DOCUMENT RESUME

ED 029 672 52 LI 001 515

By-Squires, Donald F.

An Information Storage and Retrieval System for Biological and Geological Data. Interim Report.

Spons Agency-Office of Education (DHEW), Washington, D.C. Bureau of Research.; Smithsonian Institution, Washington, D.C. Information Systems Div.; Smithsonian Institution, Washington, D.C. Museum of Natural History.

Bureau No-BR-7-1159

Pub Date Jan 69

Grant-OEG-1-7-071159-4425

Note-37p.

EDRS Price MF-$0.25 HC-$1.95

Descriptors-Administration, *Biological Sciences, Computer Programs, Costs, Electronic Data Processing, *Geology, Information Needs, *Information Retrieval, Information Storage, *Information Systems, *Museums, Natural Sciences, Orientation, Systems Development

Identifiers-*Smithsonian Institution

A project is being conducted to test the feasibility of an information storage and retrieval system for museum specimen data, particularly for natural history museums. A pilot data processing system has been developed, with the specimen records from the national collections of birds, marine crustaceans, and rocks used as sample data. The research includes design of a computerized system, requisite programming, and analysis of the requirements for a data processing system sufficient to meet the needs of a large natural history museum. Such a system is intended to be generalized and applicable to the requirements of other smaller museums and to networks of museums. Study of the costs of museum data processing by conventional methods and by computerized systems are being considered together with the organizational requirements of museum administrations to meet the impact of data processing technology. Recommendations cover the inclusion of specimen records from the Division of Meteorites, completion of programming for the inverted files, design and programming of the Systems Executive-Retrieval Generator and the query system, user experience trails, management analysis, experimentation with techniques of interrelating the scientist and the data files, and orientation and instruction of users. (Author/JB)

INTERIM REPORT

Project No. 7-1159

Grant No. OEG-1-7-071159-4425

AN INFORMATION STORAGE AND RETRIEVAL SYSTEM FOR BIOLOGICAL AND GEOLOGICAL DATA

Donald F. Squires

Director
Marine Sciences Research Center
State University of New York
Stony Brook, New York 11790

Associate for Natural History Information Resources
Office of Systematics
Museum of Natural History
Smithsonian Institution
Washington, D.C. 20560

January 1969

U. S. Department of
Health, Education and Welfare

Office of Education

Bureau of Research

Contents

Interim Report

Project No. 7-1159

Grant No. OEG-1-7-071159-4425 (095)

An Information Storage and Retrieval System for Biological And Geological Data - The First 18 Months

Donald F. Squires
State University of New York

Washington, D. C.

January 1969

The research reported herein was performed pursuant to a grant with the Office of Education, U. S. Department of Health, Education and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

PREFACE

The project described in the following pages was commenced in response to an urgent need on the part of the museum community. The impact of the research is underscored by the numbers of conferences and symposia being held about the world in response to the stimulus of this project. While the direct research results are of significance in themselves, a more important but subtle change may be occurring in response to the important shifts of role now being conceived within the museum community. The results of the latter together with direct research results have been the subject of a number of papers in professional journals as recorded in the bibliography. More will be forthcoming in the second phase of the research project.

Undertaken as a joint project, the research reported on could not have been carried out without the support of the Museum of Natural History and the Information Systems Division of the Smithsonian Institution. We wish to acknowledge the support and assistance given by Dr. Richard Cowan, Director, Museum of Natural History and Mr. Nicholas Suszynski, Director, Information Systems Division. Principal contributors to the project other than the authors of this report have been Mr. Dante Piacesi, Mr. Richard King, Mr. Howard Baldur and Mr. Kenneth Ebbs (no longer associated with the project)and Drs. Raymond Manning, George Watson and William Melson.

NumeriClature, a technique for computerization of biological nomenclature was developed through the particular efforts of Creighton, Ebbs and King. Global Reference Code, a new approach to geographic data storage and retrieval was conceived by Creighton and Piacesi and implemented by them and Mr. Warren Minami.

SUMMARY

Museum collections are rich resources for study and research by educators and students as well as the lay public. In their role as educational es' ablishments, museums may contribute more fully to the educational process by making both museum collections and the information about the specimens in the collections contained in data files (which may be chaotic and incomplete) more accessible and available to the community of scholars. Rising costs of collection maintenance and increasingly enormous new collections being deposited have prevented many museums from being able to fulfill this function. The research undertaken through the support of this grant is designed to test the feasibility of an information storage and retrieval system for museum specimen data, particularly for natural history museums.

A pilot data processing system has been developed to test the applicability of data storage and retrieval techniques to data associated with museum specimens. Sample data are specimen records from the national collections of birds, marine crustaceans and rocks. Data recorded in the field and in the laboratory are prepared in machine-readable form as a part of the specimen documentation process, read into a computerized system of storage, and retrieved according to the requirements of the student, researcher, or other requestor. The principal objective of the system is to make museum information more readily accessible.

