Procedures and Problems in the Incorporation of Data from Floras into a Computerized Data Bank[1]

CLIFTON KELLER and THEODORE J. CROVELLO
Department of Biology
University of Notre Dame, Notre Dame, Indiana 46556

Abstract

In a very real way, the published floras of various regions of the world are a summary of much of the work of past plant taxonomists. The purpose of our current research was to demonstrate that the value of this storehouse of data can be enhanced and serve additional uses if it is put into a form that can be searched and rearranged easily. The procedures that we developed to incorporate these data into a computerized data bank were described. This was not a simple process and we encountered problems such as missing data and the use of different terms in different floras to describe the same character state.

Regional floras and manuals summarize the work of previous taxonomists. They have been used to distinguish taxa and to identify unknown plant specimens. These publications may be mere checklists or they may be thorough and exhaustive compendiums of biological information that also include ecological, biogeographical and other types of data. Nevertheless, they are static and have missing data.

In taxonomy, computers have been used mainly for numerical taxonomic studies (NT), for information retrieval (IR), for specimen identification and for automatic key construction. Crovello (4) provided a recent review.

We feel that computerization of floras will integrate these uses, producing both a computerized data bank and a basic data matrix for detailed analysis of taxa. Transformation of data in floras to a form acceptable by computer should have the following advantages: 1) data on each taxon will be easier to retrieve and to update; 2) collation of data from several floras or with data from specialists, or herbarium specimens will be easier and more reliable; 3) elimination of synonymized phraseology and other editing processes should be easier; and most important 4) data in floras will become more valuable since they can be used for ecological, phenological and numerical taxonomic studies.

The purposes of this paper are: to describe the procedures that we have developed to capture information on genera of the family Brassicaceae from three floras, to indicate the problems that we have encountered, and to suggest some possible solutions.

Materials and Methods

The data were obtained from the descriptions of genera found in three recent floras. Those of Fernald (6) and of Gleason (7) cover the

[1] This project was supported partially by National Science Foundation, Office of Science Information Services, Grant GN-878, to Theodore J. Crovello.

northeastern United States, while the third reports on the European continent (9). Information was taken from both the prose descriptions of each taxon and from the keys to genera.

In overview, a coding chart for characters, and character states was devised. Data from prose descriptions and keys found in floras were encoded onto punched cards. Additional codes were added to the coding chart as they were needed. By use of the computer, the data then were checked for errors, systematized, reorganized, and transferred from punched cards to printed lists and magnetic tape. Printed lists were invaluable proofreading aids and the magnetic tape was necessary for efficient sorting and collation with similar data sets, as well as for preparation of basic data matrices useful in taximetric comparisons, evaluation of characters, key construction, and dissemination of information among workers. Relevance tables (3) and scatter diagrams of the basic data matrices were constructed to provide additional insights regarding characters and taxa. The following paragraphs outline this procedure in more detail. The methods are presented in a stepwise fashion.

Data Accumulation

1) **Develop a list of characters**—This is done by a preliminary, partial study of which characters have been used in the floras. For example, stem habit.

2) **Develop a list of character states**—Again this is done by a perusal of parts of several floras. For example, for the character, stem habit, three of its character states are: a) erect; b) suberect; 3) procumbent.

3) **Develop an abbreviated code for each character**—For simplicity we organized characters first by organ or type and then by a serial number. Thus A means it is a stem character and A1 is the character, Stem Habit.

4) **Develop an abbreviated code for each type of character state**—We found that many characters had the same kinds of character state. For example, the characters number of basal leaves and number of cauline leaves both use the actual number of leaves present as the state of the character in each taxon. This fact permits us to reduce the number of different character state types and codes considerably. Table 1 gives an example of part of our coding charts.

5) **Transform each piece of data in a flora, both in keys and in prose descriptions, to the code that describes the state of each character in each genus**—For example, procumbent stems in a genus would be transformed as A1 3. That is, referring to Table 1, it has information on a stem character (Type A), in particular, stem habit (Character A1). This information is that it is procumbent (Character State Type 1 and Character State 3). Thus A1 3 is equivalent to the prose statement, "stem habit is procumbent." Table 2 gives examples of how coded data from prose descriptions and from keys appear.

