



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The Hunt Institute Archives contains hundreds of thousands of pages of historical content, writing and images, created by thousands of individuals connected to the botanical sciences.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time and social context in which these materials were created, some of the collections contain material that reflect outdated, biased, offensive and possibly violent views, opinions and actions.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 Archival records are historical documents, and the Hunt Institute keeps such records unaltered to maintain their integrity and to foster accountability for the actions and views of the collections'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

Home

420-4383

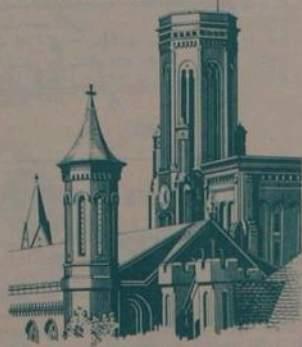
Office

381-5801.

Stan Shuler



Smithsonian Institution



ART

AVIATION AND SPACE EXPLORATION

HISTORY AND TECHNOLOGY

NATURAL HISTORY

ZOOLOGICAL PARK

All Mall buildings open at 9 a.m. and close at 4:30 p.m. September through March. Except where noted otherwise, the buildings remain open until 10 p.m. April through August. Closed Christmas Day. Admission free to all buildings. Limited guide service available by appointment.

Washington, D.C.

A
Guide
Map
to
the

Museum of History and Technology



UNITED STATES
NATIONAL MUSEUM

SMITHSONIAN
INSTITUTION

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D.C. 20560

June 26, 1969

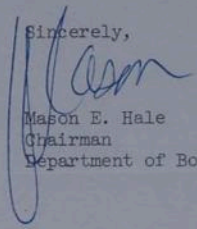
Dr. David J. Rogers
Professor of Botany
Colorado State University
Fort Collins, Colorado 80521

Dear Dave:

At this very late date let me formally thank you for your attendance at our brief conference on data processing in herbaria that we held here on 5 May. I believe this meeting accomplished much in demonstrating the diversity of efforts now being made in using EDP, as well as the tremendous amount of work ahead to make EDP a standard herbarium operation. I think it is too early to set up a formal committee, and in any event it should be done under the aegis of ASPT.

I will try to compose a summary of the ideas expressed at the meeting, however, as a means of keeping us in touch and send it to you, hopefully by September.

Sincerely,



Mason E. Hale
Chairman
Department of Botany

Dave: I appreciated very much your coming with Bob. The hindsight view of the meeting is not optimistic — we have a long way to go.

FNA still on fence! — now friction in the high ranks of Editorial Comm. — I tell more later.

Application of Electronic-Data Processing
to Museum Data Problems

--A Progress Report-- (NMNH)

Natural history museums have long been responsible for gathering, organizing, storing, and making available information about the natural world in which we live and which effect so profoundly by our activities. At this point in man's history, with the accumulation of vast quantities of documentary materials (specimens and records) and the growing awareness of the necessity to establish a balance between man and his environment, the tasks of these data centers can no longer be accomplished by traditional manual methods. The choice is clear: use of advancing technology for data-handling, or scientific oblivion for museums, and consequent loss of data critical to programs directed toward environmental understanding and management. The Smithsonian's Museum of Natural History, as one of the world's largest and most important centers for biological information, has chosen to study the feasibility of applying electronic data-processing to its problems. It will continue to give the highest priority to expanding these efforts as rapidly as resources permit.

Administratively, this program is in the Museum of Natural History and is conducted in close coordination with the Smithsonian's Information Services Division.

The current plan for the program is concerned with developing four major capabilities: (1) Making available the environmental and geographic information attached to natural history specimens; (2) more efficient management of specimen caretaking (storage, loans, etc.); (3) linking of

increasingly unmanageable bibliographic data to the specimen file; and

(4) enhancement of the research function by more efficient means of gathering, collating, and interpreting research data.

During the last two years, substantial progress has been made along several fronts.

(1) Vertebrate Zoology (Seabirds and Mammals).

Scientists of the Bird Division have devised a world-wide classification of most groups of birds. The records associated with 3,000 older collections and 2,000 new collections of seabirds already have been entered in the system. Improved specimen-cataloging techniques have been developed that make obsolete the time-honored practice of laboriously entering data by hand. The product of the program is a capability to respond to questions related to what specimens we hold, what species of seabirds occur where and when, etc., all with little human effort or expense other than the considerable cost of entering the data originally.

The numericlature for mammals is also essentially complete and data collection and introduction into the computerized system has been started. The first problem to be attacked by these methods is the collection and electronic storage of information about mammal specimens which can then be placed in inactive storage to make room for more actively used collections. Data on specimens being collected in a current study of African mammals and their ectoparasites are also being entered.

(2) Invertebrate Zoology (Crustacea, Cephalopods, and Nematodes).

These projects served as the beginning for much of the planning and development of techniques for data-processing techniques in the Museum. Now, having worked out methods for cataloging of stomatopod collections and their associated data, more than 3,000 lots, representing 20,000 to 25,000 specimens, have been entered in the system. In addition, 25,000 names of other groups of crustacea have been added to the system as the point of departure for adding information on other kinds of crustacea. Standard classifications for including cephalopods and free-living nematodes have also been developed and data entry for these groups has begun.

(3) Mineral Sciences (Petrology).

The Petrographic Reference Collection, containing more than 2,000 specimens (each with description, thin-section, and many with chemical analysis), is an important research resource for scientists outside as well as within the Museum. A procedure has been developed for capturing, storing, and retrieving information on many parameters, including name of the rock, collecting locality, catalog number, location in the Museum collections, and a complete chemical analysis.

In addition, data from oceanic rocks and New England rocks will have been added to the system by June 1969. About 200 records are being added each week.

The geological community will be informed of the availability of this valuable and useable resource for current research by distribution of catalogs of the rocks together with a description of our data retrieval capabilities.

(4) Botany (Type Catalog and Flora North America).

The value to science of the more than 60,000 type specimens are held by the National Herbarium will be greatly enhanced by the program underway. Detailed information from each type is being entered in the data system. More than 2,500 older types have already been entered and this activity continues, together with inserting similar data from types of newly described taxa. The project includes a provision for obtaining the same kind of data from herbaria in other institutions, so that ultimately a national or even international registry of types may be available to researchers.

With support from grants, a major program for the production of a Flora North America (FNA) is underway in the Museum, although the project draws on the full intellectual capacity of the botanical community. At the heart of the FNA is a developing comprehensive information system based on the most advanced computer science and technology. The core of the evolving system is a dynamic, multi-access, taxonomically structured, electronic data bank that will incorporate information from the literature, from existing collections, and from new field and laboratory studies. In addition to providing the capability of producing constantly up-dated floras of the North American Continent, the system will be the basis for conventional local floras, ecological floras, and the means for compiling, correlating, and editing systematic works of many sorts.

A pilot automated bibliography has been developed based on more than a hundred botanical works important to taxonomists working on the FNA. The

system now provides answers to queries such as: What papers have been published on genus "X" between 1940-1960 in which cytological data are an important consideration? Longer questions based on author, date, subject, geography, taxonomic group, and journal can be framed and specific bibliographic references can be retrieved from the data bank. This information system is structured on a list of the genera of vascular plants of North America.

Other accomplishments in this project include the development of computer programs for mechanical construction of taxonomic keys and descriptions, automatic identification of new collections, and editing of copy for the Flora.

Having made the decision to consider data-processing applications as a major priority in FY-70 and beyond, the Museum will continue to install systems evolved from the pilot studies of the last two years. It will also expand these activities, within the available resources, both in terms of taxonomic groups and in diversity of objectives. For example, the highly useful but time-consuming and expensive publication of catalogs of various groups, especially entomological ones, can be accomplished relatively economically by use of EDP techniques and in this form are more easily kept current with research progress.

Plans for a future interconnected, interactive network of data-banks in the systematic research centers of this country and around the world have advanced. For example, last year a magnetic tape containing specimen records for a large number of collections was sent to the British Museum

(Natural History). In just the past few months the American Museum of Natural History has joined the shared-time computer system used by researchers in the Smithsonian's Museum of Natural History. By exchanging information about their respective computer programs, the scientists of the two museums are now able to use each other's data and methodology directly from the system.

The first "giant-steps" have been taken away from potential scientific oblivion toward continuing responsible and responsive participation in the process of scientific inquiry.

Bob Read + return -

SMITHSONIAN INSTITUTION

WASHINGTON, D.C. 20560

March 12, 1969

Dr. David J. Rogers
Professor of Botany
Colorado State University
Fort Collins, Colorado 80521

Dear Dave:

Thank you very much for your kind offer to send us your TAXIR.
I should like not only to have your programs but also I would enjoy
experimenting with them, comparing computer run times (and costs)
between your system and what we are developing here and having some
fun while I am at it.

I have a problem however, in that your system will not fit our
H-1250 and to use it on our CDC 6400 in Boston I have to pay for
the machine time and the telephone just like any other user.

Even tho the actual experimentation once the system is installed
may not be too costly, the installation of it (making it fit in the
teleprocessing environment) will run into hundreds or perhaps even
thousands.

Until such time as I can find someone in the S.I. complex who wishes
to underwrite it, the best I can offer is to accept your programs
and keep them for the possible future use.

Kindest personal regards.

Sincerely,

Nick

N.J. Suszynski, Director
Information Systems Division

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D.C. 20560

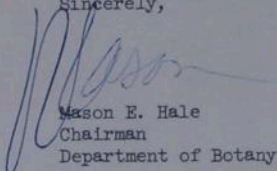
March 13, 1969

Mr. Robert C. Brill
Department of Biology
Taximetrics Laboratory
University of Colorado
Boulder, Colorado 80302

Dear Bob:

I was very pleasantly surprised to receive the packet of printout, as well as your comments on problems and programming. We are now studying in more detail the types of printout that you created and will write a fuller letter to you and Dave a little later.

Sincerely,



Mason E. Hale
Chairman
Department of Botany

28 February 1969

Dr. Mason E. Hale
Department of Botany
Smithsonian Institution
Washington, D.C. 20560

Dear Mason:

We have taken the tape of sample data you sent, extracted a few of the more interesting descriptors, converted them to our input format and built you several sample data banks a la TAXIR. Under separate cover we are returning your original tape and enclosed here you'll find some annotated TAXIR output to look at.

The biggest headache, delay and expense was the conversion of data. Our CDC 6400 was not able to read your Honeywell generated tape. Reason: your tape is organized into 152-character records and our machine can't swallow that big a record. After some experimenting, we finally found a machine that could stomach the tape, a nearby IBM 360. We programmed the 360 to read the tape and write a 2nd tape with twice as many 76-character records. Then we programmed the 6400 to read this 2nd tape, select your item number and 6 descriptors, organize them into TAXIR Define Item statements and write these onto yet a 3rd tape. This 3rd tape served as input to the TAXIR system.

Let it be noted that a great portion of the contents of your tape are, from our point of view, meaningless noise. We realize that these are signals to Reggie's programs which identify your data, but they are unnecessary for TAXIR and our conversion program had a lengthy job separating the wheat from the chaff. You'll note that our Define Item statement demands only the information you wish preserved, separated by commas and terminated by an asterisk. In TAXIR it is the Accession Module statement (in the Accession Module) and the Book Module statement (in the Book Module) which impose the desired structure on the raw data and take the place of Reggie's codes.

In addition you will notice that the conversion was not mechanically perfect. During the trip from Honeywell to IBM to CDC the character "-", for example, became transformed to "V".

All in all, this data conversion part of the operation ran us \$254.66 for computing costs alone (not including 1¢ per page of output printed), not to mention programming and other labor costs, and was a pain in the neck. If we agree to a collaborative effort on the type

Gave Dave just before Christmas

collection, let's work out a more convenient and inexpensive data preparation method. This is the bottleneck. Total computing costs for building your TAXIR banks, querying, printing of books, etc. was \$8.44 (not including 1¢ per page of output printed). None of these costs will be charged to you. I include them to show where the trouble areas are.

In closing let me point out that there are comments scattered throughout the enclosed output for your attention. We would very much like to get your reaction to these results.

Sincerely,

Robert C. Brill

RCB:gm

Enc

3 March 1969

Mr. Nicholas J. Suszynski, Jr.
Information Systems Division
Smithsonian Institution
Washington, D. C. 20560

Dear Nick:

Comes the time when I should arrange with you about the possibility of putting TAXIR into your hands. We would like very much to offer it to you but feel that it is entirely your decision whether you wish to go to the bother of getting it established there.

You are certainly aware that we received a tape before Christmas with some of Mason Hale's type materials on it. We have spent some time in conversing that tape to the format for TAXIR and have just sent the tape back as well as a selected sample of the information on it, processed by TAXIR. I am enclosing a copy of the letter of explanation to Mason for your benefit. Perhaps you would be interested to see the output.

Since it will probably take some doing to put the programs onto the operation of the 6400 in Cambridge with your tie-line to them, I wonder what you would find it ~~most~~ feasible to do. I think that you would need the advice and direction from us to make the system operable. We would probably have to work with Dan Piacesi or Reggie Creighton for some few days to get things all in hand.

Before we procede any further, I would like to hear what your reaction and desires are.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D. C. 20560

December 18, 1968

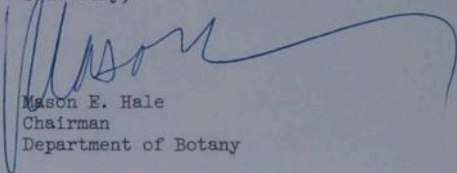
Dr. David Rogers
Department of Biology
University of Colorado
Boulder, Colorado 80302

Dear Dave:

We enjoyed your visit here very much. It is not often that we can talk shop with a visiting botanist on data processing in herbaria. Much as we may have been skeptical of your efforts in computerizing the bryophyte collection, we realize that you are embarking on a very large-scale and perhaps controversial problem. When you have built up a data bank of, let's say, 10,000 entries and try to program this for retrieval of information on geography and other descriptors, we will be very much interested in seeing how serious the problem of missing data will be. In our type project these problems are not quite as serious because we don't expect to retrieve, for example, on county or perhaps even state very often. As I said, we hope to have a more-or-less complete description of our project and initial sending of cards within a month or two. You will be the first to receive them.

Thanks again for a very stimulating visit.

Sincerely,



Mason E. Hale
Chairman
Department of Botany

Squires, Don F

Home - 942-0905

Smithsonian 628-1810

office

144-5952

(May 67)

Richard Cowan - X 5954

December 31, 1968

Dr. Mason E. Hale, Chairman
Department of Botany
Smithsonian Institution
United States National Museum
Washington, D.C. 20560

Dear Mason:

I have two good letters from you to answer, and had already started to write when your second arrived--so, I'm going to go ahead and comment on the second one at the end of the first. If this letter sounds disjointed--it is!

I, too, enjoyed my visit there. It was pleasant, among other things, to attend the seminar where I heard some talk about botanical problems (both Stone and Stern), away from the jargon of computers. But most of all, I enjoyed our visit where, at the interface (more jargon) with information retrieval systems, we talked about the problems of why we want to put such and such information into the computer's memory. Your questions about the need for putting in, say 30,000 bryophyte records, is very valid, and requires me (and all the rest of us) to think very hard about these needs. I think that we all believe that the label information on the specimens is valuable, and now we've got to say why it is important to go to the bother of extracting it onto some sort of machine-readable format. I have no doubt that we can justify this type of effort, because the information residing on the specimens is not readily available.

As of this moment, the label data, which describes much of the important knowledge about locality, geography, ecology, form habit, flowering time (or in the case of cryptogams, reproductive times), simply cannot be extracted without tremendous effort, by searches which are extremely inefficient. Using a well-designed computerized system, all the various pieces of information become available, in any format desired for any particular use.

You were worried about missing information. In our experience, this is not nearly so formidable a problem as it may seem before one has actually gone into a large data bank. What one specimen may lack, another, closely related specimen, has. By pooling the information from all the specimens, one quickly overcomes the lacks of individual specimens. I believe, after a moment's reflection, you can see that that is what we do now, without a computer. But also consider that by this process of lifting the label data from the specimen, we gain knowledge about

Dr. Mason E. Hale
December 31, 1968
Page 2

what more information needs to be gathered, thus precisely pointing to new research needs. Add to this benefit the saving of curatorial and research time which an IR system provides, and we gain more of the valuable knowledge of the botanist to do that which he should bend his talents towards anyway.

I was much impressed by the way you have decided to handle the exchange of type-specimen information with other institutions. This should be a tremendous aid to us all. I only hope that the cooperating institutions will come through with their part of the effort. I'm sure you won't get 100% cooperation, but I hope there will be sufficiently large response to get good returns. There's nothing like a working computerized system to convince people that they should get out of the 19th century.

On our offer to come and help with wiring up our TAXIR system, forget any worry about funding our trip. Chances are we'll be in Washington for some reason later this winter, and we'll just charge our expenses to some other source. But we do want you to have the TAXIR programs, and want to save any headaches the ISD might have.

We're working on your tape of types now and hope to have something useful as soon as we can transfer from your format to our define-item format--shouldn't be too long.

Thanks again for our conversations. I think the sherry was extremely beneficial, something we all need to make the wheels turn better.

Happy New Year,

David J. Rogers
Professor of Biology
Project Director

December 23, 1968

Mr. F. Thomas Galloway
Office of Science Information Service
National Science Foundation
Washington, D.C. 20550

Dear Tom:

It was a pleasure to visit you and Harold Olsen recently. It is certainly valuable for me to get the "big picture" of problems, and your conversations with me certainly helped to get rid of tunnel vision, which I fear becomes a problem with any of us as investigators.

With respect to one project which we discussed, the Flora of North America, I hope that I was able to shed some light on things. After talking to you, I spent some time with Stan Shetler and believe that he is coming closer to a good and useful working arrangement with some sort of data processing manager. He, like so many other botanists, doesn't have a good grasp of computing milieus but I think he is an honest broker, and with help from some of the rest of us, will do a good job. The project as a whole, both from practical and scientific standpoints, is well worth the effort and since we have no other person willing to shoulder the responsibility he has, we must make him go, and he can, I sincerely believe.

Best regards to Harold Olsen, and best wishes for a Happy New Year.

Sincerely,

David J. Rogers
Professor of Biology

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D. C. 20560

December 24, 1968

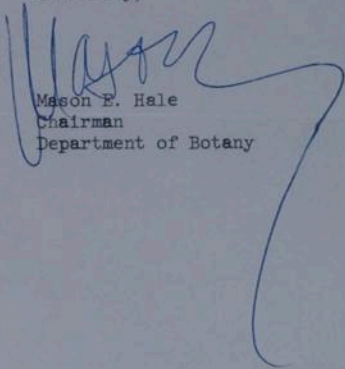
Dr. David Rogers
Department of Biology
University of Colorado
Boulder, Colorado 80302

Dear Dave:

I assume that you picked up a tape of our data bank here and now are adapting it for experimental use in your own system. Please give us any pertinent results that might have some effect on our own program here. We certainly do not have all the answers, and at this stage we try to hide behind that expression, "pilot project." If things later cave in, we can always use this out.

One of the people in ISD said that you had offered to come here and implement the taxir system. I think the only problem in accepting this offer was the lack of any travel money here to bring people in from the outside. As I have told you before, we in the biological end of these data processing programs are quite ignorant of problems in programming and other aspects of running the machines. We have little choice but to leave this to the people across the Mall.

Sincerely,



Mason E. Hale
Chairman
Department of Botany

DEC 20 1968

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D.C. 20560
16 December 1968

AIR MAIL

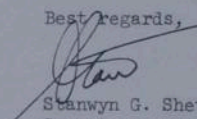
Dr. David J. Rogers
Department of Biology
University of Colorado
Boulder, Colorado 80302

Dear Dave:

It was good to see you last week and to cross swords again over the problems of data processing as they apply to museums and floras. We must keep this lively debate going until solutions to the major problems have revealed themselves to us a lot more clearly than at present.

I am interested in knowing how the family lists that I gave you prove useful, if at all, and about any mistakes you turn up. Keep in mind that they are in effect rough drafts that we have gone through on the way to preparing the preliminary generic list for Flora North America. Your suggestions on how such compilations can be put to further use by appropriate enhancement is welcomed.

Best regards,



Stanwyn G. Shetler
Secretary, FNA Project

7 January 1969

Stamryn G. Shetler
Smithsonian Institution
United States National Museum
Washington, D. C. 20560

Dear Stan:

I had intended to reply earlier but got caught with the "bug" and got badly behind in all things. I did enjoy our exchanges though if you think we crossed swords there, you should someday get involved in disagreements that occur in our own shop. We seldom draw blood, but get plenty close at times.

I hope I have been an honest broker for FNA in various descriptions around the country. There exist a number of pockets of discontent who misunderstand the intention of FNA and wherever I can I try to boost the projects score. When arguments arise over personnel or ideas, with respect to it, I simply say "Have you got anything better," which most frequently cools them off.

The family lists are useful to us in terms of checking out the ones that we are using for a pilot study on chromosome number reporting for phanerogams. We have, of course, to worry about families not included in your list and all those families whose reports do not come from the U.S. The tropical families clearly give us the devil Henry Fleming is in charge of these lists and I have asked him to give a more detailed analysis of the family names than I give you here. I think that when we have finished our investigation a list of family names in the various systems will have been compiled and this will be a part of a data bank in TAXID and control vocabulary of the family names according to the different systems will be an automatic byproduct. Since the different family names in the different systems exist in the data bank we shall be able to manipulate these lists in a number of useful ways.

You might give Mason the word that we are at work on the tape of types which he was kind enough to give us. The damned computing machinery is not very smart here so we have had to get a Honeywell computer in Denver to read the tapes. That should be done this week. Once done, we should then be able to play around with the data.

I expect to give a seminar at NSF on 7 February which will probably be no news to you. The seminar will say things about TAXIR in a very superficial manner which you know already but I hope that the NSF is sending an invitation to you.

I hopetto see you then.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm

25 October 1988

Mr. Dante Piacesi
Information Systems Division
U. S. National Museum
Washington, D. C. 20560

Dear Dant

Just a short note to say the obvious: (1) sorry I'm late in sending you the similarity-clustering-package, but it did get off in this morning's mail; (2) as per usual, we haven't had the time and/or funds to invest in an Optimal Updating Machine (i.e., another Programmer-trainee) so the flowcharts you're getting do not reflect a complete correlation with the coding. However, since the coding has been liberally endowed with "comments" the program flow ought to be readily discernable, our lack of updating notwithstanding.

But enough - at least for a short note. Hope things are going well with you. Happy OCTOBER 31 and/or DECEMBER 25 depending on what base you read this letter in.

John Wong, Jr.

JW:gm

November 4, 1968

Dr. Mason E. Hale, Chairman
Department of Botany
Smithsonian Institution
Washington, D. C. 20560

Dear Mason:

I think what you need now is something I could use very much myself - about six fingers of bourbon. I sense from your letter some frustration about the whole ADP operation, where it looks so simple at the outset, only to turn into a bucket of worms further along in the operations. If I have sensed your letter correctly, you are up against the data structuring and preparation problems, one of the most frustrating parts of all the work. There are so many things to get in gear at once that it would try the patience of a Chinese philosopher. Not the least of the difficulty is that which you're experiencing now - the transfer to some machine-readable format from the presently existing data files, specimens, or what-have-you! If it makes you feel any better to know that we're in the same boat, then let it be known that we are up against these same types of problem. If you're having to console a bunch of skeptics at the same time you're trying to make some headway, this only adds to the frustrations. Some of your people are probably just waiting to say "I told you so, now go back and pick up your quill-point pen."

I certainly understand when you want to get a full-time botanist to handle the data-preparation work. Somebody knowledgeable about the botanical significance of this or that item of information, or how to interpret some particular twist to the wording on specimens or cards is very important. But there may be some help for you on the way in the procedures we're developing now without the help of a full-time botanist. Gil Herish is developing exhaustive flow-charts of the data gathering system, instructions which can be read and followed by anyone capable of SCM operations, with built-in checks or stops for points of difficulty. Such a system allows the operator to proceed with easily read information, but stops him at the point of some problem. The problems are analysed for the type of difficulty, and a flow of the system directs the problem to someone capable of resolving it. Sometimes these difficulties are systems-oriented types, and should be answered by a person familiar with the system, and sometimes the problems are curator-oriented, and probably have several levels of complexity, some of which require only a check by another technician, some requiring attention from you or others.

The point of such a technique is to reduce the burden on any one person, such as yourself, and to allow a continuous flow of work rather than grinding to a halt at some stage when you are not available. Of course, some of this flow of work is based on the TAXIR system which has, as one of its modules, the "Data Vet" which is built in to help with the necessary checking of the data after it has been made machine-readable. You don't have to have the TAXIR system to get such a data checking process, but it sure is helpful.

I expect to be in Washington in December, and I hope I can look over your work, and maybe even have a helpful suggestion or two. Do you plan any extensive trips away from home during that month? Let me know, if you do plan to be away, and I'll try to arrange a time when we can get together. My schedule for coming to Washington is rather flexible.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D. C. 20560

October 28, 1968

Dr. David Rogers
Department of Biology
University of Colorado
Boulder, Colorado 80302

Dear Dave:

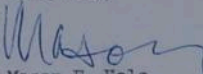
My apologies for being so late in sending thanks for the two-day ADP session. It was impressive and, while strongest on the research aspects, leaves me with many new ideas and problems. As I said, we shouldn't talk of competing systems. There are too many problems to solve in too many places.

I've spent the last month monitoring--singlehanded it seems anyway-- three people on the type project. The magnitude of the job leaves me discouraged. However, we do have an excellent SCM paper punch operator, who may well pull us through. The cost is much higher, I feel, than anything you face now. What we need here is a full-time botanist to oversee the operation (and that shouldn't be me!)

In any event, by the end of December when the dust settles, we will have some cost figures and reasonable projections on whether to continue and how fast.

As far as Stan's all-embracing project is concerned, I have little, if anything, to do with it. He is not meshing well with Gil Daniels, who likes to assume the role of senior statesman and consultant in botanical ADP. They were all here this week to lay grounds for computerizing 15,000 generic names à la BPH, a futile exercise, in my opinion.

Sincerely,


Mason E. Hale
Chairman
Department of Botany

Seq. No. 2
Family Amblystegiaceae
Name Amblystegium brachyphyllum
Author Card. Ther.
Citation Minn. Bot. Stud. 3:123. 1903.
Collector Holzinger, J. M. Coll. No. --
Collected 15 Jul 1901
Locality USA: Yellow Medicine Co.: Granite Falls:
Remarks Equals Leptodictyum riparium var. brachyphyllum
Cat. No. US 57396 Isotype

Seq. No. 1

Family Amblystegiaceae

Name Amblystegium boreale

Author Dixon

Citation Bryol. 36:5. 1933.

Collector Eckstein, H. C.

Coll. No. 40

Collected -- Aug 1881

Locality Spitsbergen: : Green Harbour:

Remarks

Cat. No. US 57397 Holotype

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D. C. 20560

May 31, 1968

Dr. David J. Rogers
University of Colorado
Boulder, Colorado 80302

Dear Dave:

Let me first apologize for not acknowledging the letter you so kindly sent in response to my first statement of our computerized type project. I appreciated your comments very much and expression of interest. Inevitable delays in programming and my own too busy schedule have prevented implementation as fast as we had wished.

We now have a program tape for recording data on 5 X 8 cards, as shown in the enclosed model. Actual typing from our own files has begun, and over the summer a temporarily employed student will make a major effort to verify and prepare type cards.

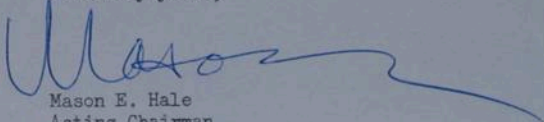
The first printout of selected types should begin by late fall, at which time a copy will be sent to you. Programming the computer is the main obstacle. This is a laborious, expensive, and time-consuming project, but once achieved, data can be processed in a matter of hours.

Responses from some 25 institutions indicated that very few botanists can prepare types directly but many would be able to fill in data sheets. A model data sheet has been prepared and will be distributed this fall when bugs are worked out of our system.

There was some criticism of how much data would be required. Actually, the amount suggested (down to longitude and latitude) was the absolute maximum. In many cases only country of origin will be known, and that is all that need be entered. In no case did I anticipate a taxonomist actually looking up coordinates for a specific locality.

Thanks again for your interest and patience as we attempt here to get a very large and difficult project rolling. I will try to keep you informed of our progress at more frequent intervals.

Sincerely yours,



Mason E. Hale
Acting Chairman
Department of Botany

Enclosure

MAR 1 1968

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D.C. 20560
Department of Botany
27 February 1968

AIR MAIL

Dr. David J. Rogers
Department of Biology
University of Colorado
Boulder, Colorado 80302

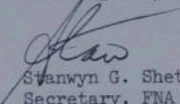
Dear Dave:

Thanks for the carbons of your letters to Hale and Chambers. It was nice to see you again and to have a chance to continue discussions about potential interaction. We will look forward to seeing Bob and George when they come.

Your letter to Chambers is appreciated. This kind of support makes life easier for both of us, I believe.

Hope you've recovered from the unmitigated torment of "Bonnie and Clyde." Now I notice that the movie is everywhere at neighborhood theaters for \$1.25--half the price! Anyway, thanks so much for your contribution to our evening. I feel a little sheepish everytime I think of your investment and hope you will let me return the favor in Boulder sometime.

Best regards,


Stanwyn G. Shetler
Secretary, FNA

*P.S. By the way, Elaine certainly did bare
her microbiology down cold when I
got back!*

22 February 1968

Dr. Mason Hale
Department of Botany
Smithsonian Institution
Washington, D. C. 20560

Dear Mason:

Thank you for your hospitality and very fine dinner Monday.

We hope to have Bob Brill and perhaps George Estabrook as well to spend time with you on March 5, 6, and 7. Details of their arrival, etc. will be forthcoming. I hope that you will find their work useful and that we can make a start on the data in the type collection.

Somehow or other I did not get the Xerox copies of the programming techniques which Ken Ebbs had planned to give me. Could you ask him to send them on as soon as possible so that Brill and Estabrook can get a feel for "what's cooking"?

I am sending a copy of this to Stan.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gn
Cc: Dr. Stanwyn Shetler

12 February 1968

Mr. N. Kenneth Ebbs, Jr.
Office of the Deputy Director
Museum of Natural History
Smithsonian Institution
Washington, D. C. 20560

Dear Ken:

Thanks for your cordial letter of the 6th. I have a call in right now to Stan Shetler concerning a proposed visit to you folks on the 19th. Ask Stan About details.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm

R. Creighton
January 4, 1968

General Considerations-Catalog of Types

The file will be maintained on magnetic tape in sequence by binomial. One and only one record will exist for each binomial. Input of a record whose binomial record is already established in the system will cause posting of the following information from the input record to the established record: Institution acronym, catalog or reference number, kind of type. The initial system will accommodate up to ten such "type pointers".

A 5 x 8 index card, called "CATALOG OF TYPES-DATA COLLECTION RECORD", will be the source document for the system. This card will be so designed as to facilitate either keypunching ^{and} paper tape input of the data contained thereon.

Periodic printouts of the data in six different sequences is desired. These sequences are as follows:

1. By Generic-Specific Name. —
2. By Family; within a family by genus; within a genus by species. —
3. By Author's last name. —
4. By "where collected" —

Country;

within country by state;

within state by county or specific reference.

5. By Collectors last name.

6. By Institution acronym.

Four processing programs plus two sort programs will be required in addition to the utility card to tape program to accommodate the card input aspects of the system. An estimated four man months of programming effort will be required after all forms design, system's and procedural documentation have been accomplished.

