New York

Dear Dr. Rogers:

In the early spring Harvard University Press will publish a translation (by Gerd von Wahlert) of Franz Schwanitz's THE ORIGINS OF CULTIVATED PLANTS. We hope you will review the book for Science.

As far as I can determine this is a translation of DIE ENTSTEHUNG DER KULTUR PFLANZEN published by Springer in 1957 (so far I have seen only the announcement that Harvard would publish the translation). We did not review the 1957 edition.

The enclosed card is for your reply. If you can review the book for Science, we will send you the review copy as soon as it is available, and we would like to have a review of approximately 450 words four or five weeks later.

Sincerely,

Sarah Dees
Book Reviews, Science

SD:lph
Enclosures

Encl card returned
10/15/65

Review for Science

This is at least the second book to appear with the same title. The first was written by Alphonse DeCandolle, and appeared in 1886, but the similarity between them ends approximately there. Franz Schwanitz's "The Origin of Cultivated Plants" (transl. by Gerd von Wahlert, Harvard Univ. Press, 1966, vi + 175 pp.) is an effort to explain how cultivated plants have evolved, and how they differ from wild species. DeCandolle's efforts were directed towards description of as many cultivated species as possible, and towards determining the place(s) from which these species arose. Both books are worthwhile.

Schwanitz follows more in the tradition of Darwin's efforts, to discover how (mostly food) plants differ in their genetic and other mechanisms from wild ones, and how these mechanisms arose ~~(not to make a documentary of our cultivated plant species)~~. The book owes a debt to the groundwork laid by Darwin, DeCandolle, and by the Russian plant breeder and geneticist, Vavilov, all generously acknowledged by Schwanitz.

Gigantism is the characteristic of cultivated plants primarily responsible for their usefulness to man--the enlarged roots of cultivated carrots contrasted to the small, woody tap root of its wild relative (Queen Anne's Lace); the large, fleshy fruits of tomatoes (Lycopersicon esculentum) versus the putative small, seed-filled fruit of wild relative, (L. pimpinellifolium). Gigantism of certain organs, a phenomenon common to many cultivated species, is ~~derived~~ caused by several different genetic mechanisms: mutation in some, hybridization in some, and polyploidy in others. Gigantism may result from increased cell size or from increased numbers of cells in the useful part. These variations, once established, are kept going by man--they seldom have any competitive ability if not nurtured, weeded, watered, etc. And the influence of environment (which must include man as a factor) is another critical ~~role~~ part of the picture for the development of cultivated species. Perhaps in

environmental modification primitive man played his biggest role in developing useful plants. Certainly he did not have a program of breeding towards a desired goal. But by chopping down competitors, keeping live stock and by generally messing up the natural habitats, he made great strides in the development of most of our cultivated species. But Schwanitz does not put it this way. To him, plant breeding is as old as agriculture (in Chapter 4, the history of plant breeding). His definition of plant breeding is much broader than I would care to make it, since to me, breeding involves much more knowledge of the biology of the organisms than the primitive people had available to them. What he must mean is a sort of selection process, where chance hybridization occurred, or a mutant suddenly appeared with some desirable quality, and these variants were kept going by some observant primitive farmer.

Whatever interpretation is made, however, this is an informative book, and useful for the interested reader. A short list of general references, mostly from the German literature, is appended.

David J. Rogers
Colorado State University

FEB 3 1966

THE UNIVERSITY OF WISCONSIN PRESS

P.O. Box 1379, 807 West Dayton Street, Madison, Wisconsin 53701 Telephone 608/262-1116

February 1, 1966

Dr. David Rogers
Department of Botany
and Plant Pathology
Colorado State University
Fort Collins, Colorado

Dear Dr. Rogers:

This is to acknowledge that the manuscript described below has reached this office safely.

Yours sincerely,

The University of Wisconsin Press

Author of ms: Jonathan D. Sauer

Title of ms: "Plants and Man on the Seychelles Coast"

THE UNIVERSITY OF WISCONSIN PRESS

P.O. Box 1379, 807 West Dayton Street, Madison, Wisconsin 53701 Telephone 608/262-1116

February 1, 1966

Dr. David Rogers
Department of Botany
and Plant Pathology
Colorado State University
Fort Collins, Colorado

Re: Jonathan D. Sauer,
"Plants and Man on
the Seychelles Coast"

Dear Dr. Rogers:

Thank you for your report of January 26 on the Sauer manuscript. You have answered all my questions; and the Committee will be grateful, I know. This report will be of real assistance to them in their further consideration of this manuscript.