6) **Punch the coded data onto 80-column computer cards.**

TABLE 1. *Part of the coding charts used to extract data from floras.*

Organ Codes	Organs	
A	Stems	
B	Rhizomes and Roots	
C	Basal Leaves	
Character Codes	**Characters and Character State Type**	
A1	Habit /1/	
A2	Longevity /2/	
A3	Stem Simple /3/	
Character State Type	**Character State Codes**	**Character States**
1	1	Erect, Scapiform, Scapose
1	2	Suberect
1	3	Procumbent

TABLE 2. *Part of the raw data from one flora, in coded form.*

Raw Data From Description (the $ indicates the beginning of a new genus).

```
    $1 DRABA /L15 2/L17 I 4 6 /LIO 1 /L17 3/M2 1/
    $2 BERTEROA /L15 3 /L7 5 /L18 10 /H7 2 /H9 2 /RI 2/A1 1/
    $3 LOBULARIA /L15 3/L7 11 7 /M17 1 2 /H7 2 /H10 1 /M13 1/
```

Raw Data From Keys (the first and last number on each line are couplet numbers)

```
    I/D3 2 /D10 43 25 /                             2
    2/L4 128/L15 3                                  3
    3/L19 3 9 1 /L168 3 /L17 1 /M9 1                4
    . . .                                           .
    . . .                                           .
    . . .                                           .
    8/A27 3/A60 3 /L7 36 27 4/M19 1 TAXON (1 Draba)
    8/A29 3/A62 1 /L7 6 /M19 92                     9
```

Data Processing

Because of the number of calculations and rearrangements of data required, the following Data Processing steps would not be practical without a digital computer. All of the necessary programs were written by the authors in the PLI computer language. All data processing was done on The University of Notre Dame's IBM 370/155 computing system. In the following paragraphs the terms given in parentheses in capital letters refer to individual computer programs. This permits easy reference to them and also should indicate to the reader that this is not a simple, one-step process.

7) **Transform the coded data from keys into the same format as used in prose descriptions**—This program also automatically checks for certain coding and key punching errors (KEY-CONVERT).

8) **Rearrange data from both prose descriptions and keys into card image format**—This program also checks for certain coding and key

punching errors. Arrange data by taxon using information from one or more floras. Print these results and also write them on magnetic tape for later use (CARD-IMAGE).

9) **Integrate data from different floras into one data set but this time sorted by character and character state codes**—(COLLATE).

10) **Search the above files to produce one integrated description**—This may be of a particular genus based on several floras. Conversely, produce an efficient, non-dichotomous key by asking for data from all genera but only for those characters that an unidentified specimen has (RETRIEVE).

11) **Create a character by taxon Basic Data Matrix from the above files**—This table is the first step in further evaluation and comparison of data using the methods of numerical taxonomy (CREATE BDM).

12) **Edit the Basic Data Matrix**—This is done with the help of two programs (RELVNT, UNIQUE). They calculate the three types of relevance suggested by Crovello (3). The taxonomist then could delete those characters or genera for which there is little information. This would give more reliable numerical taxonomic results, but the results would be for fewer genera. Alternately, the taxonomist could retain all characters and attempt to increase character relevance by collecting data from preserved specimens, field studies or other data sources.

Results

The procedures that we developed allow us to capture floristic data both from descriptive phrases and keys. This is the first step in the formation of a computer data bank for genera and species of the Brassicaceae. Our procedure not only provided an unexpectedly large number of characters displayed in an integrated and understandable form, but now allows us to coordinate our knowledge, using methods of numerical taxonomy.

Another use of the computerization of floristic data is the characterization of the contents of each publication. Table 3 contains some of the summaries that are possible. From Table 3 we note that Flora Europaea has more than twice as many genera as the American

TABLE 3. *Summary of data from three floras for genera of the Brassicaceae.*

	Gray (6)	Gleason (7)	Flora Europaea (9)	Summary of Three Floras
Number of Genera	43	48	104	127
Total Pieces of Data	820	1315	1846	3981
Average Number of Data Per Genus	10.9	20.7	17.7	—
Total number of Different Characters, Each Used at Least Once	127	162	144	263

floras and has the most total data. But it is intermediate between the two American floras in the average amount of data per genus and in the total number of different characters used. Such information can help taxonomists to decide which treatments to consult the first time as well as to estimate the degree of overlapping information in different floras.

Finally, the actual extraction and coding of data from floras was done by an undergraduate student working in the herbarium. While quite intelligent, his primary interest was not botany. This suggests that much of the labor need not be performed by the already overworked professional plant taxonomist.

