Dallwitz, M.J. 1980. A general system for coding taxonomic descriptions. *Taxon* 29: 41–46. Also available at http://delta-intkey.com

A general system for coding taxonomic descriptions

1980

M. J. Dallwitz

Summary

A generalized system for the concise representation and manipulation of taxonomic descriptions is described. The system is versatile, easy to understand, and designed to minimise coding errors. The descriptions can replace natural-language descriptions both at the time of recording and in publications. The descriptions are also computer-readable, and a program is available for translation into natural language and into the formats required by some key-generation and numerical-classification programs.

Introduction

When taxonomic descriptions are prepared for input to computer programs, the form of the coding is usually dictated by the requirements of a particular program or set of programs. This restricts the type of data that can be represented, and the number of other programs that can use the data. This paper describes a new coding system, developed from that of Watson and Milne (1972). It has been named DELTA, for DEscription Language for TAxonomy. A program has been written to generate natural-language descriptions from the coded descriptions, and to extract and reformat data for input to other programs.

The system was designed primarily for easy use by people rather than for convenience in computer programing. Consequently, it can be used as a shorthand method of recording data, even if computer processing of the data is not envisaged. The data are written in free format—that is, there is no need to place data in particular columns. The characters may be assigned numbers in any order which suits the user (there is no need to group them by character types, as required by some programs). However, this order need not be adhered to when recording the attributes of a particular taxon. Thus, attributes which are unknown or considered unimportant can be omitted, and later added to the end of the list, if required. An incorrect attribute can be deleted, and the correct one inserted in the same place or at the end.

The system is capable of encoding all of the types of character commonly used for identification and classification: unordered and ordered multistate (including two-state), counts, and measurements. Intermediates, ranges, and alternatives can be represented, and distinction is made between ‘variable’, ‘unknown’, and ‘not applicable’. There is provision for comments, which can be used to indicate such things as probability, rarity, uncertainty, qualification, or amplification.

There is some redundancy in the coding system, to aid the detection of errors. Most errors have only a local effect, so that a program can continue to scan the rest of the data for other errors.

The system is versatile enough to replace the natural-language description as the primary means of recording and publishing descriptions. Some of the advantages of using it for these purposes are: (1) The coded descriptions occupy less space than natural-language descriptions. (2) It is easier to ensure that the descriptions are consistent and complete. (3) Translation into different natural languages is easier. (4) The data are easily accessible to computer programs (for information retrieval, classification, key construction, etc.).

Successful applications of the system so far include:

Grass genera, 162 characters, 261 genera.
Genera of Caesalpinioideae (subfamily of leguminous plants), 116 characters, 162 genera.
Species *of Colpochila* (genus of beetles), 40 characters, 172 species.
Species of *Orectognathus* (genus of ants), 26 characters, 29 species.
Vegetation types on Lord Howe Island, 35 characters, 26 types.

The Coding System

Basic structure of the data

Each record (line or card) consists of data, optionally preceded by a sequence number.

A sequence number is a positive, real number (e.g., 21.43). Programs reading the data check that the sequence numbers are in ascending order. A sequence number must start in the first column of a record, and is separated from the data by at least one blank. A blank in the first column of a record indicates that there is no sequence number, and sequence checking is then restarted from zero. It is strongly recommended that sequence numbers be used, as they guard against accidental disordering of the records, and can identify each record uniquely, which facilitates correction of the data.

Data are in free format. The main separators are the blank character or space and the record boundary (that is, the start or end of a record), which divide the data into 'words'. These separators are equivalent, with the minor exception that certain data must be contained on a single record. In the following descriptions, 'blank' denotes 'blank or record boundary', unless otherwise indicated.

A construction which is used in several contexts is the 'sentence'. This is a group of words terminated by a solidus (/) followed by a blank. A sentence may contain subsidiary material or comments enclosed by angle brackets (<>). The opening bracket must be preceded by a blank, and the closing bracket must be followed by a blank or by the terminating solidus. Table 1 shows several examples of sentences, including some with comments.

Table 1. Examples of character descriptions.

```
#1. striated area on maxillary palp <presence>/
    1. present/
    2. absent/
#2. pronotum <colour>/
    1. red/
    2. black/
    3. yellow/
#3. eyes <size>/
    1. of normal size <i.e. less than 0.5mm in diameter>/
    2. very large <i.e. more than 0.5mm in diameter>/
#4. frons <setae>/
    1. with setae on anterior middle and above eyes/
    2. with setae above eyes only/
    3. without setae/
#5. number of lamellae in antennal club/
#6. length/ mm/
```

Character descriptions

The taxa are described in terms of a set of characters, each of which consists of a feature and a set of states. Four main types of character are recognized: unordered multistate (UM), ordered multistate (OM), integer numeric (IN), and real numeric (RN). Multistate characters may be further specified as exclusive (EUM and EOM). A multistate character has a fixed number of states, whereas a numeric character has (in principle) an infinite number of states. Table 1 shows some examples of character descriptions. The start of each character description is indicated by a numero (#) preceded by a blank. Then follows a sentence containing the feature description; the first word of this sentence indicates the character number. The feature description is followed by sentences containing the state descriptions (for multistate characters) or by a sentence containing the units (for numeric characters). Characters 1, 2, and 3 in Table 1 are unordered multistate, 4 is ordered multistate, 5 is integer numeric, and 6 is real numeric. (For two-state characters, the distinction between unordered and ordered is arbitrary.)

