TAXON 25(1): 85-94. FEBRUARY 1976

EDP-IR IN THE NATIONAL HERBARIUM OF COLOMBIA (COL)*

*Enrique Forero*** *and Francisco J. Pereira****

Summary

The work on Electronic Data Processing-Information Retrieval (EDP-IR) in the Colombian National Herbarium (COL) is described. The importance and relevance of automation and the practical difficulties encountered are discussed. A short review of the steps involved in up-dating and correcting the data bank is also included.

Introduction

The Colombian Flora is considered to be one of the richest floras of the world. Schultes (1951) has calculated that it has about 50,000 species of flowering plants. The 150,000 specimens[1] deposited at COL comprise an extensive representation of that Flora, while the present size of the herbarium still makes it reasonably easy to consult and to computerize.

While COL's special emphasis is on the Colombian flora, it also contains a large number of specimens from other American countries and from more distant parts of the world. However, it is the emphasis on Colombian plants which has made it an *indispensable* aid for taxonomic, floristic, ecologic, or any other related research in this region. Of such value is the collection that no Monograph of the Flora Neotropica project would be complete unless the monographers consulted the material deposited at COL.

The richness of the flora represented in COL and the need for data which are easily available were the basic reasons for the implantation of the EDP-IR project in the herbarium.

Background

Prance (1972) emphazised the need for EDP at the New York Botanical Garden. The importance of automation in systematic collections has been discussed by several authors (Squires, 1966; Crovello, 1967, 1970; Rensberger & Berry, 1967; Soper & Perring, 1967; Hall, 1972, 1974; Gómez-Pompa & Nevling, 1973a, b; Shetler, 1973; Morse, 1974), while others (Crovello, 1972) have described its actual applications in herbaria and museums.

The EDP-IR project in COL is the first of its kind in South America, and is actually second only to Crovello's work at NDG, although COL must process a larger number of specimens (65,000 at NDG, 120,000 at COL). The Flora Veracruz project (a cooperative programme between the UNAM and Harvard University) has also used electronic data processing to store information about herbarium collections. This is geographically a more restricted project, as it

* Project supported by the Colombian National Research Fund – COLCIENCIAS-, Grants Co. 010-1-30-73 and Co. 010-1-56-75.

** Jefe, Sección de Botánica, Instituto de Ciencias Naturales, Universidad Nacional, Apartado 7495, Bogotá, Colombia.

*** Director, Centro de Cálculo Electrónico, Universidad Nacional, Bogotá, Colombia.

[1] Out of these 150.000 specimens, about 120.000 have actually been registered in the data bank. The remaining 30.000 are either in the process of being filed in the herbarium or on loan to other institutions around the world.

UNIVERSIDAD NACIONAL DE COLOMBIA
HERBARIO NACIONAL COLOMBIANO
CENTRO DE CALCULO ELECTRONICO

PROCESAMIENTO DE DATOS EN EL HERBARIO NACIONAL COLOMBIANO

FORMULARIO DE REGISTRO DE NOMBRES TAXONOMICOS

HOJA 1 DE 1
FECHA Junio 15, 1975
CODIFICADOR E. Forero
SUPERVISOR M. U.

GRUPO [2] SUBGRUPO [22] | ADICION [ ] CREACION [X] ACTUALIZACION [ ]

FAMILIA	GENERO	ESPECIE	N O M B R E
133	ØØØØØØ	CONNARACEAE*	
133	C1ØØØØ	CONNARACEAE*	CONNARUS*
133	C1ØA1Ø	CONNARACEAE*	CONNARUS* AREAUCANUS CUATR.

UNIVERSIDAD NACIONAL DE COLOMBIA
HERBARIO NACIONAL COLOMBIANO
CENTRO DE CALCULO ELECTRONICO

PROCESAMIENTO DE DATOS EN EL HERBARIO NACIONAL COLOMBIANO

FORMULARIO DE REGISTRO DE EJEMPLARES

HOJA 1 DE 1
FECHA Junio 15, 1975
CODIFICADOR E. Forero
SUPERVISOR M.V.