Your speed in returning your report is much faster than the wheels of the University's check-writing department. I shall have to send your check to you in about ten days. I apologize for the delay.

Again, thank you for your advice.

Yours sincerely,

Barbara Chase
kk

(Mrs.) Barbara Chase
Assistant to the Director

BC:kk
Enclosure

THE UNIVERSITY OF WISCONSIN PRESS

P.O. Box 1379, 807 West Dayton Street, Madison, Wisconsin 53701 Telephone 608/262-1116

January 17, 1966

Dr. David Rogers
Taxonomy Lab.
Department of Botany and
Plant Pathology
Colorado State University
Fort Collins, Colorado

Re: Jonathan D. Sauer,
"Plants and Man on
the Seychelles Coast"

Dear Dr. Rogers:

Thank you for undertaking the reading of the Sauer manuscript. It has been mailed to you today. Would you fill out the enclosed postal card to notify us when it arrives?

In regard to the "Reader's Report" form attached, please do not allow it to restrict you in any way. Frequently, readers ask what we want them to tell us; this form is an attempt to suggest the general kind of comment we should like to have about a normal manuscript. If you prefer some other way of organizing your comments or if the questions do not fit the manuscript we are sending you, modify or substitute questions as appropriate. Or disregard the form entirely; we have no intention of running your reply through a computing machine. In any comments you make, I should be grateful for at least a bit more than simple yeses and noes even when answering specific questions; for our purposes, the most useful statements contain comment and explanation, with citation of examples in the text where appropriate.

I have noted in our records that we should look for your report by February 14. If there is any reason to change this date, we should be grateful to you for information as far in advance as you may be able to supply it.

You may expect a fee of fifty dollars.

I enclose a label for returning the manuscript. It is printed so that the package will qualify for the appropriate postal rate. We have a general insurance policy protecting manuscripts while they are out of our office to readers, and so the package need not be insured when you mail it. I enclose stamps to cover our estimate of the postal charges. If we have not sent enough, please let me know.

Yours sincerely,

Barbara Chase

Mrs. Barbara Chase
Assistant to the Director

Enclosures

Comments - Expand on purpose of study in introd.

P1 - Biota -
Then work A. the plates

Headings - more needed ex: under next setting -
1) geological
2. surface + soils -
3 climate

In appendix 1 - can author suggest why he did not include the authorities for the plants named? There may be a reason, why but I would suggest they be added.
also sci names ital, common names Roman

In Section on Pl 9 introductions, (+ Appendix 2) no mention of Manihot.

Fig 11 Much too small - very diff. to differentiate. also the illustrations in the legends (the boxes) are too small.

P65 depauperate for unhealthy?

SCIENCE

Author DAVIDSON, R. A., & DUNN, R. A.

Title COMPUTER SIMULATION OF CERTAIN FORMS OF EVOLUTIONARY CHANGE: A PRELIMINARY REPORT

Comments: This paper fails to be satisfactory on several counts. (1) It does not stand alone, but depends upon unpublished papers. (2) Description of input procedures and interpretation of output without any description of the operation of the model does not allow the reader to judge the validity of the argument. (3) The authors do not define their terms, but leave them to the reader to guess. For example, there is no definition of "evolutionary cycles" though this term is one of the assumptions (number two); no definition, or example, is given for procedures to (words reversed) "operationally approach a change in taxa" (assumption three). (4) The points (page 5) which "bear more directly upon mathematical details" are simple facts of statistics. Why put them here? (5) Instructions to the operator (page 6) are much too arbitrary for the reader to get any grasp of the problem. Furthermore, it is stated that the operator selects μ and σ for each character, but the last sentence on the page indicates that μ and σ are determined by the program. (6) Comments on the conclusions (bottom of p. 8 and top of 9). A summary of discriminant function analysis should be given because the principal conclusion of the paper is that this (i.e., discriminant function analysis) is a good tool for making some particular analysis not described by the paper here presented. These concluding comments are almost meaningless without at least a brief description of the techniques which this entire discussion is designed to support.

It is not clear how the computer program results can corroborate the results of any other method used to analyze real data. The program operates in direct consequence of (1) statistical assumptions of normality and independence of variables and (2) biological assumptions that the result of evolution is a change in mean and standard deviation of the evolved groups from the ancestral group. To conclude that a discriminant function derived logically from these assumptions is successful for data logically generated from these assumptions is not an empirical observation but a necessary consequence of the design of the experiment. Hence, this observation does not substantiate in an experimental way the further application of the discriminant function to nongenerated data. This, of course, does not necessarily mean that the technique involving the discriminant function is not good. As this technique is nowhere explained, it is impossible to criticize it.