Taxon descriptions

A taxon description consists of one or more ‘item descriptions’, each of which describes one form or variant of the taxon (usually one item per taxon is sufficient). An item description consists of the taxon name followed by a set of attributes. The start of an item description is indicated by a numero (#) preceded by a blank. The taxon name (together with any subsidiary information) is a sentence.

Example

Taxon name and comment.
#Archaeoglenes nemoralis <Ford, 1968: 161. Type locality:
Pablo Valley, Oahu, Hawaii. Holotype, female (Bishop Museum)>/

An attribute is coded as a word consisting of a character number, together with the character values (state numbers or numerical values) which apply to the taxon being described. There are special symbols ‘V’, ‘U’, and ‘-’ representing ‘variable’, ‘unknown’, and ‘not applicable’, respectively. These are called ‘pseudo-values’. The simplest form of an attribute is

$$c,v$$

where *c* is a character number and *v* is a character value or pseudo-value.

Example

With the characters defined in Table 1, the codes
1,V 4,3 5,- 6,8.5
represent
Striated area on maxillary palp present; or absent. Frons without setae. Number of lamellae in antennal club not applicable. Length 8.5 mm.

The general form of an attribute is

$$c<e_0>,r_1<e_1>/r_2<e_2>/\ldots r_n<e_n>$$

where *c* is a character number, r_i is a value or combination of values (see below), ‘/’ is a separator denoting ‘or’, and ‘$<e_i>$’ is optional extra information. Blanks are permitted within e_i. r_i takes one of the forms

$$v$$
$$v_1-v_2-\ldots v_m$$
$$v_1\&v_2\&\ldots v_m$$

where *v* is any character value or pseudo-value, v_j is any character value (not a pseudo-value), ‘-’ is a separator denoting ‘to’, and‘&’ is a separator denoting ‘and’.

Example

The codes
l,1/2<rare> 2,2/2&3<striped> 3,1-2 6,7-8.5
represent
Striated area on maxillary palp present; or absent <rare>. Pronotum black; or black and yellow <striped>. Eyes normal to very large. Length 7 to 8.5 mm.

Table 2 contains more examples of taxon descriptions, taken from real data.

Control phrases

If the data are to be read by a program, other information required by the program may be indicated by means of control phrases. A control phrase consists of up to four words, and must be preceded by a star (*) preceded by a blank. Table 2 shows some examples.

Table 2. Example of a deck to convert taxon (item) descriptions to the format required by the key-generating program KEY. Unless otherwise specified, character types are assumed to be UM, and multistate characters are assumed to have 2 states. The KEY STATES control phrase specifies how numeric characters are to be converted to multistate characters for the purpose of constructing a key.

```
*HEADING. Lord Howe Island vegetation types/
*REMARK. Revised 5/10/78.

*NUMBER OF CHARACTERS 35
*MAXIMUM NUMBER OF STATES 8
*MAXIMUM NUMBER OF ITEMS 30

*CHARACTER TYPES 1,RN 3,RN 4,OM 7,RN 9,RN 11,RN 12,OM 13,OM 14,IN 17,OM
   19,OM 20,OM 21,OM 22,OM 23,OM 25,OM 27,OM 32,OM 34,OM

*NUMBERS OF STATES 2,5 4,4 5,7 6,5 8,2 10,5 12,3 13,4 14,0 15,2 16,8 17,3
   18,5 19,4 20,4 21,4 22,4 23,4 24,5 25,3 26,3 27,5 28,2 29,5 30,6 31,6
   32,5 33,3 34,3 35,4

*KEY STATES 1,300/300 3,45/45 7,700/700 9,10/10 11,1.5/1.5-7/7 14,1/2/3/4

*TRANSLATE INTO KEY FORMAT

*ITEM DESCRIPTIONS
*SEQUENCE NUMBERS

1.01 # Lowland mixed forest
1.02 <rainforest subformation. Cleistocalyx - Linociera alliance>/
1.03 1,30-375 2,3/5<also on N and S> 3,0-40 4,2-3 5,2/5<occasionally on
1.04 top slope> 6,5 7,740 8,1 9,20 10,2 11,7-13 12,2-3 13,2 14,3 15,2
1.05 16,1&2&4<shrubs uncommon> 17,1 18,1&2&3 19,4 20,4 21,4 22,1 23,4
1.06 24,1 25,2 26,V 27,3 28,2 29,1 30,V<usually smooth>
1.07 31,1<some spurs, planks and stilts> 32,2-5 33,1 34,1 35,1/3

2.01 # Bubbia howeana - Dracophyllum fitzgeraldii association
2.02 <gnarled mossy forest subformation. Bubbla - Dracophyllum alliance>/
2.03 1,750-810 2,V 3,0-45<usually low> 4,V<high on cliff edges> 5,1/7 6,5
2.04 8,2 9,- 10,- 11,5 12,1 13,1<very dense> 14,3 15,2 16,2&3&4&5&8<and
2.05 tree ferns> 17,1 18,1&5<mosses abundant> 19,1 20,4 21,2 22,4
2.06 23,4<both vascular and nonvascular> 24,2 25,1 26,2 27,4 28,2 29,2&5
2.07 30,V 31,1&6 32,1-3 33,1 34,1 35,V

etc.

*END
```