ADICION [ ] CREACION [X] ACTUALIZACION [ ]

GRUPO [2] SUBGRUPO [22] FAMILIA [133]

	GEN	ESP	C	No.	H	PAIS	DPTO	MPIO	HAB	LUG	NOMBRE GEOGRAFICO	ALTURA	AB	FV	U	E	COLECTOR	No. DE COLECCION	FECHA DE COLECCION DIA	MES	AÑO
1	C1	ØA1	Ø4	1	Ø	18	12	20	99	16	RICO			99	2	9	1 J. GOMEZ	16	20	Ø1	955
2																					
3																					
4																					
5																					

TAXON VOLUME 25

includes only the State of Veracruz (Mexico); however, it aims at being very complete for that restricted area. The UNAM has also attempted to computerize herbarium data of their pteridophyte collection (Scheinvar at al, 1973).

The relevance of these projects is being proven, and now INPA (Instituto Nacional de Pesquisas da Amazonia, Brasil), of the Conselho Nacional de Pesquisas (CNPq) is preparing to mount EDP-IR for their herbarium collections at Manaus (INPA) and Belem (MG).

Importance of the project

The major asset of a computerized herbarium is that it makes more easily available the hidden wealth of information which is found on the individual specimens. Grouping and re-grouping of data can be effected rapidly with minimum detrimental handling of specimens, and in minimum time, making the work of the taxonomist easier and more efficient, and thus helping research. This point is especially valid when we consider the rapid growth of herbaria and the lack of trained research personnel to put all the available data to use. However, not all problems of taxonomy will be solved by the computers; therefore, it is necessary to stress the need for training of competent young taxonomists who will continue to study the world's flora.

Each institution in any given country should analize its priority areas for computer work. In Colombia, and specifically in COL, it was found that priority should be given in: (A) Acquainting the Computing Center personnel with botanical work; (B) Informing botanists of the immense possibilities that computers can offer for the furthering of taxonomic work; (C) Producing catalogues of plants which will be basic to future botanical research while avoiding unnecessary duplication of efforts both in the field and in the laboratory; (D) Increasing the production of botanical publications directed to the general public; (E) Acting as an information storage and retrieval center for the data scattered in other national institutions, and (F) Facilitating the curatorial and administrative routines in the herbarium.

Objectives

With these general purposes in mind, we set certain specific objectives. They are, in actual fact, based on a statement of Crovello's (1972) that herbaria have an even greater value for systematists, ecologists, etc. than they may have realized, because of the "ability of the computers to retrieve, rearrange and sort" data in many ways. Our objectives are:

1. To create a data bank which will include the information found on the herbarium specimens (locality, collector's names and numbers, dates, state of the specimens, kind of specimens, scientific names) or related to them (group, subgroup, family, herbarium in which deposited);
2. To prepare catalogues arranged according to various items (taxon, locality);
3. To prepare catalogues of type specimens;

Fig. 1. (above) Form used for Taxonomic Names, showing the way information is recorded. *Fig. 2.* (below) Form used for recording specimen data, showing the way information is recorded: Group 2=Phanerogams; Subgroup 22=Angiosperms Dicotyledons; Family 133; Connaraceae; Genus C 10 = *Connarus*; Species A 10 = *araucanus* Cuatr.; Kind of specimen 4 = Isotype; No. of duplicates = 1; Herbarium 0 = COL; Country 01 = Colombia; Departamento 81 = Arauca; Municipio 220 = Cravo Norte; Habitat 99 = Without information; Geographic locality 16 = Caño; Name of Geographic locality = Rico; Altitude = Without information; Abundance 9 = Without information; Life form 2 = Shrub; Uses 9 = Without information; State of the Specimen 1 = Flower; Collector's name, number and date of collection.

4. To create a data bank of the geographic distribution of taxa; and
5. To facilitate curatorial work in the herbarium (incoming loans, gifts).

System analysis

The work on system analysis began in January, 1974. A first report, with a large number of questions addressed to the botanical co-investigator, was produced by the Computing center in July, and a final report was completed by September. The system was divided into three main steps: Data accumulation, Additions and Up-dating, and Production of Print-outs.