Implementation of paper tape inputs would require approximately two additional man months of effort.

The following sequence of computer programs appears appropriate:

1. Card to tape.
2. Sort by record number to tie lines together.
3. Validate and format inputs into transaction records.
4. Update established records from input transactions.
5. Permute records according to the six sequence categories.
6. Sort by permutation key.
7. Print six-way listing.

CATALOG OF TYPES
DATA COLLECTION RECORD

RECORD NO.
(15-80)

		1	2
FAMILY NAME (2)			010 (3-24) 27 letters
GENERIC NAME (1)			020 (26-47) 27 letters
SPECIFIC & SUBSPECIFIC NAME (1)			030 (49-70) 22 letters
AUTHOR (LAST NAME, SURNAME) (3)		B	040 (3-24) 27 letters
LIT. CITATION (PUB, VOL, PG, DATE, ILLUST.)			(26-47) 22+
LIT. CITATION CONTINUED			050 (49-70) 22+
WHERE COLLECTED (COUNTRY-STATE,)		C	060 (3-17) 15 Country 070 (18-24) 7 state
(STATE CONT'D, COUNTY OR SPECIFIC)			(26-33) 8 state 080 (34-47) 14 Country or open
(COUNTY CONT'D, LATITUDE-LONGITUDE)		X	(49-59) 10 County or open 090 (100-70) 10 D D M M S S D D M M Y Y 179° 59' 59" letters
COLLECTORS (LAST NAME, SURNAME) (5)		D	100 (3-24) 27 Last name?
(LAST NAME, SURNAME)			110 (26-47) 27 surname? 2 sec.
COLLECTION NUMBER/DATE COLLECTED			120 (49-56) 9 not enough? (57-61) 7 10 130 (62-70) 9 D D M M M Y Y Y Y day month year
INST. / CATALOG OR / KIND OF ACRONYM / REFERENCE NO. / TYPE (6)		E	140 (3-7) 5 Inst. 150 (8-16) 9 not enough. 160 (17-24) 8 type
INST. / CATALOG OR / KIND OF ACRONYM / REFERENCE NO. / TYPE			170 (26-30) 5 Inst. 180 (31-39) 9 color ref. 190 (40-47) 8 type
INST. / CATALOG OR / KIND OF ACRONYM / REFERENCE NO. / TYPE			200 (49-53) 5 Inst. 210 (54-62) 9 color ref. 220 (63-70) 8 type
REMARKS OR NOTES		F	230 (3-24) 22 one set of comm.
REMARKS (CONT'D)			240 (26-47) 22 sec. " " "
REMARKS (CONT'D)			250 (49-70) 22 third " " "
REMARKS (CONT'D)			260 (3-24) 22 fourth " " "

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D.C. 20560

JAN 31 1968

24 January 1968

Dr. David J. Rogers
Taximetries Laboratory
Department of Biology
Armory 101
University of Colorado
Boulder, Colorado 80302

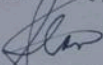
Dear Dave:

This is a belated, joint thank you to you and Bob Brill for sending me copies of your letters to Don Squires. I am very excited to think that our project here on computerizing a type catalogue might provide the kind of data bank that would support the research on your new IR system. Mason has begun to roll on this project, and you should receive a letter from him soon concerning developments thus far. If he has not already communicated with you on this matter, I feel confident in saying that he will be most eager to cooperate in every way in supplying data from our data bank. As far as the Flora North America Project is concerned, we see nothing but gain from cooperating with you in the development of programs for information storage and retrieval.

Mr. Shivanard R. Patil of your Department is applying for a Smithsonian Postdoctoral Research Associateship to work with me next year. Confidentially, can you give me an evaluation of his training and abilities?

It was good to see you and Bob again in Mexico City. Let us keep in touch.

Sincerely,



Stanwyn G. Shetler
Secretary, FNA

cc: Brill
Hale
Squires

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D. C. 20560

JAN 25 1968

January 23, 1968

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

Dear David:

The Department of Botany, Museum of Natural History of the Smithsonian Institution, is proposing to begin a computerized printout catalogue of type specimens and associated literature citations. The primary objective will be to store and retrieve information on the location of type specimens, similar to a computerized union catalogue, which would be made available to systematic botanists engaged in monographic or revisionary studies. Easy rapid access to this kind of information will materially simplify the work of monographers by reducing time now needed to track down the location of types and by distributing more equitably the problems of preserving and loaning valuable type specimens. Each institution would also be in a better position to verify the authenticity of their presumptive types.

Input for this system will begin with the type collection and catalogues maintained in the United States National Herbarium, now numbering about 60,000 entries. Printout of these data as accumulated will be sent to cooperating institutions so that they can ascertain the presence or absence of isotypes or comparable material in their own collections. The input from these institutions would be added to the tapes. Finally, data on types not in the Smithsonian but housed at other institutions would be assembled in a standard format and added to the data bank.

A statement of our general program and a sample data collection sheet (for a 3-card train or SCM paper tape) are enclosed. After you have reviewed these, would you consider the following questions?

1. Do you now maintain a separate type collection or otherwise identify type material in the herbarium? Can you give a rough estimate of the number of these?
2. Do you have a card file with pertinent identifying information on each type (e.g., literature citation, collector, number)?
3. Do you have the capability of preparing paper tapes or punched cards of your holdings (that is, in addition to those not already in the U.S. National Herbarium)? Or would you be willing to fill out a standardized data collection form for your types and send them to a data automation center?

The usefulness of a printout catalogue will increase with each new entry. I can envision updated printouts of holdings of each institution, as data are added, as well as printouts for individual genera on request. I would like

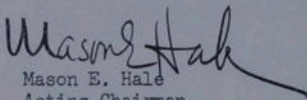
to emphasize at this early stage that we will not make taxonomic judgments, although there will be provision for pertinent remarks. The catalogue will include basically only the information necessary to identify a type, regardless of whether the name is presently accepted, recognized under a different genus, or relegated to synonymy.

At the present time we are programming an SCM Typetronic 2816 for paper tape input, later to be put on magnetic tapes, and hope to hire an SCM operator with the expectation of beginning punching within six weeks. The completion of the US material will take a year or more, depending on the amount and availability of funding. The progress of the whole project and its usefulness to the botanical community will determine how far it can be expanded.

The Flora North America Project, now based in the Department of Botany, is cooperating in many ways to achieve our goals, from making available a Typetronic machine to sharing of computer programs and related services. A printout catalogue of American types will obviously be of enormous value to both the authors and editors of the Project.

This letter is being sent to a limited number of institutions in order to test initial reactions. I would appreciate your comments and suggestions on how it could be modified or improved and how much you can contribute.

Sincerely yours,


Mason E. Hale
Acting Chairman
Department of Botany

Enclosure

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D. C. 20560

FEB 12 1968

February 9, 1968

AIRMAIL

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

Dear Dave:

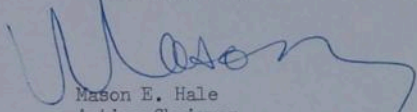
Don Squires just sent me a copy of the letter he sent to you. I feel it is realistic but a little on the pessimistic side. As far as Botany is concerned, the data will accumulate as fast as practical, even if I have to do it myself.

Don also seems unclear about a type data center. Canada would hardly be an appropriate repository, firstly because they are much too small, secondly because one type center per continent will be enough. There simply wouldn't be any exchange of data in the sense Don envisions for crustacea holdings.

We do have a problem here, as at any place, in funding, but I'm convinced that we can go a long way with resources now at hand.

Let me know when your trip is planned.

Sincerely yours,



Mason E. Hale
Acting Chairman
Department of Botany

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D.C. 20560

February 7, 1968

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

Dear Dave:

I am two letters behind in my correspondence with you; this has been caused by my peregrinations hither and yon. I wrote quickly to the draft board regarding George's imminent induction and, characteristically, have never received any response. Please keep me up-to-date on what happens there, for this could be quite serious to you. The creation of a data bank for your studies is a very serious problem. In designing our own proposal, I felt that the inclusion of four full-time persons for source data automation was bit extreme, but Nick, in his greater experience, persuaded me that we must go for that number, and I am indeed pleased now that he succeeded in talking me into it. We find that we are processing between 100 and 150 specimen ^{records} a week from each of our source data automation units. We are very hopeful that this number will increase as procedures become more standardized, but we do not anticipate the rate of input becoming greater than 200 specimen records per week. As you can see, development of 60,000 specimen records is a two-year project on that basis, if everything goes well.

Development of a comparable data bank in botany would, I would judge, be approximately five or six years away unless considerable additional manpower becomes available to the Department of Botany. Such additional support for this project is not apt to develop within our present budgetary framework, and I cannot reprogram manpower from my own project. I have been able to put Ken Ebbs onto the job of organizing the program tapes for Mason Hale's project, and we hope very much that they will begin to develop a data bank within the next month. Mason's approach to this is a very broad one, particularly in view of the isotype problem, and he will involve herbaria from many parts of the country in the project. However, the rate of inclusion of the materials will not be appreciably faster than the machine operator for at least the first six months of the effort. I agree with Number 2 of your January 5 letter. Undoubtedly, because of the isotype system there is greater probability of botanical type records having a common central storage bank. We are very hopeful that the British Museum will join us in this effort and that together with the National Museum of Canada we can create a three-way repository for information about botanical types.

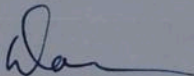
My urging of the Ortner material upon you was based upon the fact that it is complex in character, and, even more important, it is relatively small and could, conceivably, be processed in either card or tape form within a relatively short time. (I notice now that the January 5 letter is from Bob Brill, so I will apologize to Bob for not writing directly to him and ask that you share this letter with him.) Although you anticipate being operational by June, I doubt very much that even the Ortner collection could be put into machine-readable form in that time span. I might note that Stan Shetler seems to be primarily concerned about North American taxa, while Mason Hale is taking the broader viewpoint of the U.S. National Herbarium as a whole. I have encouraged Mason to consider the types as only part of a larger project and, when we design formats and so forth for types, that these be equally applicable to other materials in herbaria which may again eventually be automated.

So, in summary, the problem is not accessibility to the data, but in getting specimen data into a machine-readable form, whether it be on cards or on tapes. We are proceeding as rapidly as possible in as many areas as we can to accumulate material in machine-readable form, but I think any hope of getting 60,000 records together in a single discipline in the next two years is very, very unlikely without external supplementary funds. It might be possible for you, acting as an independent organization, to develop through the National Science Foundation a project to put two technicians and a single typetronic unit into the Herbarium, thus expediting the inclusion of data. If this seems like a reasonable solution please let me know, and we can talk further about it.

I haven't heard anything about the IFIP meeting in Edinburgh next August and so can't tell you much about it. If you find out anything please let me know.

With best regards to you, George, and Bob.

Sincerely,



Donald F. Squires
Deputy Director
Museum of Natural History

9 January 1968

Dr. Donald Squires
Natural History Museum
Smithsonian Institution
Washington, D. C. 20560

Dear Don:

This letter sort of backs up the one from Bob Brill. We are indeed excited to hear that the data concerning the types of your botanical collection are being actively looked to for their information retrieval aspects. As Bob suggested we would be very pleased if a copy of the data bank could be made available to us as the tapes are produced. We would very much like to try out our new IR system with your data bank on the types.

We will be happy to hear if it is feasible to let us have copies of whatever tapes or other input you have designed for this work. If there is any work necessary on our part let me know and one of us will get there to do it. It was good to see you in Mexico.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm
Copy: Stanwyn Stetler

P.S. Have you heard about an IR meeting sponsored by "IFIP" in Edinburgh next August? If so, tell me ☐ If it is worth going to see or to participate.

4 January 1965

Dr. Donald Squires, Deputy Director
Museum of Natural History
Smithsonian Institution
Washington, D. C. 20560

Dear Don,

When during the final discussion at the Mexico City symposium the topic was raised of the Smithsonian's type collection of plants, it occurred to me (and I immediately mentioned it to you) that the curatorial information covering this collection would make an excellent data bank for the TAXIR system. Several considerations struck me at once as pertinent.

1) Earlier we had discussed Ortnier's "morgue" as a candidate data bank for our system. I mentioned that with our compressed characteristic function technique these data would very likely fit entirely in core storage. You rightly pointed out that while the resultant retrieval speed would be impressive, a larger data bank that would force us out onto tape or disk would be a more valid test of our system. The 60,000 specimens in the type collection (as compared to Ortnier's 8,000 specimens) are of this more desirable magnitude.

2) Judging by the responses of Drs. Thomas and Gomez Pompe it seems likely that the size of the collection will increase from contributions by other museums and that other type collections, although not added physically to the Smithsonian collection, would be added to the same IR system, so that workers could ultimately interrogate a single system for finding types wherever they may be located. All of this would be on top of the normal sporadic contributions made by collectors of new species. The collection then is, and when automated will be even more so, in a state of expansion. This property of the collection should enable us to test the ability of our system to update the search files. While not underrating the importance of the Ortnier ^{material} or the frequency with which it may be consulted, is it not true that it is quite static?

3) As I think I may have told you, the XPOP software we plan to use is not ready for our machine as we had expected. We are going ahead with a preliminary working version of the system coded in FORTRAN. While this first version will not have the design flexibility that XPOP will later provide, it will incorporate our compressed characteristic function technique, will retrieve efficiently, and what is more to the point, will be ready for use sooner than our original plans anticipated. Our target date for an operating

system is June of this year. We must check the details with Nick Suszynski but I believe that after checkout on the CDC 6400 in Boulder we could install the system fairly easily on the CDC 6400 at Cambridge, Mass., to which you have a tie-line, thus putting control of the data bank in your shop. I have no idea how Stan Shetler feels about this, perhaps he has already formulated other plans for the type collection, but it is perhaps well to consider that the TAXIR system probably represents the fastest method of automating this important collection.

One technical problem does exist. We will have to gain access to equipment which can convert the data from your SCM paper tapes to IBM punch cards and then we will have to write a ^{small} program to make the data conform to our input format. How serious this data conversion will be can only be determined by members of our group talking to Shetler and Suszynski.

If the above notions gain acceptance with you and your staff, please let us know and we'll make arrangements with you to set up collaboration. Incidentally I felt that the Mexico City symposium was quite productive. Despite some wide differences in viewpoint and a serious problem of communication, I think the conference will prove to be of great importance for the future of taxonomy.

Sincerely,

Robert C. Brill

RCB:gm
CC: Nicholas Suszynski
STANWYN SHETLER

4 January 1965

Dr. Donald Squires, Deputy Director
Museum of Natural History
Smithsonian Institution
Washington, D. C. 20560

Dear Don,

When during the final discussion at the Mexico City symposium the topic was raised of the Smithsonian's type collection of plants, it occurred to me (and I immediately mentioned it to you) that the curatorial information covering this collection would make an excellent data bank for the TAXIR system. Several considerations struck me at once as pertinent.

1) Earlier we had discussed Ortner's "herbarium" as a candidate data bank for our system. I mentioned that with our compressed characteristic function technique these data would very likely fit entirely in core storage. You rightly pointed out that while the resultant retrieval speed would be impressive, a larger data bank that would force us out onto tape or disk would be a more valid test of our system. The 60,000 specimens in the type collection (as compared to Ortner's 3,000 specimens) are of this more desirable magnitude.

2) Judging by the responses of Drs. Thomas and Gomez-Pompa it seems likely that the size of the collection will increase from contributions by other museums and that other type collections, although not added physically to the Smithsonian collection, would be added to the same IR system, so that workers could ultimately interrogate a single system for finding types wherever they may be located. All of this would be on top of the normal sporadic contributions made by collectors of new species. The collection then is, and when automated will be even more so, in a state of expansion. This property of the collection should enable us to test the ability of our system to update the search files. While not underestimating the importance of the ^{curatorial} data or the frequency with which it may be consulted, is it not true that it is quite static?

3) As I think I may have told you, the XPOP software we plan to use is not ready for our machine as we had expected. We are going ahead with a preliminary working version of the system coded in FORTRAN. While this first version will not have the design flexibility that XPOP will later provide, it will incorporate our compressed characteristic function technique, will retrieve efficiently, and what is more to the point, will be ready for use sooner than our original plans anticipated. Our target date for an operating

system is June of this year. We must check the details with Nick Suszynski but I believe that after checkout on the CDC 6400 in Boulder we could install the system fairly easily on the CDC 6400 at Cambridge, Mass., to which you have a tie-line, thus putting control of the data bank in your shop. I have no idea how Stan Shetler feels about this, perhaps he has already formulated other plans for the type collection, but it is perhaps well to consider that the TAXIR system probably represents the fastest method of automating this important collection.

One technical problem does exist. We will have to gain access to equipment which can convert the data from your SCM paper tapes to IBM punch cards and then we will have to write a ^{small} program to make the data conform to our input format. How serious this data conversion will be can only be determined by members of our group talking to Shetler and Suszynski.

If the above notions gain acceptance with you and your staff, please let us know and we'll make arrangements with you to set up collaboration. Incidentally I felt that the Mexico City symposium was quite productive. Despite some wide differences in viewpoint and a serious problem of communication, I think the conference will prove to be of great importance for the future of taxonomy.

Sincerely,

Robert C. Brill

RCB:gm

CC: Nicholas Suszynski

STANWYN SHETLER

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D.C. 20560

FEB 9 1968

c/o Office of the Deputy Director
Museum of Natural History

6 February 1968

Dr. David J. Rogers
Taxometrics Laboratory
University of Colorado
Boulder, Colorado

Dear Dr. Rogers:

I would like to thank you very much for sending the package of reprints on the development of the cluster analysis techniques. Both the methods of character representation and of cluster analysis are fascinating, and should have wide appeal among any classical taxonomists who have had some working experience with computers. During two partial weekends, during which I had trouble putting the papers down to do anything else, I did come up with a number of questions on I/O considerations and possible concept extensions. However, I have gathered in talking with Mason Hale that you will be visiting here soon, so will hope to bring up these questions in person.

I will look forward to seeing you during your visit, and hope you will have time available for some discussion.

Sincerely,

N. Kenneth Ebbs, Jr.
N. Kenneth Ebbs, Jr.

11 October 1967

Dr. Donald F. Squires
Museum of Natural History
Smithsonian Institution
Washington, D. C. 20560

Dear Don,

Herewith is a start of a summary of our impressions after our visit with you and the various staff members. The impressions overall are that you have a good group of people and we were pleased to learn from them as much as we did.

We have not named the individuals concerned feeling that this would not necessarily improve what we say and you know about them. This paper is strictly for your perusal and will possibly be out of date tomorrow. I hope however that it serves some useful purpose for you.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm

General Problems

In this discussion, a "data bank" consists of specimens to be curated or on which studies are to be made. The size of the data bank plays a role in the considerations given to IR problems by the various curators. Either the size overwhelms the curators or causes them not to feel any real need for a system different from what has been used for the last 150 years. The entomologists fall in the first category, and the reptile people are in the latter bunch.

Sophistication of knowledge about computers does not necessarily play a role in whether a curator wants machine-oriented IR systems or not, particularly when talking about the job of keeping records for the data bank. Peters, the herpetologist, says that an IR system for curating his snakes is a waste of good time and money; but we must also hasten to add that Peters is one of the most ardent supporters of computer use when working on the process of classification. He essentially wants a "scratch-pad" IR system when he is working up his classification, but no system for telling you how many accessions he has made this year, what the number of loans were, the number of bottles refilled with alcohol or other such types of information. The opposite approach is taken by another crowd - the bird people - on curatorial IR systems. They want one; but they also want the scratch-pad type as well. That's a good group! I don't know what motivates the differences between the herpetologist and the ornithologists, since both of them have relatively (to the entomologist) small data banks.

Your largest data bank, and the one for which the least sophistication exists, is your entomological group. Foote (USDA man) seems the most vocal spokesman for an IR system. I do not recall the names of the other chaps, but most had just about heard of a punched card system, and had not gone any further with it.

There was not too much that Fleming, Brill, and Estabrook could say about the Oceanographic Sorting Center. Theirs is an inventory problem, and any system that satisfies a museum curator will probably satisfy them. I don't know whether we generated any ideas from talking to them, or whether we produced anything useful to them. Perhaps something was accomplished.

The National Oceanographic Data Center seems to be in about the same shape today as when I visited them about four years ago. The only difference is that at that time, they were just designing a card system, and today they are using the card system to put away information that is going to swamp them sooner or later. I have no particularly useful comments to make about this work.

Nick Suszynski has the biggest job of all. I think you have assigned the work to a fellow who is very capable, and if anyone can succeed in getting along with a bunch of prima donna biologists, he should be able to. Whether he can manage to educate the biologists in one generation is a doubtful question. There are so many areas in which the various people need knowledge about the what, why, and how of computer use that I doubt that it can be accomplished on an informal basis. I would strongly recommend to Nick that he set up some sort of "information" course for these biologists. Even the people who are doing the pilot projects with Nick are not fully cognizant of the need for good programming and the need for a very careful reworking of their own data.

From our own experience, we know that the techniques to get data properly structured so that your IR system can work efficiently are non-trivial. While you want the computer system to allow as near natural language input as possible, you still have to be more precise and consistent in structuring your "characters" (taxonomic meaning of the word) than ever before. This is where you will have the most trouble with the older workers - they simply will not bother to follow someone else's instructions. Taxonomists, who have made a fetish of consistency in some directions, are probably the world's most untidy group when it comes to consistency in their actual classification and data procedures. If you don't agree, just look at the record books of any curator for the last 15-20 years.

On Nick's own grounds (computing knowledge) I'd say that he and his programming staff are as efficient as any group we've seen. It is certainly a pleasure to see how they're getting on. Realistically, I want to bring up one problem that may arise in the future. That is, will Nick or Creighton, or the others, be willing to embellish their own procedures with one that we might produce? We think that within our allotted time (two years), we'll have a system going for part of your work, but it may be that for reasons unrelated to the system, these won't be picked up by your computing staff. We know enough about the psychology of systems analysts and programmers to

know that this could be something that they would feel to be a slap in the face. I hope that we won't be a threat to them, and that is the reason I bring this up with you now. We will do our best to play down anything that we might do to cause them to feel that we are taking over, or are trying to push a system onto them, but the best thing we can do is keep the communication going.

With respect to the general capabilities of various parts of your show, I think you have about the only useful reactions we can make in the above paragraphs. There are probably more people there working on computer methods than at any other similar institution in the world. Biologists in general pick up things that they term "mathematical" or "hardware" later than a bunch of engineers would. I'm sure that this is nothing you don't know already. But they have not had much chance, in either their formal education or in later professional work, to get with the right sort of people to get training they really need. For example, if you ask almost any biologist or mathematician what sort of math the biologist ought to have, the answer will be "statistics." But this is not the case for workers in systematics and evolution. The types of math they need have not yet found their way into most college curricula at the levels at which a biologist might have access to them. Strangely enough, the kind of things our kids are being taught in primary and secondary schools come closer than anything in the usual geometry, trigonometry, calculus route. The biggest gun in numerical taxonomy, Sokal, will tell you that statistics is "the answer to cancer" and while I think one ought to know some statistics, this should not be the only course the biologist gets.

We had a pleasant letter from Mason Hale after Estabrook had talked to their Thursday luncheon club. Hale said that George's talk was a very pleasant break from the "Sokalitis" that most of them had come to expect from the numerical taxonomists. The point of bringing up Hale's comment is that we (George and the others here) have discovered that there is a real interest amongst taxonomists for newer methodologies, as long as you do not preface your remarks with the words - "You are a stupid, orthodox, classical nincompoop." You do not get much cooperation that way, but then you'd better have something to say to the taxonomist, even after you agree that he might be fairly intelligent. To this end, we have worked up a course, called taximetrics, which goes into methodologies, not only on how math and the computer

works, but also on a method of structuring taxonomic data. This course has been very successful, and I think that most professional taxonomists would benefit by taking it. Maybe you should get Stan Shetler, or Cheatham, or Ortner, or Bridge, or some of the others to come out and spend a couple of months working with us. We could give them the benefit of the course in shorter periods than needed for grad. students, and let them work on projects of their own interest.

We know that Ortner wants to work directly with us. Even though Ortner is not now working with Nick, unless you clear this with Nick, we won't touch it. Ortner might be another pilot project for the IR system. We would be glad to have his collaboration, provided this is agreeable to you.

3 October 1967

Dr. Mason E. Hale
Department of Botany
Smithsonian Institution
United States National Museum
Washington, D. C. 20560

Dear Mason:

Thanks for your kind words in your recent letter. Reprints as requested are enclosed. Anybody that wants more, I am afraid, will have to get a Xerox copy of these. Perhaps one of the most important of our contributions is the paper by Estabrook and me published in BioScience entitled, "A General Method of Taxonomic Description for a Computed Similarity Measure." This last paper has attracted a very large audience and we have only this and one other copy of the reprint. I have reordered but do not have them yet. Incidentally, Henry said he left all the reprints he had.

Let me thank you for your time and effort with our group there. It is a pleasure to see things moving under your hand.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm

OCT 3 1967

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D. C. 20560

September 27, 1967

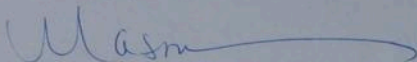
Dr. David Rogers
Department of Biology
University of Colorado
Boulder, Colorado

Dear Dave:

Better late than never, so please accept my thanks for your visit here several weeks ago. The highlight was George's talk at our luncheon seminar, where I understood, mostly for the first time, exactly what you are doing. Your work is a most needed antidote for Sokalitis, a disease that has alienated many taxonomists.

Would you send 10 or 20 reprints of the articles you have available (we can easily pay if needed)? Henry promised to leave some and everybody is asking where they are.

Sincerely yours,



Mason E. Hale
Acting Chairman
Department of Botany

29 September 1967

Dr. Donald F. Squires
Museum of Natural History
Smithsonian Institution
Washington, D. C. 20560

Dear Don,

I'll buy three (3) only Margaritas in Mexico. The reason for the three - I don't think it is safe for more.

We are working toward a summary report for you. Again, the time spent with you and your colleagues was a real payoff and you'll be hearing about it soon.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm

SEP 28 1967

TO : Dave Ryer

FROM: D.F.S.

The bill is \$3.60.
You have the following
options
1) Cash on the front end
2) Pay it off in
drinks in Mexico City.
3) Make a down-payment
on a Washington - Denver
round-trip.
Terms 30 days net.
Cheers.

29 September 1967

M. Nick Suszynski
Smithsonian Institution
Washington, D. C. 20560

Dear Nick,

Thanks a heap for the copy of Ahumada's retrieval system.
Thanks once more for the time you and your crowd took up
with us.

I think that we carried back some very good ideas, and ~~the~~
though it is a bit early, I think we'll have a ~~present~~ system
even sooner than expected. I look forward to seeing you in
Mexico City.

Best regards,

David J. Rogers
Professor of Biology

DJR:gm

20 September 1967

Dr. Donald F. Squires
Deputy Director, Museum of Natural History
Smithsonian Institution
United States National Museum
Washington, D. C. 20560

Dear Don,

This is just a note of thanks for your extremely fine cooperation. Details of our interviews will be collated and we will later give you a summary of our visit. It was a very valuable contribution to our developing I.R. system.

I just remembered that I used your or your secretary's private phone number for our telegram. If you could tell me how much we spent I will belatedly pay off for it.

Please accept our thanks in toto for your and your wife's hospitality and we look forward to seeing you in Mexico in December.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm

29 August 1967

Dr. Donald Squires
Smithsonian Institution
Washington, D. C. 20560

Dear Dr. Squires:

Dr. Rogers, who is off at the AIBS Meeting in Texas this week, asked me to send you copies of his Information Retrieval Questionnaire for distribution at the Institution. (I am mailing these separately.) He would like to have these reach people genuinely interested in information retrieval and willing to go along with the use of computers.

He and his colleagues from the Taximetrics Lab expect to be in Washington the week of September 11 - 15. He will have to leave Wednesday afternoon, but the others will stay through Friday.

Sincerely yours,

(Mrs.) Geneva Metcalf
Secretary

Taximetries Laboratory

Armory 101

1 August 1967

12
Camp

Dr. Donald F. Squires
Deputy Director, Museum of Natural History
Smithsonian Institution
United States National Museum
Washington, D. C. 20560

Dear Don,

Glad to have your proposal. We took the liberty of Xeroxing it - that is why we are returning your copy so fast. It has lots of goodies in it, and ~~we~~ want longer to see it.

In case you did not get a copy of our proposal I am sending (with yours) a preliminary write-up which says things about the work but not about the time, the staff, or the money. We're out of the complete proposal.

More to you next month about our anticipated visit.

Sincerely,

David J. Rogers
Professor of Biology

DJR:gm

SMITHSONIAN INSTITUTION
UNITED STATES NATIONAL MUSEUM
WASHINGTON, D. C. 20560

JUL 28 1967

July 20, 1967

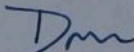
Dr. David J. Rogers
Department of Biology
University of Colorado
Boulder, Colorado 80302

Dear Dave:

Enclosed is a copy of the proposal for our data processing program which has been funded by the Office of Education. I would appreciate it if you would return it to me following your study, for we are desperately short of copies.

We look forward to seeing you this fall and showing you some of the progress we have made.

Sincerely yours,



Donald F. Squires
Deputy Director
Museum of Natural History

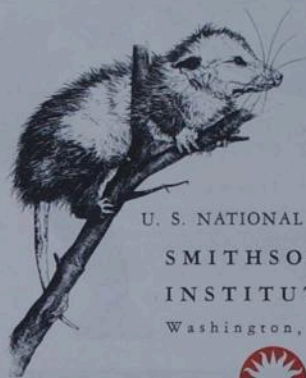
Enclosure

AUG 1 1967

Copy made by Xerox for us
Original returned

A guide to the

MUSEUM
of
NATURAL HISTORY



U. S. NATIONAL MUSEUM
SMITHSONIAN
INSTITUTION
Washington, D. C.



*Langley's Mail
Wed evening.*

Taxinetrics Laboratory

November 14, 1966

Dr. Sidney R. Galler
Assistant Secretary (Science)
Smithsonian Institution
Washington, D.C. 20560

Dear Sid:

I enclose herewith two copies of the following:

1. a preliminary write-up and description of our information retrieval system.
2. a preprint of a paper "A General Method of Taxonomic Description for a Computed Similarity Measure."

These two will, I hope, give you and the others some background before you get here. The first is a part of a proposal we are preparing for the Office of Science Information Services, NSF. In it you will find a rather sketchy account of the computer methods to be done, as well as the design of the input to the computer.

The second paper is more intended for the biologist than for the computer expert, though the latter will find it helpful in the considerations needed to structure the input. While this paper describes a method of data structuring for another of our projects, the clustering method which really came to fruition through your efforts, the main point in sending it now is to give background reading for the method we intend to develop more fully for the IR program.

Would you please route these to Squires, et al.?

It is our intent, during your visit, to flesh out the whole procedure with the computers and with the input. We will spend some time on December 1 with the structuring of the input with sufficient biological discussion to be meaningful to you and Squires particularly. Following this, we will then give the details of computer activities, and this portion should take up the rest of the time.

Dr. S. R. Galler

- 2 -

November 14, 1966

We have you scheduled for the Sigma Xi lecture on Friday, Dec. 2.
Looking forward to your visit.

Sincerely,

David J. Rogers
Professor of Botany

DJR:ch

Enc.



SMITHSONIAN INSTITUTION

Washington, D.C. 20560
U.S.A.

October 20, 1966

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

Dear Dave:

Thank you for your letter of October 13, 1966. I was pleased to learn that December 1 and 2, 1966, will be suitable for our visit to Colorado. I will send you our itineraries as soon as we have our reservations made.