The research includes design of a system, requisite programming, and analysis of the requirements for a data processing system sufficient to meet the needs of a large natural history museum.

Such a system is intended to be generalized and applicable to the requirements of other smaller museums and to networks of museums. Study of the costs of museum data processing by conventional methods and by computerized systems are being considered together with the organizational requirements of museum administrations to meet the impact of data processing technology.

INTRODUCTION

Museum collections are rich and important resources for study and research, but increasing costs of collection upkeep have forced many educational institutions to abandon the study collections amassed during years of collecting on the part of faculty and students. As a result of this loss, students in many universities do not have the first-hand access to museum information as had been the case when the university museum was not only a "place to visit" but was also a resource for study. Faculty researchers must now travel to distant museums to select, and/or study materials, or to simply record the data associated with the specimens. The research undertaken in this project contributes directly to the educational process by facilitating the use of these important resources by making the information housed in museums more accessible. There now exist only increasingly expensive and inadequate means to meet this requirement of museums.

Museums are also faced with the problem of the increasing accumulation of data associated with individual specimens, and the necessity for still greater detail to be recorded if the collections are to serve a useful function in today's and tomorrow's science. Numbers of specimens being housed by museums is sharply increased not only in response to the thrusts of today's larger science, but because of the emphasis in environmental studies, an area of investigation closely akin to "natural history" for which study museums were created.

To test the feasibility of applying data processing techniques to museum collections and the information associated with the specimens, a pilot project was designed. Sample data are specimen records from the national collections of sea birds, marine crustaceans, and rocks. The objective of the project is to make the information about these specimens more accessible through data processing, not to meet the task of filing specimens. Data recorded in both the field and the laboratory, including the documentation of the circumstances of the collection of the specimen, the nature of its environment, and also the results of subsequent study of the specimen are prepared in machine readable form, read into a computerized system of storage and retrieved according to the requirements of the student, the researcher or other requester.

In defining the descriptors of these collections, it was recognized that the questions posed to the data bank developed in the project would invariably contain either a unit of the taxonomic hierarchy by which biological classifications are structured, and or some geographic designator. Thus, these two characteristics became the primary cross-reference subject fields, with secondary fields including parameters of vertical distribution (depth, altitude or geologic age), name of collector and date of collections. All other data would be structured as subsidiary files. By the means of unique catalogue numbers for each specimen or groups of specimens, data are associated directly with the specimen.

Unfamiliarity with museum collections may raise the question as to the importance of the primary and secondary characteristics. It is important to fix the specimen in the three dimensional space

of the earth through geographic and altitudinal reference axes. In the instance of geological collections, geological age is the analogous characteristic to the vertical dimension, expressing the stratigraphic succession. The biological (and to a lesser degree, petrologic and mineralogic) classification schemes are hierarchically arranged, greatly structured tree-like systems, developed and maintained under elaborate international regulations which permit us to develop a truly universal reference framework not dependent upon language, politics or other variables. Collector and date of collection while of importance historically, also serve to relate the specimen to a fourth dimension of time. This has become particularly important as man irrevocably modifies his environment. Increasing use of museum collections for pre-pollutant baseline analyses has emphasized the importance of the date of collection as a reference point.

In accordance with the axiom "capture data at the earliest opportunity" we have utilized punched paper tapes as the input medium, permitting flexibility in the input operation and long record length. CDC (formerly SCM) Typetronic 2816 and Friden 2301 Flexowriters are used so that the creation of a label (the document which is intended to be permanently associated with the specimen) or other paper record (file card, transactions record, etc.) starts the information processing. Using this tape to drive other printers, a variety of file documents are created, while the master tapes serve as input to the computer file. This aspect of the project has been fully depicted in a previous publication (Squires, 1966).

Caught between increasing size of collections and increasing maintainence costs, many museums have been forced to retrench severely. As a result, a number of important effects upon museum behavior has been noted:

1) Museums, already badly behind in documenting their collections, are facing larger backlogs of cataloguing than ever before. (Table 1).

2) Museums are being called upon to supply more information to the academic community than ever before, for the use of collections is not diminished, but rather has increased. However, costs of handling collections within museums have been increased because of the need for investigators to borrow collections. (Table 2).

3) Students, because they are not now in close proximity to collections, are not as aware of the important resource museums constitute, and may duplicate collecting efforts wasteful of both manpower and funds.

4) Students who do find their way to collections are often discouraged by the problems of manually extracting information from the collections and are dissuaded from following a line of research or a potentially rewarding career in systematic studies of museology.

Great impetus has been given to the general scientific problem of the quality of the environment as the result of a growing public awareness of the consequences of pollution and the general deterioration of our habitat. Constructive legislative actions of responsible political figures have resulted in a major national effort to understand

Table 1

Growth of the natural history collections of the Smithsonian Institution.