The data recording for 120,000 specimens was done from December 16, 1974 until May 31, 1975. On June 2nd, the 3500 specimen mycological collection deposited at ICA (The Colombian Agricultural Institute, Bogotá) was incorporated into the system.

Since February the data are being punched onto IBM cards and at the same time work is being done on the requested programs. The data are transferred from the registration forms (Figs. 1 & 2) to IBM cards, and from these to two 2401 magnetic tapes (which will serve as a back-up system) and, also, to a 2314 disk where the various programs and dictionaries will also be recorded.

Two preliminary programs, called CHEREJEM and CHERTAXO, are already in operation for computerized control of the quality of the information, before it is recorded on the tapes and on the disk.

Data processing

The data recorded from each specimen are shown on Table 1. Figures 1 and 2 illustrate the forms especially designed for this project.

Since the scientific names plus the name of the family usually take up between 20 and 60 (or even 70) characters, a separate form (Fig. 1) was designed to register this information, which is in turn, punched onto separate cards. Because we were only able to use one IBM card per specimen, it was necessary to replace words with codes in certain items for the data bank. However, these codes will again be substituted by words in the print-outs so that they can be used by persons not familiar with the project. The codes used for the families are the family numbers in the Herbarium. Codes for genera and species were assigned by the student-assistants as they registered the information. Genera are arranged alphabetically within each family and species are also in alphabetical order within each genus.

The programs are being written in COBOL, and will be used in the National University's IBM 360/44 computer. The initial print-outs to be produced are:

1. Taxonomic names (Families, genera and species) and the codes assigned to them (Table 2);
2. Itineraries (Table 3);
3. Type specimens. This print-out will facilitate the validation of types. It will now be easy to remove the types from the general collection and place them in separate herbarium cabinets. The information can also be fed into any international type register in the future (Table 4);
4. Inventory of specimens deposited in COL, divided by groups and subgroups;
5. Standard print-outs will be sent to specialists interested in a given taxonomic group (Table 5);
6. Geographic distribution of taxa (Table 6);
7. Catalogues of plants arranged by geographic areas (Departamentos, municipios or more precise location) or by altitude above sea level (usually accompanied by a circumscription of the area by Departamentos, Municipios, etc.) (Table 7).

Problems

Some of the commonest problems encountered and their resolution follow:

1. Determination of specimens: Constitutes one of the main inconveniences. No taxonomic

decisions were made during the process of data accumulation. These will have to be made by the users.

2. Spelling of scientific names: We tried to standardize this as much as possible. It is not intended that the users accept the print-outs blindly. They must make corrections and report them to the person in charge, in order to introduce the necessary changes in the data bank.
3. Specimens on which information is scarce: These have been recorded as they appear in the herbarium. The use of some of the print-outs mentioned above should help to solve some of these problems.
4. Handwriting on labels: It was attempted to familiarize the assistants with the hand writing of most of the collectors whose specimens fall into this category, with fairly good results.
5. Variation in the information on locality: This presents a serious problem. It was considered necessary to reduce the information to a minimum but to be reasonably flexible and accurate at the same time. Information on geographic coordinates as well as names of small rivers, creeks, "fincas" or very small towns or "veredas" were eliminated, unless they constituted the only available information, in which case they are put down under "Geographic locality". For specimens collected in Colombia, information on Departamento, Municipio and Geographic locality was recorded (Geographic locality includes the names of important rivers, mountains and mountain chains, "páramos" or, as already explained, other data when none of these are available).
6. Habitat: Due to the wide diversity of terms used by plant collectors to describe habitats, it was necessary to limit this information to those most commonly used. A list of 45 different habitats was prepared, and additional terms were used for the algae. It was considered that this list was reasonably complete.
7. Bibliographic references: Were excluded from the start because they do not appear together with herbarium specimens. It will be the work of the users to fill in this information on the print-outs when needed.

Several of the above difficulties were pointed out by Crovello (1972) for the NDG project. However, the resolutions adopted differ somewhat in certain cases.