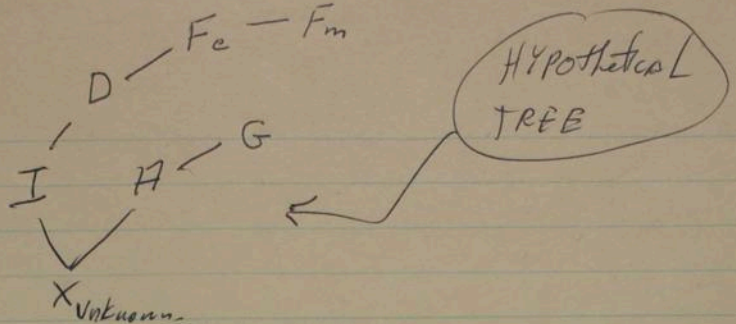
Paragraph one of the paper is not a satisfactory introduction to its contents. The last sentence of the paragraph is incorrect. Individuals are proposing various models, and perhaps some statement of those at work in these endeavors should be cited, if applicable to this paper. Certainly, systematics is not "in the process of developing" etc.

Overall Evaluation

- ☐ Excellent, merits rapid publication
- ☐ Publish if space is available
- ☐ Belongs in specialty journal
- X ☐ Should not be published anywhere

Advisor's Name David J. Rogers Date June 23, 1966

12-5M-0166, KP



Paper outlines the computer program which simulates the kind of evolutionary change found in Froelichia. This program avoids the necessity of having to actually measure real data which actually is known to take time and "lack flexibility" i.e. drugs.

Assumptions for Program

1. The base taxa are well defined

2. Taxa arise during EVOLUTIONARY CYCLES this term is nowhere defined.

3. I don't know what it means to approach a change.

I think 3 says that it is possible to measure the difference between object (taxa) numerically by different methods.

(I don't see this as an ASSUMPTION)

4. Characters describe the Genotype of an Organism. [What is a Genotype Estimate? How do you measure ~~[illegible]~~ the accuracy of its estimate? What does probabilistically have to do with it]

5 is clear. Who needs this assumption?

6. this is all he really wants. Characters are Random variables with Normal distributions. This assumption supports a statistical approach.

7. Standard Evolutionary trauma is allowed.

Math Points.

1 ~~[illegible]~~

$$\frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{\frac{x-\mu}{\sigma}} \, dx = P_r(-t \le R_v \le t)$$

mean
Standard dev } these are the only two variables.

1 + 2. } These } statements are simple facts well known to anyone who has taken an introductory course in Probability or Statistics.

Page 6 does not explain well the preliminary steps.

e.g. the operator determines the following values

① n
② p
③ and an array of pairs

$(y_1, \sigma_1), (y_2, \sigma_2), \ldots (y_p, \sigma_p)$

the following extra values are required

① the constant but arbitrary number of locations.

② the array of pairs.

$(b_1, c_1), (b_2, c_2), \ldots, (b_p, c_p)$.

Where do these values come from?
Third line from the bottom does not make sense.

~~Page~~ Page 6 Statement (3) says that μ and σ are determined by the operator - last sentence Page 6 says M and σ are determined by the program — which is it?

P7 ¶P ~~start last time~~ 2nd Sentence — this Multivariate normal etc etc ... where $\rho = \bar{\rho}$. Seems to make no sence.

note (end of 2nd P does not make sense in context).

P7 and 8 meanig of $\bar{\rho}$ transformations is not clear — how they are used is not clear the effect — what cards are being drawn from which deck and why. What does $\pm$ mean in the transformations.

The program operates in direct consequence of 1. statistical assumptions of normality and independence of variables and 2. biological assumptions that the result of evolution is a change in mean and standard deviation of the evolved groups from the ancestoral group. To conclude that a descriminant function derived logically from these assumptions is successful for data logical generated from these assumptions is not an empirical observation but a necessary consequence of the design of the experiment. Hense this observation does not substiantiate in an experimantal way the the furthur application of t the descriminant function to nongenerated data. this of course does not necessarily mean that the technique involving the descriminant function in no good. As this technique is nowhere explained it is impossible to critisize it.

premise: for data which satisfies assumption Q the descriminant function is applicable.

premise: Data is artificially generated to satisfy assumption Q.

conclusion the descriminant function is applicable to the generated data.

