Bothalia 11, 3: 349–353 (1974)

Progress in the computerization of herbarium procedures

J. W. MORRIS*

ABSTRACT

Herbarium automation projects at Cape Town (A. V. Hall), Notre Dame (T. J. Crovello), Harvard (L. I. Nevling), Ottawa (J. H. Soper), Brisbane (S. L. Everist) and the British Antarctic Survey (D. M. Greene) as well as the proposed system at New York (G. T. Prance) are described in detail. It is found that data are coded for projects involving small numbers of specimens while, for large systems, data are entered uncoded where possible. It is noted that not one automation project has failed and that all users were enthusiastic about the future of such operations.

The need for a large system in South Africa is outlined and the uses to which it could be put are listed. The system planned for use is briefly described.

INTRODUCTION

Herbaria are data banks in that they store information about plant specimens in a retrievable form, in much the same way that reference libraries store information on their shelves. The specimens and associated labels in herbaria constitute an inventory of botanical information and a history of the plants from the times of the first collectors until the present. With the traditional method of filing specimens it is, for practical reasons, possible to retrieve only limited kinds of information. Theoretically it is possible to compile from specimens a generally valuable map of the distribution of each plant species, to prepare floristic lists for any given region, to plot changes in distribution and abundance which have taken place and to collate a vast amount of information about the habitats, flowering time and other properties of the plants. The major difficulty, however, has been the time needed to retrieve information from each label and to extract the elements needed for any particular purpose. With the advent of electronic data processing (EDP) and, more important, the growing availability of computers to botanists, much more in the way of retrieval is now possible. Since Crovello & MacDonald's (1970) pioneer survey of biological data-processing much progress has been made. The present survey is not nearly as complete as that by Crovello & MacDonald, as far as coverage is concerned, but, where appropriate, more detail of each application is given· This survey is concerned only with specimen data, and taxon data banks, such as that envisaged by the Flora North America Program (Shetler, 1971), are not considered.

Following a survey of herbarium automation, motivation for and the main objectives of a proposed data bank for South African Herbaria are set out, and the means by which these objectives can be obtained are considered.

REVIEW OF SOME OPERATIONAL AND PROPOSED SYSTEMS

Hall (1972 a & b) in Cape Town was the first known curator in South Africa to apply EDP to a herbarium collection. His main design criteria were flexibility of output and simplicity of input; the programme to be compatible with a wide range of computers and to be operational on fairly small ones. He found that as a number of elements on specimen labels were commonly repeated for many specimens, most elements could be coded before capture with one- to four-digit numbers, or alpha-numeric codes. Captured data, in the sense used in this paper, are those words and phrases accessioned by the computer's storage device in such a form that they may be retrieved in a variety of ways by programming the computer. The words and phrases may be entered in full, or represented by code symbols or numbers as done by Hall. He considered it easier to edit a list of all possible elements once, give elements codes and then use the codes when accessioning. The work of repeatedly punching and editing precise citations written out in full suggested little reward for much effort. Data were entered by standard computer card, taxon by taxon, each taxon being headed by one card with the taxonomic information for the species, followed by one specimen card for each herbarium sheet. In addition to the data fields listed in Table 1, Hall coded seven hierarchical supra-generic levels. Eleven programme modules, written in FORTRAN IV (except for the plotter routines), control transactions within his bank. The programme was initially written for an IBM/1130 16K computer with 500K disc store and a plotter, but has been expanded recently (Hall, 1972b) for operation on a Univac 1106 131K computer under remote terminal control. Retrieval by virtually any combination of coded elements is possible.

Hall's system was designed primarily as an aid, at the specimen level, to research in systematics. Likely benefit from such a bank include rationalisation of specimen collection by pin-pointing poorly-collected areas and taxonomic groups. Important data can readily be assembled on the travels of the major early collectors and the places they visited, and of herbarium type-specimen holdings. The task of writing floras of specially circumscribed areas will be assisted (A project with this aim has been initiated; A. V. Hall, pers. comm. 1973). Hall considers that, for the immediate future, the chief users of a data banking system such as he described, are likely to be those making regular use of certain special collections. Other users are those concerned with the balanced expansion of collections of critical groups, perhaps as part of a monographic study. The costs involved prohibit data-banking on a major scale at present (Hall, 1972a). The software and hardware used by Hall limit the size of his data bank to its present order of magnitude.