A final problem consists in the availability of processed information to other researchers. A set of regulations is being worked out at the present time. Some print-outs will be offered free of charge while others will have to be sold. Information on this subject will be given to those interested.

Additions and up-dating

In a new phase of the project it is proposed to develop programs which will enable the computer to add new specimens to the data bank, to make changes in the information already recorded (new identifications on material recorded as "INDET"), and to transfer sets of data from one place to another within the bank (from one species to another, from one genus to another, from family to family, from "indet" to determined). This will allow us to keep the data bank up to date and thereby to increase its usefulness.

A dynamic system has already been established in the Herbarium to allow for these new steps to be taken and to eliminate any interference with other normal herbarium practices.

Acknowledgments

Dr. Alvaro Fernández-Pérez, Past Director of the Institute of Natural Sciences, gave us full support during the initial stages of the project. The understanding of Prof. Polidoro Pinto, Acting Director, and other members of the Institute's staff is much appreciated. We are grateful to Ing. Emilio Latorre, System Analist, and Ing. Gloria Giraldo, Programmer. Student assistants Myriam Uribe (Supervisor), Henry Bernal, Alicia Cuervo, Consuelo Díaz, Josefina Espina, Lupe Galeano, Constanza Galvis, Bethsabé León, Myriam Moreno, Beatriz de Ruíz, Esperanza Torres and Gabby Torres performed

the data accumulation. Ing. Pablo Buriticá and Ing. Edgar Orjuela have cooperated in the addition of the ICA mycological collection to the data bank.

Our very special thanks go to COLCIENCIAS for financial support over a two year period

Dr. G. T. Prance read the manuscript and made helpful suggestions. Mrs. Anne E. Prance corrected a draft of the English version.

References

CROVELLO, T. J. 1967 – Problems in the use of electronic data processing in biological collections. Taxon 16: 481-494.

CROVELLO, T. J. 1972 – Computarization of specimen data from the Edward Lee Greene Herbarium (ND-G) at Notre Dame. Brittonia 24 (2): 131-141.

CROVELLO, T. J. & R. D. MACDONALD 1970 – Index of EDP-IR projects in systematics. Taxon 19: 63-76.

GÓMEZ-POMPA, A. & L. I. NEVLING, Jr. 1973a – La Flora de Veracruz. *En* A. Gómez-Pompa & A. Butanda C. (Ed.): El uso de computadores en la Flora de Veracruz, Pág. 2. U.N.A.M. México.

GÓMEZ-POMPA, A. & L. I. NEVLING, Jr. 1973b. – Ordenación de datos para la descripción de especies para la Flora de Veracruz. *En* A. Gómez-Pompa & A. Butanda C. (Ed.): El uso de computadores en la Flora de Veracruz, pp. 34-41. U.N.A.M. México.

HALL, A. V. 1972 – Computer-based data banking for taxonomic collections. Taxon 21 (1): 13-25.

HALL, A. V. 1974 – Museum specimen record data storage and retrieval. Taxon 23 (1): 23-27.

MORSE, L. E. 1974 – Computer-assisted storage and retrieval of the data of taxonomy and systematics. Taxon 23 (1): 29-41.

PRANCE, G. T. 1972 – The urgent need for computer access by the New York Botanical Garden. New York.

RENSBERGER, J. M. & W. B. M. BERRY 1967 – An automated system for retrieval of museum data. Curator 10 (4): 297-317.

SCHEINVAR, L. A., GÓMEZ-POMPA & L. ALONSO 1973 – Sistema automático de recuperación de información para el Herbario Nacional del Instituto de Biología de la UNAM. *En* A. Gómez-Pompa & A. Butanda C. (Ed.): El uso de computadores en la Flora de Veracruz, pp. 106-125. U.N.A.M. México.

SCHULTES, R. E. 1951 – La riqueza de la Flora Colombiana. Rev. Acad. Col. Cienc. Ex. Fis. y Nat. 8: 230-242.

SHETLER, S. G. et al 1973 – An introduction to the Type specimen register. Smithsonian Institution Contributions to Botany (FNA Report 70) 12: iii-v, 1-186.

SOPER, J. H. & F. H. PERRING 1967 – Data processing in the herbarium and museum. Taxon 16: 13-19.