Year	Increase	Total Specimens
1964	---	47,731,000
1965	1,234,000	48,965,000
1966	1,281,000	50,246,000
1967	988,000	51,234,000
1968	937,000	52,161,000

Table 2

Summary of scientific visitors to the Museum of Natural History, Smithsonian Institution, during 1967.

Department	Visitors for one day	Visitors for more than one day	Total Visitors
Anthropology	425	1,420	1,845
Botany	75	125	200
Entomology	185	221	406
Invertebrate Zoology	80	117	197
Mineral Sciences	250	986	1,236
Paleobiology	130	367	497
Vertebrate Zoology	170	790	960
Total Visitors	1,315	4,016	5,331

the nature of the environment and the management of the factors which control it. There is a growing awareness of the importance of museum collections and the information associated with museum specimens in the documentation of environmental changes in both direction and magnitude. Growing use of collections in response to this national program creates new problems for the museum curator. New combinations of data are desired at a time when the museum profession because of reduced workforces has generally retreated into maintaining only narrowly structured reference files principally for the use of systematic biologists. Renewed awareness of the importance of collections as a part of the information resource of the biological sciences in particular, has come late to a museum profession still proudly conscious of the pre-eminent role of museums 60 years ago(Crompton,1968).Pressure exerted through the professional societies by committees of the American Institute of Biological Sciences, the American Geological Institute and other national organizations, has forced the scientific community to be more aware of the informational problems faced in the sciences and of the alternative technologies available to them in attacking these problems. Treatment of the informational resources of the museums in a complementary fashion must proceed hand in hand with these other projects, or the museums will be perilously isolated from the forward motion of science (Squires, 1969).

One may not assume from the above that museum collections may be vestigial. Rather, the scientific community, faced with the literature explosion, with enormous data gathering potentials, and the other aspects

of what is often referred to as the information overload, has declared the museum collection problem to be the responsibility of the museums to solve. This problem can best be summarized as one of increasing collections being less well documented - information loss is probably at its greatest now - with the resultant effect of the collections having progressively lessened significance and utilization in the broadest scientific context. Examples of the potential reversal of this trend may be cited in the interest now exhibited in the Smithsonian's collections by other Federal agencies who are aware of this important information storage and retrieval project (Squires, 1968).

To solve the "musuem problem" in isolation from other informational problems of today would be relatively easy and completely useless. To apply inventory control techniques to museum collections is not difficult, but does not reach to the crux of the problem. Our experience over the past 18 months in the development of this project indicates that museum personnel may not fully understand the dimensions of their own despair, so fully submerged in the bewildering backlogs are its curators. Museum specimens are an informaticnal resource as are books in a library. If the task is only finding the book (specimen), simple techniques can be applied; rather, we feel the task to be one of ascertaining the full content of the book indexed in such a fashion so that selected information may be retrieved. Here the analogy ends, for unlike a book which is written, bound and shelved, specimen data is continually enriched by subsequent study and the importance of the specimen increases almost geometrically with its utilization.

In the initial period of this project, the investigators have become increasingly impressed with the amount of data not being recorded, due to the inability of existing systems to capture data and because of the complete breakdown of existing retrieval systems (the curators). We feel that the data acquisition and registration by the working scientist will improve when he can utilize an adequate information storage and retrieval system, because he will know that these data can be recorded in a fashion which will be useful to him in the future. Museum curators will better manage the data acquisition procedures in their own units so that more information is obtained from personnel, both in-house and from other collectors. Finally, and perhaps most significantly, greater and better use will be made of museum collections in education and research endeavors through the availability of the information contained in museum collections.

METHODS AND RESULTS

Historical Perspective:

The present research project is a direct outgrowth of a series of discussions begun in 1963 when the Director, Museum of Natural History, Smithsonian Institution, appointed a committee, with the author of this report as its Chairman, to develop a general understanding of the potential of data processing for the museum community. The discussions of this committee led in 1965 to a contract with Federal Systems Division, IBM, funded by the Office of Systematics, Smithsonian Institution, to study the feasibility of the codification of biological names. The results of this project indicated that codification was not required and suggested that biological nomenclature be treated as alphanumerically. In the same year the Management Services Division of Peat, Marwick and Mitchell & Co. made a comprehensive study of the utilization of computers by the Smithsonian Institution, submitting a report embodying specific recommendations for the development of a data processing system in the Museum of Natural History. Following these reports, the Museum of Natural History embarked upon a program of research, in conjunction with the newly founded Information Systems Division of the Smithsonian Institution, with the objective of developing a comprehensive proposal. The present research was initiated in July of 1967 through the support of the Office of Education, Department of Health, Education and Welfare.