The label data from 65 000 specimens have been captured by Crovello (1972) in the first, large, operational herbarium computerisation project in the Greene Herbarium (NDG). A Friden 7102 flexowriter was used over a period of two years to code the data fields listed in Table 1 and provide machine-readable paper-tape input for each specimen. The computer in use is an IBM 370/158 and all programmes are written in PL/1, although IBM's SORT/MERGE routine is used as well.

In a detailed motivation for computerization, Crovello mentioned the richness of NDG in type material and the taxonomic and environmental importance of the pre-urbanization era during which most specimens were collected. As Greene, the principal collector, left no notebooks, information as to where and when he collected was not available but can be retrieved from a data bank based on his

* Botanical Research Institute, Department of Agricultural Technical Services, Private Bag X101, Pretoria, South Africa.

Table 1. *Summary of label-data captured for various projects* (*P=proposed, U=captured unabbreviated, C=captured in code form, T=typed but not captured*).

Author (date)	South African Herbaria	Hall (1972)	Crovello (1972)	Gómez-Pompa & Nevling (1970 & 1973)	Soper (1969)	G.T. Prance	Everist (1972)	Greene (1972)
Herbarium Code		BOL	NDG	MEXU	TRT/ CAN	NYBG	BRI	
A. *Administrative data fields*								
Accession number	P		U	U	U	P	U	U
Herbarium name	P	C			C		C	C
B. *Taxonomic data fields*								
Family	P	C		U		P	C	
Binomial/trinomial and authority	P	C	U	C	U	P	U	C
Identifier (sometimes with date)	P		U	U	U	P	U	C
Annotator, annotation date and notes	P		U					
Type status	P	C	C			P		
C. *Collection data fields*								
Collector, number and date	P	C	U	U	U	P	U	C
Country, State or Province	P		C	U	U	P	U	C
Locality description	P	U	U	U	U	P	T	
Quarter-degree reference system or latitude & longitude	P	U		U	U	P	U	U
Altitude	P	U	U	U		P	U	
State of the specimen (SOS)	P		C	U	U	P	C	C
D. *General*								
Abundance	P	C	U	U				
Habit (and height)	P		U	U		P	T	
Ecology		C	U					
Habitat (including soil and vegetation type)	P		U	U	U	P	T	
Uses			U	U		P		
Other notes	P		U	U	U	P	T	

specimens. As it is a closed collection, data accumulation for the project would not continue indefinitely but could be a pilot project for the use of EDP in biology. Naturally, corrections and new annotations on specimens being returned from loan will be input as received. Greene's collection was too large for manual retrieval, but not so large as to make the project unwieldy and too expensive. A complete list of types in the Greene herbarium would aid taxonomists at other herbaria who were seeking types. Crovello concluded his motivation by stating that the special attributes of the Greene herbarium and realisation of some of the ways that the application of EDP to this collection would assist interested biologists, led him to the conclusion that the project was so rich in actual and potential results as to justify the effort.

Gómez-Pompa & Nevling (1970 & 1973) are collaborating in the preparation of a machine-aided Flora of Veracruz (Mexico). Their first attempt at data-capture involved recording all information from herbarium labels so that it could be computer-retrieved. Elaborate "dictionaries" were drawn up for coding the information (for example, 130 flower-colour classes in two languages). The attempt was abandoned after a trial run as it took too long to code a herbarium specimen (L. I. Nevling, pers. comm. 1972). The present system uses a standard collector's label as base (Table 1). In addition, associated species, ambient weather conditions, flower colour, local names and whether a plant is annual or perennial, are recorded, where available, in free format. Data are punched onto computer cards and herbarium labels are considered a useful by-product of the system. The computer being used for this project is a Burroughs B6500 and the programmes are written in ALGOL. Many of the computer programmes have, in the absence of more highly-trained programmers, been written by students. The data have been used to produce a provisional check-list of the flora. It is planned to retrieve other data from the data bank and to correlate it with environmental data. Distributions are mapped by means of a Calcomp plotter. Very useful suggestions for curators intending to use EDP for the preparation of a flora are given by Gómez-Pompa & Nevling (1973).