SQUIRES, D. F. 1966 – Data processing and museum collections: A problem of the present. Curator 9 (3): 216-227.

TABLE 1

SUMMARY OF SPECIMEN DATA RECORDED AT COL, AND OF THE NUMBER OF CHARACTERS ASSIGNED TO EACH ITEM

Group (Cryptogams or Phanerogams): 1 digit
Subgroup: 2 digits
(Algae, Fungi, Mosses, Lichens, Ferns, Gymnosperms, Angiosperms Monocotyledons, Angiosperms Dicotyledons)
Creation, Addition or Up-Dating: 1 digit
Family: 4 digits in Form No. 2 (Fig. 2); Max. 22 in Form No. 1 (Fig. 1).
Genus (GEN): 3 digits in Form No. 2; Max. 22 in Form No. 1.
Species (ESP): 3 digits in Form No. 2; Max. 50 in Form No. 1.
Kind of Specimen: 1 digit.
(Type, Holotype, Cotype, Isotype, Lectotype, Type Var., Paratype, Any other type, Regular specimen).
Number of duplicates deposited in the herbarium (No.): 1 digit
(When more than 9 duplicates are present, the collection is recorded twice).
Herbarium (COL, ICA, MEDEL, INPA, ANT, NY, etc) (H): 1 digit
Country (PAIS): 2 digits
(American countries were assigned individual codes. Other continents are represented by separate codes).
Departamento (DPTO): 2 digits
(For "Departamentos" and "Municipios" we used the codes already established by the DANE-Departamento Administrativo Nacional de Estadística-)
Municipio (MPIO): 3 digits
Habitat (HAB): 2 digits
Geographic Locality (LUG): 2 digits
Name of the Geographic Locality (NOMBRE GEOGRAFICO): 16 digits
Altitude (ALTURA): 3 digits
Abundance (AB): 1 digit
(Common, Rare, Dominant, Very rare, Very common, Abundant)
Life Form (FV): 1 digit
Uses (U): 1 digit
(Medicinal, Poisonous, Construction, Ornamental, Food, Spice, Hallucinogenic, Cattle Feed, Narcotic)
State of Specimen (E): 1 digit
(Flower, Fruit, Flower-Fruit, Steril, Fertil, Photo, Drawing)
Collector's Name (COLECTOR): 15 digits
Collection Number (No. DE COLECCION): 7 digits
Collection date (FECHA DE COLECCION): Day: 2 digits; Month: 2 digits; Year: 3 digits

TABLE 2

FORMAT OF COMPUTER PRINT-OUT OF TAXONOMIC NAMES AND CODES

CONNARACEAE (COD. 133)

Código Género	*Código Especie*	
B 10		BERNARDINIA
B 10	C 10	BERNARDINIA COMANS (CASAR.) SCHELLENB.
B 10	999	BERNARDINIA INDET
C 10		CONNARUS
C 10	A 10	CONNARUS AGUSTIFOLIUS (RADLK.) SCHELLENB.
C 10	A 20	CONNARUS ASIATICUS WILLD.
C 10	P 10	CONNARUS PANAMENSIS GRISEB.
C 10	999	CONNARUS INDET
999		CONNARACEAE INDET

TABLE 3*

FORMAT OF COMPUTER PRINT-OUT OF ITINERARIES. FIRST OPTION. (ARRANGED BY COLLECTION NUMBER)

ITINERARIO H. H. SMITH

LOCALIDAD	*No. DE COLECCION*	*FECHA*
MAGDALENA, SANTA MARTA	1	ABRIL
MAGDALENA, SANTA MARTA	2	ABRIL
MAGDALENA, SANTA MARTA	3	MAYO
MAGDALENA, SANTA MARTA	4	SIN INFORMACION
MAGDALENA, SANTA MARTA	5	MAYO
MAGDALENA, SANTA MARTA	6	SIN INFORMACION
MAGDALENA, SANTA MARTA	7	MAYO

FORMAT OF COMPUTER PRINT-OUT OF ITINERARIES. SECOND OPTION. (ARRANGED BY COLLECTION DATE)

ITINERARIO J. J. TRIANA

LOCALIDAD	*No. DE COLECCION*	*FECHA*
MARIQUITA	1925-4	MARZO 1853
SIN INFORMACION	1925-2	MARZO 1853
SIN INFORMACION	2542-1	MARZO 1853
BOGOTA	1925-5	ABRIL 1853
BOGOTA	2021-1	ABRIL 1853

* This is only an example. The data do not correspond to reality.