The need for better information handling techniques for museums is clearly recognizable. During the first 18 months of this project, the broad outlines of a data processing system envisioned as responding to general requirements has been developed and implemented. In this period, input procedures were standardized for the cataloguing process utilizing punched paper tape units. Systems and programs were completed making possible the entry of data thus prepared into the computer, and its manipulation through the initial stages of file structure. Query capability was developed for an intermediate Work in Process file, i.e. a temporary holding file for input data pending verification and assignment of appropriate code data. Consideration of, and preliminary implementation of formats of output and protocols by which stored data will be recalled and presented to the scientist has been undertaken. Involved in the techniques, has been the development of a satisfactory numerical expression for classical Linnaean taxonomic binomials and hierarchy and the automated procedures by which such numericlature is assigned. A totally new concept for the manipulation of geographic terminology was developed when anticipated techniques failed to materialize. The new technique, Global Reference Code, provides totally automated input once a distinct expression of locality is established.

1. Data Preparation and Input

Experience of the past 18 months indicates that our understanding of the input problem was deficient, and our experience with punched

paper tape input machines was not adequate to plan manpower requirements for data preparation. (For a full discussion of source data automation see Squires, 1966). Experimentation with various alternatives during the summer of 1968 has indicated that greater efficiency of personnel and machines is obtained when the tasks of data preparation and of data input are separated. Both functions have been found to have lesser educational requirements than predicted, provided that adequate supervision is given to personnel. Further, the special necessity for generating a large data base more rapidly than would be done under normal circumstances (in order to provide a substantial body of information for manipulation in the computer and against which meaningful queries may be made) dictates an ususually large staff.

It was originally intended to have a cataloguer prepare all specimen records and type them, but we find it to be more efficient to have one cataloguer preparing records and a typist entering them. Nine or more technicians (level 7 through 9) were engaged in preparing data for three machine operators in the course of their normal work. The cataloguers supplied by the project are more effectively occupied in collating data, assembling materials, and in checking the data submitted by the others, and in general supplementing existing museum cataloguing and technical support. The effectiveness of this arrangement is demonstrated by an increase in input processing from 100 specimen records per week to as many as 100 per day under the new system. Other benefits include better reliability of input record, less operator fatigue, more complete specimen records (museum technicians having more experience, prepare more complete records than less experienced cataloguers). Further, turn around time for input

data is reduced by having the project cataloguer available to proof-read input records.

We now appreciate that greater numbers of persons than originally projected must be associated with the input of data if a data base of any size is to be generated. Although efficiencies have been made in data preparation through the use of punched paper tape systems, these efficiencies do not of themselves generate the volume of data required with only the previously existing workforce.

Several areas of the input operation have been found to be particularly troublesome: 1) Data preparation - existing museum informational files have proven to be notoriously incomplete and surprisingly inaccurate. The result is increased cataloguer time devoted to the preparation of specimen records than had been anticipated and a far more critical requirement for proofing of input statements. The technique of input preparation provides an initial opportunity for the cataloguer, or machine operator, to visually scan the record, but additional proof-reading is required once the punched paper tapes have been read by the computer. Present procedures for the construction of a holding file "Work in Process", calls for input of the punched paper tape through a reformatting program which makes several fundamental checks against the quality of data. Specific among problems for cataloguers are the translation of linear numerical units into standard expressions, (i.e. weight from ounces and pounds into metric units, or into decimal equivalents of pounds, or fractions of ounces, etc.) Conversion from non-metric to metric units is done by the computer, but it is necessary for input personnel to restrict the expressions to a series of conventions. 2) Compilation of numericlature files. The decision to handle input of biological and geological names as alphabetical entities required the development of a hierarchy of classification and a thesaurus through which each generic and specific name

Table 3

Specimen data recorded for pilot project collections

Crustacea	Birds	Petrology
Catalog Number	-	-
Generic Name	-	Petrographic name
Sub-generic Name	-	
Species Name	-	
Subspecies Name	-	
Author Name	-	
Locality (in four levels)	-	-
Latitude and Longitude	-	-
Collector	-	-
Collector's numbers	-	-
Date collected	-	-
Depth	Altitude	Depth/Altitude
Number of specimens	Collection code	Collection code
Sex	Description code	Number of specimens
Preservative	Donor's name	Donor's name
Collecting gear	Preparator's name	Chemical analysis (by percentage, element)
Identifier	Age of specimen	
Nomenclatorial type	Sex	Trace elements
Publication information	Fat	Radiogenic isotopes
	Skull ossification	Non-radiogenic isotopes
	Reproduction anatomy	Petrographic description
	Soft part color	Modal analysis
	Molt condition	Thin section(s)
	External measurements	Density
	Stomach contents	Associated rocks
	Ecological notes	Geological age
	Parasites	Mineral composition
	Related Specimens	Exhibition information
	Nomenclature type	Texture
	Disposition data	Analytical methods
	Publication information	References

combination would be approved for correct statement and by which familial, ordinal, class and phylum units would be appended. The name and associated classificatory information is then translated into numerical units by the computer (Numericlature).