Programs

General descriptions

Two programs are currently available.

CONFOR (Dallwitz 1979). This program converts the data into other formats. It allows masking of both items and characters. The options currently available are:

Write binary file. The binary file contains the same information as the original data, but in a form which can be read more quickly by a computer. If several runs are required on the same set of data, the cost can be reduced by writing a binary file in the first run, and reading from it in the subsequent runs. *Translate into DELTA format.* This option can be used to tidy up the data by removing excess blanks, putting the attributes in order of character number, and generating record sequence numbers. Also, the characters may be renumbered, in order to maintain a logical order when new characters have been added. *Translate into natural language*. This produces natural-language descriptions. *Translate into KEY format*. This converts the data to the format required by the key-forming program KEY (Dallwitz 1974). *Translate into MULTBET format.* This converts the data to the format required by most of the Lance and Williams classification programs (e.g., Lance and Williams 1967).

PABTRAN (Higgins 1979). This program translates peekaboo cards into KEY or DELTA format.

Technical details

The reading of the character and taxon descriptions in DELTA format is carried out by subroutines, which can be taken from the program CONFOR and used in other programs. Alternatively, programs can use the binary output produced by CONFOR. This is read simply by unformated READ statements, and requires no decoding or error checking. However, because of the generality of the coding system, there may still be a considerable amount of processing required to simplify or transform the data to the form needed for a particular application.

The programs are written in ANS FORTRAN (American National Standards Institute 1966) (except for the parts of PABTRAN which read and write binary cards). The storage and comparison of characters (symbols) is not defined in the standard, and is handled as follows. Characters are stored in integer variables, 1 per variable or array element, left justified with blank fill; that is, they are set in DATA statements as 1H*x* and read and written in A1 format. They are compared by means of the usual FORTRAN .EQ. and .NE. relations. This may not be allowed on some machines. However, the method of character representation can easily be changed, by changing DATA statements (in two subroutines), a reading subroutine, and a writing subroutine.

The *User's guide to the DELTA system* (Dallwitz 1979) includes a 48x microfiche containing complete listings of CONFOR, PABTRAN, and KEY. The programs are also available on magnetic tape. When requesting a tape, please specify code (EBCDIC or ASCII), density, and block length.

Acknowledgements

I am grateful to many colleagues for supplying diverse data and ideas, which have guided the development of the coding system and the associated programs. I particularly wish to thank Mr. L. Watson, of the Australian National University, Canberra.

References

American National Standards Institute 1966. American National Standard FORTRAN, X3.9–1966. American National Standards Institute, New York.

Dallwitz, M.J. 1974. A flexible program for generating identification keys. *Syst. Zool.* 23: 50–57.

Dallwitz, M.J. 1978. User's guide to KEY: a computer program for generating identification keys. CSIRO Aust. Div. Entomol. Rep. No. 4, 14 pp. + microfiche.

Dallwitz, M.J. 1979. User's guide to the DELTA system: a general system for coding taxonomic descriptions. CSIRO Aust. Div. Entomol. Rep. No. 13, 71 pp. + microfiche. [Actually published in April 1980.]

Higgins, J.P. 1979. User's guide to PABTRAN—a computer program for translating a character-taxon matrix from peekaboo cards into KEY or DELTA format. CSIRO Aust. Div. Entomol. Rep. No. 6.

Lance, G.N., and Williams, W.T. 1967. Mixed-data classificatory programs. I. Agglomerative systems. *Aust. Comp. J.* 1: 15–20.

Watson, L., and Milne, P. 1972. A flexible system for automatic generation of special-purpose dichotomous keys, and its application to Australian grass genera. *Aust. J. Bot.* 20: 331–352.