In addition to Dr. Squires and Mr. Suszynski, I would like to bring along Mr. Reginald Creighton, Manager of the Information Storage and Retrieval Project at Smithsonian. I hope this meets with your approval.

The subject of my speech for the Friday luncheon to the Sigma Xi chapter will be a summary of the science activities at the Smithsonian.

I am looking forward to seeing you again.

Kindest personal regards.

Sincerely yours,

Sidney R. Galler
Assistant Secretary
(Science)

- Taxonomy Laboratory

October 13, 1966

Dr. Sidney R. Galler
Assistant Secretary (Science)
Smithsonian Institution
Washington, D.C. 20560

Dear Sid:

We will be pleased to have you, Dr. Squires and Mr. Sussynski here on December 1 and 2. Let us know when you have your flight plans and I will meet you at the Denver Stapleton Airport. I will also arrange for accommodations.

Sometime between now and the 10th of November, could you give me three sentences (short) which will tell the Sigma Xi group the nature of your talk? Or would you rather I say to the organizer that you will give us a summary of the science activities at the Smithsonian? Any projection equipment required?

Looking forward to seeing you.

Sincerely,

David J. Rogers
Professor of Botany

DJR:ch



SMITHSONIAN INSTITUTION

Washington, D.C. 20560
U.S.A.

October 10, 1966

Dr. David J. Rogers
Professor of Botany
Colorado State University
Taxonomy Laboratory
Fort Collins, Colorado 80521

Dear Dave:

Thank you for your letter of September 29, 1966. Of course, realizing your altitude is higher and your air is cleaner and the fact we may all get sleepy and tired at the end of one day -- convinced me that we should make it a two day trip!

In line with your suggested dates, the best time for all of us is December 1 and 2. Dr. Donald F. Squires, Assistant Director, Museum of Natural History, and Mr. Nicholas J. Suszynski, Director, Information Systems Division, will be accompanying me on this trip. This should give us enough time for the tour and an opportunity for me to speak at the Friday luncheon meeting of the local chapter of Sigma Xi.

I look forward to seeing you again and to a most interesting and stimulating discussion.

Best personal regards.

Sincerely yours,

Sidney R. Galler
Assistant Secretary
(Science)

September 29, 1966

Dr. Sidney R. Galler
Assistant Secretary (Science)
Smithsonian Institution
Washington, D.C. 20560

Dear Sid:

We are, to say the least, very pleased to have your letter of September 23. The opportunities which your visit will undoubtedly open up are some of the most exciting that I can imagine.

From our standpoint, any date convenient to you in the months of November or December is fine. Let me right away put one or two strictures on this time, as follows. We would be very pleased if you could arrange a time for our meeting such that you could speak to the University-wide, Friday luncheon meeting of the local chapter of Sigma Xi. This Sigma Xi chapter is one of the most active in the country, and it is not uncommon to have 250 to 300 members of our faculty at the meeting. I am sure that this group would welcome an opportunity to hear of the Smithsonian's scientific program. The dates we have open for this meeting are: November 4, December 2 and 9.

To get a clear picture of our anticipated IR system, a period of about 12 hours is needed. To do this in one day is a bit too concentrated, and would allow little time for any discussion, questioning, etc. I recommend, therefore, a two-day session on a Thursday and Friday.

Another factor to be considered in time arrangement is the altitude differences between here at 5,000 ft., and Washington's sea level. Not all people are affected by this change, but some are, and one may get a little tired at the end of the day. Variation in individual response is hard to predict. But remember—we have clean air in these parts.

If your schedule cannot be arranged to coincide with the meeting dates of Sigma Xi, please do not let this stand in your way. If you cannot make this conveniently, then any time in November or December would be fine. Non-stop flights from either Dulles or Friendship to Denver are quite convenient. We will meet your plane in Denver.

Dr. S. R. Galler

- 2 -

September 29, 1966

If you, or others of your faculty who come along, would care to spend an extra day here, we'd be glad to show off some of the spectacular aspects of the local scenery.

Very sincerely,

David J. Rogers
Professor of Botany

DJR:ch

P.S. It would help us if you could give us an early preliminary okay or refusal for the Sigm Xi lecture.



SMITHSONIAN INSTITUTION

Washington, D.C. 20560
U.S.A.

September 23, 1966

Dr. David J. Rogers
Professor of Botany
Colorado State University
Taxonomy Laboratory
Fort Collins, Colorado 80521

Dear Dave:

I was delighted to learn from your letter of September 8, 1966, of the exciting progress that you are making on your information system project. Also, I am most appreciative of your invitation to me to visit you at the Colorado State University to be briefed on the biological, mathematical, and programming techniques.

I am delighted to accept your invitation. The next problem is to discover a date of mutual convenience. May I suggest that you present me with several dates that would be suitable for you. In the meantime, I will alert the several members of our faculty who are keenly interested in information systems for biology in the hope that it would be convenient for them to accompany me on my visit.

Needless to say, Dave, I am personally delighted with your progress. I know that it has been a long and hard road for you and I am pleased that we are able to remove one or two pebbles from your path.

Kindest personal regards.

Sincerely yours,

Sidney R. Galler
Assistant Secretary
(Science)

- Taxonomy Laboratory

September 8, 1966

Dr. Sidney R. Galler - *Asst. Secretary (Science)*
Smithsonian Institute
Washington, D.C. 20560

Dear Sid:

We have been concentrating our attention for the past several months on the procedures for an information system in biology. It has been our method of operation, as you know, to combine the efforts of biologists, mathematicians, and programmers in the attack on the classification problems. Because this is the case, it is my firm belief that methodologies which we have either developed or will have developed shortly for information retrieval, should be of immense value to you. I would like to invite you to hear about these methods and would suggest that if it is at all possible that you visit us here along with your programmer and anyone else with primary concern for the IR system. I suggest that you come to see us because we can give you the benefit of the biological, mathematical, and programming techniques more completely than could be done if I alone came to visit you. If I visit with you, I would not be able to give you the intimate details of the computer system which we have hopes of adopting.

Let me know if such a meeting is of interest to you, and if you can possibly spare the time to visit us here. If this is not possible, then I will certainly want to say as much as I can say knowledgeably to you and your group there.

The method that we hope to put to work is still in need of research to flesh out the bones of the idea. We will shortly prepare an application to Burt Adkinson's group at NSF for support of this research. Let it be said that we as a group do not intend to build any mammoth information center here in our shop. Our goal rather is to do the groundwork for a general system which other interested people may use without the necessity of developing whole procedures for themselves. First, we are adopting a compiler which is best described as a "program builder" (it has already been designed and tested for another purpose by another group). With this compiler as the skeleton, we will develop a program precisely defined for biological information retrieval. Much of the work on this aspect has already been done. I will develop for my own interests a retrieval system for Manihot which illustrates the applicability of the general

Dr. S. R. Galler

- 2 -

September 8, 1966

system for biologists. We will have an essentially natural language method of query to the system.

There are several claims made in the above paragraph for the generalities of our approach, and to show that our claims for these generalities are not completely "blue sky", is the reason why I invite you and your colleagues to visit us. I hope you can make it. If not, let me know and I will plan to stop by to see you.

Sincerely,

David J. Rogers
Professor of Botany

DJR:ch

Dr. S. Galler

February 7, 1966

N.J. Suszynski

Retrieval System for Biological Nomenclature

Recent discussions with SI staff, your correspondence with Colorado State University, meeting with Dr. Sidney Holt of F.A.O. and Dr. Frank Perring of Monks Wood Experimental Station indicate that such a system is most needed. I would like to recommend an approach which Dr. Wallen and I discussed.

This system if developed would provide as a start two listings:

1. Alphabetical Index (for each and/or both kingdoms).
2. Indented Hierarchical Alphabetical Index (for each kingdom).

The computer techniques for such a system are not completely foreign to us since the industry has been using them for a number of years in "Engineering and Manufacturing Parts Explosions".

As you can visualize these techniques are used in planning manufacturing processes and in scheduling production. In some cases complex final assemblies result with up to 50 or more levels of explosion. In the case of Taxonomic breakdown we have seven levels. On the other hand, we have many more "parts", but that is a problem which will be handled by providing adequate computer programs and sufficient storage devices. We can develop a numbering sequence using the computer for each level to facilitate computer updates of the Master File as new "parts" are introduced.

Copies to: Mr. Bradley, Drs. Cowan, Humphrey, Squires, Wallen, Taylor

Briefly, I am suggesting the following procedure for capturing this data:

A taxonomist will list all of the existing names in each class, order, family, genus, etc. At the completion of this effort we will have six lists; each will be keypunched, and each will be read into the computer, sorted and a sequential number will be assigned to each name in each group, such that in each group the first name within it will begin with number one and the last one will end with some number (n). Another computer program will be developed which will generate a composite key consisting of all hierarchical combinations of the above six numbers. These composite keys will be punched out on tabulating cards and taxonomist will select only those that are legitimate, rejecting the rest. At this point retained cards will be reproduced, each card duplicated in groups of several hundreds. Each participating taxonomist will receive a deck of identical cards to enter onto them all of the species in that particular taxonomic group. When completed, these cards would be returned to SI, the nomenclature of species as written on the card will be punched into the same card and the information will be introduced into the Master File. This approach facilitates doing taxonomic work all over the world, perhaps even in the countries with excess currencies.

I am confident we can develop this computerized system and I will take the responsibility for implementing it. The more difficult and truly monumental task is that of coding and transcribing the amount of data available. I understand that the current estimate for the animal kingdom is something like one million species. Each and every one has to be checked for correctness of coding and entered on the punched card to be converted into computer media (preferably discs). If we were to assume that it will take an average of 5 minutes to enter one species on such a punch card then we are faced with 5 million minutes or 76 man years of work for the animal kingdom, and I am afraid that even this projection is grossly underestimated because there will be species that will take perhaps days to classify.

UNITED STATES NATIONAL MUSEUM
MUSEUM OF NATURAL HISTORY
MUSEUM OF HISTORY AND TECHNOLOGY
ASTROPHYSICAL OBSERVATORY
CANAL ZONE BIOLOGICAL AREA
NATIONAL ZOOLOGICAL PARK
RADIATION BIOLOGY LABORATORY
NATIONAL AIR AND SPACE MUSEUM



SMITHSONIAN INSTITUTION

Washington, D.C. 20560
U.S.A.

NATIONAL GALLERY OF ART
NATIONAL COLLECTION OF FINE ARTS
FREER GALLERY OF ART
JOHN F. KENNEDY CENTER FOR
THE PERFORMING ARTS
NATIONAL PORTRAIT GALLERY
INTERNATIONAL EXCHANGE SERVICE
SCIENCE INFORMATION EXCHANGE

November 22, 1966

Dr. Dave J. Rogers
Professor of Botany
Colorado State University
Fort Collins, Colorado 80521

Dear Dave:

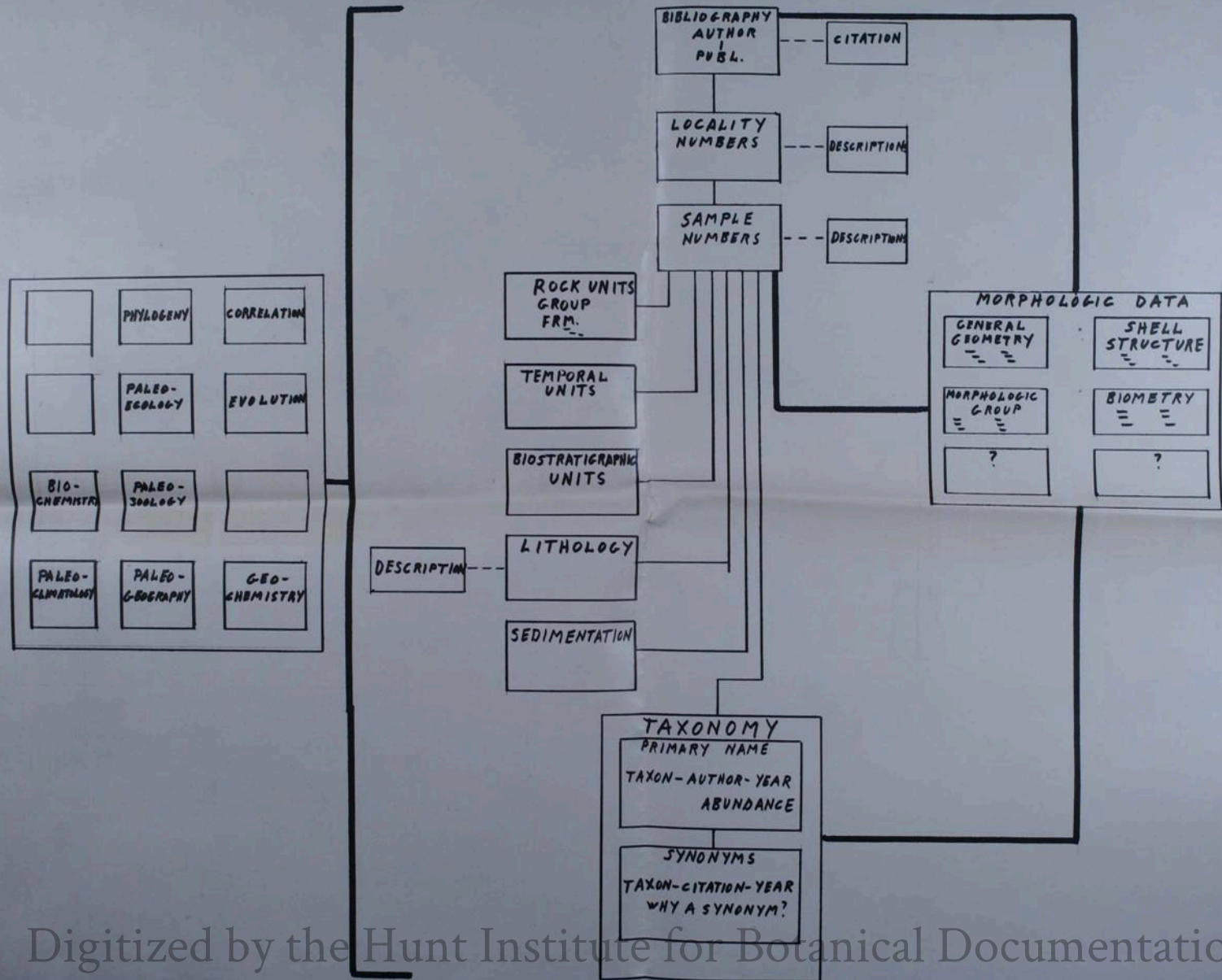
I appreciate receiving your letter of November 14, 1966 with two copies of "A General Method of Taxonomic Description For a Computed Similarity Measure," and "Preliminary Description of the Information Retrieval Project from the Taximetrics Laboratory of Colorado State University." I found both reports to be extremely interesting and certainly full of background reading for the forthcoming trip. I will route copies to Dr. Squires, Mr. Suszynski, and Mr. Creighton.

I hope all is well with you. I look forward to visiting you and to seeing what exciting things you have been doing.

Kindest personal regards and best wishes for a joyous Thanksgiving.

Sincerely yours,


Sidney R. Galler
Assistant Secretary
(Science)



PROPOSAL FOR RESEARCH AND RELATED ACTIVITIES SUBMITTED
TO THE U.S. COMMISSIONER OF EDUCATION FOR SUPPORT
THROUGH AUTHORIZATION OF THE BUREAU OF RESEARCH

Title:

An Information Storage and Retrieval System for Biological and
Geological Data

Applicant Organization:

Smithsonian Institution, Washington, D. C. 20560

Initiator:

Initiated by

Donald F. Squires
Deputy Director
Museum of Natural History
Telephone: 144-5952

Transmitted by:

Transmitted by

James Bradley
Acting Secretary
Smithsonian Institution
Telephone: 144-5101

Duration of Activity:

18 months from date of award

Total Federal Funds Requested:

\$292,927

Date Transmitted:

MAY 3 1967

Title:

An Information Storage and Retrieval System for Biological and Geological Data

Principal Investigator:

Donald F. Squires

Contracting Agency:

Smithsonian Institution

Total Federal Funds Requested:

\$292,927

Duration of Activity:

18 months from date of award

Summary:

Museums must modernize their record-keeping systems in response to the acceleration of today's scientific effort and resulting enormous increases in size of collections if they are to continue providing information to the scientific community, including students, university faculty, and research institute scholars. Because collected specimens are the core of biological and geological information, a computerized system of collection management is essential not only to efficient utilization of museum resources but also as an interface with conventional libraries and other subject-oriented facilities. The Smithsonian Institution has assumed leadership in this area and is developing a comprehensive system.

Any information to be of value must be correct and must be provided when needed. To the extent that it is factual, it presents the researcher with answers to specific questions; to the extent that it is timely, it is available when needed. The Smithsonian Institution is achieving a major design objective by developing a system which will be easy to use by people having no knowledge of programming techniques. It is called the Smithsonian Institution Information Retrieval System. This System is flexible in that it handles fixed and variable length records (in increments of the logical record size) as well as bit strings containing characteristics significant of a given form of life or object.

The communication between the user and the System is accomplished through "COBOL" which is a subset of English language. The System will be very easy to use: the user states his requests in the form of "IF" statements: IF SPECIES IS X-US Y-US AND LOCATION IS NEW JERSEY THEN PRINT. Thus the user indicates relationships which will satisfy his request and this constitutes the query language. He need not concern himself with file structure, input and output commands, etc. These functions are performed by the System in conjunction with "COBOL".

The development of the proposed system will permit greater accessibility of fundamental resource materials of specimens and related data to students at all levels as well as senior scholars.

Description of the Proposal

(a) The Problem

Urgent requirements for increasing communications in science have led to great efforts to improve library techniques and bibliographic resources and particularly to establish information networks. It is the thesis of this proposal that specimens are a form of library resource and that the retrieval of information about specimens and based upon specimens is as fundamental to knowledge as the traditional library. Specimens are three-dimensional representations of data contained on the printed page. A growing opinion in science suggests that the informational content of a three-dimensional object is far greater than previously thought and can be more fully utilized. Bibliographic treatment of specimens, as discussed later, certainly suggests this. A specimen may be thought of as the kernel about which information accretes in successively more abstract (less specimen directed) layers. Specimen collections are therefore an important library resource and their informational content should be carefully integrated into any informational system.

Museum collections are utilized by many persons outside the museum community. Although there are over 5,000 museums in the United States, only a limited number of these (perhaps 100) are repositories of research collections. Examination of the records of these museums demonstrates that the vast majority of persons studying their collections are university personnel, all levels of faculty and student investigators.

Museums do more than provide access to their collections; collections are loaned for study to students and scholars outside museum walls, both nationally and internationally. The Smithsonian's Museum of Natural History made 22,520 exchanges of specimens with other institutions during 1966 and loaned more than 153,165 specimens; more than 3,000 scholars visited the collections, spending one day or many months in their research. Over half of these visitors were students. These scholars, in addition to museum and research institute personnel, are the potential users of a museum information system.

Museum collections are not then quantities of specimens stored in dark, dusty rooms, unused and unwanted; nor are they relics of man and his environment, the castoffs of our technology. The collections are the substance of man's knowledge, the objects forming a three-dimensional library basic to the educational process and, as such, may be borrowed, visited and examined by students and professionals in the continuing process of expanding our knowledge of our environment, our world, our universe.

Study of museum collections requires knowledge of their contents and access to the vast store of information associated with the collections. Museums have generally coped with the increasing size of collections, the filing systems for the specimens themselves providing for the retrieval of particular objects. Museums have been less successful in making available information about specimens, for the manual cross-referenced files have not been adequate to keep pace with rapidly growing collections.

Object collections should be the core of a biological information system, for they represent the concrete, the reality of the information of a system even when that information is abstracted several degrees from the specimen. For example, information about a specimen of plant may primarily consist of the name of the plant, its identification number, and the locality from which it was collected. Information about the collector, the date of the collection, the means by which the museum may have acquired the specimen, and perhaps even some particulars about the environment of the plant in the living state are common museum records. The requirements of modern science and new directions of research dictate that even more information must be stored; for instance, details of the environment from which the plant was collected, its associated insect fauna, associated plants and animals, season during which it was collected, approximate age of the specimen, and other possible characteristics.

Second-level information is still based upon the specimen, and may include bibliographic citations to references in which that plant specimen was described, studies in which the specimen was used as a voucher specimen, as, for example, in toxicological studies or in the search for natural products. In these instances, there would be references to pharmacological literature or to chemical analyses. Or it may have been a voucher specimen for an ecological study in which the behavior of a certain mammal species was described and in which that plant species was found to be a food item. Perhaps the specimen was a voucher specimen in a study of the distribution of a noxious weed.

Tertiary levels of information are concerned with the concept of the species of which the specimen is only an example. Most of the records at this level will be bibliographic and will deal with systematics, ecology, physiology, biochemistry, pathology, economic utilization, etc. The ability to relate the concept as it occurs in literature to the reference resources (specimens) is not now possible. Using a tertiary-level system, a student reading of the occurrence of tumors on blue crabs could immediately relate this information to museum specimens of tumors, the factual basis for record, as well as to the distribution of blue crabs and hence to the availability of live material.

If a general information system is to be devised, museum collections should be an integral part of such a system. Museums must therefore develop a methodology for handling the several levels of information: 1) primary information about individual specimens; 2) references to individual specimens and the body of information accumulating about individual specimens; 3) references to bodies of data incorporating representative specimens. Such information systems should have the primary goal of interfacing with other systems not primarily object oriented.

In addition to its role as host to visiting investigators, the Smithsonian Institution has developed a number of working relationships with universities. To encourage and assist students utilizing the facilities of the museum, a number of predoctoral fellowships have been established. In order to facilitate the relationship between these students and their parent universities, the Institution has negotiated educational agreements with 14 universities, making possible direct supervision of student dissertations by the staff of the Museum of Natural History. The Institution also encourages its staff to participate in university affairs by giving courses at both the undergraduate and graduate levels.

(b) The Project

The Smithsonian Institution's Museum of Natural History, in its role as one of the largest repositories, has a responsibility to assume leadership in the development of a program for the utilization of an information system in both national and international natural history museums. In the development of such a system, the Museum will assure that better means of communication are established between museums, universities, and other centers of research and that proper interfaces are developed between such a system and subject-oriented information systems. In the belief that a general information net will eventually be established, the Museum will consider its role in such a network and will place particular emphasis upon development of appropriate relationships with it. The Smithsonian Institution is represented on appropriate interagency committees considering establishment of such systems, and, among others, COSATI has been briefed on the proposal of the Museum of Natural History.

In the three year period for which support is requested, it is planned to test a general data storage and retrieval system with particular emphasis upon three subject areas representing different degrees of sophistication of knowledge and areas in which information flow is of particular concern. These areas are: 1) oceanic birds; 2) marine

crustacea; 3) marine rocks. All relate to the present national impetus in oceanographic research and are areas in which specimens and information about the specimens are in particular demand because of a large number of active research programs.

Although the initial tests are involved with specimen-related data and will deal with primary level information, the general system will be capable of dealing with secondary and tertiary level information as well.

(c) Work Done to Date and Use to be Made of Findings

In 1963 the Director, Museum of Natural History, established a scientific staff committee which developed a general knowledge of data processing in the the course of meetings, consultations, and training courses.

In 1965 the Smithsonian Institution, through its Office of Systematics, contracted with Federal Systems Division, IBM, to study Codification of Biological Names*, and to recommend appropriate ways in which this problem could be attacked. This contract terminated with the conclusion that the codification structures proposed were not feasible and that biological names not be codified, but treated as alphabetic words in a data processing system. In this same year, Mr. Suszynski became Director of the Smithsonian's Information Systems Division. Mr. Suszynski's arrival coincided with initiation of a study by the Mangement Services Division of Peat, Marwick, Mitchell & Co of the Smithsonian's plans for a Computer Center and of its various programs in data processing and their requirements, at the request of the Smithsonian. The terms of this proposal follow the recommendations of the IBM and PMM studies and conform with the policies of the Institution in the development of a data center.

The Smithsonian Institution has kept informed of activities in data processing in the United States and abroad, and has maintained contact with persons involved in the development of biological data retrieval systems. In early 1964, copies of a report treating simple data processing systems were circulated to staff, Museum of Natural History, and to other museums in the United States and abroad. Replies were generally unenthusiastic, and response from the American Museum community has not changed since. Discussions of data processing at the 1966 meeting of Directors of Systematic Collections were inconclusive. We believe however, that through demonstrations of the operation and benefits of a data retrieval system, other organizations may be encouraged to implement them.

*Copies of the unsolicited proposal forming the basis for this contract are available for review.

Interest internationally is now greater. In July, 1966, agreements were reached between the Smithsonian Institution and the Natural History Division, National Museum of Canada, on the exchange of personnel and complementary programming. This will result in faster development of the system and will make possible direct interchange of data between the two institutions. The Natural History Museum of Canada is also developing the possibility of assistance to Mr. Sydney W. Gould, Connecticut Agricultural Experiment Station, New Haven, Connecticut, in the development of "geocode" a vital part of the system discussed here.

Discussions in October 1966, with the British Museum (Natural History) indicated that although that institution is not yet ready to move into data processing, it will follow with considerable interest the developments here. Expressions of agreement with the concept of compatibility indicate that future potential for interchange of data is substantial. A representative of the staff, British Museum (Natural History) will spend the month of August 1967 with the Information Systems Division, Smithsonian Institution, to enable him to report on possible implementation in England.

Negotiations for data exchange are proceeding with the University of Mexico, Commonwealth Scientific and Industrial Research Organization, Australia, and the Department of Scientific and Industrial Research, New Zealand. Cooperative efforts are being made with the University of Toronto in the exchange of botanical and geological data. In the United States, we have followed closely the interests of the University of Kansas, the Los Angeles County Museum, the Field Museum of Natural History and the American Museum of Natural History. The Smithsonian Institution was recently briefed on the new data retrieval language under development at the Colorado State University under the direction of Dr. David Rogers.

To meet the problem of making information about the Museum collections available in a timely and accurate manner, a small pilot project provides information previously unavailable with respect to marine birds of the Pacific Ocean. This practical experience led to the establishment of the basic design of the Smithsonian Institution Information Retrieval System, subsequently referred to as the "System."

The results of the proposed project will be disseminated among the scientific and museum community through workshops, conferences, publications and information services. The first international conference on data processing in natural history museums, botanical gardens and herbaria, jointly sponsored by the Smithsonian Institution and the Universidad Nacional Autonoma de Mexico, will be held in Mexico City on December 18-20, 1967. Other conferences of this sort will be scheduled as frequently as seems warranted by progress made in data processing in museums.

A series of workshops on the proposed system is planned. Initially, these will be at the technical level. The first is planned for the Fall of 1967, when staff members of the National Museum of Canada, Natural History Division, and the British Museum (Natural History) will come to the Smithsonian Institution for training purposes. During the Summer, 1967, an associate from the Universidad de Mexico will work at the Smithsonian.

Also receiving consideration is a project whereby the Smithsonian Institution, with the cooperation of the Association of Directors of Systematic Collections, will sponsor a series of workshops for administrative and scientific personnel to better inform them of data processing techniques and of the ways in which cooperative use of the System being developed by the Smithsonian can be effected. Similar programs to be held under the auspices of the U. S. National Museum Act, are under discussion with the Director, U.S. National Museum.

A series of publications are planned as a sequel to "Data Processing: A Problem for the Present" (CURATOR, Vol. 9:3, pp. 216-226). The Editor of this journal has assured the Principal Investigator that further articles on this subject will be welcomed. The Principal Investigator is also preparing a chapter on data processing techniques for a standard reference on Museum Registration Techniques.

Most universities are so closely aligned to museums that they will be aware of progress through the above sources of information. For dissemination of information about the proposed research to the persons outside the museums and particularly to the non-academic community, great reliance will be placed upon information services such as the Office of Education's ERIC and the Smithsonian Institution's Science Information Exchange.

Object Collections are a fundamental resource for scholarly investigations in the fields of science, history and art. Increasing costs of maintenance of collections are such that many universities no longer maintain reference collections, but rather have aligned themselves with museums in cooperative educational agreements in order to provide these resources. Accessibility of collections and specimen related data, particularly in natural history museums, is now a more vital requirement to education at all levels than ever before. This requisite is evidenced by greatly increased demands for collections and data by either loan or visits to museums by undergraduate students as well as senior scholars.

Proposed Plan of Work (See flow charts in Appendix)

The proposed project divides into two major subsystems:

1) source data preparation and 2) system design and programming.

Source data preparation is performed by a "cataloger" who unites specimen data, collection information and the specimen itself through a label attached to the specimen. Smith Corona Marchant electronic punched-paper tape hardware is being used to record the data.^{1/} The operation of the equipment is presented in the flow chart (figures 1 & 2). Although experience with the equipment has not been sufficient to establish sufficient records of time savings, the new procedure is estimated to have shortened by 60% the time required to record a specimen. This source data automation illustrates the primary objectives of the effort: 1) To preserve scanning files and such other records as may be required for collections management and for curatorial work at a lowered cost per unit; and, 2) to develop an input to a computerized system of information storage and retrieval while preparing the most basic of the source documents, the specimen label.

Systems design and programming are in progress for the first stage of the evolutionary pattern of development for the System. Since classes of input data vary widely from one scientific discipline to another, an early decision was made not to use one of the available generalized retrieval systems; but to develop an operating and compiling system, which on the basis of queries submitted to it, will be able to compile a custom program for these requests only. (The generated program will be tailored to the data content and format of a file pertinent to the request) This will eliminate time lost in scanning numerous files and in going through time consuming computer instructions not directly pertinent to a given set of requests for information.

^{1/} Data Processing and Museum Collections: A Problem for the Present
Volume IX/3 1966 Pages 216-226 Curator - M. N. H. Donald F. Squires

Initially the pilot system will respond to interrogations where the species name or other hierarchy names are specified. It will then be enhanced to process location oriented interrogations where species is not identified.

Lastly, response to parameters independent of location and species will be implemented.

To prevent the recurring problem of hardware incompatibility which causes reprogramming of the developed systems, Smithsonian Institution uses "COBOL" as a query language in conjunction with the System. (The Department of Defense has set the requirements for "COBOL" and virtually all computer manufacturers include "COBOL" capability conforming to DOD standard) and its subroutines will be written in "COBOL" thus facilitating installation of the System anywhere on any computer having "COBOL" compiler (subject to minimal configuration in terms of memory size and peripheral devices).

TABLE I

Typical Interrogations

Locate, list and count all species in genus X.
Determine number of species Y in given collection and state source.
Does collection contain specimens of species Y from locality A?
Advise re data on specimens from locality A.
What species occur at Station C of the Argus cruise?
What species are associated with Species Y?
List distribution of species Y.
What has been identified from accession No. 100001?
List species in sequence by catalogue number.
What was collected by Smith (collector No. 5555) in locality A?
Name of identifier of species Y?
Give temperature ranges where species Y and Z are found.
On what kind of substrate does species Y occur?
List synonymies of species Y and Z.
What parasites are found on host Z?
On what hosts are parasites Z found?
What isotopic data are on file for specimen X?
What is chemical composition of minerals Y and Z?
What geologic data are on file for specimens Y and Z?
What types in genus X are in collection?