TABLE 4

FORMAT OF COMPUTER PRINT-OUT OF TYPE SPECIMENS

GRUPO: FANEROGAMAS SUBGRUPO: ANGIOSPERMAS DICOTILEDONEAS
FAMILIA: CONNARACEAE
GENERO: ROUREA
ESPECIE: ROUREA ANTIOQUENSIS CUATR.
PAIS: COLOMBIA CLASE DE TIPO: ISOTIPO
No. DE EJEMPLARES: 1 ESTADO: FLOR
DEPTO: ANTIOQUIA MUNICIPIO: VALDIVIA LUGAR GEOGRAFICO: RIO CAUCA
COLECTOR: R. D. METCALF No. DE COLECCION: 30045 FECHA DE COLECCION: 17-02-1942

TABLE 5*

FORMAT OF COMPUTER PRINT-OUT TO BE SENT TO SPECIALISTS

FAMILIA: MELIACEAE

COLECTOR	*NUMERO*	*DISTRIBUCION*
A. FERNANDEZ P.	2835	COLOMBIA: GUAINIA
J. M. IDROBO	1435	COLOMBIA: LA MACARENA
J. F. BRETELER	3523	VENEZUELA
E. FORERO	1525	COLOMBIA: CHOCO

TABLE 6*

FORMAT OF COMPUTER PRINT-OUT OF GEOGRAPHIC DISTRIBUTION OF TAXA

FAMILIA: CONNARACEAE GENERO: CONNARUS

ESPECIE: CONNARUS PANAMENSIS GRISEB.

PAIS:	*DEPARTAMENTO*	*MUNICIPIO*	*LUGAR GEOGRAFICO*
COLOMBIA	CHOCO	QUIBDO	RIO ATRATO
COLOMBIA	ANTIOQUIA	TURBO	
PANAMA			ZONA DEL CANAL
PANAMA			DARIEN

* This is only an example. The data do not correspond to reality.

TABLE 7

FORMAT OF COMPUTER PRINT-OUT FOR THE PREPARATION OF CATALOGUES OF PLANTS FOUND IN A GIVEN AREA (Explanation in Text).

PLANTAS DEL DEPARTAMENTO DEL META

GRUPO: FANEROGAMAS SUBGRUPO: ANGIOSPERMAS DICOTILEDONEAS
FAMILIA: CONNARACEAE

GENERO: CNESTIDIUM ESPECIE: CNESTIDIUM RUFESCENS PLANCH.

MUNICIPIO: SAN JUAN DE ARAMA
LUGAR GEOGRAFICO: RIO GUEJAR, SIERRA LA MACARENA

HABITAT: BOSQUE HUMEDO

ALTURA: 500 FORMA DE VIDA: BEJUCO, LIANA. USOS: SIN INFORMACION

COLECTOR:	No. DE COLECCION.	FECHA DE COLECCION:	ESTADO:
J. M. IDROBO	1263	22-01-1951	FRUTO
W. R. PHILIPSON	2121	18-01-1950	FRUTO

GENERO: ROUREA ESPECIE: ROUREA GLABRA HBK.

MUNICIPIO: SIN INFORMACION
LUGAR GEOGRAFICO: SIERRA LA MACARENA, RIO GUEJAR

HABITAT: BOSQUE, BOSQUE DENSO

ALTURA: 550, 800 FORMA DE VIDA: LIANA USOS: SIN INFORMACION

COLECTOR:	No. DE COLECCION:	FECHA DE COLECCION:	ESTADO:
J. M. IDROBO	750	20-12-1950	FRUTO
W. R. PHILIPSON	2208	24-01-1950	FLOR