Data processing for an automated flora of Ontario was started in 1963 by Soper. His intention was to study, firstly the distribution and affinities of the flora of Ontario, and then to extend the project to the whole of Canada (Crovello & MacDonald, 1970; Soper, 1969; Soper & Perring, 1967). Although this was probably the first herbarium EDP project to become operational, it is still only a pilot study (T. J. Crovello, pers. comm. 1973).

One of the main reasons Soper cited for indexing a herbarium was that specimens suffered appreciably from use and were largely irreplaceable. Although the actual specimens were required, on many occasions they could be spared all or most of their handling by first consulting an index. Label data captured by Soper are given in Table 1. Initially, a Friden flexowriter was used for coding data, but presently he is using an IBM 2741 typewriter on-line to the computer for data capture (Morris, 1973). An even more sophisticated procedure is being developed (McAllister *et al.*, n.d.). Apart from the production of labels, output will consist of distribution maps drawn by a plotter, catalogue cards and tabulated listings of label data. Label data considered essential, desirable or optional for capture are given by Beschel & Soper (1970).

G. T. Prance (pers. comm. 1972) has motivated strongly the provision of computer facilities at the New York Botanical Garden. The system envisaged will eventually automate the entire Garden, from salaries, and journal mailing lists to garden-plant accessions. As far as herbarium labels are concerned, Prance will start with clearly-defined sub-sets of data rather than trying to work systematically through the entire herbarium. This approach will provide manageable-sized projects and obtain useful results fairly quickly. It will be possible to obtain specimen labels as a by-product of data capture done to reshuffle data to assist with phytogeographical and ecological research. Data to be captured are indicated in Table 1.

Everist (1972) is using a Ricoh 5000 paper-tape punching typewriter with two paper-tape readers and two punches primarily for preparing herbarium labels and obtaining machine-readable output as a by-product. Label data, as set out in Table 1, are typed. Duplicate labels are produced automatically by feeding the paper-tape through the typewriter as many times as required. Data common to more than one consecutive specimen are pre-punched and automatically inserted.

Labels of all the specimens of one family in the herbarium have been encoded, as a trial, in addition to the data captured for new specimens. They have made slow progress on the specimen backlog and conclude that a magnetic tape encoder or computer card punch would be more satisfactory than the typewriter. Their main difficulty is finding geographical co-ordinates for old specimens. Everist concludes that the results are already justifying his expenditure. Within a few years he expects meaningful answers to many questions on distribution, floristic and habitat data and flowering times for much of the flora of Queensland.

The ultimate necessity for the preparation of a catalogue during taxonomic and phytogeographical studies was pointed out by Greene (1972) of the British Antarctic Survey. In common with many other institutes, he faced staff shortage, limited support facilities and rapidly growing collections. He decided to set up a data bank and retrieval system to expedite the administration of his material, consisting of about 30 000 specimens. Required from the system were: (a) a catalogue of all data for every specimen with facilities for incorporating new data and updating records; (b) immediate availability of all data in a variety of sequences; and (c) a rapid method of duplicating information, such as preparing herbarium labels. As it was considered too expensive to have all these tasks carried out by a computer, a paper-tape typewriter was used for (a) and (c) and an ICL 1900 computer was used for storing the data (Table 1) and performing retrieval tasks. A normal-language card-file and a coded computer file of specimens have been made.

Greene's data bank is proving to be really valuable in speeding up the production of regional floras, and the chore of checking duplicate herbarium labels and index cards is eliminated.

The only data fields captured by all systems (Table 1), to sum up extant and proposed systems, are the binomial/trinomial and collector's name fields. Other commonly-recorded fields include accession number, family, identifier, state of the specimen (SOS), country, state or province, locality description, degree reference or co-ordinates, altitude, habitat and other notes. The detail and form of recording vary from one system to another. For example, some workers record genus and species separately and others together, some record only the state of the flower in the SOS field, while others note flowers, fruit, seeds, leaves and roots in this field. An important finding from the literature survey is that the two projects with fairly small inputs (Hall and Greene) pre-code information while organisations with larger data sets prefer to enter data uncoded as far as possible.

Six herbarium data-capture projects are known to be in operation already and at least one more (New York Botanical Garden) is due to start soon. Although problems have been encountered with all the projects, it is encouraging to note that not one has failed and, furthermore, that all officers with systems, when approached, were enthusiastic about the future of herbarium automation.