This conclusion is logical not empirical and hence not an experimantal result.

July 8, 1965

Mr. Theodore J. Crovello
Department of Botany
University of California
Berkeley 4, California

Dear Mr. Crovello:

We finally are getting around to answering some of the points that are pertinent concerning the paper that you submitted to BIOSCIENCE. I should preface any remarks that I make by letting you know my own dissatisfaction with the large amount of published information on what has been unfortunately designated as "numerical taxonomy." My dissatisfaction stems from my own experience over the years with the various statistical models proposed and the general lack of a good theoretical base for the proposed methodologies.

First of all, it seems that most of the methodologies, and I must say that these include my own up through the BIOSCIENCE paper, assume a sound theoretical knowledge of taxonomy. This is, as I am sure you are aware, an assumption entirely without base. We have no basic rules for classification. We have no basic rules for characters as input for classification. We have no definitions of the hierarchical categories. Without these rules and without these definitions no real work can be done from good and sound mathematical models, for, unless you can tell a mathematician what your rules are, he can not possibly formulate models to reflect these rules.

Some indications of this sort are given very well by Edgar Anderson. He has expressed his dissatisfaction on a number of occasions in several publications with the application of "bookbook" statistics to problems that were never intended to be used with these statistics. I concur in his reactions and if I have made any advancement over Dr. Anderson's last stated case, it is that I recognize my own shortcomings as a mathematician and my own lack of sound rules for classification. I have at least attempted to collaborate with various mathematical experts--we now have one on our staff--whose training is pointed toward the interpretation of biological rules to sound mathematical models. I find few other examples where the biologists acknowledge their shortcomings in this field.

With this loose preamble, let me quote to you statements made by the mathematician on our staff with reference specifically to your paper. This should tell you only that we are not attempting to stifle your interests in more rigorous methodologies for taxonomy but maybe to point out some directions that we feel to be important in this new and fascinating field. The following statements are made by Mr. Estabrook, our mathematician:

1. True claim of paper:

It is possible to subject a biological measurement to the affine transofmration: $T(x) = ax + b$ where a = the reciproval of the maximum measurement in some set of measurements and b = $(a^{-1}$ - the minimum measurement in the same set)(a). The assumed upper and lower bounds for the transofmration for this set of measuremants would be 1 and 0 respectively.

2. Claims not necessarily true:

"...provides for a more efficient operation by removing the step of coding completely". Aside from the fact that no precise definition of the notion "coding" has been made, any reasonable definition of the concept "coding" that occurs to me indicates that coding and its problems are not only removed but not even effected in many cases

"The character state values of a character for each OTU are used as the basic data source." Not only does this affine transformation not remove coding (i.e., the establishment of character states), but it requires it for input. The effect of this transofmration is merely to change the scale onto which the measurements are placed. No matter what ruler was used to make the measurements, it is still necessary to designate character states where they were formerly required.

"And because each character now varies in its states from 0 to 1, they are all of equal weight...." Nowhere in this paper is a definition of the notion "character weight" given; this makes it difficult to talk about the concept. However, even if weight is to be taken as casually synonymous with "importance" or "effect on the results of the process," the fact that two characters vary from 0 to 1 does not imply that they are equally important or that they effect the results in an equal manner. In some instances the range of a character (or the absolute value of the state (Crovello seems to think this is important since he requires that the transformed characters range not only over an equal interval but have maximum and minimum values of 1 and 0 as well)) can influence the value of, say, a similarity measure for example, but I would like to suggest that considerably more important factors would be the distribution of objects into the several states of the character, the number of states in the character, the standard deviation (as well as the mean) of the character, etc. Even if the notion of weight were defined, until it is stated how the characters are to be used in the taxonomy process, the above statement continues to be unfounded. There are many similarity measures, for example simple Matching and Pearson Lee Regression, which are invariant under affine transformations on the raw data. If "has equalweigt" is to mean "varies over the same interval," then the statement (less the comment re Adansonian Principles) is true, but I'm sure this is not the trivial conclusion Corvello meant to draw.

3. Relevance of paper:

Affine transformations can be used only when raw data is numerical (or two state), i.e., orderable. Many biological descriptions are not numerical or orderable. These could not be transofmred even if the transformation were considered desirable. This transformation does not necessarily solve the "coding problem." For many analyses, it does not even address the coding problem. There is no profit in publishing trivial FORTRAN IV programs.