Although the nomenclatorial thesaurus is open ended, and the classification scheme alterable, the problem of compilation of such taxonomic lists and the assignment of names to heirarchical catagories was unexpectedly time consuming. For the non-biologist it should be noted that for many groups of organisms "lists" of the names do not presently exist. 3) Assignment of geographic designators. Although initially it was hoped that the translation of geopolitical names into some computerized scheme which would permit facile handling within the system and multiple routes of access to the file for retrieval purposes, such a technique did not materialize. Global Reference Code (Piacesi and Creighton, in Press) most closely approaches the ideal, but requires the manual assignment of latitude and longitude to each new record of locality, with automated codification thereafter.

Because of the work requirements indicated by the foregoing, a minimum of two persons in addition to the cataloguers and the machine operators is required for the scope of work undertaken in this project. These additional persons have specific responsibility for programming of source data automation machines, maintaining numericlature on a current basis, assignment of Global Reference Codes to new geographic assignments and providing general supervisory support and coordination for data preparation. Work in each of these areas is greatly assisted through the computer analysis of the Work in Process File which indicates those records needing assignments of codes, and automatically assigns codes to those informational utilities it has already handled as well as automatically performing validation of most data

items. As the numeriClature and Global Reference Code thesauri are built up, fewer of these records will require manual alteration.

Through the first 18 months of the project 11,867 lots of specimens were processed representing specimen records on about 23,000 individual specimens. This task was accomplished through utilization of 9 cataloguers employed through the project, two of whom were principally engaged in source data programming, supervision and global reference coding. Numerous part-time personnel were utilized to provide numeriClature and Global Reference code data. Although more precise accounting of effort involved in the project will be developed for a final report, the following summary indicates the nature of the manpower requirement for the productivity indicated through the first 18 months.

2. Systems Design and Programming[1]

A number of constraints were placed on the original design of the system. Among the more important of which was the necessity for the system to be expandable to large capacity, for the present holdings of the Smithsonian's Museum of Natural History alone are in excess of 50 million specimens. There are over 3 million known species of plants, animals and rocks and minerals, each having an average of 10 synonymous names, for a total of 30 million names. Although at the outset we could reckon on recording an average of 14 data catagories for each specimen, this number did, and is still rapidly expanding, with the result that the record length associated with each specimen has increased dramatically to over 30 data catagories.

[1] A full and more technical description of the system is being published by the Senior Systems Analyst (Creighton, in press).

Table 4

Manpower requirements during first 18 months of pilot project

	Specimen record preparation	Specimen record entry	Preparation of Numeriçlature Files	Preparation of Global Reference Codes	Source Data Automation Programming	Supervision and other miscellaneous tasks
Grant supported man-months	---	58.0	22.0	4.5	7.0	16.5
Smithsonian contributed man-months	21.0	7.5	---	---	2.0	---
Production achieved	---	11,867 processed 8,000 backlogged 19,867 lots	40,389 names	1965 localities	---	---
Productivity	---	303.3 lots per man-month	1836 names coded per man-month	436 codes assigned per man-month	3.0 man-months per division	---

A significant requirement was that the system not require large numbers of new personnel, but be, once developed, adaptable within the museum structure and staffing. Fortunately museum technical and professional personnel are of high caliber and the techniques of data processing are rapidly learned, although as yet we have not made as dramatic a breakthrough in increasing the efficiency of specimen data handling as had been anticipated. The system, if it is to be utilized, must be compatible with existing card files and other visual records, none of which will be replaced within the present (and perhaps next) generation of museum curators. Further, because of the immense literature and the classification schemes deeply imbedded in both literature and the philosophy of practitioners of systematic biology, it was important that the system adapt <u>to the user</u> as completely as possible. Thus the requirements for the input of names, not codes, the flexibility of structural classifications of those names, and the necessity for dealing with geo-political terminology and other geographic references rather than an arbitrarily selected grid reference scheme.

Although important strides are being made in the technological problem of machine interfacing, we have been from the beginning concerned with facilitating communication between a non-computer oriented scientist and the data processing system and between various types of computers located at different centers. For these reasons, COBOL (Common Business Oriented Language), a subset of the English language was selected as the query language for the system. In this language the user states his requests for data in the form of "IF" statements: "if X-US Y-US <u>and</u> location is New Jersey then perform selection through retrieval." Thus the user indicates logical relationships between data catagories which will satisfy his request and this in turn con-

stitutes the query language. Although sacrifices were made to utilize COBOL, it remains a language which may be used on a variety of computers.