TABLE 2

Types of Data Recorded on Specimens1. Directory Items

Kingdom	Genus
Phylum	Species
Class	Author and date
Order	Nomenclature Code
Family	

2. Data Base Items

Nomenclature Code	Field number
Location (lat., long./UTM grid coords.)	Accession number
Geographic occurrence (country, state, town or ocean as appropriate)	Catalogue number
Date collected	Number of samples (by sex, if approp.)
Collector	State of samples
Cruise name	pH value
Station number	Depth of water
Original name (coded)	Substrate
Time of collection	Habitat
Temperature	Citation reference
Current (direction, velocity)	Scientific name (4)
Salinity	
Geologic age (formation, stage, system)	Isotopic data
Formation name	Mode of preservation
Thin sections available (rocks)	Related flora/fauna
Texture	Parasites
Mode (mineral name, abundance)	Diseases
Trace elements	Hosts

Primary access to the data is through the binomial. (Binomial is a Latin word-pair used to identify uniquely each species of the plant and animal kingdom) "Floating" levels of phylogenesis are cross referenced, from the class level down to the binomial, permitting access at higher levels such as family, order, etc. (The structure is described as "floating" because it is organized to permit periodic modification or reconstruction).

Secondary access to data is through geo-political location references. A hierarchical location code is employed to integrate the diverse location reference methods (latitude, longitude; Marsden, name of territory, etc.) used for data input into a unified language permitting a recall of data at a variety of area levels above that of data input.

Processing a request for recall of data involves two operations: general discrimination and specific search. The general discrimination phase, using the phylogenetic and/or location references cited on the request delineates to the specific search the location of the data pertinent to the request. In effect, it eliminates most of the data base from consideration. All of the remaining data will have some of the characteristics which will satisfy the request, presumably one or more will satisfy all of them.

The proposed approach is economical of computer time because retrieval is accomplished by disqualification of large portions of data based on examination of abstracted portions of this data in an index-like list of descriptors. A computer search of the descriptor short file or directory yields the storage location in the long file (text file) of the records requested. Located text data is then examined for precise correspondence to the full range of request criteria. Selected data records are associated with nomenclature and/or bibliography data as desired and printed for distribution to the requestor.

The System design follows the following basic tenets:

1. All data is entered in prescribed formats and is searchable.
2. No abbreviation or translation to coded form is required (of the system user). The computer translates natural language input into whatever code is required and retranslates output.
3. User training is minimal.
4. Indices will be used for processing economy, e. g.
 - .. phylogenetic
 - .. geo-political
 - .. catalog number, and
 - .. stratigraphic indices

Requests for data are originated by the user in coded simple statements such as IF NAME IS X-US AND LOCATION IS PANAMA, THEN PERFORM RECORD-SELECTION.

The user having mastered a restricted subset of the procedural portion of the query language, communicates directly with the computer eliminating the need for programming analysts serving as translators. The users statements, in computer modified form, are employed in the general discrimination phase (rough search) of processing, then are processed again (unmodified) as actual procedural statements, together with program segments selected from a library tape, in "compile-and-go" mode for the final search. For each search, the unique program generated is tailored to the specific parameters of the request batch.

Since the procedural portion of the query language, (user requests) is truly machine independent, no retraining for the users is expected in the event a different computer is employed.

A list of nomenclature above the generic level is in taxonomic sequence enabling assignment of a "Numericlature Code" by the computer. The code, together with the name, enables the system to achieve many objectives, among which are:

1. Validation of names submitted as input for specimen records.
2. Validation of names used in retrieval queries.
3. Association of names with higher taxa.
4. Association of synonyms with preferred names.
5. Identification of homonyms.
6. Identification of the magnetic disc or a reel of tape containing the data.
7. Association of bibliography, etc. with taxa name.

The interrogator using the computer assigned code can post subsequent findings, enter synonyms and derive filing techniques for photo slides, figures, specimens, etc. The "Numericlature Code" is structured to allow inclusion of new taxa or further refinement of nomenclature by a suffix code, initially assigned the value zero

(permits subsequent inserts and revisions without disturbing any nomenclature not involved in the modification).

A geographic code will be used to relate different methods of expressing location, such as: political name, Marsden, latitude-longitude, etc., so that a query will retrieve on all data no matter how entered. A four character code, each character corresponding to a matrix, within a matrix, representing successively decreasing portions of the area of the globe will be employed. The system now planned for use is the "Geo-Code" of Mr. Sidney W. Gould which has been adopted by the Biological Records Center, Nature Conservancy (England) for use in Flora Europea and data processing in conjunction with the International Biological Program.

Data will be stored as a variable number of fixed length records in a header-trailer fashion. As many trailer records as necessary to record all data relating to a given specimen are permitted. The system is open ended.

The system is designed to provide for continuing expansion through the various stages of this project. At least 30,000,000,000 characters of data in the final data bank will be needed to accommodate 3,000,000 species; each species containing 100 specimens. It is assumed 100 characters will be required to describe each specimen.

Both current technology and cost considerations prevent current on-line access to the whole of data. Only those portions of data which are pertinent to a given batch of requests will be brought on-line and directly accessed to preclude examination of irrelevant data.

A sufficiently large disc-pack computer system is therefore required with some tape units for peripheral processing.

The user of the System obtains maximum economy by stating as specifically as possible some level of phylogenetic name and/or some level of geo-political area reference. Conversely, broad questions will generate many files and large number of answers with the associated increase in cost.

In addition to the dual access retrieval system, a simple retrieval subset will permit access to data by way of the catalog number.

Queries relating to a given expedition for example, could be expressed in terms of the interval of catalog numbers assigned to that expedition and thereby recalled.

The approach suggested for this project has been implemented on a smaller scale by the Smithsonian Institution in the development of the Ecological Study of Birds for the Pacific Ocean. In this computerized system, searches in the form of Boolean Algebra (AND, OR, AND NOT operators) are directed against the data base consisting of bird sightings, marine mammals and the environmental information. Information selected out as a result of these queries can be presented in detail as well as in a summary form. In addition, various correlation analyses which involve sightings versus environmental data have been produced. (The Pacific Ocean Information Storage and Retrieval Project is mentioned only to indicate demonstrated feasibility of the proposed system.)

Basically the System will be composed of four elements:

1) Request parameter cards, 2) Several directories, 3) Master specimen inventory file, and 4) Series of computer programs, all operating under the control of the executive program.

1) REQUEST PARAMETER CARDS

The requester writes simple conditional statements identifying the binomial of interest. For example: if information on the American Pelican is desired, the request would be written as follows:

IF SPECIES EQUALS PELICANUS-AMERICANUS THEN
PERFORM RECORD-SELECTION.

All requests of a given requester will bear his routing code. The initial system will service up to 25 routing codes and several hundreds of request parameter cards for each search.

2) THE DIRECTORIES (only Binomial index is discussed)

The directory will be contained on discs. It will include: a list of all the binomial names for specimens found in the Specimen Inventory File; a numerical identifier for that name (to provide hierarchal reference); synonyms, homonyms and reference to bibliography.

The directory will also contain nomenclature and identifiers for name of kingdom, phyla, classes, orders, families and genera. The directory serves to:

- a) Validate requests; requests for species not a part of inventory will be deleted, with appropriate commentary to requester. Only answerable requests will be processed.
- b) Provide the system with comparison parameters for extracting data.
- c) Identify which data files are to be searched; only inventory files germane to the requests will be processed.
- d) Permit a requester to use names, which he already knows, rather than numbers which he must memorize or reference.
- e) Permit programs to communicate with data in code, succinctly and efficiently, without encumbering the requester.
- f) Provide a means for periodic printing of systematic lists without recourse to the voluminous master specimen inventory file.
- g) Facilitate optimum batching of requests.

Coding of the Binomial Directory is already in progress on a limited scale at the Smithsonian. The trial formats for nomenclature, numeromenclature and bibliography appear to be satisfactory and can be furnished upon request.

3) MASTER SPECIMEN INVENTORY FILE

The magnetic tape master specimen inventory file will contain at least one record for every item (specimen) of inventory. The input record will be created on paper tape automatically as labels for the specimen are typed. The record will identify binomial, where collected, by whom, accession number, characteristics, ecological and environmental parameters, and additional information.

Whose
system?

The file will be arranged in phylogenetic sequence, categorized into appropriate taxa groups to best facilitate retrieval of data.

4) THE COMPUTER PROGRAM SYSTEM

This phase of the system generates retrieval program using three basic ingredients:

- a) the binomial directory (discussed above).
- b) the request parameter cards (discussed above)
- c) a library of precoded source language program segments.

A library of program segments provides the mandatory statements for housekeeping, subroutines and subroutine linkage and points of interface with the request parameter statements.

Based on the request parameter cards, pertinent directory records are selected which identify the disc or the tape containing wanted data. Requests containing binomial names not on the directory tape are rejected. The directory records selected identify to the system the numerical code for the binomial, thereby providing the hierarchical capability (inherent in the numericlature coding structure). The retrieval program is thereby equipped with the requester's extraction statements and with the directory's constants which relate names to data fields and data files. It remains, only to include from the program segment library tape, those segments which relate to the particular requests to be processed. This done, the retrieval program is completed. This source language program will compile in a few minutes. The time required to effect retrieval of data is a function of the ability of the program generation module to exclude low yield data files from the retrieval phase. It does this by examining all requests and selecting only those that produce the highest frequency of reference to the fewest number of data files. Non-qualifying requests are deferred to subsequent processing cycles or are processed by a special cycle depending on priority code and deferment history.

Each time that a low yield request is rejected, its priority code is automatically increased by one; thus guaranteeing that eventually even a very low yield request is processed.

It is to be emphasized that the Program Generation Module presents to the system for compilation a source language program tailored to the specific requests to be processed. It includes only those directory items and program segments which are required to process requests selected for processing. It directs the computer operator to mount only those data files containing desired data records.

As a unique retrieval program is generated for each batch of requests the "Compile and Go" mode is used.

Assuming five data files in the master file with a data base of 500,000 specimens, through-put time for a stack of requests having 25 independent requesters should be in the vicinity of 4 hours of computer time.

The selected records are sorted and printed. Sorting is performed on the distribution code (grouping all of the requests belonging to an individual requester). A new page would be ejected on the high speed printer when a new distribution code is sensed. Data listings are decollated and distributed via inter-office mail. Future distribution might employ remotely located terminal devices.

MASTER FILE UPDATING

Programming will be developed to update the master specimen inventory files as frequently as desired. New specimens will be introduced into it. Previous entries may be deleted from it. Fields of information may be changed. New species, families, etc. may be incorporated as desired. Appropriate safeguards and programmed audit procedures will afford protection against errors and illogical or unauthorized modification or introduction of data.

BINOMIAL DIRECTORY UPDATING

Programming will be developed to provide for updating of the directory as required. Once compiled, the file should need little attention other than occasional updating.

PROGRAM SEGMENT LIBRARY TAPE UPDATING

Segment on the library tape will be modifiable as required.

Time Frame

The system will be designed in modular fashion to permit continuing expansion through all phases of the project. Development and implementation of the pilot project will take three years. Systems design may be accomplished in two years. Some subsets are already in progress. The activities to be carried on and the required applied time are set forth in the following chart:

Anticipated Steps in the Development of the Proposed System

	% Completed			
	12 mo.	18 mo.	24 mo.	36 mo.
1. Staff orientation	100			
2. ADP coordination of the gathering and preparation of data for the binomial dictionary file 52,800 initial species (200 per day)	100			
3. Methodology (formats etc) for preparation of specimen and bibliographic data (for the categories of the pilot project)	100			
4. Programming to digitize the relative location of taxa in the phylogenesis	100			
5. Programming to update and maintain taxa lists	100			
6. Key punching of 52,800 taxa names	100			
7. Programming to convert paper tape to magnetic storage	100			
8. Programming to validate taxa names submitted with data	100			
9. Programming to examine input data for correctness and completeness (other than taxa name) bibliographic entries and specimen data (for the categories of the pilot project)	100			
10. Programming to format raw data on magnetic tape to a searchable format	100			
11. Programming to update a master tape with accessions	100			
12. Programming to sort magnetic tape to enable updating, printing of catalogs etc.	100			
13. Programming to print catalogs of specimen data by taxa	100			
14. Programming to search magnetic tape (in a non-index linear fashion) to process initial queries	100			

		% Completed			
		12 mo.	18 mo.	24 mo.	36 mo.
15.	Programming to print catalogs of bibliographic data by taxa	100			
16.	Programming to cross index taxa names with magnetic data files containing specimen with files containing bibliographic data	25	100		
17.	Taxa search optimization through use of the binomial dictionary (both specimen and bibliographic data)	0	100		
18.	Gathering and preparation of data for the geo-political cross index directory	50	100		
19.	Key punching of geo-political designators	25	100		
20.	Programming to codify geo-political expressions to enhance searchability			100	0
21.	Programming to validate geo-political expressions submitted as data and to post the codified representation	0	25	100	
22.	Programming to print catalogs of specimen data by Geo-Code	0	0	100	
23.	Programming to print catalogs of bibliographic data by Geo-Code	0	0	100	
24.	Programming to search magnetic tape (in a non-index linear fashion) to process location oriented requests	0	0	100	
25.	Programming to develop a Geo-Code cross index	0	0	25	100
26.	Geo-location search through use of the Geo-Code cross index	0	0	0	100
27.	Programming to establish a catalog number cross index	0	0	0	100
28.	Programming for searching of catalog number intervals	0	0	0	100
29.	Programming to establish a stratigraphic cross index directory	0	0	0	100
30.	Programming for searching by stratigraphic parameters	0	0	0	100
31.	Programming of system executive compiler "retrieval program generation module"	0	0	100	

The costs for the second eighteen months of this project are estimated as follows:

I	Direct Costs:		
		Federal	Local
	A. Personnel compensation:		
	Principal Investigator 10%		\$ 3,489
	Information Systems Director 20%		6,309
	MNH Staff Scientists 50%		9,528
	Senior Systems Analyst 100%		28,004
	Program Analyst 40%		9,256
	Program Analyst 40%		10,568
	Program Analyst 100%	\$ 22,459	
	Program Analyst 100%	22,459	
	Program Analyst 100%	19,057	
	Cataloger 100%	13,395	
	Cataloger 100%	13,395	
	Cataloger 100%	11,199	
	Bibliographer 100%	9,255	
	Typist 100%	<u>10,206</u>	
	Total Personnel Compensation	\$121,425	\$67,154
	B. Personnel Benefits	12,142	6,715
	C. Travel & Transportation of Persons	7,500	
	D. Supplies and Materials		1,500
	E. Communications	300	
	F. Other Services:		
	Research operations		7,500
	Computer time	<u>81,400</u>	
	G. Equipment		<u>3,500</u>
	H. Total Direct Costs	\$222,767	<u>\$86,369</u>
	I. Overhead	<u>84,651</u>	
II	Total Cost	<u>\$307,418</u>	<u>\$86,369</u>
III	Cost Sharing	78.1%	21.9%

Personnel and Facilities

- a. Personnel - See Appendix II
- b. Facilities

It is planned to have a Honeywell 1200, 5 disc-6 tape system, or its equivalent, available for work under this project. The source data automation equipment will be located in the Museum of Natural History. This will consist of Smith Corona Marchant punched paper tape equipment, tape equipment, flexowriter, and keypunching equipment. Office space and supplies will also be available for use under this project.

BUDGET

Initiator: Donald F. Squires

Duration of Activity: 18 months

Institution: Smithsonian Institution

Beginning Date: Date of contract award

I Direct Costs:

	Federal	Local
A. Personnel Compensation:		
Principal Investigator 10%		\$ 3,214
(D. Squires, date of employment - 12-18-61)		
Information Systems Director 20%		5,821
(N. Soszynski, date of employment - 11-3-65)		
MNH Staff Biologists (4):		14,511
(100% of 1 man-year for 12 months, 50% of		
1 man-year during 12-18 month period)		
Senior Systems Analyst 100%		25,871
(R. Creighton, date of employment - 7-17-66)		
Program Analysts 100%, 2 @ \$20,695	\$ 41,390	
(to be hired)		
Program Analyst 100%	17,560	
(W. Minami, date of employment - 12-11-66)		
Program Analyst 40%		8,540
(J. Crockett, date of employment - 8-14-66)		
Program Analyst 40%		9,740
(W. Wheeler, date of employment - 3-12-67)		
Bibliographer 100%	8,527	
(to be hired)		
Typist 100%	9,403	
(to be hired)		
Catalogers 100%, 2 @ \$12,340	24,680	
Cataloger 100%	10,318	
(all to be hired)		
Total Personnel Compensation	\$111,878	\$ 67,697
B. Personnel Benefits	11,188	6,770
(FICA, TIAA Retirement, Disability, &		
Life Insurance, Medical Insurance, Tuition		
Refunds, Travel Accident Ins.)		
C. Travel & Transportation of Persons:	7,500	
(To coordinate work with museums in the U. S.		
and abroad)		
D. Supplies & Materials		1,500
E. Communications	300	
F. Other Services:		
Computer time, 814 hours @ \$100/hr.	81,400	
Key punch operations		7,500
G. Equipment:		
Typetronic Systems for Data Preparation		17,500
H. Total Direct Costs	\$212,266	
Overhead expense - 38% of Total Direct Cost	80,661	
(Overhead is audited by the Defense Contract		
Audit Agency, Silver Spring, Md.)		

II Total Costs \$292,927 \$100,967

III Cost Sharing 74.3% 25.6%

Source Data Automation in the Department of
Invertebrate Zoology

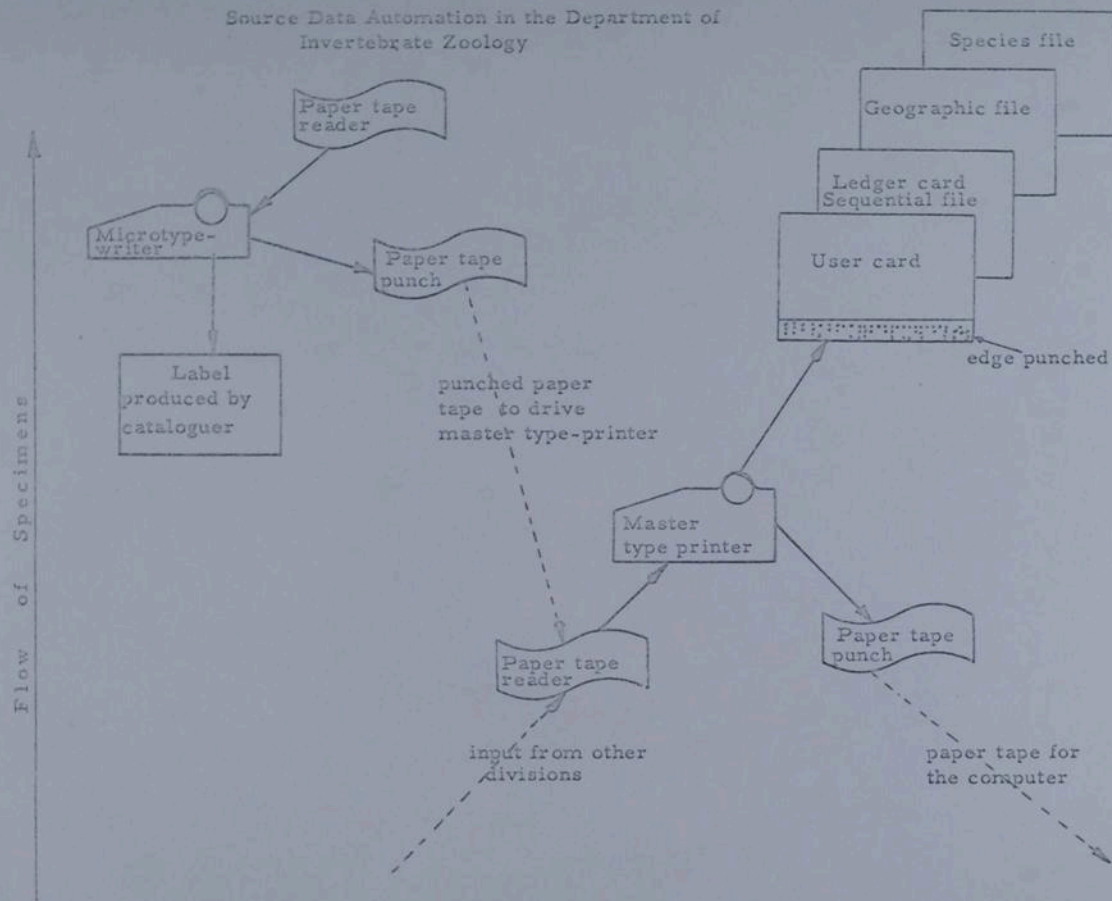
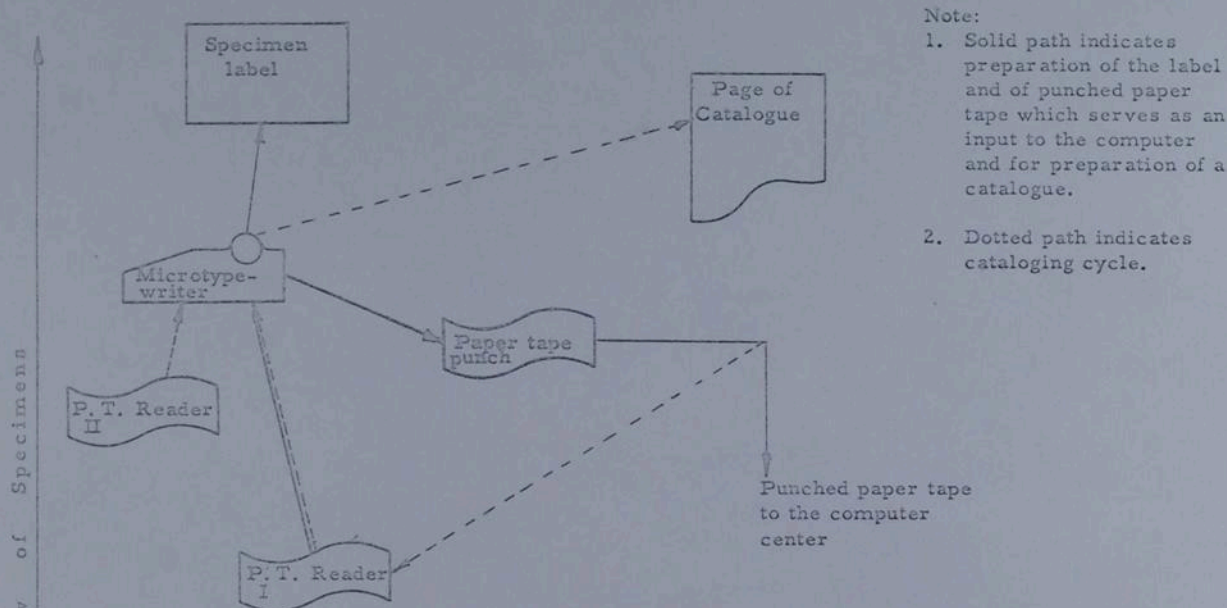


Figure 1

Flow Chart of Source Data Automation in Division of Birds



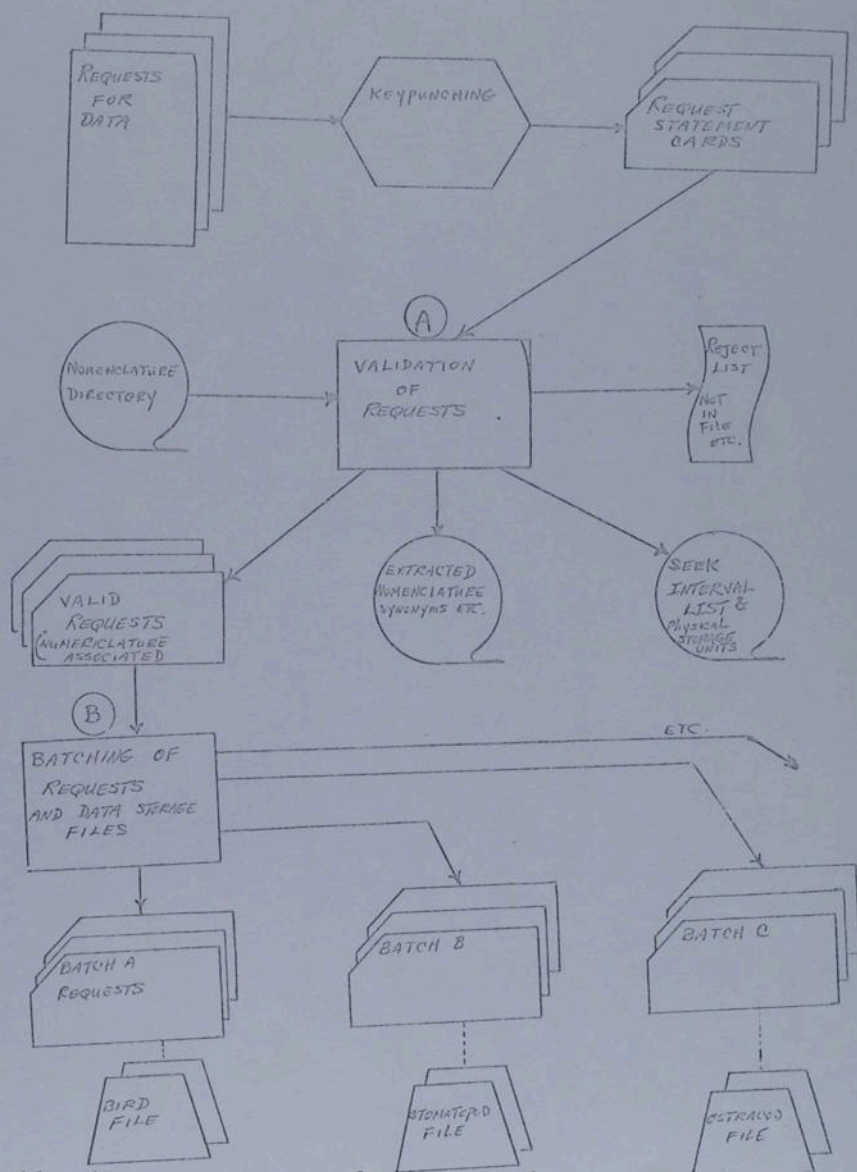
P. T. Reader I has a control loop for the control of the microtypewriter's carriage during the preparation of label.

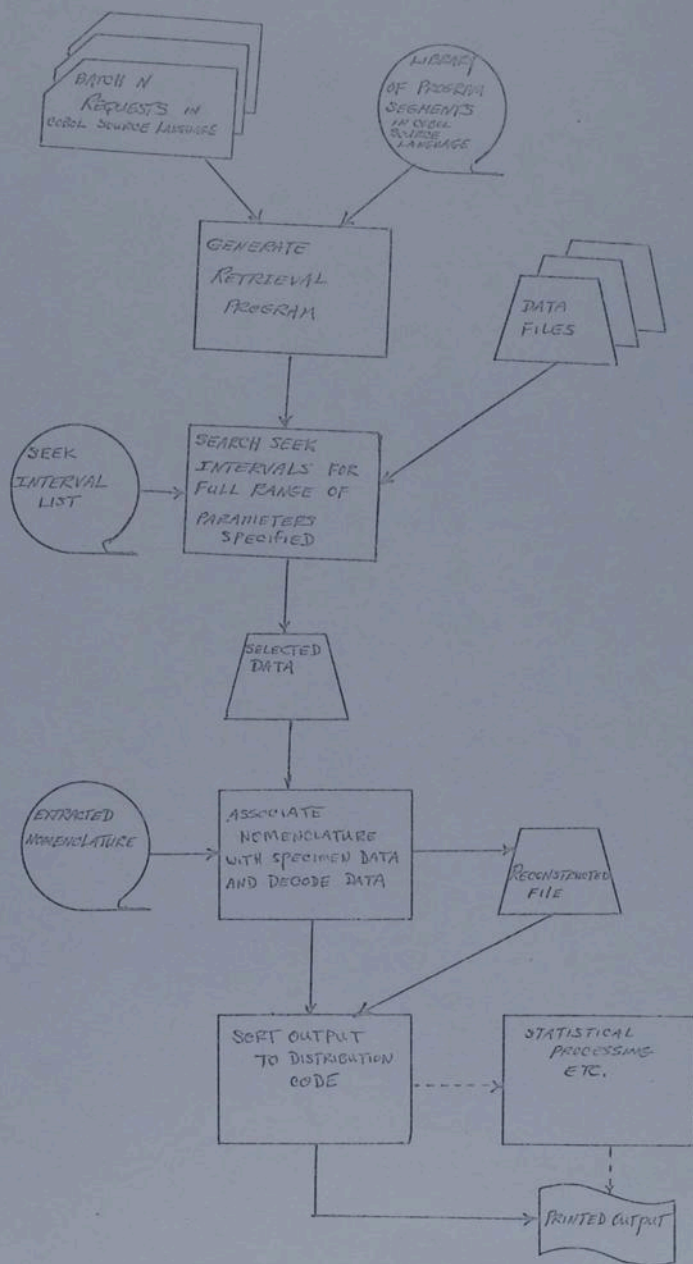
P. T. Reader II contains reiterative loop which reproduces data repeated frequently in a cataloguing sequence. The punched paper tape is fed back through the system to produce a standard catalogue in microtype. During cataloguing cycle, paper tape punch is disengaged.

Figure 2.

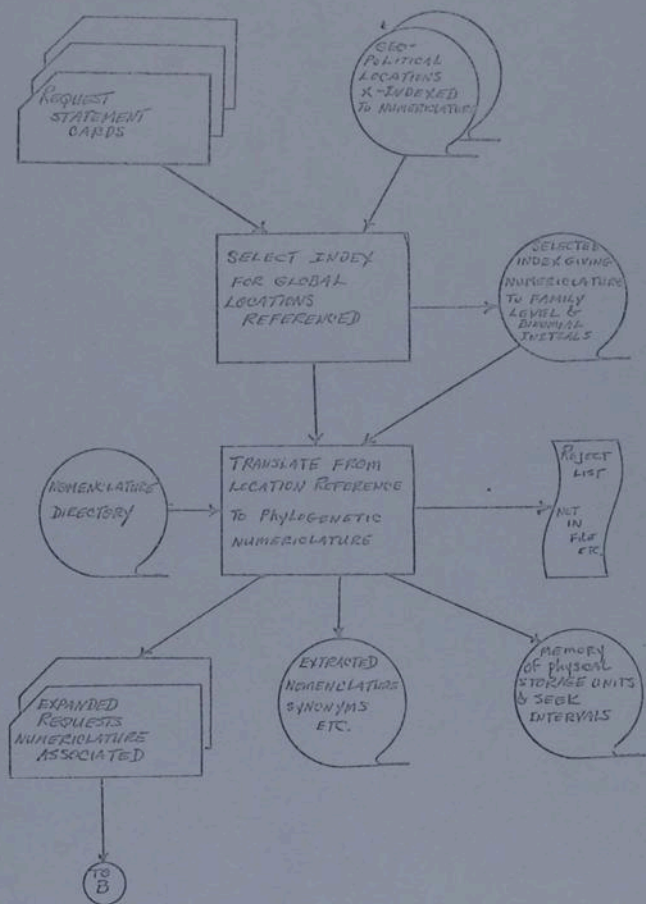
APPENDIX I

QUERY PROCESSING 1. BATCHING

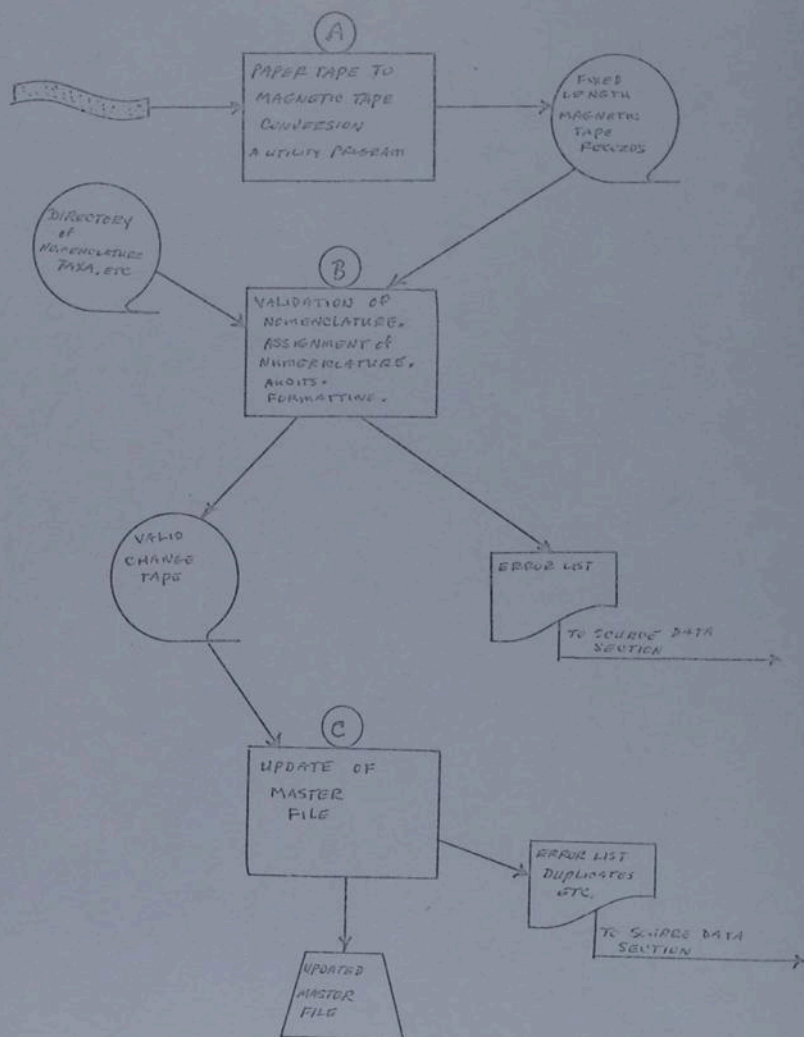




PROCESSING REQUESTS WHERE PHYLOGENETIC NAME IS NOT GIVEN
BUT A GLOBAL LOCATION IS GIVEN



MASTER FILE UPDATE



APPENDIX II

Personnel

Dr. Donald F. Squires, Deputy Director, Museum of Natural History

10% of time will be devoted to this project, part of the responsibilities of this position is the development and implementation of a data processing system. Involvement will be in liason and coordination of scientific staff of the Museum and other scientific organizations with systems-development. General scientific guidance will be given.