THE NEED FOR A SYSTEM FOR SOUTH AFRICA

Soper & Perring (1967) state that a satisfactory EDP system should be a by-product, or bonus, of the initial need to label accurately all specimens kept in the herbarium. This statement holds for South Africa but we also have other motivation. Crovello (1967) considers that EDP will never replace the taxonomist but, rather, it will do away with some of the routine tasks that an intelligent scientist should be relieved of. The writer sees this, too, as the aim of our system, but would take Crovello's statement further and contend that EDP brings within the range of possibility answers to a number of taxonomic, ecological and environmental questions in need of answering.

Labels for the 20 000 specimens accessioned annually by the Botanical Research Institute are typed and the specimens filed by families at present. By typing labels in such a way that machine-readable copy is produced as a by-product, no extra cost per label is involved. Although the equipment is more expensive than an ordinary typewriter (as little as R50 per month hire), shortcuts in the form of abbreviations and constant fields, entered automatically by the computer, increase the speed of accession to balance out the increased cost to a greater or lesser extent

Two or twenty top-quality copies of a label are equally easy to produce and, furthermore, all are identical.

In addition to the production of neat, multiple-copies of specimen labels, there are other uses for data captured in machine-readable form in the process of making labels. An important use is for the production of check-lists of large or small areas. Examples are: (a) Highveld Agricultural Region—for the present ecological survey; (b) Lesotho or Natal—for a National or Provincial list; (c) a quarter-degree grid square—a list is needed for the square in which Pretoria is situated for the Pretoria Flora at present in preparation; (d) poorly-collected areas like Tongaland or the Richtersveld, to give a plant collector an idea of what is known of the flora before setting out on a general collecting trip; and (e) endemic distribution data for assessing conservation priorities of different specific areas.

By using the computer to count the number of specimens and number of species collected in each quarter-degree square, poorly-collected-from areas should be generally evident and attention can be devoted to collecting from these areas.

On a recent field trip the author needed to know the distribution of *Terminalia sericea* in South Africa to test a hypothesis. A distribution map immediately produced by the computer would have spared him a search through the specimens in the herbarium. Although it can be argued that the effort of drawing one distribution map by hand is not very great, the advantage of being able to draw distributions automatically as required, for any of the 10 000 most common plants in South Africa, should be obvious.

Once the specimens of the early collectors, such as Burchell, have been accessioned it will be easy to print out an itinerary of the places they visited. A similar service to present-day collectors would produce collectors' registers, compiled by computer.

With a quick method of data capture a great deal of information which is presently for all practical purposes "lost" becomes usable. For example, "spot" identifications, voucher specimens and sight records could be added to the data bank and used in it, admittedly with a lower, but defined, degree of confidence. At present, this valuable information is discarded.

In addition to their value within the Botanical Research Institute, the herbarium data will be available for inclusion in the proposed agricultural data bank of the Department of Agricultural Technical Services.

SYSTEM PLANNED FOR THE BOTANICAL RESEARCH INSTITUTE, PRETORIA

For data capture we propose to use cathode-ray tube (CRT) screens with keyboards linked by telephone line to the Burroughs B6700 computer run by the Department of Agricultural Technical Services. A computer programme will flash the headings, like GENUS:, SPECIES:, COLLECTOR:, DATE:, on the screen and the operator will key in data directly from the label. A number of short-cuts designed to streamline accessioning will be used. For example, extensive use will be made of abbreviations which the computer will expand, and conversions of British standard units of measurement to metric values will be carried out by computer. All data fields that are repeated on more than one consecutive label will be inserted automatically after the first entry. Preliminary editing of the data will be done by computer and will be followed by manual checking.

We are presently negotiating for the use of SELGEM to satisfy our information retrieval needs. SELGEM was developed by the Smithsonian Institution (Creighton & Crockett, 1971; Creighton *et al.*, 1972) for collections similar to our own. Unfortunately, we are not even able to consider the GIS package advocated for use by Krauss (1973) as it is incompatible with the computer available to us. Other current data-banking systems were considered. In particular, the comprehensive Cambridge system developed by J. L. Cutbill (Cutbill & Williams, 1971), and TAXIR developed by D. J. Rogers (Rogers, Flemming & Estabrook, 1967), were investigated.

Various CRT screen lay-outs are being experimented with at present and we are considering what information is worth capturing from the specimen backlog and how new accessions will be treated. Standards for each data field are being drawn up.