Through the first 18 months of the project, programs to edit and audit data were developed and general systems concepts were established. Programming is completed through a "Work in Process File", a data holding file in which input is edited and numeriClature and global reference codes are added. This stage is the final step before data are admitted to the "Data Bank". Processing beyond the Work in Process File requires complete validation of input data both by computer and the releasing division. In order to manipulate the Work in Process File, which will usually contain records of about 10,000 specimens, a limited query capability was developed. At this stage we are able to read input punched paper tapes containing specimen record data, to store this information in a holding file, and to perform limited queries upon that file.

Examples of the types of queries which have been performed to date, as specific searches, include:

1) If Country equals "United States" and state equals "Alaska" and locality equals "Bristol Bay" and breeding status equals "breeding" perform search through retrieval. Yield: a listing of these bird specimen records (printed in full in the Work in Process format) satisfying these requirements.

2) If genus equals "Gonodactylus" perform selection through retrieval. Yield: a printing of the specimen records of all species of Crustacean.

3) If collector-name equals "Ecklund, C.R." perform selection through retrieval. Yield: a listing of specimen records of specimens collected by C. R. Ecklund.

4) If genus equals "Dacite-0148"[1] and island-group equals "Mariana Islands" perform selection through retrieval. Yield: specimen records of Dacite-0148 from Mariana Islands.

The function of the initial stage is to accept input data, reformat from the various forms in which originating units submit the data, to edit and to add new information. Through the Numericlature Input Thesaurus, a numerical code[2] for handling names internally within the system, is assigned. A Numericlature Output Thesaurus relates author and date citations of the name and links synonymous names with the valid name. Input data are also examined and an Abstract Search Index prepared which permits, at a later phase, rapid searching of the text of specimen records. This later device, termed "Zipmode" search is especially important in providing economy of operation when a question is asked which involves association of a specific data field (which may or may not be empty) with a name.

Main storage for the system is the Data Bank in which specimen records are ordered by Numericlature (or, put another way, taxonomically) and within each species, sequentially by catalogue number. Specimen records are stored in their complete text. It is not necessary for the computer to search each data field entry when examining this file for particular types of records, for example, bird specimens for which molt data are recorded, for the computer inventories specimen records for which information is present in this data field by means of the Abstract Search Index. It then selectively searches those specimen records (identified by catalogue number) for which molt data are recorded.

[1]Rock terminology has been developed parallel to biological nomenclature. "Generic" equivalents are rock names adapted from Troger's Classification.
[2]The structure of this code is described in Creighton (in press) and will be the subject of an expanded paper now in preparation.

Because taxonomic sequence is utilized for the main file structure, it is important to clarify the methods used to construct this sequence. No fixed system of classification exists, and indeed, the science of systematic biology is far from ready to create a 'final' scheme. Rather, the taxonomies used are fluid and change with new discoveries, and are often the subject of considerable disagreement between researchers. We have adopted the principle that just as curators will order specimens on shelves by some arbitrarily selected scheme, so should they utilize the same arbitrariness in selecting a classification for the data bank. We have followed the general principle of utilizing a published, comprehensive classification wherever possible so that other users of the system can be aware of the groupings involved. Any scientist adopting the system can change the higher taxa assignments of any and all species as he so desires, constructing a classification to his own needs and concepts. This flexibility is vital to the system in order that it not prematurely harden the arteries of systematic biology.

Retrieval from the data bank is accomplished as follows:

1. A request. The requester writes simple conditional statements identifying the taxonomic unit of interest and other restricting parameters. 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. The computer will search the Nomenclature Input Thesaurus and associate both the numericlature code for the name Pelicanus americanus and the address for the specimen abstract record in the Abstract Search Index. A negative response from the Nomenclature Input Thesaurus means that the name has not yet been placed in the system or that the input

name was misspelled. A negative response from the Abstract Search Index would indicate that although the name had been included, no specimen records have been incorporated into the data bank. If records are present, they are than selected from the Data Bank.

2. A request containing a geographic term. The requester would write the additional qualifying term containing the geographic terminology. For example: if information was desired on the American pelican from Key West, the request would be written as follows:

If species equals Pelicanus americanus, and locality equals Key West, then perfrom record selection. The query cards containing the geographic term will initiate selection of the appropriate Global Reference Code polygon functions from the Location Query Thesaurus. The abstracts of the taxonomic group selected from the Abstract Search Index by the numericlature search will then be matched in Global Reference Code to nominate those records which will be retrieved.

3. A request containing a collection date. This particular parameter was selected to provide the third dimension to the collection, a characteristic of specimens which is of increasing importance. Should the requester be interested in Pelicanus americanus from any locality, but only those which had been collected in 1855, then the request would be written as follows:

If species equals Pelicanus americanus, and if date collected equals 1855, then perform record retrieval. Expression of date collected, entered in time-span query cards would be utilized to select those record abstracts which had survived the first screening.