Born December 19, 1927, New York, U. S. A. ; A. B. Cornell University, 1950; M. A. University of Kansas, 1952; Ph. D., Cornell University, 1955; Assistant and Associate Curator, Department of Fossil Invertebrates, American Museum of Natural History, New York, 1955-1962; Associate Curator and Curator in Charge, Division of Marine Invertebrates, Museum of Natural History, Smithsonian Institution, 1962-1964; Chairman, Department of Invertebrate Zoology, 1965; Deputy Director, Museum of Natural History, 1966 - .

Author of 48 scientific publications in invertebrate paleontology, invertebrate zoology, marine biology, marine geology.

Mr. Nicholas J. Suszynski, Jr., Director, Information Systems
Division, Smithsonian Institution.

20% of time will be devoted to this project in the development and
implementation of the overall system and especially as it applies
to the data organization and search strategies.

Born March 29, 1929, Lomaz-Poland, U.S. Citizen; B.S.I.E.
Syracuse University 1956; Programmer, Sylvania Electric, Camillus,
N.Y. 1956-1957; Computer Systems Analyst-Mathematician, Sperry
Rand Univac, St. Paul, Minnesota, 1957-1958; Programming
Supervisor, Minnesota Highway Dept., St. Paul, Minn. 1958-1960;
Manager Analysis and Programming, General Electric Co., Johnson
City, N.Y. and Bethesda, Md. 1960-1965; Director, Information
Systems Division, Smithsonian Institution, 1965 - present.
Author of three technical papers on simulation and compilers. Author
and co-author of computerized systems for simulation of other
computers; bridge and roadway design; Integrated Manufacturing
Information System; Real Time Command System; Data Reduction;
Numerical Control of Machine Tools and Management Information Systems.
Member of Association for Computing Machinery (organized N.Y.
Southern Tier Chapter), member of the Executive Board and Program
Chairman; Special Interest group for Business Data Processing (ACM);
Finance Co-chairman 1964 Spring Joint Computer Conference
(American Federation of Information Processing Societies)

Mr. William R. Wheeler; Manager, Scientific and Administrative
Applications, Information Systems Division, Smithsonian Institution

50% of time will be applied to this project.

Born June 4, 1937 U. S. Citizen; B. S. Mathematics, Gannon College 1959;
M. S. Mathematics, Stanford University 1961. Mathematical-Analyst,
Lockheed Missiles & Space Company 1959-1961; Mathematician with
U. S. Army 1961-1963; Senior Mathematical Engineer, Lockheed
Missiles & Space Co. 1963-1964; Senior Systems Engineer, General
Electric Co. 1964-1966; Manager of Scientific and Administrative
Applications, Smithsonian Institution 1966 - present.

Directed the development of a SIMSCRIPT compiler for the GE-635.

Technical consultant and instructor for computer sales. Responsible
for the development of the Agena satellite wiring diagram system.
Developed systems to simulate missile trajectories and shock-wave
phenomena. Specialist in scientific and commercial computer
applications.

Mr. Wheeler is a member of the Mathematics Association of America and
the Association for Computing Machinery.

Mr. Reginald A. Creighton, Manager of Information Storage and Retrieval projects, Information Systems Division at Smithsonian Institution
100% applied to this project.

Born 8 December 1933, U.S. Citizen, B.S. Business Administration

Mt. St. Mary's College 1958;

Chubb & Son Underwriters, Systems Research Programmer, Short Hills, N.J. 1952-1956; Ox Fibre Brush Co. Inc., Methods Analyst, Frederick, Md. 1957-1958; National Security Agency, Computer Programmer, Ft. G.G. Meade, Md. 1958-1959; Hq. U.S. Army Data Processing Center, Computer Programmer, Governors Island, 4, N.Y. 1959-1961; Bureau of Yards & Docks, Computer Programmer, BU DOCKS, Washington, D. C. 1961-1962; General Electric Co., Lead Programmer, Bethesda, Md. 1962-1964; General Services Administration, Chief of Operations Branch, Wash., D.C. 1964-1965; Independent Consultant for the ISR techniques at Smithsonian Institution, Wash., D.C. 1966; Smithsonian Institution, Manager ISR Projects, Wash., D. C. 1966 to present.

Author and co-author of computerized systems for telemetry simulation modules, real-time warning system, and in-flight information retrieval systems for Air Force satellites. Developed ISR system for Army personnel records. Other programming includes accounts receivable, payroll, statistical analysis, inventory management and command and control modules for satellites. Authored various technical proposals and systems descriptions, training manuals, etc. Member of the Association for Computing Machinery. Total of 12 years experience at systems design and programming.

Mr. James J. Crockett, Program Analyst, Information Storage and
Retrieval Projects, Information Systems Division,
Smithsonian Institution

80% applied to this project.

Born 27 September 1935, U.S. Citizen.

Course work in mathematics, University of Maryland;

Central Intelligence Agency, Tabulation project planner, Wash., D. C.
1956-1959

Naval Ordnance Laboratory, Senior Systems programmer,
White Oak, Maryland 1959 - 1966;

Program Analyst, Information Storage Retrieval projects,
Smithsonian Institution 1966 -

Author of:

Computerized system for Wind tunnel data reduction,
Mini response ranging and a retrieval system for N. O. L.'s
technical library. Authored NOLTR 64-125; NOI Retrieval
system for the 7090. Member of the Association of Computing
Machinery

Total of 10 years experience at systems design and programming.

Mr. Warren N. Minami, Computer Specialist, Information Storage and
Retrieval Project, Information Systems Division,
Smithsonian Institution

100% applied to this project

Born April 21, 1938; U. S. Citizen; B. S. in Industrial Management
with credits for a minor in Biology, San Jose State College,
California, 1962; 24 credits completed in February toward
Master's Degree at American University

United States Air Force, Management Engineering Officer, Andrews
AFB, Md. 1962-1965; Documentation Inc - NASA Scientific and
Technical Information Facility, Management Analyst, College
Park, Md. 1965 - 1966

Established improved methods for handling scientific and technical
data for ISR by the NASA research community; assisted in documenting
systems designed for, and procedures used in, the manual handling of
data; briefed and oriented foreign and domestic dignitaries on the
operation of the NASA ISR system; conducted training sessions for the
employees on the NASA ISR system; and performed statistical analyses
of the data gathered and advised management of any significant findings.

APPENDIX III

Other Information:

This proposal has not been submitted to any other agency.

The content of this proposal has been discussed with Miss Roena Swanson and Mr. Eugene Kennedy of the Office of Education.

None of the personnel assigned to this project have any current or uncompleted projects under the sponsorship of the Office of Education.

The outputs of the System as proposed will not contain any tabular outputs. The output will consist of a free-flowing response to questions of the researcher. The answer will be in sentence form. This has not yet been developed and therefore, it is not possible to attach samples of the work to be undertaken.

The Smithsonian Institution is an "Establishment" which is under a Board of Regents. The Institution proper, as distinguished from a number of Government Bureaus that have been committed to its administration by the Congress, is a private corporation under the guardianship of Congress.

The Institution therefore assumes a dual role, being both private and Governmental. It performs research projects for other Government agencies as an educational institution under the cost principles of the Federal Procurement Regulations, Section 1-15.3.

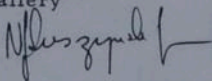
UNITED STATES GOVERNMENT

Memorandum

TO : Mrs. Virginia Purdy
Assistant, Historian
National Portrait Gallery

DATE: March 1, 1967

FROM : Nicholas Suszynski



SUBJECT: The Catalog of American Portraits

Attached is a proposed skeleton for the Catalog of American Portraits. It can be fleshed out by including additional requirements, (in terms of the data to be stored, or the use to which this data will be put).

This proposal was assembled to indicate how certain activities can be aided by the use of a computerized system. For this reason copies of it are being distributed to other Smithsonian museums and galleries as noted below. They may see some of their needs fitting this or a similar approach.

If this description points a way to accomplish some of your work with less effort or makes for a better dissemination of your information, I will feel that it has served its purpose.

cc: J. Bradley, S. R. Galler, F. Taylor, R. Cowan, R. Multhauf,
C. Nagel, D. Scott, J. Pope, J. Magruder, D. Squires, S. Bedini



Buy U.S. Savings Bonds Regularly on the Payroll Savings Plan

THE CATALOG OF AMERICAN PORTRAITS

(General discussion of a procedure for updating and maintaining a quarterly catalog of American Portraits)

In any data processing activity at least three functions have to be accomplished;

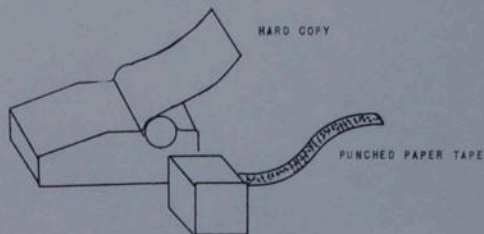
- I. Input data has to be prepared in machine readable form.
- II. A set of programs has to be developed to maintain old data plus the incoming data according to the requirements governing the use of this data.
- III. Reports or publications are produced on an as needed or cyclical basis.

INPUT DATA PREPARATION

At present, National Portrait Gallery maintains a visual card file containing one 8"x10.5" card for each portrait. Each card contains the following information:

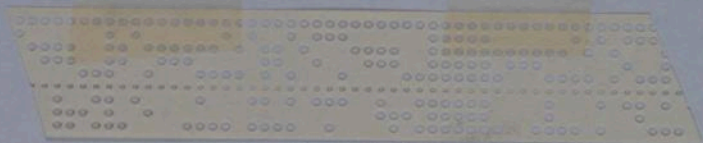
1. CATALOG (Inventory) NUMBER
2. SITTER DATA
 - a) Sitter
 - b) Dates
 - c) Occupation
 - d) Region(s)
3. ARTIST DATA
 - a) Artist
 - b) Dates
 - c) Region(s)
4. DESCRIPTION OF PORTRAIT
 - a) Medium
 - b) Dimensions
 - c) Inscription
 - d) Condition
 - e) Depiction
 - f) Head To
 - g) Hands
 - h) Length
 - i) Pose
 - j) Background
 - k) Clothing
 - l) Other
5. PROVENANCE
 - a) Date of Portrait
 - b) Successive Owners, Addresses, Dates

These 8"x10.5" cards are used to answer day-to-day questions from the public and provide a ready reference file for curatorial purposes. These cards could be prepared on a typewriter with a paper tape punch attachment. They would be prepared as the new accessions are brought in.



Proposed typewriter-paper tape punch unit

National Portrait Gallery can begin its automation with no additional effort if cards and a punched paper tape are generated in a single typing. Even if we were not to proceed any further, at this point we would have the benefit of capturing the data in a machine readable form. (This is what the ADP jargon calls "Source data automation.")



Sample of punched paper tape

Several manufacturers make an electric typewriter with an attachment which punches appropriate code configuration for each key stroke of a typist. By taking this approach, data is typed only once at the time when the visual record is made; in addition, computer readable input is generated automatically without the additional keypunching, (a photograph of such equipment is attached on the last page).

COMPUTER PROGRAMS

Necessary programs have to be developed to do the following:

1. Edit and audit incoming records from the paper tape. (By editing we mean the rearrangement of the data in a form needed for processing. By auditing we mean checking the data for correctness and completeness of information.)
2. Establish the initial data bank and provide the capability to update this Master File with new entries, as well as the capability to modify previously entered information, or when needed, to eliminate a record altogether.
3. Several sort programs to establish necessary sequences of records for printing of outputs as well as for sequencing of data prior to updating of the Master File.
4. Print programs for forming and printing of catalogs and, on request, reports.
5. Search program to satisfy such requests as:

"Provide me with a list of all painters and their sitters who lived in the Eastern United States prior to 1900."

REPORTS AND PUBLICATIONS

It is possible to generate a catalog of American Portraits on a quarterly basis with little difficulty (a sample is attached). This is based on an assumption that for the first few years the catalog would not exceed 10,000 items. It is assumed that each catalog item will require two lines of printing so that the whole catalog will consist of 20,000 print lines.* At forty lines per page this will generate a catalog of 500 pages in one hour of computer's time. At a later time when the volume of data increases to the point where printing of quarterly issues is too expensive, it may be feasible to do a new printing on a yearly basis and to issue a supplement on a quarterly basis.

In addition to the catalog and periodic supplements it will be possible to query the Master File and provide computer generated searches which will be responsive to general questions.

.....

*The user's need will determine format and this will be developed at a later time. The catalog can be listed in any one of several of the following sequences:

1. a) Alphabetically by sitter
b) Alphabetically by painter
c) Alphabetically by geographic location
d) Alphabetically by owner
e) Alphabetically by pose
etc.

2. Numerically by catalogue number.
3. Any combination of the above, i. e.:

Alphabetically by sitter, within a painter, within a owner

COST

Cost can be estimated on the basis of the three functions mentioned on the first page, that is:

- 1) Input data preparation, 2) Systems Development and Programming and 3) Computer time to produce the reports.

INPUT DATA PREPARATION

One SCM 2816 or equivalent typewriter with a paper tape punch and a control device for the carriage return.....\$5,440.
Inasmuch as the typist or a clerk is required to produce visual 8" x 10.5" records, paper tape is generated as a by-product of this operation and no additional clerical effort should be associated with the preparation of the computer input.

SYSTEMS DEVELOPMENT AND PROGRAMMING

Three months of GS-13 and 9 months of GS-11 effort, approximately \$10,000. (Eventually it is hoped that programming and systems development will be provided to the various bureaus of SI according to their needs without charges. At this time, however, our budgeting problems prevent us from offering it on this basis.)

COMPUTER TIME

Computer time will be available at no cost. It will be provided as a part of services from ISD.

DATA USED IN KEY TO NEOTROPICAL SNAKE GENERA

PREPARED BY JAMES A. PETERS

THE FOURTEEN GROUPS OF CHARACTERS LISTED HERE ARE THOSE USED IN THE IDENTIFICATION OF NEOTROPICAL SNAKE GENERA USING COMPUTER KEYS. THE DATA SHOULD BE RECORDED IN THE SAME SEQUENCE AS LISTED HERE. THE DATA IS PUT INTO THE PROGRAM IN LINE 5000, FIRST TYPING THE WORD "DATA," AND THEN ENTERING THE 14 CHARACTERS, SEPARATING EACH FROM THE OTHERS BY A COMMA. THE LINE BELOW IS A SAMPLE DATA LINE:

5000 DATA 2, 2, 17, 1, 2, 2, 1, 0, 3, 0, 136, 2, 2,

CHARACTERS:

1. CONDITION OF BODY SCALES.

- 0 = VENTRALS AND DORSALS SAME SIZE
- 1 = VENTRALS FEEBLY ENLARGED, DO NOT COVER BELLY
- 2 = VENTRALS MUCH ENLARGED, DORSALS IN HORIZONTAL ROWS, SUBEQUAL
- 3 = VENTRALS MUCH ENLARGED, DORSALS IN HORIZONTAL ROWS, VERTEBRALS MUCH LARGER THAN PARAVERTEBRALS
- 4 = VENTRALS MUCH ENLARGED, DORSALS DIAGONAL, SUBEQUAL

2. FACIAL PIT BETWEEN EYE AND NOSTRIL.

- 1 = PRESENT
- 2 = ABSENT

3. MIDBODY SCALE ROW COUNT.

RECORD THE ACTUAL NUMBER OF ROWS AT MIDBODY

4. SCALE ROW REDUCTIONS.

- 1 = SAME NUMBER OF ROWS THROUGHOUT BODY
- 2 = SCALE ROW COUNT ABOUT ONE HEAD LENGTH BEFORE ANUS AT LEAST TWO LESS THAN AT MIDBODY

5. ANAL SCALE.

- 1 = SINGLE
- 2 = DIVIDED

6. KEELING OF SCALES.

- 1 = SOME OR ALL SCALES KEELED
- 2 = ALL SCALES SMOOTH

7. LOREAL PLATE.

- 1 = PRESENT
- 2 = ABSENT
- 3 = TWO OR MORE PLATES IN LOREAL POSITION

8. ANTERIOR TEMPORAL PLATES.

RECORD ACTUAL COUNT ON ONE SIDE, IN WHOLE NUMBER, UNLESS AREA HAS LARGE NUMBER OF SMALLER SCALES, NOT PLATES, IN WHICH CASE RECORD AS "10".

9. SCALE PITS.

- 0 = ABSENT OR NONE
- 1 = SINGLE PIT PRESENT
- 2 = TWO PITS PRESENT

10. COLOR PATTERN OF BODY.

- 0 = UNICOLOR
- 1 = STRIPED OVER ENTIRE LENGTH OF BODY
- 2 = COMPLETE BANDS OR RINGS AROUND BODY
- 3 = OTHER THAN ABOVE

11. CONDITION OF TEETH ON MAXILLARY BONE.

- 1 = NO GROOVES, NO DIASTEMA, LAST TWO NOT ENLARGED
- 2 = NO GROOVES, NO DIASTEMA, LAST TWO ENLARGED
- 3 = NO GROOVES, A DIASTEMA, LAST TWO ENLARGED
- 4 = GROOVES OR CANALS ON ANTERIORMOST TEETH
- 5 = GROOVES OR CANALS ON POSTERIORMOST TEETH
- 0 = DOES NOT FIT ANY OF THE ABOVE

12. NUMBER OF VENTRAL SCALES.

RECORD ACTUAL COUNT FOR VENTRALS.

13. CONDITION OF PREFRONTAL PLATES.

- 0 = ABSENT
- 1 = ONE SCALE (I.E., FUSED)
- 2 = TWO SCALES (I.E., NORMAL CONDITION)
- 3 = THREE SCALES
- 4 = PREFRONTALS FUSED WITH INTERNASALS

14. SUBCAUDAL SCALES AND TAIL.

- 1 = SUBCAUDALS SINGLE, TAIL TIP NORMAL
- 2 = SUBCAUDALS DOUBLE, TAIL TIP NORMAL
- 3 = POSTERIOR SUBCAUDALS FINELY DIVIDED, TIP NORMAL
- 4 = TIP OF TAIL WITH RATTLE

IF ANYONE WISHES TO TEST THE PROGRAM, HE CAN SEND THE DATA ARRANGED AS ABOVE, WITH IDENTIFYING NUMBERS IF MORE THAN ONE SPECIMEN, TO ME AT THE U. S. NATIONAL MUSEUM. OR, IF YOU HAVE ACCESS TO A TIME-SHARED COMPUTER USING THE "BASIC" LANGUAGE, I CAN SEND YOU A TAPE OF THE PROGRAMS FOR USE IN YOUR MACHINE.

(THIS STATEMENT IS ITSELF A COMPUTER PRODUCT. THE "BASIC" LANGUAGE NOW INCLUDES A PROGRAM CALLED "FORMAT***", WHICH IS USED IN THE PREPARATION OF MATERIAL TO BE REPRODUCED. THE PROGRAM ADJUSTS MARGINS, MAKES THE RIGHT HAND JUSTIFICATION SEEN HERE, AUTOMATICALLY ADJUSTS SPACING, BOTH OF LINES AND PAGES, AND CENTERS HEADINGS. EDITING IS VERY EASY, BECAUSE ERRORS CAN BE REMOVED BY RETYPING ONLY THOSE LINES IN WHICH THEY OCCUR. THE COMPUTER THEN INSERTS THOSE CORRECTED LINES WHERE THEY BELONG.)

SMITHSONIAN INSTITUTION
JUNE 6, 1967

THE ROLE OF TIME-SHARE COMPUTING IN MUSEUM RESEARCH

JAMES A. PETERS

Division of Reptiles, U. S. National Museum
and

BRUCE B. COLLETTE

Bureau of Commercial Fisheries Ichthyological Laboratory
U. S. National Museum

Hidden behind the dinosaur skeletons, Corot paintings, woodburning locomotives, Chinese jade, and other familiar sights of public and private museums are the research laboratories where work on the museum's collections takes place. As is true of most other large museums, the Smithsonian Institution resembles an iceberg: the public displays represent the one-ninth visible at the surface and the collections and research represent the submerged eight-ninths. Some of the earliest users of statistics in biology were museum investigators, faced with the problems of variation within and between populations, and species of plants and animals. The visitor to the research areas of a museum can expect to see desk calculators in almost every laboratory, and if he were to check the publications issuing from those laboratories he would find that statistics have played an important role in the decisions made and presented to the scientific community. It would seem, then, that the computer should have been immediately incorporated into museum research, because of the ease with which it can handle tasks that consume days and weeks of time for the average investigator.

This has not been true. Although a few workers have been able to use computers in the solution of highly esoteric problems which have been widely touted as panaceas for the problems of the field of systematics, by far the majority of us have continued to use our desk calculators and ignore the potential of the computer. This has not always been a matter of lack of interest. One of the reasons we have hesitated to go into extensive use of computers is the magnitude of our problems. We are often involved in the calculations of certain statistical values for comparatively small samples representing different species, different populations within those species, different sexes within the populations, and different age groups within the sexes. In the analysis of our material we are likely to want to make preliminary comparisons, study the results, recombine and substitute data on the basis of those results, add new material as it becomes available, and so on. In batch-processing on a central computer, this procedure does not work--or, at least, we have not found it practical. It was much easier to turn repeatedly to the desk calculator, which gave us the flexibility we needed. Occasionally, when we did find ourselves with large quantities of data for processing, perhaps of a magnitude that made the desk calculator impractical, we turned to the electronic computers, only to find a new source of frustration. We seldom had programs for the

type of statistics we wished to obtain, and few programmers understood what our data meant biologically. On these occasions we have temporarily shelved problems because we could not get programs written to give us the necessary answers. Finally, of course, the cost of computer services was almost automatically prohibitive to most museum workers, many of whom operate on small maintenance budgets, supplemented when possible by research grants from private corporations or governmental agencies.

These problems disappeared for the research workers at the Smithsonian recently when a connection to a multi-access computer was installed. At the time of this writing, three teletypes have been available for 6 months, and staff use has skyrocketed. The size of the problem is no longer a handicap. We do not hesitate to run the smallest samples as well as the largest. The small samples, which we never would have considered for computer analysis because of cost, are handled so much more rapidly through time-share computing that the savings in man-hours more than equal the minimum expense of the charges for time used. Since the data are stored on paper tapes, new material can be quickly added and the analysis run again. The development of time-share computing has brought three major benefits to the staff. The first is the release of time formerly spent on the desk calculator, permitting more time for evaluation of results;

the second is the increased magnitude of the problems that can be undertaken; and the third is that the value of characters can be tested during a study, rather than after it has been completed.

Perhaps equally significant as the time-share facility itself, to those engaged in museum research, is the availability of simple languages for programming. Scientists at the Smithsonian have been able to learn to write programs in "BASIC" very rapidly. Several staff members were already familiar with FORTRAN and ALGOL, and the transition for these individuals to time-share computing was practically instantaneous. For those of us without such previous background, BASIC has provided an independence of action never before experienced in our relationships with computer operations. Starting cautiously with simple standard statistical calculations such as standard deviations and correlation coefficients, we have begun to discover the enormous potentials that lie in time-share computing for the relief of some of the more time-consuming museum duties. We have developed a series of programs, many of which use

the same format for data input, thus allowing us to run several different analyses with a single paper tape of data. We are moving more and more rapidly in the direction of computer use for the solution of all kinds of statistical problems encountered in museum work that have not been approached in this way. And we are beginning to visualize the magnitude of the changes that can result from widespread use of multi-access computers by museums around the country. The latter is perhaps the most exciting possibility of all.

Use of Time-Share Computing at the Smithsonian Institution

A review of the various ways in which the multi-access computer is being used at the Smithsonian not only gives some indication of the value of time-sharing at the Smithsonian Institution but also shows the diversity of the research being carried on there. Utilization falls into three principal categories at the present time--data processing, data storage and retrieval, and text editing. Future development of several other possible uses is planned.

The first category, data processing, is unquestionably the most significant and active area of utilization at present. Most of the data processed come from studies on the biology of plant and animal species; the comparatively simple analyses are designed to show the existence or absence of differences between samples of populations

from all over the world. Both of us have used the computer for the calculation of a series of values (standard deviation, standard error, mean) for Hubbs-Perlmutter diagrams (Figure 1), which can then be plotted directly from the teletype readout. One of us (Collette) is using programs for regression analysis, analysis of covariance, and plotting of $\bar{x}$ and $\bar{y}$ co-ordinates in a taxonomic study of the needle-fishes. This work provides a basis for distinguishing previously unrecognized populations and species, and makes it possible to allocate a newly collected specimen to the proper species by checking its position on a computer-generated graph (Figure 2). Maurice E. Salmon, of the Conservation-Analytical Laboratory, Museum of History and Technology, has written several programs to facilitate data reduction in that laboratory. The first dealt with the characterization of the spectra of X-ray spectrographic tubes, and involved about 200 observations per tube spectrum, including the angle of diffraction, the observed intensity, and the intensity incident on the sample. Salmon calculated that hand reduction of the data would require 50 to 60 hours; the time-share computer solution was completed in less than four hours. A second program was developed by Salmon to determine and apply sets of empirical coefficients to correct quantitative X-ray fluorescence data for matrix effects, which involved the solution of n simultaneous equations for n unknowns. Nicholas Hotton, of the Division of Vertebrate Paleontology, is analyzing a

series of 50 measurements of skeletal parts of modern lizards in the hope that information gained concerning intra- and inter-specific variation can then be extrapolated to analysis of large series of dicynodont reptiles collected in South African Permian deposits. Paleontologists are not customarily so fortunate as to uncover fossils in large series, but occasionally it happens, and Hotton hopes to establish some basis for determining specific limits in fossil reptiles through his analysis. W. T. Chase, of the Research Laboratory, Freer Gallery of Oriental Art, used the computer for weight conversions of 120 Chinese bronze ceremonial vessels. He also used the computer to calculate the d-spacings, from X-ray diffraction measurements taken in a 2-radian powder pattern camera, using a program he wrote himself. Martin A. Buzas, of the Division of Invertebrate Paleontology, is processing data consisting of proportions expressing the relative abundance and diversity of species of Foraminifera in samples from the Arctic to the Gulf of Mexico, by using the Shannon-Weiner information function. Foraminifera are one-celled animals whose cast shells are important both in their intrinsic interest to science and as valuable indicator species for oil-bearing layers of the earth. Klaus Rützler, of the Division of Echinoderms, Department of Invertebrate Zoology, employs a program that provides information on the zoogeographic similarities and differences between sponge samples. Henry Setzer, Division of Mammals, uses the computer in the analysis

of the mammals of South Africa, where his work is supported by a United States Army contract because of its significance in work on arthropod-borne viruses and their relationship to human infectious diseases. Ernest A. Lachner and Martin Wiley, of the Division of Fishes, have used the computer in analyzing changes in body proportions and number of breeding tubercles in hybrids and parents of a genus of minnows. In addition to these major examples, perhaps a dozen other scientists at the Smithsonian have used the time-share service for data reduction.

The second category of Smithsonian use of time-share computing, data storage and retrieval, is particularly useful in routine identifications. In the Museum of Natural History, with its collections of over 50 million specimens of plants and animals from all over the world, staff scientists are inundated by specimens of all kinds for which identification is required. This situation exists, of course, in every large museum in the world, and is a problem shared by the entire museum community. Often the identification is a very simple procedure, involving nothing more than a glance by the specialist who studies the particular kind of organism presented, who uses the built-in "computer" which he has been cramming with data throughout his professional life. But it is still possible, in many groups of organisms and in many parts of the world, to collect large amounts of material that require many hours of work and study for accurate identification. When such a

collection arrives at the museum, a series of events begin that culminate in a name being assigned to each specimen. These events include, first, a review of the literature both on the area concerned and the major groups or organisms included; second, the use of "keys," a standard tool for identification, to arrive at tentative names for the specimens; and third, the comparison of the specimens with material already in the collections as well as with previously published descriptions to confirm and validate the tentative assignment of a name. Obviously, the first and third steps depends upon access to an adequate library and collection, and these facilities have been developed at the Smithsonian over the years. The assignment of a tentative name, however, is a different matter. If a previous worker has studied either the geographical area or the species concerned, it is probable that he has published papers which can be used for this second step. Frequently, however, either no such work has been published, or, if it has, no keys have been included, and one must start from the available literature and construct his own identification keys. Published keys go out of date fairly rapidly as a consequence of additional study and publications, and it is difficult to keep up with the changes except within narrow limits. The use of a key is sometimes rather tedious, since keys are almost invariably constructed on a dichotomous "either-or" basis, and one must proceed from the beginning through a long

series of alternatives to an identification (although eventually the key becomes so familiar that the early steps can be omitted). The identification procedure appeared to be an area where storage of such sequential information in a computer memory bank, followed by challenge of the memory with data drawn from individual specimens from new collections, might be practical and time-saving. One of us (Peters) decided, therefore, to see if it might be possible, by using BASIC programming, to translate into machine language a standard dichotomous key used in the identification of snakes from Central and South America and permit the computer to select and reject alternatives important in making identifications. It was soon found that the simplest procedure was to permit the computer to select from a series of alternatives rather than from only a pair, and 14 groups of characteristics useful in the identification of snakes were encoded. After the input of these 14 encoded items, the computer prints out a generic name in a few seconds of computer time. Occasionally the computer lists two names for genera that cannot be distinguished on the basis of the series of 14 characters, and it provides additional information that will differentiate between them. In rare instances the machine prints two or more names for which the difficulty of individual recognition is too great for easy encoding in the program. The user must then return to the literature for additional information on determination of those genera.