Problems

Mechanical problems such as key-punching errors are readily found and corrected either by computer or by careful proofreading. But others are more difficult to detect. Characters or character states, particularly those from keys, may represent only a small sample of the taxon under investigation. Another problem is synonymized phrases. This occurs when different words are used to describe the same character state, *e.g.*, leaves hairy *versus* leaves pubescent. Observation of the outputs from COLLATE and UNIQUE should serve as guides to detect such errors. Since COLLATE sorts the data by the characters and character state codes, data regarding a given character and character states are in close proximity. This simplifies recognition of codes which differ. UNIQUE tells how many times a character and character state is used by each of the sources. Thus when several sources recognize a particular character as important and another author fails to do so, chances are good that he has not overlooked this character but has used a synonymized phrase. RELVNT may be helpful in location of such phrases. Particular attention to those characters with low relevance may be rewarding since low character relevance frequently is a result of one of the above types of error, or it is a character that should be noted because of its particular diagnostic importance. Missing data and relative reliability of data within and among floras are two other recognized problems which may be solved in part by our methods, since it is able to encompass data from many sources. Such reinforcement of the state of a character in a genus by combining estimates of its value from several floras should increase accuracy. Additional problems recognized by us but not yet dealt with adequately include: a) modified character states, *e.g.*, *usually* long, or *moderately* long; b) measurements given as ranges rather than as more meaningful statistics; c) vague statements, *e.g.* south to Georgia and Mississippi; d) "not" phrases in keys, *e.g.*, without combination of characters described above; and e) different author's concepts of what each taxon encompasses. This is not too serious for genera, but it may prove frustrating at the species level.

Discussion

A major reason for undertaking this project was to demonstrate that the huge amounts of information currently present in our published floras has an even greater value than currently thought. We believe that, as for herbaria (5), intelligent use of the computer can help us to realize this increased value of floristic work. A second reason for pursuing this project is our belief that such easily obtainable data, when coupled with simple numerical taxonomy, can be an efficient way for new workers to gain insights into relationships among all of the taxa of a group. Although inaccuracies doubtless exist in the data we captured, most of the information should be reliable. As workers new to the study of the Brassicaceae, we feel that the preliminary taximetric analysis that we shall perform will give us sufficient insights into simultaneous generic relationships to justify the effort expended in the extraction of these data from the floras.

Baker (1) stressed the need for a flexible key to be used for larger and incompletely known genera. He adopted an edge-punched card system that was both rapid and effective. But computers provide even greater flexibility and efficiency since: a) they are not limited by the 168 numbered holes of his punched card; b) they are not limited to sorting on key characters, since their format yields readily to statistical approaches; and c) their files are easier to update and to reproduce so that information can be shared with other workers.

The Morse (8) program package for computer-assisted identification requires a Basic Data Matrix and the storage of keys which appears to be more of an end product of a specialist's research rather than the flexible tool needed as the Basic Data Matrix is being created. We feel that capture of data from floras is an important first step which should increase efficiency in construction of final keys by Morse's program.

Character and character state selection is of greatest importance especially when information is to be shared by several workers. Selection of single word descriptions (*e.g.*, leaves *mostly* entire) allows us to reduce the probability of errors, but arrangement and comparison of all the various ways in which a statement may be given make standardization extremely difficult.

Except for an unpublished pilot study in *Carex* by the *Flora North America Program* and for data from several species of northeastern United States genera used by Morse (8) in his computer-generated keys, we know of no other computer assisted studies of this kind by others. Crovello (2) earlier extracted data on *Salix* species from a California flora. Because he also had extensive data on these species collected by himself, he could estimate the reliability of that floristic data. It proved less reliable, but still useful.

Our next step is to build a reliable data base of taxa of the Brassicaceae to use as a standard against which we can compare the floristic data discussed in this paper. If data from several floras does prove reliable, we hope that our methods may be of use in the *Flora North America Program,* where perhaps 50% of the genera will not be treated by specialists in those particular taxa. Our procedures may prove to be an efficient way to generate a "first draft" of the character by species Basic Data Matrix necessary in the production of a sound *Flora North America.*

Literature Cited

1. BAKER, H. A. 1970. A key for the genus *Erica* L. using edge-punched cards. J. So. African Bot. **36**:151-156.

2. CROVELLO, T. J. 1968. The effect of missing data and of two sources of character values on a phenetic study of the willows of California. Madrono **19**:301-315.

3. ———. 1968. The different concepts of relevance in a numerical taxonomic study. Nature **218**:492.

4. ———. 1970. Analysis of character variation in ecology and systematics. Annu. Rev. Ecol. and Systm. **1**:55-98.

5. ———. 1972. Computerization of specimen data from the Edward Lee Greene Herbarium (ND-G) at Notre Dame. Brittonia **24**:131-141.

6. FERNALD, M. L. 1950. Gray's Manual of Botany. American Book Co., New York, N. Y. 1632 p.

7. GLEASON, H. A. 1963. The New Britton and Brown Illustrated Flora. Hafner Publishing Co., New York, N. Y. 655 p.

8. MORSE, L. E. 1971. Specimen identification and key construction with time-sharing computers. Taxon **20**:269-282.

9. TUTIN, T. D. (ed.). 1964. Flora Europaea. Volume 1. Cambridge University Press, Cambridge, England. 464 p.