At this step in the query process, the essential elements of what, where and when have been met. More specific requirements relating to special data fields will be met by first searching the Abstract

Search Index which will indicate those records containing data in specific fields, and then through a match of the data involved in the specimen records. The Abstract Search Index is expressed in the form of dichotomous yes-no choices recorded against the various data catagories, and serves principally to eliminate those records in which information is lacking in the requisite fields.

Specimen records which have met the requirements of the search and which have been located in the three major files, for which further restricting statements have been issued, met and retrieved, will then be copied onto an output magnetic tape in preparation for printing as output. Before the final printing is accomplished, the requester will have the opportunity to specify the printing format (for example, see Manning, 1969) and also to associate additional taxonomic information including synonyms, authors, dates of publication of the species, and such other taxonomic remarks which are incorporated from the Nomenclature Output Thesaurus.

Examples of interrogations presumed to constitute the genre of questions of the future are more complex and include;

Print the geographical distribution of the genus Gonodactylus collected at depths greater than 20 fathoms, listing the species alphabetically.

What species of the family Squillidae collected between 50 and 100 fathoms occur together in the Gulf of Mexico? Arrange list as cross-index listed alphabetically by genus and by species.

List all records of the Dendroica caerulescens collected above 3,000 feet, in May, June or July from the Appalachian Mountains. Arrange by months and alphabetically.

List all records of the orders Procellariiformes and Pelecaniformes which have recorded soft part colors or external measurements. Arrange list by catalogue numbers and cross-reference to an alphabetical listing.

List all records of Basalts from Mexico with SiO_2 content of between 42% and 45%.

List all records of Dunite with up to 5% modular chromite associated with amphibolite.

An essential part of the original work proposal was not only the association of specimen related data with the specimen, but also bibliographic citations of specimens with the specimen associated data. In making projections of the development of a data base of bibliographic citations of sufficient extent to permit experimentation in the manipulation of citations, our thinking was particularly focused upon the catalogues of the type collections (which we now recognize are too limited in scope and size, but seemed several years ago to be enormous because of our limited data handling capability). The vast majority of specimens lack bibliographic citations, or if they have cited in the literature, the fact has not been recorded with the specimen data. As a result, as the data base grows, the proportion of specimen records containing bibliographic citations decreases relative to the total data base and the number of returns per query decreases. Experience indicates that the expense of searching literature for missing specimen citations would be prohibitive and therefore will not be done for crustacea, birds and petrographic collections. As an alternative, we have enlarged the scope of the original project by the inclusion of meteorite specimen data in order to provide an adequate sample of specimen related bibliographic citations.

Meteorites are a special type of rock. The data requirement of the Division of Meteorites is not different from the Division of Petrology, and the input programs and the tape read programs for petrologic specimens may be fully utilized by the Division of Meteorites. Because there are a limited number of specimens of meteorites (approx-

imately 4,000 known specimens) each has been intensively studied. A succession of increasingly sophisticated technology applied to the study of meteorites has resulted in the accumulation of a large literature about each specimen. Each meteorite may also have been broken into several specimens and distributed through a number of museums, with each of the pieces receiving differing amounts of scientific scrutiny. Much of this literature has been brought together in a notable catalogue entitled Catalogue of Meteorites, by Hey (1966), which provides an easy mechanism for obtaining, in a single place, bibliographic citations to specimens. These citations may therefore be entered as a part of the specimen record as the specimen is catalogued.

RESULTS

Results of the first 18 months of work may be briefly summarized:

1) Complete turn around of data entry was accomplished from the stage of data preparation through production of punched paper tape, reading of the tape by the computer, assembly of the Work in Process File, and printout of that file.

2) Limited query capability was generated for the Work in Process File.

3) Completion of the systems for the automatic validation and assignment of numeriClature and Global Reference Code.

4) Entry of 25,000 specimen records into the system.

Accomplishments toward the goal of completing the system are measurable, but less tangible is the effect of the system upon the museum community. To learn what reactions we should expect, a group of five museum directors representing large, small, private and university museums were brought together for a three day briefing.

The purposes of the meeting were to present the system, even in its limited capability, and to test the educational value of an orientation program and its effectiveness in communicating the results of the research to date to the museum community. Each director was asked to prepare a report upon the meeting, specifically addressing four questions. The responses to three of these questions are summarized below (question 4 dealt with the briefing program and is not relevant here):

Question 1: Is the general design and performance of the data processing system something other museums, particularly smaller museums, should be aware of? In how much detail?

Summary Answer: Greater awareness of the progress of the project is needed in the museum community, and we must increase the number of publications describing it. Specific recommendations suggest that a kit on "How-to-get-Started" might be issued through the auspices of the U.S. National Museum, and that the capability of further enhancing this publication with an implementation team would be desirable. The system is considered to be a desirable adjunct to museums on all scales.

Question 2: What is the general impression of the system and of its potential utility?