It will be immediately recognized that this new program tends to reduce greatly the time and effort spent in the use of keys for identification. If such a machine key is available, it also reduces to a large degree the time required for literature search (the first step, above). It should be strongly emphasized, however, that the computer does not relieve in any way the final step--confirming the identification--because any identification obtained from either published or machine keys must always be considered tentative until verified by comparison with literature descriptions and previously identified specimens. We might add parenthetically that some of the biggest headaches in systematics come from investigators who do not follow through on the confirmation step before they publish information based on their use of keys.

The third principal category of use of multi-access computing by the Smithsonian staff, text editing, is based on a special program, that permits the preparation of text material for review and correction. It incorporates specific printing instructions, which include such things as right- and left-hand margin adjustment and justification, centering of headings, indentation, line spacing, and automatic pagination. The BASIC language makes correction of errors extremely simple, since only the program instruction line containing the error must be retyped, and the computer automatically substitutes the

corrected line in its proper place. We have used this program to produce letter-perfect copy, which is then photolithed for distribution. The only handicap at present is the teletype itself, since it provides copy only in capital letters. We expect an early solution to this difficulty, however, if the future proves that this is a practical way to produce manuscript material for large-scale reproduction and distribution. The program has already been used to provide material for distribution at lectures on machine programming of systematic keys for specimen identification.

The Future of Multi-Access Computing in Museums

Insofar as we know, the Smithsonian is presently the only museum in the country with a direct connection to a commercial time-share computer. Several university museums have in-house time-share operations available through the computer facilities of the university, but none of them have actually installed the teletypes to make the facilities directly available to the research scientist in his laboratory. We know of one investigator who tried BASIC programming and found it too restricted for the magnitude of his computer-generated "numerical taxonomy," but it is our experience that most systematists today are not engaged in that level of computer use. On the contrary, there are many workers who use no statistical methods in their taxonomic work, and who have developed a distinct distrust of all computer work as a

consequence of the extravagant claims made by the proponents of numerical taxonomy. We believe that multi-access computing will be of greatest advantage to this group as well as those workers like ourselves who have devoted many long hours to the computation of measures of distinctiveness between samples. For the first group, it provides a painless introduction to the computer, followed by an easy access to information and data never before used in the analysis of their organisms and the relationships thereof. For the second group, it means relief from hours and days of work, plus the opportunity to do some analyses previously considered impractical because of the time required. The competitive aspects of the time-share computing business, the reliability and ease of access to the computer, the truly low cost, and the rapidity with which innovations, equipment, and programs are made available to users, all combine to make multi-access computing a genuine blessing to museum research.

Time-share computing is not, however, an un-mixed blessing. We are already beginning to discover some of the shortcomings, including the limitations on storage capacity of the memory disk; the comparative slowness of the teletype in input of data and output of results; the restriction on the length of an individual program, with the attendant limitations on the total amount of data that can be processed in an individual problem; and an occasional busy signal on the teletype line. These problems, we suspect, are going to be common to all users

of multiple-access computers, however, and as a consequence, will be among the first subjects for attack by commercial suppliers. We firmly expect all of the benefits of a truly cut-throat competition to come to the nation's museums!

We also expect a comparatively rapid development of computerized keys for identification, as described above. The evolutionary sequence that development of this aspect of museum "computerizing" might undergo can be predicted. First of all, it is to be expected that specialists, once they gain their initial access to a time-share computer, will develop machine keys focused on the organisms currently under their intensive study. These keys will be tested for accuracy and validity, and then translated back into a series of dichotomous choices for publication in revisions or monographs. Since, however, a key submitted for publication immediately begins to go out of date because of continued investigation, the specialist will find it useful to maintain his computer key, inserting corrections and additions as they are published. We expect that workers in other institutions will soon begin to take advantage of the up-to-date information stored in a colleague's computer, by asking for a taped copy of the key for use in their own machines, rather than trying to use a published and already slightly obsolescent key. Soon there should be rather widespread exchange of computer tapes between museums with compatible systems, and this exchange eventually will result in an over-loading of the memory

systems of the computers available to the individual users. If the amount of such exchange developed a sufficiently large volume of memory bank material, it might then become practical for many museums in the United States to consider forming a common network, all with access to a mutual store of statistical programs, keys, and other such repeatedly used systematic information. Under these conditions, the specialist would continue to maintain the computer key for the organisms he knows best, and all other workers would have immediate access to that key by calling it out of the common storage. We further predict that ultimately there will no longer be much purpose or reason in publication of the systematic information that could be stored, because of its ready availability to workers in an up-to-the-minute state. We even expect that this development might lead to extensive revision in publication policies of museums in general, in some ways cutting down on the flow of publications, and in other ways, permitting a general improvement in the quality of the published materials as a consequence of eliminating repetition. Eventually, field and laboratory data used by taxonomists could be stored in paper tape files (similar to the storage of original data by the American Documentation Institute) and made available in a form ready for computation by anyone wishing to check the results of a previous investigator. The consequences of these changes in mode of action would be, to say the least, far-reaching in the museum world.

Although it is not part of the computer, the teletype machine and its telephone connections might also be of future significance to research workers in museums. One of the sources of delay in our work today is the necessity for correspondence concerning material in another museum. A manuscript may be delayed for a month or more while one awaits a single vital piece of catalogue information or the supplying of a minor yet important item of data concerning a specimen requiring re-examination by the local curator. The teletype would provide transfer of this information in a matter of seconds, and would facilitate the transmittal of lists of specimens, numbers that might be copied incorrectly when received by telephone, and other such data and facts. Editors could request that their authors verify citations or supply missing material from manuscripts, rather than be faced with the choice of searching out the material themselves or waiting for an answer to a letter. Although this aspect of time-share computing could be made available to museums without the computer connection, it is highly unlikely that a network for information transmittal would develop independently of the much more important facet of computer access because most museums would not underwrite the cost of a teletype simply for the time it would save in intermuseum communications. It can be achieved only as a byproduct of time-share computing, and thus becomes an important aspect of it.

Conclusions

It should be clear that we do not share the pessimism concerning time-share computing that has developed in some areas of the computer business and that has been expressed in some recent articles in DATAMATION. We have found the facility immensely useful, and have high hopes for the future development of the industry and its application to our activities. If the perhaps extravagant predictions of some supporters of multi-access computing are realized, and the teletype connecting to a central computer becomes as commonplace in business as the cash register or typewriter, we hope that universities and museums will be found in the forefront in the development, rather than following along behind commercial development. Our potential uses are likely to be much more diverse and manifold than those of the average business, and it will be unfortunate if we do not develop them as rapidly as feasible.

Legends

Figure 1. Hubbs-Perlmutter diagrams of the number of predorsal scales in five species of Strongylura. The horizontal line represents the range, the vertical line the mean, the black rectangle two times the standard error on each side of the mean, and the open rectangle one standard deviation on each side of the mean.

Figure 2. Scattergram of head width on body length for two populations of Platybelone argalus from the Atlantic Ocean. Open stars represent specimens from the islands of Annobón and Fernando Póo in the Gulf of Guinea. Closed stars represent western Atlantic specimens. The regression curve is a second degree polynomial based on western Atlantic specimens.

Introduction

The attached notes are intended to outline the problems which exist in the storage and retrieval of paleontologic data. They are not complete in detail and have not benefitted from wide circulation and criticism. Nevertheless, they should serve as points of departure for system analysis, the purpose for which they have been written.

Paleontologists deal with several different kinds of data, wholly aside from the subject matter itself.

I. Most important in terms of volume is 'sample-oriented' data. Under this heading would come most information in systematic papers in which taxa were described from a number of localities and samples.

II. Probably most important from the point of view of useful information is 'synthetic' data; statements of opinion, conclusions, and discussions of topics which may be totally unrelated to particular fossils, samples, or localities. Synthetic data, when given in conjunction with sample-oriented data must be recoverable within both sample-oriented and synthetic search systems.

III. 'Morphologic' data pertains to the description of individual taxa and taxonomic categories. It will generally be logically independent of, though associated with, sample-oriented and synthetic data, although in some cases sample-orientation will be an important attribute of the morphologic data (for example, unusual specimens from a certain sample).

IV. 'Curatorial' data will pertain to the physical position of specimens in collections. This is an important house-keeping body of information and will be closely tied to sample-oriented data.

Most of the information which will be entered under the broad headings of I through III will come from the published record, but unpublished information of various sorts will also enter under them.

I. Sample-oriented data

A. Bibliographic data ... will include

1. author's name
2. year of publication
3. citation..title, journal, volume number, pages.

Notes: Each author-publication pair is unique and will serve to index the data obtained from the particular publication. The CITATION serves to completely identify this pair but will not be used to enter the main body of the file.

B. Geographic location

1. data given by the author
 - a. narrative or map description of localities ... for example township and range, mileages.
 - b. locality numbers (often)
2. data to be entered in primary file
 - a. latitude and longitude locality descriptor, or other standard method to be adopted Survey-wide.
 - b. locality numbers, beginning with 1 for each author-publication pair.

Note: Both of these items of primary input must be cross-indexed to each other and to the author-publication pair.

3. data to be entered in secondary file
 - a. author's narrative or map description of the locality.
This will supplement the primary data but will not be used

to enter the main file.

4. Potential uses of locality data

- a. to create lists of all localities from a particular area, with the area defined by latitude and longitude.
- b. as the frame of reference for sample information and the various types of information to be related to samples.

C. Stratigraphic position of samples

1. Data given by the author

- a. locality from which the sample came. Note: A locality is a geographic point and a sample is material or information collected at that point so that one or many samples may come from a single locality. (Sample numbers are to be sequential for each locality and none will be the same as the locality number except by coincidence.)
- b. distance of the sample above or below a particular horizon. This is usually a point in a stratigraphic column but may span a greater interval, especially in wells.

2. Data to be entered in primary file

- a. sample number (assigned at time of input and cross-referenced cross-referenced to locality number and author-publication number, thus making it unique.

3. Data to be entered in secondary file

- a. Original sample number and details of stratigraphic position, in words, or feet, or meters.

4. Potential uses of sample data

- a. samples will be compared with each other on the basis of fossil content, lithology, environmental attributes, and probably most of the other parameters which are sample-oriented.

Note: The combination of author-publication, locality, and sample specifies a single entity and is basic to all other sample-oriented information.

D. Rock units

1. Data given by author

- a. group
- b. formation
- c. member
- d. bed, tongue, lentil

2. Input in primary file

Separate positions must be maintained for each of these categories. Within each category the particular rock units must be coded. If something other than the complete alpha name is used then a secondary file for printout of the alpha name will be needed.

E. Sedimentology

1. The author may describe the sediment in detail in terms of:

- a. sorting and related parameters
- b. mineralogy
- c. structure and fabric

At present inclusion of such information in reports on paleontology is limited, and therefore the categories of potentially frequent use are not clear, so provision must be made for expanding this area with time.

F. Lithology

1. Data which may be given by the author
 - a. basic rock type (such as shale, marl, sandstone)
 - b. qualifiers and modifiers (such as fissile, dark brown, fine grained)
2. Input in primary file
 - a. basic rock type and one or two modifiers, perhaps color and bedding.
Both input and output should be in alpha.
3. Input in secondary file
 - a. complete lithologic description

G. Temporal and biostratigraphic terms

1. Data provided by the author
 - a. period, epoch, and age
 - b. system, series, stage, zone
2. Input in primary file

Each of the terms, if given, must be entered. Both sets of terms are heirarchical, and the terms will be applied to individual samples or groups of samples.

Note: The format used here should be the same as or compatible with that to be used by the Survey as a whole; this is also true of other interdisciplinary categories such as Rock Units, Lithology, and Sedimentology.

H. Taxonomy

1. Data provided by the author

- a. the hierarchical series: phylum, class, order, family, genus, species, subspecies, variety of forma.
- b. the name of the original author, and the date of first publication, of the genus, species, subspecies and family.
- c. synonyms for the particular genus-species combination being described.

2. Input in primary file

- a. the primary taxonomic designation, in terms of the hierarchy given above (or as much of it as is provided by the author).
- b. individual synonyms under the primary name. Each synonym is in the same format as the primary name, but each is subordinated to the primary name.

The system must be able to reflect:

- i. that the name is used as a synonym by author X, to whom it is cross-referenced.
- ii. that the author has related all synonyms placed under a primary name to each other, as well as to the primary species.

iii. why each synonym was put in synonymy (if this can be determined)

3. Requirements and difficulties

By definition each synonym is a primary name to someone else. Somehow we must thread together what each author calls primary and secondary species (or genera, or subspecies) so that we may:

- a. ask for all synonyms cited by all authors for any particular taxon.
- b. ask for all species or subspecies or genera into which a particular species has been placed by the processes of synonymy.
- c. ask for a complete list of all species (for example) which have been placed in a particular genus as primary taxa, as synonyms, or as both.
- d. it should be possible to interrogate the file by sorting for:
 - i. any taxon, either singly or in combination. For example; family X, or family X, genus Y, species Z Jones.
 - ii. a particular taxon-author combination: genus Y Smith, or genus Y species Z Jones.
 - iii. for relationships between taxa and other information in the system, such as lithology or age.
- e. A further refinement at the taxonomic level is the inclusion, often-times, of abundance data pertaining to individual taxa in individual samples. These data may be given in absolute numbers where the sample is standardized in weight or volume,

or it may be in relative terms based on estimates of abundance. These data, where given, must be annexed to the taxon-sample-locality-author chain. A means for encoding all of the various methods of representing abundance must be developed.

II ... Synthetic Data

A large variety of data categories which are interpretive rather than factual are almost always present in paleontologic papers. Some of these categories are:

- correlation
- paleogeography
- paleobiology
- phylogeny
- paleoclimatology
- evolution

It is difficult to anticipate or enumerate all of the actual and potential categories. Therefore the indexing system must be both open-ended and capable of revision. Personal freedom is orders of magnitude greater for expressing interpretive as opposed to factual conclusions, and thus subdivisions under major categories like those listed above will be even less well-defined. It may be best to rely on alpha summaries of the information rather than to try to code everything.

For example under 'Phylogeny' it might be best to use 'development of Globigerina from Hedbergella in latest Cretaceous' than to use code.

There will be inter-relationships between the various categories of synthetic data, and also with sample-oriented data. For example, because correlation involves both rock and biostratigraphic units it will have to be related to these categories. The correlation may be made with respect to specified localities or samples only so that connection to these categories must be established.

III Morphologic Data

A. All fossil organisms are described morphologically by verbal descriptors which are supposed to impart a succinct mental impression of the feature described. Some descriptions may contain dozens of such descriptors and several times that many modifiers. Because of the large number and continuing increase of descriptive terms an open-ended glossary, capable of revision, must be established. Each term in the glossary will have to be defined with respect to how it is used in the system.

I know of no one in the P and S Branch who is presently interested in putting morphologic data into a storage and retrieval system, but it would be desirable to build the capability to do so into the system.

B. Biometry

Many paleontologic papers include measurements, numerical summations, statistical results and other such data which many users will wish to record and recall. Such data will be related to taxa, which in turn will be related to particular samples, localities, and other data categories. The variety of this numerical data is great and will evolve, so that it will be necessary to allow each specialist to specify his categories, define the importance or sequence of the particular measures he deals with, and add, delete, or change the emphasis of these measures in the future.

IV. ... Curatorial Data

The name of the depository in which a particular specimen is stored, and the floor, the case, and the drawer in which it is to be found are what must be indexed under this heading. Much more than an indexing system is implied here, however, because present practices of sample storage and the current state of the collections will have to be assessed before any system can be proposed.

For specimens which have been described in publications, or otherwise specified in unpublished reports, notebooks, etc., it must be possible to cross-index them with the appropriate sample-oriented data.

Conclusion

In broad outline these are the kinds of information which are needed in a paleontologically oriented storage and retrieval system. Each requirement undoubtedly places restrictions on the possible overall solution, from the technical point of view. Specifics and numerous examples of each of the requirements can be supplied, but this is probably best delayed until at least preliminary systems analysis has provided a set of hypothetical solutions for consideration.

Of overriding importance to everything that has been outlined here is the requirement that any proposed system be easy to use by the average scientist, without requiring either much of his time or elaborate expertise. Ideally it should be possible for a reasonably intelligent secretary to do nearly all of the input work.

James F. Mello

SEP 28 1967

A BIBLIOGRAPHIC DATA RETRIEVAL SYSTEM
FOR BOTANY

Sergio Ahumada R.
Centro de Calculo Electronico
Universidad Nacional Autonoma de Mexico

25 August 1967

This paper is presented as the report to the Office of Education and Training for the activities developed during the Smithsonian's Summer Research Participation Program for 1967.

Thanks are due to the Director of the Information Systems Division, Mr. Nicholas J. Suszynski, Jr., for the help and advisement that without hesitation at any time provided to the author.

Thanks are given too to Mr. Reginald A. Creighton for his immense interest, advisement and good will in this work.

For Mrs. Mildred Gins my sincere thanks for her patience and kindness without limits.

For Mrs. Nancy Steil my great thanks for her understanding and help in typing the originals.

INDEX

	<u>PAGE</u>
I PROPOSALS OF THE SYSTEM.....	1
II MAIN CONSIDERATIONS IN THE DESIGN OF THE SYSTEM.....	3
FLORA MONITOR ORGANIZATION.....	9
III SOURCE DATA AUTOMATION.....	15
GUIDE FOR CODING THE DATA COLLECTION SHEET.....	23
MAGNETIC TAPE RECORD LAYOUT.....	31
IV QUERIES TO THE SYSTEM.....	42
FIELDS SUSCEPTIBLE TO BE QUERIED.....	43
THE FORM TO QUERY THE SYSTEM.....	47
OTHER POSTRETRIEVAL PROCEDURES.....	51
V INPUTS TO THE SYSTEM.....	53
INPUT PROCEDURES.....	54
PARAMETER INPUT CARDS.....	58
VI OUTPUTS FROM THE SYSTEM.....	62
PARAMETER OUTPUT CARDS.....	64
VII CONCLUSIONS	67
VIII BIBLIOGRAPHY.....	68

A Bibliographic Data Retrieval System

S. AHUMADA

I PROPOSALS OF THE SYSTEM.

As stated in the draft of the National Science Foundation (NSF) for Flora North America Project, [1], an automatic system for bibliographic data manipulation should be implemented in the first stage of the Project. The system designed through the following pages will intend to cover all specifications mentioned in phase 1 for EDP considerations of the cited proposal.

The system covers from the necessary steps to collect the information, via paper tape; to feed the computer, in this case the H-1200 of Smithsonian's Information Systems Division, with the punched tape and to convert it to a magnetic tape record; the necessary processing stages to update, edit, retrieve information from the files so formed and the way of disseminating the obtained information.

Due to the general idea of the utilization of the paper tape as the input media shown through the Proposal, [1-a], seems as convenient to take advantage of the same media to disseminate the retrieved information.

In this form data will be accepted in the system in paper tape form, using any code, will be converted through the utilization of the necessary translation tables formerly provided to the system, to a magnetic tape for the rest of the processing.

The retrieved information, optionally, will be punched in paper tape in the same original code of a specific requestor, providing in addition of a print-out with the answers, a media to make any convenient reproductions for further dissemination. This will have the advantage of utilizing different sorts of paper tape punching equipment already installed providing compatibility with the central system.

A Bibliographic Data Retrieval System

S. AHUMADA

The print-out of reference tables, such as subject index, taxa names, etc., listings by geographical areas, taxa, author, etc., will be performed through the retrieval option described below. The retrieval time will be minimized and an appropriate qualification has been given to the retrieved information, trying to improve both the data contents of the system and the way of querying it.

II MAIN CONSIDERATIONS IN THE DESIGN OF THE SYSTEM.

Due to the experimental stage of the bibliographic system under consideration seems to be convenient to implement a set of computer programs that providing the expected functioning in its first stages, will be easy to modify until its final appearance during the 15 year period of FNA. This makes the present system to have certain characteristics which make both the user and the programmer to take advantage of the utilization or programming of the system in its present form and to make easier for both the utilization of future improvements. As a time shared system has been mentioned as the most probable one for future operations the routines developed for this system may be the basis upon which the other ones could be developed. The considerations in the user's point of view are:

- A. A source data automation procedure has been developed which under proper utilization, accomplished by the following of the instructives for filling the data collection sheets, will make very easy and with a minimum in personnel to submit data. Special care has been taken in trying to manipulate data from various sources, each with a possible different paper tape punch equipment through the utilization of parameter data that will inform the central system of the characteristics of each source. The purposes and restrictions in their utilization and a guide for the operator have been provided too.
- B. The queries to the system may be coded by the users themselves as COBOL statements, [2], closely resembling the English language.
- C. Identification codes for institutions, requestors, additional information concerning the type of output desired, etc., may be written again in English-like language providing an easy tool for the dissemination of the information.

A Bibliographic Data Retrieval System

S. AHUMADA

- D. The output forms are self-explanatory and enough marks have been added to make easier their reading.
- E. Identification codes for distribution of the retrieved information were provided in the paper tape output making easy to reproduce the obtained bibliography.
- F. A notification will be sent to the bibliographic staff for FNA when the answer to a query is not satisfactory, or when there is no answer at all, increasing the possibility of improving the information contained in the system, guaranteeing for the users the best possible service.

The former capabilities will make the user to code completely both his data and queries and the utilization of "work options" will make the requestor to indicate the type of processing desired.

In which respects to the programmer:

- A. The system is conceived as a Master System, under the direction of the COBOL Monitor, which will perform in essence the functions of monitoring and supervising the functioning of the necessary procedures or subroutines to accomplish a determined set of functions selected by the user.

This modular design will permit the modification, deletion or increase of any subprogram, which determine a definite function or set of functions, without changing the whole system, making not only easier the labor of the programmer but being a more general approach for the programming, debugging and testing of the system itself.

- B. All the programs will be coded in COBOL, except those needed for the operation of the paper tape devices, which will be coded in assembly language, EASYCODER, permitting the completion of the whole system in a brief lapse of time. Documentation effort for the programs will be diminished too.

A Bibliographic Data Retrieval System

S. AHUMADA

- C. Advantage has been taken of the organization of the COBOL compilers D & H for the H-200 series, [3], permitting the organization of the programs in the following form:

<u>Name</u>	<u>Type</u>	<u>Language</u>
FLORA MONITOR	Resident Program	COBOL
UPDATE	Non-resident Program	COBOL
RETRIEVAL	Non-resident Program	COBOL
EDITING	Non-resident Program	COBOL
.		
PAPER TAPE	Non-resident Program	EASYCODER

The terms "resident" and "non-resident" should be understood as in [3] i.e.: The program that always remain in main memory at execution time, and those programs that are read from a binary run tape and stored in an "overlay" area in core when needed, through the utilization of a GO TO or a PERFORM COBOL statements in the resident section, respectively.

As was established in the former reference for non-resident segments, their Data Division should be the same as the one in the resident segment from which they are called and their priority numbers should be greater than 49 because the FLORA MONITOR, resident program in this case, will set up the necessary indicators for initialization in each one. The only difference in the established way of coding the whole set of programs and a conventional one is the assignment of priority numbers greater than 49 and the requirement of the same Data Division in the non-resident segments, and on the other hand the control in the operation in the whole system is much better accomplished without more programming effort, which may even be distributed by those segments for performing a simultaneous work that will speed up the whole process.

As it is possible to see a better utilization of the memory, 65,000 characters in the H-1200, will be performed both in core utilized by the programs and by the data.

Seems to be convenient to manage the most error recovery procedures through the utilization of own-coding routines, being in a position in which it is possible to direct in an appropriate form the behavior of the system, and just taking advantage of the error exits in the COBOL Monitor either due to the impossible - recovery nature of the errors found, or the development of an unnecessary programming effort, or when the amount of memory required could make very expensive the development of special routines.

Special effort should be made in the re-initializing sections due to the general complexity of the system and the various incidents that may occur through a long time processing operation, guaranteeing that at least all the work done in segments proceeding the one in which a fatal error has occurred will not be lost, saving computer time and providing the operator with enough information on the action that should be taken for the recovery. In figure 1, page 7 a drawing shows the general organization of the system based upon the mentioned concepts and in figure 2, page 8 it is shown the aspect of the memory at working time.

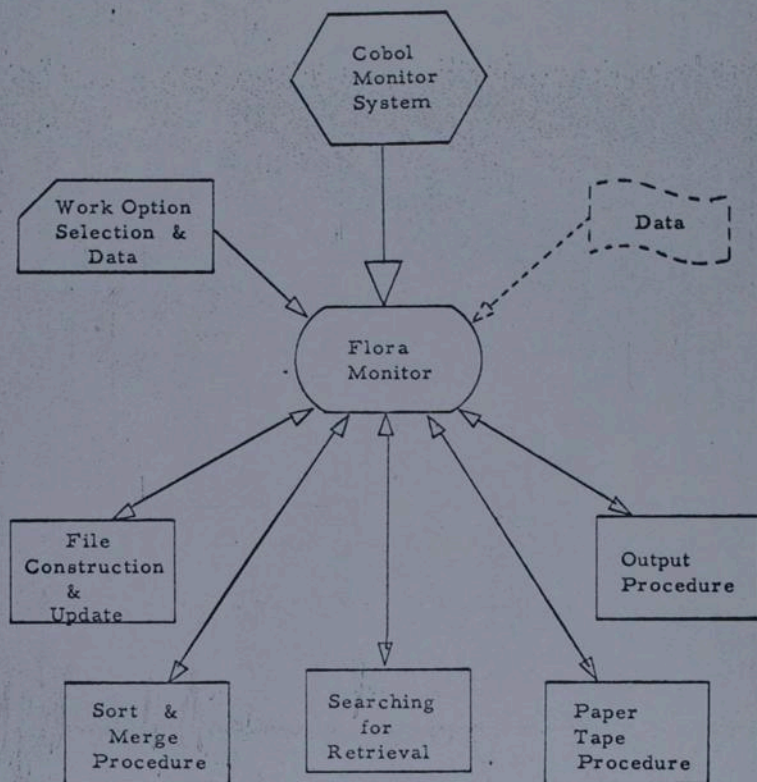


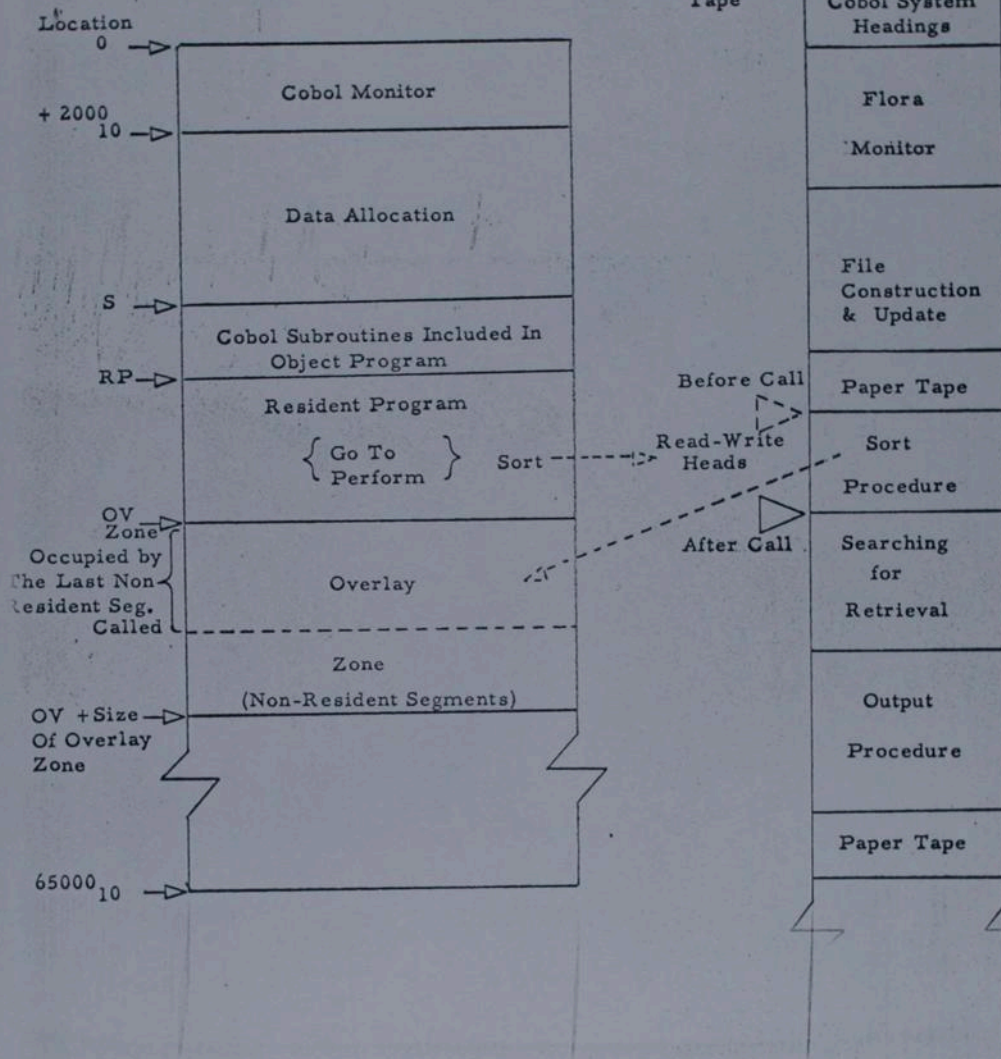
Figure 1. General Organization of the FLORA MONITOR.

A Bibliographic Data Retrieval System

S. AHUMADA

Cobol's Binary
Run Tape

MEMORY DISTRIBUTION

Beginning Of
Tape

FLORA MONITOR ORGANIZATION

The main functions of the resident program and its relations with the other procedures are described. The coding in COBOL is not shown and only general suggestions in that sense are given.

NAME OF THE SECTIONMAIN OPERATIONS

NEWS-SECTION

The collection and processing of basic data for the functioning of the system:

- 1) Working options

- 2) Equipment needs

APPOINTMENTS-SECTION

The processing of data concerning the utilization of non-resident segments:

- 1) Procedure selection based upon working options choosen.

- 2) Establishment of initial conditions and parameters for the non-resident segments functioning through the creation and updating of a "procedures table containing:

- a) Name of procedure.

- b) Indicative of "under present operation" or "not needed" at a certain moment.

- c) Indicative of "successful compliance".

- d) Indicative of "fatal error that occurred."

- 3) Actual calling for non-resident segments.

- 4) Controlling non-resident segment's operation.

REFERENCES-SECTION

The access to reference tables such as Subjects, Geographical terms, etc., and code translation tables for paper tape operations; loading them from magnetic tape and storing them in main memory for their future utilization by other procedures.

COMMUNICATIONS-SECTION

The informing to the operator of:

- 1) Options selected and procedures needed for successful accomplishment.
- 2) Procedure under present operation.
- 3) Error presence and type of recovery procedure.
- 4) Input for re-initializing the system through operator's console.

ERROR-RECOVERY-SECTION

The management of error conditions in the procedures:

- 1) Notification of the type of errors found.
- 2) Indicatives for correction or re-initializing.

QUERIES-SECTION

The storage and functioning of the user's queries:

- 1) Including in this section the actual set of queries and submitting the FLORA MONITOR as a source program [2] .
- 2) If the option is not for queries this section will not be utilized and the FLORA MONITOR will be loaded as object program from the binary run tape.