4. Suggestions:

Take what you read about methodology less seriously. The "coding problem" is not solved, and many of the published discussions of it make considerably less sense than this one. Address coding from a biological point of view and then consult a mathematician for the methodology. Ask yourself what constitutes a sound basis for biological comparison. How can it be measured? How should the information in the observation of the basis for comparison be represented or summarized biologically? What bases for comparison should be used to form the input for the numerical (or any other kind) analysis of your own group of organisms? When you can explain what you want and what you mean biologically (which is the real difficult part of numerical taxonomy--biological not mathematical), then a mathematician or statistician can communicate with the comptting machine which can in turn PERHAPS make SUGGESTIONS on how to do your taxonomy.

Sincerely yours

David J. Rogers
Curator of Quantitative Taxonomy

DJR:MDF
Enclosure

Department of Botany
University of California
Berkeley 4, California
June 20, 1965

Doctor David Rogers
The New York Botanical Garden
Bronx 58, New York

Dear Doctor Rogers:

I hope this letter reaches you before you have mailed an answer to my last. I have just finished reading Kendrick's Taxon article and it brings up a nice coding problem which his method doesn't completely solve.

The problem is what to do when one has a quantitative character with a mixture of states like the number of utricle layers in Halimeda. In Salix, an example would be stamen number. In some species there is one, others have two, others have three, others have four, others have five, others have three to five. One might code as follows:

Code #(or letter)	State (stamen #)
1	1
2	2
3	3
4	3-5
5	4
6	5

Then using only this one character to keep it simple, an OTU with 3 stamens would not have the same similarity value when compared to one with 3-5 stamens as would one with 5 stamens when compared with the one with 3-5 stamens. (when Kendrick's method on page 153 is used). A simple coefficient which would just tally the number of states in common would be fine for this case, but as pointed out by Kendrick among others, such a simple coefficient can not discriminate between mismatches between adjoining states and those between states at opposite ends of the range. For this reason the latter is undesireable.

I would appreciate reading your comments on this situation.

Sincerely yours,

Theodore J. Crovello

Department of Botany
University of California
Berkeley 4, California
June 15, 1965

Doctor David J. Rogers
Curator of Quantitative Taxonomy
The New York Botanical Garden
Bronx, New York

Dear Doctor Rodgers:

Today I received notice that my article submitted to Bioscience had been rejected. I was disappointed, of course.

As you suggested to Doctor Leisner, I would appreciate it if you would "give just criticism directly" to me. The second paragraph in the letter from Doctor Leisner contains your recommendations and criticisms. I would like to comment on them now and would welcome an expansion of your views in return.

1. "...the paper itself is not a contribution." The contribution this paper sought to make was (a) to point out errors in Cain and Harrison's (1958) method; (b) to present a procedure which would correct such errors, given the criterion of equal weight for all characters; (c) to give a bit more publicity to an alternative method for obtaining equally weighted characters. It made no pretense to be a treatise on coding. If this paper accomplished the above three objectives, I think it could be called a contribution, and would be proper material for the "Research Notes" section of Bioscience, for which it was intended and so stated. Furthermore, the contributions listed above that I wanted to make with this article have (since my paper was submitted to Bioscience) been made recently by Sheals(1965) and will be made by Gower(in press).

2. "...the author does not understand all of the problems involved in coding..." Again, please comment on your statement, keeping in mind that discussion of coding in the paper was limited to the simplest example. But even if one has characters like the number of utricle layers that are best handled by coding, they can still be subsequently handled after coding by condensation.

3. "...some of the statements he makes are clearly out of line." Which ones? Why? Out of line with respect to what?

4. "His literature citation indicates he has not read as much as he should have read." I have one comment and one request. I hope you do not think I have read only the literature cited. Since the paper was intended as a research note I deliberately restricted literature cited to those works that were directly made reference to in the paper. My request is that you tell me what I "should have read".

5. "...he does not know exactly what is supposed to be done." From this comment you obviously do know, so I would appreciate being told what is to be done, since my three years of research may well be worthless if I am unaware as to what is supposed to be done.

The one change I would make would be the addition of the following to precede the last paragraph of the paper.

> If some characters must for some reason be coded, condensation can still accomodate them along with noncoded characters. If the coded states of a character can not be arranged in a sequence, then one is probably dealing with a compound or multiple character (Davis and Heywood 1963) that should be dissected into its unit characters.

Would you accept my paper as a contribution with this addition?

Thanking you in advance for your comments and criticsms, I await your reply.