Summary Answer: The system is viewed as potentially greatly increasing the value of collections and their utilization in education and research. Serious questions are lodged against the utility and value of putting data on older collections into a data processing framework. The ability to retrieve specimen related data not merely retrieval of information on the presence of a specimen meeting certain requirements is important. Specific recommendations arealso made that there

be a concerted museum effort towards improving the quality of specimen related data.

Question 3: What information should we particularly develop as a result of our research? What will you want to know? What kinds of data should we accumulate to demonstrate performance and to provide needed justifications for other museums to get started?

Summary Answer: It is clearly apparent that detailed cost studies are desired. These studies should include analyses of direct costs of the project, and also analyses of man-power requirements for implementation. It is, also, clearly apparent that comparative costs showing the effectiveness of electronic data processing as opposed to traditional data processing methods must be developed and that these analyses must take cognizance of the less tangible factors such as benefits to systematic research as practised in museums and to other types of research both in and outside of museums.

RECOMMENDATIONS

The original scope of work envisioned a three year pilot project. This is on schedule now. However, evolution of thinking, experience with the accumulated data base, the progress of data handling technology, and new scientific requirements have indicated new directions for data and systems design. In summary, the following tabulation indicates the general scope of the activities projected for the second period of the grant.

1. Continuation of input of specimen records from the Divisions of Birds, Crustacea, and Petrology to a working data base. Over 11,000 lots of specimens have been entered thus far. At an anticipated rate of 430 records per week, the expected data base for experimentation will be about 45,000.

2. Enlargement of the scope of work to include specimen records of the Division of Meteorites in order to provide a better data base of specimen-related bibliographic citations for data manipulation studies.

3. Completion of programming for the inverted files.

4. Design and programming of the Systems Executive-Retrieval Generator.

5. Design and programming of query systems with the benefit of the inverted files.

6. User experience trials with questions representing present and projected types of request for information are asked of the data base and individualized presentations of the responses are constructed.

7. Collection management analysis to indicate the feasibility of the information storage and retrieval system developed from a management viewpoint including cost analysis of alternative procedures.

8. Experimentation with various techniques of interrelating the scientist and the data files, including various output techniques, various dissemination processes.

9. Orientation and instruction of museum administrators, university faculty, scientists, and others, in the use of the system.

We confidently expect this project to result in the development of an information system feasible for all types of natural history collections, and that implementation of the system will be within the financial capability of the museum complex, both private and university. As the system is more widely adopted, information

flow through the scientific community will be enormously enhanced.

Dissemination of information about the project, its status and potential, will be accomplished through articles published in technical journals, in scientific periodicals, and in popular magazines. Members of the Smithsonian staff associated with the project attend symposia and meetings in which information handling is discussed and give talks before groups interested in modern museum methods. A trial program of instruction and training in data processing systems was held, and on the basis of this experience, a continuing series of such conferences is planned for the next year. To each of these the Museum Administrator and one of his staff are invited. We hope, through such pairing, that both the administrative and implementation personnel will become familiar with the problems and promises of the work. The sessions include fundamental instruction in data processing systems, utilization of the equipment in both entering and querying the system, and discussions of implementation.

From the onset of the program we have been vitally interested in the development of compatible information systems in other museums and universities. In carrying out this phase of the project, magnetic tapes containing Crustacea specimen records were transported to the British Museum (Natural History) for experimental data manipulation purposes. The National Museum of Canada has been a close correspondent in the development of the project and plans a very similar operation for their own museum, and eventually other museums in Canada.

Bibliography

Creighton, R. A. (in press) The Smithsonian Institution's Information Retrieval System (SIIRS). Proc. 6th Annual Colloquium on Information Retrieval.

Crompton, A. W. (1968) The Present and Future Course of our Museum. Museum News, vol. 46, no. 5, pp. 35-37.

Galler, S. R., J. A. Oliver, H. R. Roberts, H. Friedmann and D. F. Squires (1968) Museums Today. Science, vol. 161, pp. 548-551.

Hey, M. H. (1966) Catalogue of Meteorites. British Museum (Natural History), London.

Manning, R. B. (1969a) A Computer Generated Catalogue of Types: A By-product of Data Processing in Museums. Curator, in press.

_____ (1969b) Automation in Museum Collections. Proceedings of the Biological Society of Washington. In press.

Piacesi, D. and R. A. Creighton (in press) An Approach to the Geography Problem in Museums. Proc. 6th Annual Colloquium on Information Retrieval.

Squires, D. F. (1966) Data Processing and Museum Collections: A Problem for the Present. Curator, vol. 9, no. 3, pp. 216-227.

_____ (1968) Collections and the Computer. Bioscience, vol. 18, no. 10, pp. 973-974.

_____ (1969) Schizophrenia: The Plight of the Natural History Curator. Museum News, vol. 48, no. 7, pp. 18-21.