During programming development stage, care should be taken in the sense of minimizing the memory requirements for this monitor in order to increase the size of the non-resident programs and to dispose of enough locations to store the most possible data, fact that will minimize too the time required in input-output operations. In figure 3, page 12 , a general flow diagram of the monitor's operation is shown.

A Bibliographic Data Retrieval System

S. AHUMADA

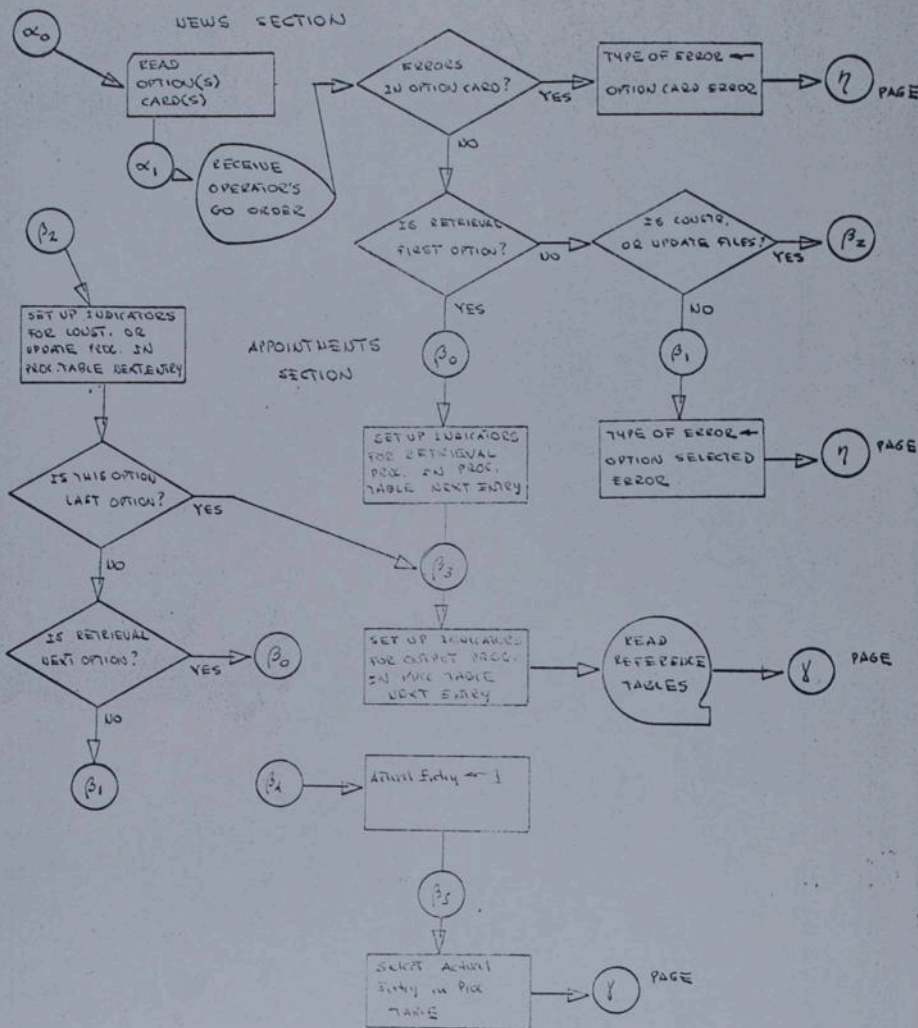


Figure 3. General Flow Diagram of the FLORA MONITOR, PART 1 OF 3

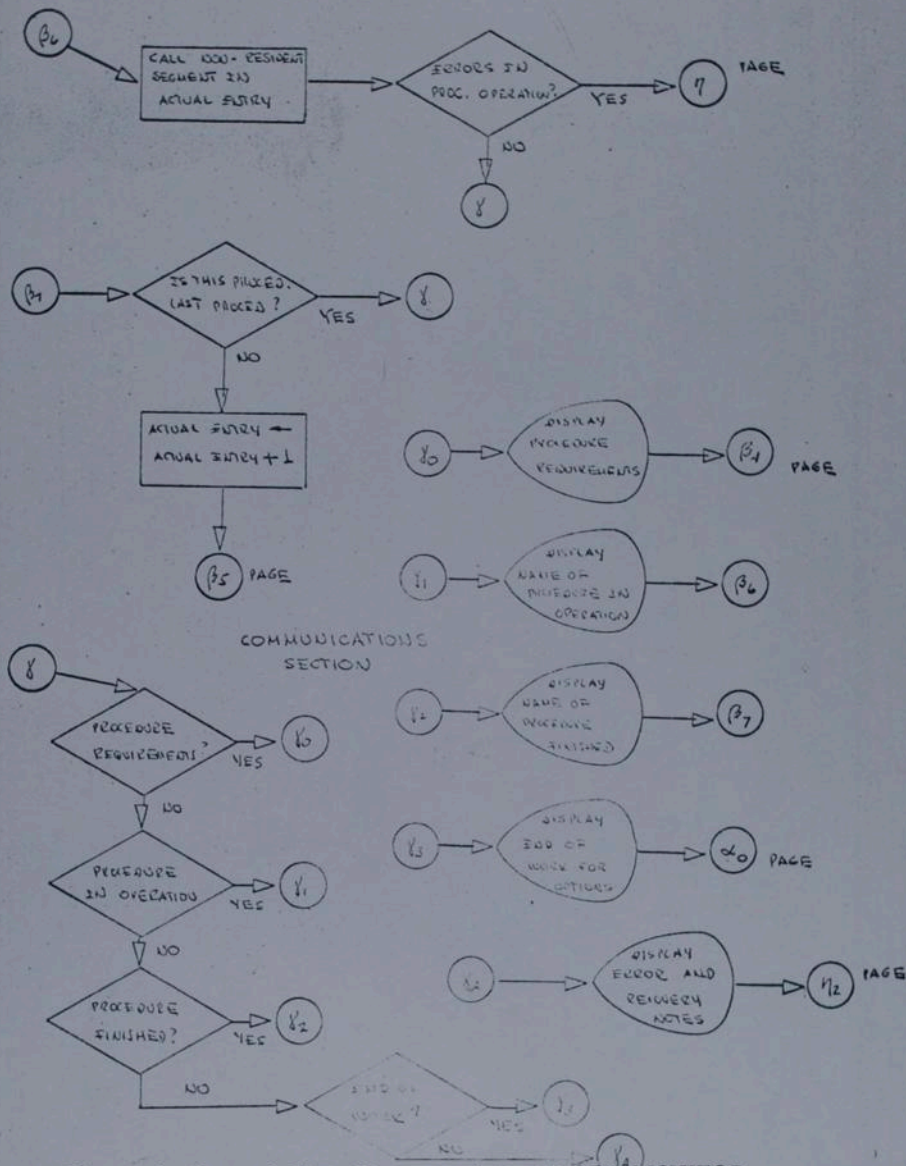


Figure 3. General Flow Diagram of the FLORA MONITOR.
PART 2 OF 3

A Bibliographic Data Retrieval System

S. AHUMADA

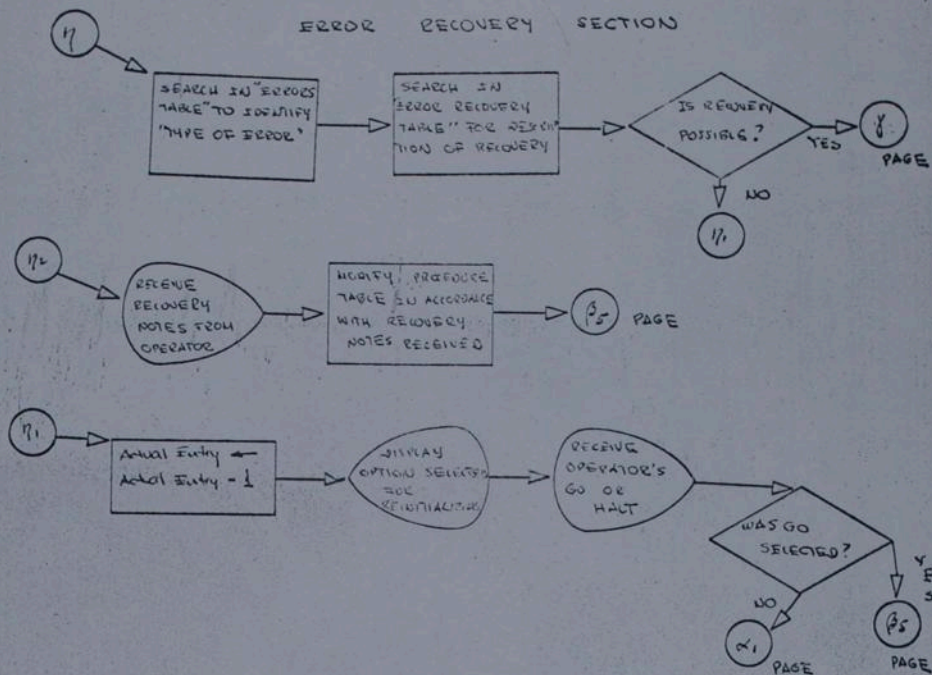


Figure 3. General Flow Diagram of the FLORA MONITOR.
PART 3 OF 3

III SOURCE DATA AUTOMATION.

The first step in the automatic procedure for bibliographic information is the automation of the original data that will form the files, from which the retrieval procedure will extract those matching with a query.

In this chapter a description of the sheet that will contain the source information and an instructive for filling that form and submitting data to the system and the magnetic tape record are shown. Through a flow diagram the general steps in this procedure are indicated and a sample of the Data Collection Sheet is included. As it has been explained before, the paper tape is the media utilized as information support. It is shown here the whole process with examples of bibliographic cards as are commonly used, the hard copy obtained in the SCM typewriter and the punched paper tape as it will appear before entering the computer. It is possible to see from this draft the bottlenecks in the data collection process and which obviously will be the steps the bibliographer will need to take more care of.

NOTES TO THE USERS-

In order to provide to the users of this system a knowledge of their participation in the processes of file construction and updating a description of the activities they will make follows.

They are concerned with the capturing of the source information for the system using the Bibliographic Data Collection Sheet. The limitations are shown, and using flow diagrams a description of the whole process is made.

Bibliographic Data Collection Sheet

The purpose of this sheet is to collect from the various sources such as books, reprints, former index cards, etc., the necessary information to form a magnetic tape record for each document, based upon which the retrieval of the information will be made.

The bibliographer will be the person that will fill the form which afterwards will be sent to the SCM Typetronic Operator for the typing of the main index card and the simultaneous punching of the paper tape that will be the input for the computer.

A general description of the appropriate form of filling that form is given and a flow diagram of the operations involved is shown in figure 4, page 18. Using samples of source documents the whole process is shown in figures 5-a to 5-c, pages 19 to 21, and the sample filled form is shown in figure 6, page 22.

- A. It is possible to consider the form divided into three main parts more or less arbitrarily, each one with various entries which define the whole document.

1. Data concerning the author(s) of the document:

Name of the first author
Dates of the first author
Name of the second author
Dates of the second author
Type of authors entry
Single or multiple authors
Dates of publication
Type of dates of publication

2. Data concerning the classification of the contents of the document:

Title
Subjects field
Taxa field
Geography field

3. Data concerning the document itself and remarks:

Publisher
Collation
Language of document
Type of document available

A Bibliographic Data Retrieval System

S. AHUMADA

Type of availability
Last place or person which has the document
Dates of validity
Translation indicator
Original language
Bibliographed issue is original
Others
Other taxa

- B. The appropriate way to fill each one of the entries follows. All entries will be written Left Justified in their respective fields and the size of the field should not be exceeded.

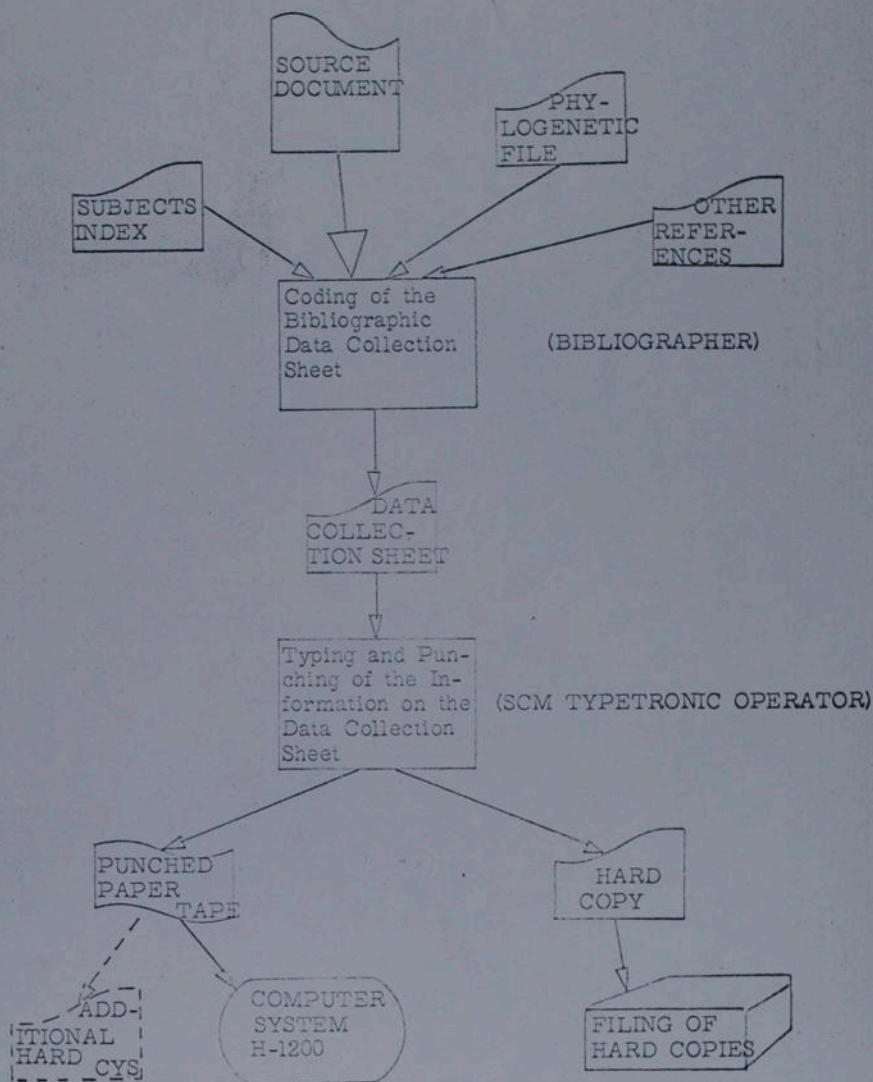


Figure 4. Bibliographic Data Collection Procedure.

A Bibliographic Data Retrieval System

S. AHUMADA

English author.

Only one date of publication.

Subject mentioned in title.

Braun, E. Lucy. 1955. The phytogeography of
unglaciated eastern United States and its
interpretation. The Botanical Review
21: 297-375. (No. 6, June).

5956 Salem Road
Mt. Washington
Cincinnati 30, Ohio
Individual copies 60¢
The Botanical Review
Fordham P.O.
N. Y., N. Y.

Geography
mentioned in
title.

Name of
publication
fully spelled.

Reference to the avail-
ability of the document
in the card and last
location of document.

NO MENTIONS TO:

Taxa and Means the publication should
Other Taxa. be consulted.

Dates of author. "

Language of doc- "

ument(presumable
English)

More detailed geo- "

graphical entry.

Other subjects. "

REFERENCE IS MADE TO THE APPROPRIATE TABLES.

Figure 5-a. Example of bibliographic card and main considerations
about the information required by the system.

A Bibliographic Data Retrieval System

S. AHUMADA

NO MENTIONS TO:

Geography.	Means the original publication should be consulted.	
Dates of author.		"
More detailed subject treatment.		"
Language. (presumable English)		"
Type of document available and type of availability.		"
Dates of validity.		"
Other Taxa.		"

English author.

One date of publication.

Bowden, Wray M. 1959. Phylogenetic relationships of twenty-one species of Lobelia L. section Lobelia. Bull. Torr. Bot. Club 86: 94-108.

Subject mentioned in the title.

Taxa mentioned in the title.

Publication indicated by abbreviation.

Figure 5-b. Example of bibliographic card and main considerations about the information required by the system.

A Bibliographic Data Retrieval System

S. AHUMADA

Explicit reference to subject terms.

Only one publication date.

English author.

POPULATION DYNAMICS

Stebbins, G. L. 1957. Self-fertilization and
population variability in the higher plants.
Am. Naturalist 41: 337-354.

Subject
mentioned in
the title.

Publication
name
abbreviated.

NO MENTIONS TO:

Taxa and
Other Taxa.

Geography.

Date of author.

Language (most
probable English)

Not enough subject ref-
erences.

Type of doc. avail. and type
of availability.

Means the original publication
should be consulted.

"

"

"

"

"

Figure 5-c. Example of bibliographic card and main considerations about the information required by the system.

BIBLIOGRAPHIC DATA COLLECTION SHEET

FLORA NORTH AMERICA PROJECT

S. AHUMADA

A Bibliographic Data Retrieval System

Figure 6. Sample of Data Collected.

FIRST AUTHOR	<u>Brown, Stephen, L.</u> NAME			DATES	<u>1890</u> <u>1933</u>			
SECOND AUTHOR	<u>Hellerman, Lucy, H.</u> NAME			DATES	<u>1892</u> <u>1940</u>		<u>A</u> <u>S</u>	
DATES OF PUBLICATION	<u>1929</u> 1ST	<u>1930</u> 2ND	3RD	4TH	5TH	6TH	AUT/ED	SIN/ED
								# AND TYPE
TITLE OF DOCUMENT FULLY SPELLED	<u>Phylogenetic relationships of few species of <i>Lobelia</i> L.</u>							
SUBJECTS	<u>Phylogeny</u> 1ST	<u>V</u> QUAL.	<u>Genetics</u> 2ND	<u>V</u> QUAL.	<u>Taxonomy</u> 3RD	<u>G</u> QUAL.		
	4TH	QUAL.	5TH	QUAL.	6TH	QUAL.		
	7TH	QUAL.	8TH	QUAL.	9TH	QUAL.		
	10TH	QUAL.	11TH	QUAL.	12TH	QUAL.		
TAXA	<u>Lobelia</u>							
GEOGRAPHY	<u>USA</u> COUNTRY	<u>Indiana</u> STATE OR PROVINCE	COUNTY OR SUBPROVINCE					
	OTHER SPECIFICATIONS							
PUBLICATION	<u>Bull. Bot. Soc. of A.</u> PUBLISHER			<u>Vol. 21. Pages 218-270</u> COLLATION				
	<u>E</u> LANG. OF DOC.	<u>E</u> DOC. AVAILABLE	<u>P</u> AVAILABILITY	BEGINNING VALIDITY FINISHING VALIDITY				
	<u>O</u> TRANSLATION	ORIGINAL LANG.	<u>D</u> BIBL. ISSUE	<u>Graphs included</u> OTHER COMMENTS				
OTHER TAXA	<u>Lythra</u>			<u>Cardinalis</u>				
INSTITUTION NAME	<u>Botanical Club of A.</u>			BIBLIOGRAPHER NAME	<u>E. Adams</u>		DATE	<u>8-23-67</u>
							PAGE	<u>1</u> OF <u>10</u>

GUIDE FOR CODING THE DATA COLLECTION SHEET

S. AHUMADA	NAME OF THE FIELD	REQUIREMENTS	SIZE OF THE FIELD IN CHARACTERS, & TYPE
A Bibliographic Data Retrieval System	Name of the first author	Write the name of the only or main author of the document. For English names, the order will be: Last name__, First name__, Initial__. For foreign names the order will be: First Last name__, Second Last name__, First name__. The punctuation symbols , (comma) - (dash) _ (space) should appear in the positions shown. The symbol _ was used here instead of space which will be the one used in the actual sheet.	40 Alphabetic
	First author dates: birth death	Write in order birth and death dates for first author. Dates will be expressed in the form: Month, two characters; Day, two characters; Year, four characters.	8 characters for each date. Numeric
	Name of second author	Write the name of the second author in the same way as "Name of the first author."	40 Alphabetic
	Second author dates: birth death	Write in order birth and death dates for second author in the same way as "First author dates."	8 characters for each date. Numeric

GUIDE FOR CODING THE DATA COLLECTION SHEET

	NAME OF THE FIELD	REQUIREMENTS	SIZE OF THE FIELD IN CHARACTERS, & TYPE
S. AHUMADA	Type of authors entry	Write: "A" if the author's fields contain actual authors. "E" if the author's fields contain editors.	1 Alphabetic
	Single or multiple authors	Write: "S" if the author's fields contain all the authors that wrote the document. "M" if there are more authors not included in the author's fields.	1 Alphabetic
A Bibliographic Data Retrieval System	Dates of publication	Write from left to right and in Month, Day, Year order: Only date of the publication or, inclusive dates (1850 to 1852) and/or, exclusive dates (1850, 1952).	6 fields for 6 dates with 8 characters each. Numeric
	Type of publishing dates	Write: First character: number of same type dates from left to right. Second character (type of dates): "I" for inclusive dates. "S" for separate or exclusive dates.	2 characters Alphanumeric

GUIDE FOR CODING THE DATA COLLECTION SHEET

NAME OF THE FIELD	REQUIREMENTS	SIZE OF THE FIELD IN CHARACTERS, & TYPE
Title of Document	Write the full title of the document without abbreviations.	250 Alphanumeric
Subjects	Write up to 12 subjects, from left to right to the closest one with the document to the farthest one. Add after each name, in the 20th character of the field, a qualification of the treatment of that subject in the document using: E for "excellent" V for "very good" G for "good" F for "fair" P for "poor" Use <u>ONLY</u> the accepted names for subject kept in a table in the system (19 characters in length).	20 characters for each of the 12 subjects provided Alphabetic
Taxa	Write the name of the plant or botanic type treated mainly in the document. Use <u>ONLY</u> the accepted names for taxa kept in the Phylogenetic File. Two fields of 27 characters each were provided in which is possible to write up to binomials. Other taxa mentions should be written in the "other taxa" field.	54 Alphabetic

GUIDE FOR CODING THE DATA COLLECTION SHEET

A Bibliographic Data Retrieval System	S. AHUMADA	NAME OF THE FIELD	REQUIREMENTS	SIZE OF THE FIELD IN CHARACTERS, & TYPE
		Geography		
		Country	Write the abbreviation of the main country the document makes reference to. Use <u>ONLY</u> the standard letters provided for each country.	3 Alphabetic
		State or Province	Write the name of the state or province the document makes reference to.	25 Alphabetic
		County or Subprovince	Write the name of county or other political subdivisions.	25 Alphabetic
		Other Specifications	Write complementary geographical information.	25 Alphabetic
		Publication		
		Publisher	Write the abbreviation of the name of the publisher of the document. Use <u>ONLY</u> those abbreviations accepted.	40 Alphanumeric
		Collation	Write number of pages, chapters, illustrations, indicator, etc.	25 Alphanumeric

GUIDE FOR CODING THE DATA COLLECTION SHEET

S. AHUMADA

A Bibliographic Data Retrieval System

NAME OF THE FIELD	REQUIREMENTS	SIZE OF THE FIELD IN CHARACTERS, & TYPE
Language of Document	Write the code of the language of the bibliographed issue: E for English, S for Spanish, F for French, G for German, R for Russian, I for Italian, etc.	1 Alphabetic
Type of document available	Write the type of document available for a possible reader, using: D for document itself, R for reprint, L for library issue, M for microfilm.	1 Alphabetic
Type of availability	Write the availability of the document for possible readers, using: P for any people, R for Research personnel of FNA, G for Research personnel in general, N for not available.	1 Alphabetic

GUIDE FOR CODING THE DATA COLLECTION SHEET

S. AHUMADA

A Bibliographic Data Retrieval System

NAME OF THE FIELD	REQUIREMENTS	SIZE OF THE FIELD IN CHARACTERS, & TYPE
Remarks		
Dates of validity: Beginning Finishing	Write the beginning and finishing dates of validity of the contents of the document, using 2 characters for the month, 2 for the day and 4 for the year.	8 characters each date. Numeric
Translation	Write: O for original in the language. T for translation.	1 Alphabetic
Original language	If the document is a translation write the original language in the same form as the "language of document" above.	1 Alphabetic
Bibl. issue is original	Write the type of document that was available for the bibliographic card using the same codes than in "Type of document available" above, maybe including F for index card.	1 Alphabetic
Others	Write additional information about the bibliographed document.	20 Alphanumeric
Other taxa	Write all other taxa mentioned in the publication and not included in the taxa field. Use the same amount of characters and accept the same restrictions than in the latter field.	27 characters for each entry. Alphanumeric <u>Variable length</u>

As it is possible to see from the reading of the lines above, various sets of specific words were established. They will not only provide an easier method in the searching procedure but will uniformize the input data from the various sources. The words of those tables that will be kept in the system will validate the input data and will be used instead of the codes in the final printout of the retrieved information.

Seems to be useful to consider those tables as separate files which are susceptible to be updated and even searched by themselves, providing not only the necessary reference to the bibliographer, but additional information which may be for example frequencies for queries about a certain subject, taxa, etc.

The former lines try to induce the idea of the importance of the correct utilization of those data; a reference of the names of the tables and the generic contents of each one are given below, and the proposed entries in each one are shown in Appendix I.

<u>NAME OF THE TABLE</u>	<u>CONTENTS</u>	<u>DATA FIELDS INVOLVED</u>
SUBJECT TABLE	All given and accepted subject terms using 19 characters of alphabetic type.	Subjects.
PHYLOGENETIC TABLE	All the taxa references which are possible to apply to FNA arranged in hierarchical sequence using 27 characters for each name.	Taxa and other Taxa.

A Bibliographic Data Retrieval System

S. AHUMADA

<u>NAME OF THE TABLE</u>	<u>CONTENTS</u>	<u>DATA FIELDS INVOLVED</u>
GEOGRAPHIC TABLE	All the countries, states, counties, provinces names based upon which a Geocode will be developed, using 20 characters for each name. Seems to be convenient to add to this table the connecting codes for obtaining directly the mentioned Geo-code	Geography.
AUTHORS TABLE	All identification names for types of authors and single or multiple authors entries.	Type of authors Single or multiple authors.
DATE TABLE	All identification names for type of publishing dates entry.	Type of publishing dates entry.
SUBJECTS QUALIFICATION	All possible objectives for the treatment of the subjects in the document.	Subjects.
LANGUAGE TABLE	All names of possible languages to be found in documents related to FNA.	Language of document and original language.
DOCUMENT AVAILABLE TABLE	All possible types of documents available for a possible reader.	Type of document available and Biblio- graphed issue is original.

A Bibliographic Data Retrieval System

S. AHUMADA

<u>NAME OF THE TABLE</u>	<u>CONTENTS</u>	<u>DATA FIELDS INVOLVED</u>
AVAILABILITY TABLE	All possible types of availability of the document.	Type of availability.
TRANSLATION TABLE	All possible names for the translation entry.	Translation

The SCM typetronic operator, or any paper tape punching device operator, will start then with the reading of the Data Collection Sheet, typing and simultaneously punching paper tape. After having punched the necessary identification codes, a draft of which is shown in the Inputs to the System Chapter, page 53; will reject to the bibliographic staff of FNA all those sheets that present evident errors and will keep a record of those tapes sent to the Computer Room, indicating the source, number of entries and dates of sending.

In case the retrieved information punched on tapes is to be reproduced, using again the same source device, operators will take care of the leader trailer codes and marks of identification, and on the number of records that appear in the tape. Care should be taken too in the establishment of the margins and tabulators as applicable to that specific machine for obtaining the appropriate format. As it will be seen in the Inputs Chapter human-readable characters will appear in the tape in order to avoid misunderstanding.

MAGNETIC TAPE RECORD LAYOUT.

In the following lines the magnetic tape record that will contain the information pertinent to bibliography for FNA is described.

Two approaches have been considered, fixed length and variable length in records. The items are described in the order they will follow in the record and their main characteristics are shown. In all alphabetic fields the information will be considered left justified.

A Bibliographic Data Retrieval System

S. AHUMADA

A. FIXED LENGTH RECORD.

1. ACQUISITION NUMBER:

Will be provided by the system in the construction or updating procedures, so they will be not coded by the bibliographer. They will be consecutive numbers from 1 (one) upward and will be utilized as control numbers and as a reference in the retrieval procedure.

Size of the field: 6 characters. Type: numeric.

2. NAME OF FIRST AUTHOR:

Will contain the full name of the first, only or main author of the document. Two cases were considered:

a) English names: Will appear in the form -

LAST NAME____, FIRST NAME____, INITIAL____.

b) Foreign names: Will appear in the form -

FIRSTLAST NAME____, SECOND LAST NAME____,
FIRST NAME____.

The punctuation symbols, (comma) - (dash) _ (space) should appear in the positions shown.

Size of the field: 40 characters. Type: alphabetic.

3. FIRST AUTHOR DATES:

Will contain the birth and death dates,[9], for the first, only or main author considered in the preceding field. The order of appearance will be birth date and death date expressed in the form: Month, Day, Year.

Size of the field: 8 characters for each date. Type: numeric.

4. NAME OF SECOND AUTHOR:

Will contain the full name for the second author in the same form described for the first author.

Size of the field: 40 characters. Type: alphabetic.

5. SECOND AUTHOR DATES:

Will contain the birth and death dates for the second author as described in First Author Dates.

Size of the field: 8 characters for each date. Type: numeric.

6. TYPE OF AUTHORS ENTRY:

Will contain an indication if the names written in the author's field are actually authors of the publication or editors of the same. The letter A will indicate authors and the letter E will indicate editors. Other useful codes may be added in this field.

Size of the field: 1 character. Type: alphabetic.

7. SINGLE OR MULTIPLE AUTHORS:

Will contain an indication to know if the author or authors written in the field are the only one or ones that wrote the document, or on the contrary there are more than the quoted ones. The letter S will indicate single and the letter M will indicate multiple authors.

Size of the field: 1 character. Type: alphabetic.

8. DATES OF PUBLICATION:

Will contain six fields in which the date or dates of publication of the document will be written. The order will be: from left to right the inclusive dates of the publishing (date A through date B) or the only date; the not inclusive dates (date A, date B, date C) in which the document publishing was continued.

Size of the field: 8 characters for each date.
Type: numeric.

9. TYPE OF PUBLISHING DATES:

Will contain the type of dates written in the preceding field. The first character will indicate the number of date fields of the same type, from left to right, in the preceding field and the second character the type of dates defined. This qualification may be: the letter I for inclusive dates or the letter S for separate dates. For the remaining dates the type will be considered the opposite one.

Size of the field: 2 characters. Type: alphanumeric.

10. TITLE OF DOCUMENT:

This field will contain the full title of the document as spelled in the original, that is without abbreviations.

Size of the field: 250 characters. Type: alphanumeric.

11. SUBJECTS:

Twelve fields have been provided that will contain the subject classification of the contents of the document. From left to right, from the closest subject matter

A Bibliographic Data Retrieval System

S. AHUMADA

that matches with the document to the farthest, will be written, indicating with that order the relevance degree with the document.

Size of the field: 20 characters for each subject.
Type: alphabetic.

12. TAXA:

Will contain eleven fields in which will be written the taxonomic classification of the plants which are the main theme of the document. As the system will keep a Phylogenetic Table with the names of all possible entries, at least down to family, only the classification of lower levels than family will be coded alphabetically in the Data Collection Sheet. In the input procedure, either construction or file updating, the whole classification will be established in order to guarantee the retrieval of higher levels.