Sincerely yours,

Theodore J. Crovello

Theodore J. Crovello

P.S. In reading the above letter through once more, it occurred to me that you may interpret its tone as being antagonistic. This would be a mistake. Your recommendations (or their paraphrase) contained in Doctor Leisner's letter were blunt and straight to the point. I am merely trying to be as concise.

I respect you because you chose not to remain anonymous and I am grateful for the interest you have expressed in me and hope it will continue.

BioScience

Editors in Chief: JOHN R. OLIVE and ROBERT S. LEISNER *Managing Editor:* TOM WHEATON COWARD

Advertising: JOSEPH BOURGHOLTZER, INC., 45 NORTH BROAD ST., RIDGEWOOD, N. J. 07450 • PHONE (201) - 652-3353

10 June 1965

Dr. David J. Rogers
Curator of Quantitative Taxonomy
The New York Botanical Garden
Bronx Park
New York, New York

Dear Dave:

As your suggestion, I have forwarded your comments to Mr. Crovello with the added recommendation that he contact you directly for additional discussion.

I informed Mr. Crovello that he should not be discouraged but rather take the comments in the same spirit as they were presented and to proceed to develop a superior paper.

Being the eternal optimist, I believe we may have saved a soul for biology and computers.

Dave, I appreciate your help, and I hope that all is well with you.

Sincerely yours,

Bob

Robert S. Leisner
Co-Editor-in-Chief
BioScience

RSL:sfl

AMERICAN INSTITUTE OF BIOLOGICAL SCIENCES
3900 Wisconsin Ave., N.W. Washington, D. C. 20016 • 202 - 362-6733

June 1, 1965

Dr. Robert S. Leisner
Co-Editor-in-Chief, BIOSCIENCE
3900 Wisconsin Ave., N. W.
Washington, D. C. 20016

Dear Bob:

The paper entitled "Condensation; a method for avoiding character state coding in numerical taxonomy while simultaneously giving equal weight to all characters" by Crovello which you sent us for review should not be published. There are two reasons for this: one, the paper itself is not a contribution, and two, I do not want this paper to spoil the author's reputation at this stage of the game. I would like to encourage him to continue as a person interested in quantitative taxonomy, and I am afraid that this would do damage to him.

As I understand the situation, he is a graduate student at Berkeley, and he has no one there at Berkeley with whom he can confer. As a result, he is sort of working in a vacuum. I have been in correspondence with him on various problems of classification using the computer and would like to encourage him to continue. It is clear that he does not understand the problems involved in coding, and some of the statements he makes here are clearly out of line. His literature citation indicates that he has not read as much as he should have read, and further he does not know exactly what is supposed to be done. The fact that he would submit for publication a rather naive set of FORTRAN statements is another indication that he needs to have some good advice.

As far as I am concerned, I would like to explain to him exactly what is needed in coding, what character coding is all about, and show him specifically what he has done and how he can improve the situation. Therefore, I do not care to remain anonymous as a reviewer but would ask that you convey to him in as kind a tone as can be arranged that this paper will do him more harm than good and that we will be glad to discuss this paper with him in a letter if he is interested. Again, remember that I would like to keep this guy going; we need more people working in this area. His major research group is the genus Salix which is in very great need of a computer type of analysis.

I would like to keep the paper here as a basis for the discussion after you have corresponded with him, telling him that we have kept the paper and if he wants us to we will give it a just criticism directly to him. I hope this procedure is satisfactory; if not, please instruct me.

Sincerely

David J. Rogers
Curator of Quantitative Taxonomy

DJR:MDE

BioScience

Editors in Chief: JOHN R. OLIVE and ROBERT S. LEISNER *Managing Editor:* TOM WHEATON COWARD

Advertising: JOSEPH BOURGHOLTZER, INC., 45 NORTH BROAD ST., RIDGEWOOD, N. J. 07450 • PHONE (201) - 652-3353

20 May 1965

Dr. David J. Rogers
Curator of Quantitative Taxonomy
New York Botanical Garden
Bronx Park
New York 58, New York

Dear Dave:

Since I am unable to judge the merit of the enclosed paper, I would appreciate your comments and recommendations as to its scientific merit and whether it should be considered for inclusion in BioScience.

I believe the method described by Dr. Crovello resulted from a research project on the numerical taxonomy of the genus Salix.

Thanks for any assistance you might be able to provide.

Sincerely yours,

Robert S. Leisner
Co-Editor-in-Chief
BioScience

RSL:sfl

Enclosure

AMERICAN INSTITUTE OF BIOLOGICAL SCIENCES
3900 Wisconsin Ave., N.W. Washington, D.C. 20016 • 202 - 362-6733