CONCLUSION

The merits and procedures of herbarium automation that have been discussed by six herbarium curators who have applied EDP to their collections to improve the service given by their herbaria, have been considered for a herbarium data bank for the Botanical Research Institute in Pretoria. It is concluded that for our purposes such a data bank includes the uncoded data indicated in Table 1. Such a system can be realized practically and will provide immediate access to information needed for a variety of purposes and which was hitherto for practical purposes unavailable.

ACKNOWLEDGEMENTS

The author is grateful to the Secretary of the Department of Agricultural Technical Services for making possible the study tour which laid the foundation for this survey. Grateful acknowledgement is made for the time and resources made available to the author by all the overseas and local data bankers who were visited during this tour. For kindly commenting on the manuscript I thank B. de Winter, S. L. Everist, T. J. Crovello and A. V. Hall.

OPSOMMING

Huidige projekte vir herbarium automatisasie in Kaapstad (A. V. Hall), Notre Dame (T. J. Crovello), Harvard (L. I. Nevling), Ottawa (J. H. Soper), Brisbane (S. L. Everist), die "British Antarctic Survey" (D. M. Greene) en New York se voorgestelde sisteem (G. T. Prance), word volledig bespreek. Daar word gevind dat die inligting vir take met 'n klein aantal plant monsters gekodeer word, terwyl die inligting vir groot sisteme, so ver moontlik sonder kodering verwerk word. Daar word genoem dat geen een van hierdie automatisasie projekte misluk het nie en dat die gebruikers almal baie entoesiasties oor die toekoms van hierdie metodes is.

REFERENCES

Beschel, R. E. & Soper, J. H. 1970. The automation and standardization of certain herbarium procedures. *Canad. J. Bot.* 48: 547–554.

Creighton, R. A. & Crockett, J. J. 1971. SELGEM: A system for collection management. *Smithsonian Institution Information Systems Innovations* 2 (3): 1–24.

Creighton, R. A., Packard, P. & Linn, H. 1972. SELGEM Retrieval: A general description. *Smithsonian Institution Procedures in Computer Science* 1 (1): 1–38.

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

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

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

Cutbill, J L. & Williams, D. B. 1971. A program package for experimental data banking. *in*: J. L. Cutbill (editor) *Data processing in biology and geology*. Systematics Association Special Volume 3: 105–113.

EVERIST, S. L. 1972. *Computer processing of labels in the Queensland Herbarium*. Unpublished paper read at annual Conference of the Museums Association of Australia, Sydney, October 1972.

GÓMEZ-POMPA, A. & NEVLING, L. I. 1970. La flora de Veracruz. *An. Inst. Biol. Univ. Nal. Autón. México* 41. *Ser. Botanica* (1): 1–2.

GÓMEZ-POMPA, A. & NEVLING, L. I. 1973. The use of electronic data processing methods in the flora of Veracruz program. *Contr. Gray Herb.* 203: 49–64.

GREENE, D. M. 1972. A taxonomic data bank and retrieval system for a small herbarium. *Taxon* 21 : 621–629.

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

HALL, A. V. 1972b. The use of a data-banking system for taxonomic collections. *Contr. Bolus Herb.* 5: 1–78.

KRAUSS, H. M. 1973. The use of generalized information-processing systems in the biological sciences. *Taxon* 22: 3–18.

MCALLISTER, D. E., LEERE, A. B. & SHARMA, S. P. n.d. A batch process computer information retrieval and cataloguing system in the fish collections of the National Museums of Natural Science. *Syllogeus* 1 : 1–20.

MORRIS, J. W. 1973. *Overseas advances in quantitative ecology and computerised data banking in biology*. Unpublished Departmental Report, Botanical Research Institute, Pretoria.

ROGERS, D. J., FLEMMING, H. S. & ESTABROOK, G. F. 1967. Use of computers in studies of taxonomy and evolution. *in*: Th. Dobzhansky, M. K. Hecht & W. C. Steere (eds) *Evolutionary Biology*, New York.

SHETLER, S. G. 1971. Flora North America as an information system. *Bio Science* 21 : 524–532.

SOPER, J. H. 1969. The use of data-processing methods in the herbarium. *An. Inst. Biol. Univ. Nal. Autón. México*. 40, *Ser. Botanica* (1): 105–116.

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