The eleven subfields are:

<u>NAME</u>	<u>SIZE IN CHARS.</u>	<u>TYPE</u>
Kingdom	1	Numeric
Phylum	3	Numeric
Class	3	Numeric
Order	3	Numeric
Family	3	Numeric
Tribe	4	Numeric
Genus	4	Numeric
Section	4	Numeric
Species	8	Numeric
Binomial	54	Alphabetic
Species Author	4	Numeric

Size of the field: 100 characters. Type: alphanumeric.

A Bibliographic Data Retrieval System

S. AHUMADA

13. GEOGRAPHY:

Will contain the description of the geographical zone considered in the document. The five fields provided are:

<u>NAME</u>	<u>SIZE IN CHARS.</u>	<u>TYPE</u>
Country	3	Alphabetic
State or Province	2	Alphabetic
County or Subprov.	4	Alphabetic
Other Specifications	25	Alphabetic
Geo-Code	6	Alphanumeric

The three first fields will contain an abbreviation obtained using the Geographic Table and the Geo-Code, [6], field will be obtained in the input procedure.

Size of the field: 40 characters. Type: alphanumeric

14. PUBLICATION:

Will contain four fields in which major characteristics of the document itself are described.

<u>NAME</u>	<u>SIZE IN CHARS.</u>	<u>TYPE</u>
Publisher	40	Alphanumeric
Collation	25	Alphanumeric
Language of Doc.	1	Alphabetic
Last Location of Doc.		
Type of Doc. Avail.	1	Alphabetic
Type of Availability	1	Alphabetic
Place or Person	20	Alphanumeric

The language field will contain a letter i.e.: E for English, S for Spanish, F for French, R for Russian, G for German, etc.

A Bibliographic Data Retrieval System

S. AHUMADA

The type of document available will contain, for example: D for document itself, R for reprint, L for Library issue, M for microfilm, etc.

The type of availability field will contain: P for any people, R for research personnel, G for research personnel in general, N for not available, etc.

Size of the field: 88 characters. Type: alphanumeric.

15. REMARKS:

In five fields will provide to the possible requestor with useful additional information:

<u>NAME</u>	<u>SIZE IN CHARS.</u>	<u>TYPE</u>
Dates of validity		
Beginning	8	Numeric
Finishing	8	Numeric
Translation	1	Alphabetic
Original Language	1	Alphabetic
Bibl. issue is original	1	Alphabetic
Others	20	Alphanumeric
Other Taxa	188	Alphanumeric

The dates of validity make reference for those for which the document was written, i.e.: The evolution of the Campanulla in California between 1847 and 1850.

The translation field will contain the letter O for original or the letter T indicating that the bibliographed document is a translation.

The original language field will contain in the case of a translation the code for the original language in the same form that was used for the field "Language of Document"

A Bibliographic Data Retrieval System

S. AHUMADA

The "Bibliographed issue is original" field will make reference to the type of document that was available for the collection of the data by the bibliographer. In addition to the codes used in the field "Type of Doc. Available" the letter F may be used to indicate that the record was formed based upon a fiche or index card and not on a document.

"Other Taxa" will contain references to those cited in the document. A maximum of 28 species may be included.

Size of the field: 209 characters. Type: alphanumeric.

Total amount of characters utilized:

SIZE OF RECORD: 1095 CHARACTERS

The data that appear in the former description is obtained through the utilization of the Bibliographic Data Collection Sheet and closely resembles it and in some cases is translated to an abbreviated form by the system through the utilization of the tables shown above.

In the case in which there is not available information to fill the provided fields, they will contain zeroes or spaces in accordance with the type of information they should have, numeric or alphanumeric respectively.

It has been considered as useful to provide the input procedure with the possibility of receiving data either by punched paper tape or by cards in order to give more flexibility to the user.

B. VARIABLE LENGTH RECORD.

As it is possible to see reading the lines above, there are a certain number of fields for which due to the type of information they will contain and to the various sources

that may provide that information seems as convenient not to use a bounded set of characters, but a variable one, to record that information on tapes. An approach in that sense is shown in which follows, preserving those fields for which the type of information may be adjusted to a fixed length, and considering as variable the others.

The order of the information recording is the same and there will be shown only the fields considered as variable. The numerals correspond to those of the former approach.

8. DATES OF PUBLICATION:

Will contain the date or dates during which the publishing of the document was developed. Spaces will separate the fields and dashes or commas will indicate inclusive or not inclusive dates respectively in the source form. Eight characters, numerals, are considered enough to write a date in the order Month, Day, Year.

Minimum size of the field: 8 characters, maximum size of the field: as convenient. Type: alphanumeric

The field 9 in the fixed length layout will now be considered as part of this field, the 8th, in variable length.

10. TITLE OF DOCUMENT:

Will contain the full title of the document. Neither minimum or maximum sizes defined. Type: alphanumeric.

15. REMARKS:

The "Others" and "Other Taxa" fields will be considered as variable. The remaining fields may be adjusted to fixed length.

Size of the field: 19 plus variable length. Type: alphanumeric.

Total amount of characters utilized:

SIZE OF RECORD: Variable.

In figure 7, page 41, the drawing for the Magnetic Tape Record is shown. Although both approaches, variable and fixed length in records, were shown, a suggestion is made in the sense that seems to be much more easy to handle a fixed length record both in updating and in sorting procedures. It could be possible to generate another record of the same size in case that the actual size could not be enough to record all the information needed, for example in author's field on recording corporate author's.

The disadvantage of having non-utilized space in tapes is compensated by the time saved in updating and sorting the files and the general manipulation of the magnetic tapes.

The updating procedure, on the other hand, may be simplified if the length of the records is already known.

IV QUERIES TO THE SYSTEM.

In which follows a description of the record fields that will be queried most frequently and the form for doing so are shown. The term "relevance", expressed in the form of levels, will be used to indicate the approximation of the answer to the query. The first level will indicate the closest possible answer for that query about a certain field, while a higher number in the level will indicate that the answer is farther to the query, that is, possibly the answer was found in other than the specified field, or the field has by itself its grading, i.e.: in the field SUBJECTS, the code order by itself will indicate the proximity to the optimum level.

The output procedure after the searching one will indicate the relevance level of the answers to the queries and will sort by this item the set of responses found in order to provide the user with the best possible appearance of the information.

The relevance levels may be used by the system itself in order to obtain a description of the set of answers in terms other than a list of them, but compared with the master file in which they were found. In that way it would be possible to obtain relations as:

- A. Number of first, second, etc., level hits compared with number of records processed. Percents of the same number of hits in respect to the possible totals.
- B. A counter of hits for each one of the items requested in a single query will be kept by the system in order to provide to the user the following information:
 - 1. Number of hits in the first, second, etc., items of the query.
 - 2. Percents of the number of hits in first, second, etc., level of relevance in respect to the number of hits in first, second, etc., items requested.

The figures above, and many others, possibly may provide the user with a guide to query the system again in the case that the relevance was very low for his purposes, being in a position to decide either to reduce the restrictions in the query or to increase them.

They will guide the bibliographer also to look for material not provided yet to the system, i.e.: missing subjects, geographical areas, taxonomic representatives, etc.

The former will create the need of at least two classes of reports:

- A. The first one oriented to the bibliographic staff for FNA Project that will contain the information mentioned in A. above which may provide a useful guide for the improving of the information that is contained in the system.
- B. The second one for the user of the system which in addition to the answers will contain the information contained in B. above.

Nevertheless, the ideas exposed before are just a sketch, and they need to be improved and may be changed. A statistical procedure seems to be capable of providing a good tool to manipulate that information.

FIELDS SUSCEPTIBLE TO BE QUERIED.

Although in principle all the fields that contain the record are susceptible to be queried, either by the type of information the requestors will need, or by the characteristics of the fields themselves, the majority of the queries will be made upon:

SUBJECTS - As was established in the description of the magnetic tape record twelve fields were provided to

write the names of the subjects the contents of the document is more related with. These are written from the closest one with the document at the left most field, to the farthest one in the right most field. This classification provides twelve levels of relevance, level one in the left most position to level twelve in the right most one. The procedure will start with the verification in the Subjects Table of the existence in the file of at least one reference of the subject provided in the query. If there is not a reference, the query will be rejected immediately, saving time of search.

The searching procedure will look for matching from left to right, each record with the query and will add to the record itself the relevance level that was found in order to write it in the hits tape.

As all the information related with subjects will appear only in this field of the record, i.e.: if there is a subject mentioned in the title, it ought to be added in the subjects field and so on for other fields of the record, if a certain document does not match with the query, it will be rejected and no other search in any field will be made.

Total number of relevance levels: 12

TAXA -

Eleven fields were provided in the magnetic tape in order to code the taxonomic description of the specimens that are the theme of the document, from kingdom to species author. All queries concerned with taxa will be matched with the records in the file, based upon the level indicators and the code obtained through the Phylogenetic Data and the subsequent Phylogenetic Table for the searching. So the process will be:

A Bibliographic Data Retrieval System

S. AHUMADA

- A. Verification of the validity of the taxa name provided in the query.
- B. Transformation of the name or names provided to the corresponding codes using the Phylogenetic Table.
- C. Search for matching the obtained codes with those of the records in the file in the TAXA Field.

A hit in matching will make the record be written in the "hits" tape.

If the matching was not successful, the procedure will look for matching it in the OTHER TAXA field, being this matching of relevance two.

Total number of relevance levels: 2

GEOGRAPHY- Five fields were provided to contain geographical references of the bibliographic entries. The last one seems to be the most useful one in the searching procedure and grading of the relevance of the answers due to the grading itself of the Geo-Code.

The first step to be done is to transform, when possible, the names provided in the query to the most convenient Geo-Code, through the utilization of the Geographic Table. Afterwards the searching procedure will look for matching with the Geo-Codes in the records of the file, relevance one, and to write those records in the output tape if the matching was successful. If it was not, it is possible to select the next grade in the Geo-Code, relevance two, which will define a more general geographical area.

The procedure may continue up to the fifth relevance level, as six possible characters were provided in the Geo-Code field, each one defining an area of less extension than the former one, the fifth level of relevance will indicate, for example, the northern part of the U.S.A., level that seems to be the most generally acceptable one for searching purposes.

At the same time, a matching will be made in the other fields of the geographical entry, state or province, county, etc., using the names provided in the query, in order to guarantee the most complete retrieval, especially in those countries not included in FNA for which a less detailed Geo-Code is expected.

Total number of relevance levels: 6

DATE
OF
PUBLICA-
TION

The queries may be restricted in this field too, provided with Day, Month and Year, and with six entries that with an additional code have a wide set of possibilities. The query may indicate the range of dates, so as before, after, between and the matching will be made not only in the dates but in the code which specifies the type of dates in the record.

A match in this field will be considered of relevance one. If no match is done in this entry, the field Dates of Validity may provide a reference, of relevance two, that may be used to retrieve the information.

If neither in the Dates of Publication nor in the Dates of Validity fields a match is made, there exists the possibility of using the field Author Dates as a relevance three answer. Nevertheless, this last option may not be useful.

Total number of relevance levels: 3

From the above lines it may be seen that the queries to the system need to be bounded in respect to the possible set of answers. A set of words such as: UNLIMITED, RESTRICTED, NORMAL, etc., seems to be capable to define the output, in relation to the relevance levels mentioned above, and which meaning could be: all the information that match with the query in any relevance field; only those records that match with the query in relevance one or two will be accepted as answers, etc., respectively. A uniform concept of relevance, as applicable to this specific set of queries and data could provide a good tool for the evaluation of the system, which will interpret the answer as a whole and will provide in the output a value for each one of the obtained answers; the greater the number, the less accuracy in the answer, and will provide to the user, in addition with the figures mentioned in the first paragraphs, a guide to query the system and at the same time to improve the information it contains.

All those references mentioned above will be indicated in the output, for which at the same time seems to be useful to define with options bounding the appearance of the information in the final printout.

THE FORM TO QUERY THE SYSTEM.

It has been said in the former paragraphs that the queries to the system will be written in COBOL. This idea is covered in the appropriate form in reference [2] and only the necessary explanations will be given here. For the columns to be used refer to [3]. When the processing for a computer run includes queries to the system and not only updating options, the FLORA MONITOR will be submitted as a source program which will contain in its QUERIES-SECTION the actual statements for the retrieval. In case that the options selected do not include the retrieval one, the FLORA MONITOR will be directly loaded from the binary run tape.

This Queries Section will be used by the APPOINTMENTS-SECTION of the monitor only when an option for query the system has been submitted, option detected formerly in the appropriate

section of the monitor. The QUERIES-SECTION contains a PERFORM RETRIEVAL statement that will call the corresponding non-resident segment to the main memory, and which main functions will be to disseminate the set of answers and to take care of relevance levels and accounting of hits and ratios in which respects to the total number of elements in the file as was explained before.

The hits tape will be treated afterwards by the OUTPUT, non-resident, segment and the performing of the actual output will be made in accordance with the selected options. In figure 8, page 52, it is shown the aspect of the source program that contains the queries.

It has been pointed out formerly that some additional information is needed in which respects both to the set of obtained answers and to the output printout or papers tape punching. This information, in which respects to the queries option, is explained below. The updating options are described under INPUTS TO THE SYSTEM and the output description is given in OUTPUTS FROM THE SYSTEM. It has been considered too that queries related with the printout of the full reference tables, Subject, Geographical, Taxa Terms, etc., will be treated under the Queries option.

Parameter Cards Indicating Type of Processing Desired.

Three main considerations have been made for these cards:

- A. All the cards will be punched starting always in the indicated column. A set of words for clarifying the meaning has been added, nevertheless these words will not be recognized for the input procedure, but only those that will set up indicators for the actual processing.
- B. It has been considered as convenient to reproduce the actual query written in COBOL including it in these parameter cards in order to print it in the output form. This will add clearness in the reading of the answers, providing the original request for the data obtained. This query will not be the one that will be executed, it is just for comments in the output.

Nevertheless this duplication may be considered an inconvenience for the users, the actual coders, and may be avoided in the output procedure.

- C. As there is not properly a scanning procedure for this parameter cards, due to the economy in the system, the columns mentioned below as the ones that will contain the information should be respected. If not, an error indicator will be set up and the procedure will stop, as shown in figure 3, page 12, being this error of non-recovery type.

The actual format of the cards follows:

1. First card which describe the option selected and the solicitor of the processing:

<u>COLUMNS</u>	<u>WILL CONTAIN</u>
1-10	OPTION __ IS __
11-22	RETRIEVAL __ _ _
23-36	. __ PROPOSED __ BY __
37-77	Name and/or code for person or institution requesting data.
78-80	Sequence number. In this case will be 001.

This first card, except for the word "RETRIEVAL" is the same that is utilized in other options. The field between columns 37 and 77 is the one that will provide the user's identification either in the input or in the output, and that will be punched in human-readable form in the paper tapes for dissemination.

A Bibliographic Data Retrieval System

S. AHUMADA

The character is used here instead of space, but actual spaces will be used in the coding.

2. The second card contains information about the type of searching that will be used and if the query is for the bibliographic files or for the reference tables.

<u>COLUMNS</u>	<u>WILL CONTAIN</u>
1-10	Blanks
11-28	SEARCHING <u> </u> WILL <u> </u> BE <u> </u>
29-38	The words: RESTRICTED or UNLIMITED or NORMAL <u> </u> <u> </u> <u> </u> <u> </u>
39-57	<u> </u> UNDER <u> </u> THE <u> </u> SYSTEMS <u> </u>
58-63	The words: FILES <u> </u> or TABLES
64	. (PERIOD)
65-77	Blanks
78-80	Sequence number, in this case 002.

The terms Restricted, Unlimited and Normal are understood as in page 47 above.

3. The third and following cards will contain the reproduction of the query that will be printed in the output.

A Bibliographic Data Retrieval System

S. AHUMADA

<u>COLUMNS</u>	<u>WILL CONTAIN</u>
1-10	Blanks
11-19	QUERY <u>IS</u>
20-77	Actual query given to the system.
78-80	Sequence number, in this case 003.

For subsequent cards needed for a large query the columns 1 to 19 will contain blanks; the last card will contain *Epp* in columns 78 to 80 indicating End of Option.

For a complete example of the utilization of this cards, refer to figure 11 page 66 .

OTHER POSTRETRIEVAL PROCEDURES.

In addition to the statistical procedures suggested above for giving information about the answers found in which respects to the total amount of items in the files, seems as convenient to keep records on the most frequently queried fields.

These records will provide, after a certain time in the utilization of the system, a true idea of the actual behavior of the botanist-bibliographer-computer complex instead of the assumed one through these pages. It is quite probable that some things will be changed as a result of those mentioned figures, maybe the sorting order of the master files, maybe the size of the record and even the fields in the magnetic tape record due to the fact that the information mostly requested does not concern with some items and on the other hand the botanist is badly needed of unexistent ones. Nevertheless at the present stage it is almost impossible to predict which will be the changes and when will be made and only a general approach has been suggested.

V INPUTS TO THE SYSTEM.

The general considerations under which the procedure that will accept data to the system will be designed are mentioned in the following pages. It has been said before that paper tape will be used as a rule as the input data support and same considerations should be made in this sense.

It is possible to predict that although the headquarters of FNA Project have been set up at the Smithsonian Institution, a considerable amount of work will be done in other places both in which respects to bibliographic and editorial aspects. It has been established too that the main operations in which concerns to Data Processing will be made at the Information Systems Division through the utilization of a Honeywell 1200 Computer System at least in the first stage of the Project. This centralizing of the operations will create the need of a common media for communication from the various sources from which it is possible to receive and/or to send information concerning the Project.

It is possible to predict too that some of the Institutions participants in FNA are utilizing paper tape equipment up to this moment and that some others are in a position to acquire it either only for the purposes of the Project and/or for other needs. It seems reasonable to assume that not all of them will may dispose of the same type of equipment, with the same codes for the representation of the information.

The preceding reasoning has suggested the development of a procedure for the manipulation of different kinds of paper tapes, with different codes and levels.

This procedure will be capable of reading any paper tape, convert the information to the H-1200 code to any paper tape code and punch the translated information, mainly for dissemination purposes; all of these accomplished through the utilization of translation tables.

These tables will be permanent parts of the system and will be accessed by the Paper Tape Procedure, written in EASYCODER as was mentioned before, and will be updated and increased in number as modifications in the utilized codes and notifications of new equipment installation will be submitted.

INPUT PROCEDURE.

It is one of the non-resident segments of the system and appears as UPDATE in Main Considerations in the Design of the System. The main functions of this piece of the system will be:

- A. The acceptance of data from paper tape or cards with the corresponding error checking.
- B. The translation of the information contained in some fields as Taxa, Geography, etc., to a coded form, more efficiently handled during internal operations, mainly in the searching for retrieval procedure, through the utilization of the tables mentioned in Source Data Automation.
- C. The sorting of the "input file" joined by the major sorting keys Taxa, Subjects and Geography, which seem to be the most frequently used fields in the record, through the utilization of the sort procedures, 11 and 12, non-resident too.
- D. The updating of the master files after the sorting procedure with the error checking involved. In this section care should be taken in the development of a checking for duplications algorithm that will avoid the presence of more than one record for one bibliographic entry.

This is especially needed due to the fact that the system will handle inputs from various sources which eventually may submit the same data with the corresponding inconveniences in the output.

- E. The updating of reference tables for the system such as Subjects, Taxa, Geography, all miscellaneous descriptors of publications, etc.

Seems as convenient due to the possible size of those tables to keep a separate tape from which they may be read to the memory. This makes easier too the updating and searching under these data.

- F. The establishment of error indicators during the whole processing and the utilization of the Error Recovery Procedure when needed, and the indications to the users of duplications or invalid data submitted.

It has been mentioned that this procedure will deal with both paper tape and cards, which will make the input to have more flexibility, not depending on the functioning or not of any of those devices. Nevertheless the paper tape will be the most used media.

Acceptance of Data.

To accomplish the function of handling various types of paper tapes, through the utilization of the Paper Tape Procedure, the following considerations were made. They are based upon the idea of giving the user of this system to most possible liberty both in the utilization of his own source data automation equipment and in the internal formatting of bibliographic data that his institution is accustomed to manipulate. In reference 13 some useful ideas are shown, which are used in this paper.

The source data automation equipment is identified by a "header record", which in order to avoid misunderstanding and speed up the process will be submitted to any user by the data center, that contains the necessary codes for recognizing the type of paper tape suspected. This recognition permits the system to select the appropriate tables for translation and therefore it is able now to continue with the tape reading.

These first codes should be equipment independent and in order to have the most of generality only five levels in the tape should be punched.

From now on the computer will be able to "understand" the source data submitted. Nevertheless, there still remains the human error possibility in both the source data automation and computer operation processes. In order to avoid this trouble, which is mainly of identification of tapes sent or received or processed and their origins and destinations it was augmented to the ideas in 13 a "human readable" set of characters formed by the different codes in the paper tape punch equipment.

Each character is represented by 5 or 6 continuous frames which, depending on the code punched in each one will form a picture. The characters utilized in those five or six frames do not need to have any relation among them. For example the letter A in a pictorial appearance may be formed by the characters %***% of some paper tape punch equipment.

An ideal set of those characters is shown in figure 9, page 60, and it is possible to see that combination of frames for any character may be found for practical purposes with the available equipment.

The whole block of pictorial data will be formed by the word INPUT, which will avoid a possible confusion in which respects to the traffic of the spool of tape, and the identification of the person or institution that submits the data. This identification set will be the same used in the first Option Card in columns 37 to 77.

A Bibliographic Data Retrieval System

S. AHUMADA

A translation table will permit to process those pictorial data either for verification in the input or for punching in a similar way, except that the first characters will be OUTPUT, the identification of a requestor, before his results.

The advantages in the information dissemination process may be inferred from the discussion above.

This two parts of the record, indicative of equipment and identification of participant, will be made only once for each participant and punching device and they will be copied as leaders of each spool of tape sent to the data collection center.

The next entry in the "header record" will identify the information itself that will be received and the characters with special meaning, assigned by the user, in the tape, such as field separators, etc.

Each field will be numbered, in accordance with each participant desires, and the correspondence with the established format in the magnetic tape is made by the association with special declaratives.

These declaratives may be:

DECLARATIVEMEANING

SYMBOL __ IS __

Identifies the separator among fields in the paper tape.

START __ MESSAGE __ IS __

Identifies the character used as "beginning of data" in the paper tape.

END __ MESSAGE __ IS __

Idem for "end of data".

DISCARD __ MESSAGE __ IS __

Error corrector or delete indicator.

TAXA __ IS __

Identifies the user assigned number for the standard field TAXA in magnetic tape.

A Bibliographic Data Retrieval System

S. AHUMADA

AUTHOR__IS__

IDEM for the AUTHOR field
in magnetic tape.

Etc.

The symbol__ indicates space.

The character or number to be used follows each one of those declaratives. There is no restriction either the numbers or in the declaratives in the sense that they may refer to reference tables.

After this leader information the actual data is now punched. A similar procedure may be used in the output, formatting for each paper tape device. In figure 10, page 61 a full example shows all the concepts used through the former lines.

PARAMETER INPUT CARDS.

In the same form that were utilized for the Retrieval Procedure, certain cards containing information about the main indicators to be established for the functioning of the system are described for the Input Procedure.

The same ideas and restrictions that were mentioned before are applicable in this case.

1. First card will indicate the option selected and the participant's identification:

COLUMNSWILL CONTAIN

1-10

OPTION__IS__

11-22

The words:

CONSTRUCTING or
UPDATING__ _ _ _

A Bibliographic Data Retrieval System

S. AHUMADA

- 23-36 . PROPOSED BY
- 37-77 Name and/or code for person
or institution submitting data.
- 78-80 Sequence number, in this case
will be 001.
2. The second card defines the utilized input media,
the files that will be considered and if this option
is the last one or a Retrieval Option follows:

<u>COLUMNS</u>	<u>WILL CONTAIN</u>
1-10	Blanks
11-25	<u>INPUT</u> <u>MEDIA</u> <u>IS</u> <u> </u>
26-29	The words: CARD or TAPE
30-52	. <u>JOB</u> <u>UNDER</u> <u>THE</u> <u>SYSTEMS</u> <u> </u>
53-58	The words: TABLES or FILES <u> </u>
59-65	. <u>IT</u> <u>IS</u> <u> </u>
66-68	The words: THE or NOT
69-77	<u>LAST</u> <u>JOB</u>
78-80	Sequence number, in this case EZZ, end of option.

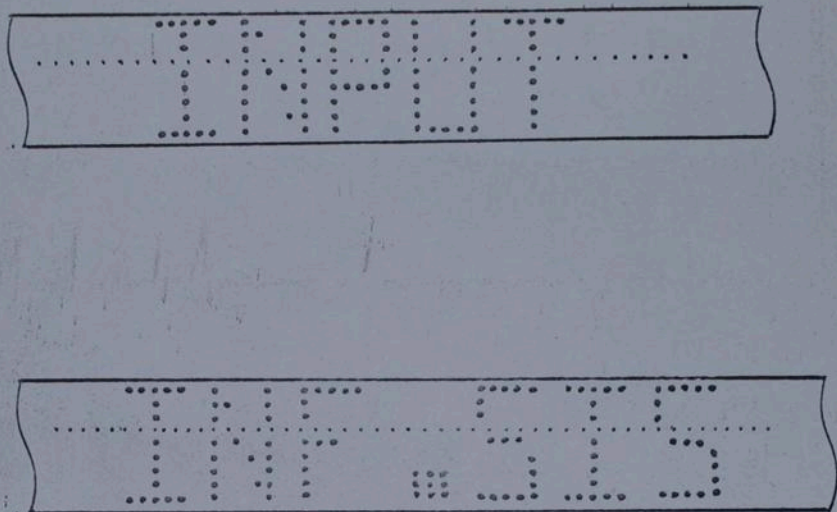


Figure 9. An example of an ideal 8 levels by 5 frames pictorial set of characters in paper tape.

A Bibliographic Data Retrieval System

S. AHUMADA

INPUT BOT. CLUB. OF AMERICA. symbol is #

start message is @ end message is % discard message is \$

title is 01 first author is 02 dates first author is 03

other taxa is 20@#01#phylogenetic relationships of ten

species of lobelia#02# braun, stephen, 1 #03#18901933#

cccc#29# inplata cardinalis #%

Upper case letters indicate pictorial characters, lower case ones, normal, punched characters in the code of the punch equipment.

Figure 10. An example of the appearance of the leader record in the paper tape submitted for updating files. The characters are supposedly punched.

VI OUTPUTS FROM THE SYSTEM.

The functions that this piece of the system will accomplish are related with the reading of the "hits" tape, the one that contains the answers to the queries and/or the reference tables requested, and to perform the final printout and/or punching paper tape of its contents.

In which respects to the maintenance of the system it will print too, the reference indicators of its behavior as was suggested in **QUERIES TO THE SYSTEM**.

The first printout will be the analysis of the behavior of the system, which is directly oriented to the bibliographers. It will be produced automatically after the retrieval procedure.

The second one, oriented to the user, will contain:

- A. The query made by the user.
- B. The set of answers.
- C. The value of the set of responses in relation to the files.

The appearance of the final information may be adjusted to the users desires too, for example in which respects to data that could be superfluous for the purposes of same person.

In other parts of this paper references were made to the treatment of the paper tape and the identification codes it contains. It will only be remarked here that the word output and the identification of the requestors, provided in the Option cards, will be punched in pictorial mode as a leader of the blocks of data retrieved.

Seems to be convenient to define with similar words to those used in the retrieval procedure the amount of data that will appear in the final printout, i.e.: NORMAL, when the whole record will be printed; RESTRICTED, when fields like Author's Dates, Others, etc., may be omitted; and other possible restrictions.

Although a sketch of the OUTPUT form is not presented in this paper it is suggested that:

- A. The organization in the appearance of data in this form will be from left to right, from the most important forms such as Taxa, Subjects, Geography, etc., to the less important ones.

This disposition will have the advantage that a descimination of the answers can be made in a very easy form, if the terms of the query correspond with those in the left end of the output form. This will make possible to sort in some instances by an order other than the normal one in the Master File, the hits tape.

If the major terms in the left end do not correspond with the expectancies although they matched with the query, there is no reason to read the whole set of indicatives. On the other hand if the expected quality in the information requested is found in the major terms, in this left to right graded sheet, details will become important and may be assessed in the usual reading order.

- B. Enough headings and lateral notes may be added to clarify even more the appearance of the information, through the utilization of characters such as arrows or asterisks denoting ends of fields and special remarks.

PARAMETER OUTPUT CARDS.

Following with the same idea than in the Retrieval and Input procedures it is shown here the corresponding ones for the Output.

1. The first card will indicate the option selected:

<u>COLUMNS</u>	<u>WILL CONTAIN</u>
1-10	OPTION __ IS __
11-22	EDITING __ _ _ _ _
23-36	__ PROPOSED __ BY __
37-77	Name and/or code for person or institution.
78-80	Sequence number, in this case will be 001.

2. The second card defines the type of output desired and the equipment to be used:

<u>COLUMNS</u>	<u>WILL CONTAIN</u>
1-10	Blanks
11-34	OUTPUT __ MEDIA __ IS __ PRINTER __
35-37	The words: AND or NOT
38-61	__ PAPER __ TAPE __. EDITING __ IS __

A Bibliographic Data Retrieval System

S. AHUMADA

62-71

The words:

RESTRICTED or

NORMAL _ _ _ _

78-80

Sequence number, in this case

E00, end of option.

In case that these cards are omitted, the system will produce a normal output in the printer without punching paper tape.

VII CONCLUSIONS.

A preliminary approach to the bibliographic system needed for FNA has been discussed.

It has been pointed out that this will not be the final system for the Project and on the other hand that its implementation and programming should be finished in the less amount of time. In this respect seems to be possible to complete the programming stage in a time not greater than 7 or 8 months.

The coding in COBOL will speed up the whole process and the EASYCODER program would take a reasonable amount of time to be finished. On the other hand the modular organization of this system may provide a natural way of work distribution which may even go to a complete parallelism.

Care should be taken in the attention and supervision to the behavior of the system by both the botanist and the programmer in order to make the necessary changes and improvements accordingly with the experience, same that will be applied in the time shared system.

VIII BIBLIOGRAPHY.

1. Draft of National Science Foundation Proposal for Flora North America Project. March 1967.
2. COBOL - A Low Cost Route Through Query Language Problems. Reginald A. Creighton. Smithsonian Institution.
3. Honeywell EDP. Software Manual. Series 200. COBOL Compilers D & H. Order No. 065.
4. Preliminary Report on the Marc Pilot Project, Library of Congress, October 1966, Washington, D. C.
5. Sydney W. Gould, "Family Names of the Plant Kingdom", International Plant Index, New Haven and New York, 1962.
6. Sydney W. Gould, "A Geo-Code Method of Identifying Geographical Areas", International Plant Index, New Haven, 1966.
7. Encyclopedia Britannica World Atlas, Encyclopedia Britannica Inc., Chicago, Toronto and London, 1947.
8. Computer Literature Bibliography 1946-1963, U.S. Dept. of Commerce, NBS No. 266, 1965.
9. Sydney W. Gould, "Authors of Plant Genera", International Plant Index, New Haven, 1965.
10. Charles Davies Sherborn, An Index to the Genera and Species of the Foraminifera, Smithsonian Institution, 1955.
11. Honeywell EDP Software Manual, Series 200, Tape Sort C(V) and Collate C(V) Order No. 207

12. Honeywell EDP Software Manual, Series 200, Tape Sort C and Collate C, Order No. 018.
13. A Generalized Data Collection Process, Reginald A. Creighton, Smithsonian Institution.
14. Honeywell EDP Hardware Manual, Series 200, Programmer's Reference Manual, Order No. 139.