Dallwitz, M.J. 1993. DELTA and INTKEY. In: Fortuner, R. (editor). Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision. Johns Hopkins University Press, Baltimore. p. 287–296.
Also available at http://delta-intkey.com.

DELTA and INTKEY

1993

M. J. Dallwitz

DELTA (Dallwitz 1980; Dallwitz and Paine 1986; Partridge et al. 1988) is a multipurpose format for generating identification keys. It is not geared, as are many formats, to the requirements of one particular type of program (e.g., Dallwitz 1974; Rohlf et al. 1981; Swofford 1984). It was designed to be easy for people to use. On the other hand, a degree of complexity was necessary to avoid loss of significant information, and the complexity has increased over the years in response to requests from users. I wrote a program called CONFOR to translate the format into natural language and into formats used by various other programs. This makes the data accessible to programs that carry out key generation, phenetic and cladistic analysis, and interactive identification and information retrieval. CONFOR also helps with maintenance of the data, such as keeping the data tidy and changing the order of the characters.

The DELTA Coding System

The DELTA format is based on ordinary text files (sequential files of ASCII characters, with records of up to 120 characters). These files may be created and modified with any text editor or word processor (we will soon be writing a new system based on random-access files, with an integrated editor). The data are in free format; that is, they do not have to be positioned in fixed fields in the records. The examples below are taken from a small subset of one of Leslie Watson's data sets (Watson and Dallwitz 1981; Watson et al. 1988). We distribute this subset with the programs.

```
#1. <Synonyms: i.e. 'genera' included in the current description>
#2. <Longevity of plants>/
    1. annual <or biennial, without remains of old sheaths or culms>/
    2. perennial <with remains of old sheaths and/or culms> <Figs 1, 2, 18>/
#3. <Mature> culms <maximum height: data unreliable for large genera>/
    cm high/
#4. Culms <whether woody or herbaceous>/
    1. woody and persistent/
    2. herbaceous <not woody, not persistent>/
#5. Culms <whether branched above>/
    1. branching <vegetatively> above <Fig. 2>/
    2. unbranched <vegetatively> above <Figs 1, 7>/
#6. <Culm> nodes <whether hairy or glabrous>/
    1. hairy <Figs 4, 33>/
    2. glabrous <Fig. 4>/
#7. Leaf blades <shape: data very incomplete>/
    1. linear/
    2. linear-lanceolate/
    3. lanceolate/
    4. ovate-lanceolate/
```

Figure 18.1. Part of a DELTA format character list.

Figure 18.1 shows a DELTA character list. Five types of character are available: text characters (e.g., *1*); multistate characters, which can be either ordered (e.g., *7*) or unordered; and numeric characters, which can take either real (continuously variable) values (e.g., *3*) or integer values. Comments enclosed in angle

brackets can be placed anywhere; they are omitted from most kinds of output. There are no restrictions on the numbers of characters or states or on the amount of text.

Figure 18.2 shows a coded taxon description. The name of the taxon is at the top. A typical *attribute* consists of a character number, a comma, then a state number (e.g., attribute 6: *6,2*). Text attributes are slightly different, consisting of a character number and text within angle brackets (e.g., attribute 1: *1*<Czernya ...>).

```
# Phragmites <Adans>/
1<Czernya Presl, Miphragtes Nieuwland, Oxyanthe Steud., Trichoon Roth, Xenochloa Roem. &
Schult.> 2,2 3,80–400(–1000) 4,1–2<often somewhat persistent> 5,1<especially when main culm
damaged>/2 6,2 7,2–3 8,1 9,2 10,1 11,3 12,2 13,5 14,1<20–60 cm long, plumose, the fertile lemmas
surrounded by long white silky hairs> 15,2 16,2 18,– 19,2 25,9–16 26,1 27,1<at least above the L1>
28,2 29,1 30,1 31,2 32,1 34,2 35,2<rounded on the back> 36,2 37,3 38,1 39,1 40,2 41,(2–)3–10 42,1
43,1<acute to acuminate or aristulate> 44,1/3<narrow-attenuate, muticous to aristulate> 45<(if
lemmas aristulate)>,1 46,3 47,1 48,1 50,1–3 51,1 52,2 53,1 54,1 55,1/2 56,3<or two in the lower
floret> 57,1 58,2 59,3 60,1 61,2 62,1 63,2 64,2 66,1 67,1 68,2 69,1 70,2 72,2 73,2 74,3 77,4 82,3
83,1&2&3&5&6
```

Figure 18.2. A description coded in DELTA format.

In more complex cases we can have several states separated by "/" (meaning or), and we can have comments associated with any of those character states (e.g., attribute 44). We can have ranges, denoted by "–" (e.g., *7*), and we can have states separated by "&" (meaning *and*) (e.g., *83*). The three separators /, –, and & can be combined within the same attribute. Ranges of values of numeric characters can include parentheses to indicate values outside the normal range (e.g., *3*).

Output Produced from DELTA Data

Natural-Language Descriptions

We can use CONFOR to translate a coded description into natural language, as shown in Figure 18.3. This was produced and typeset automatically from the data, without manual intervention. It corresponds to the data in Figure 18.2. Notice that parts of the description are in italics. These parts constitute a diagnostic description. The diagnostic characters were selected by the program INTKEY and then fed through to CONFOR, which was instructed to italicize the parts of the description corresponding to these characters.

Phragmites Adans.

Czernya Presl, *Miphragtes* Nieuwland, *Oxyanthe* Steud., *Trichoon* Roth, *Xenochloa* Roem. & Schult.

Habit, vegetative morphology. Perennial. Culms 80–400(–1000) cm high; woody and persistent to herbaceous (often somewhat persistent); branching above (especially when main culm damaged), or unbranched above. Nodes glabrous. Leaf blades linear-lanceolate to lanceolate; broad. Adaxial ligule a fringe of hairs.

Reproductive organization, inflorescence. Plants bisexual, with bisexual spikelets. Inflorescence paniculate; open (20–60 cm long, plumose, the fertile lemmas surrounded by long white silky hairs); not comprising 'partial inflorescences' and foliar organs. Spikelet-bearing axes persistent Spikelets not in distinct long-and-short combinations.

Female-fertile spikelets. Spikelets 9–16 mm long; compressed laterally; disarticulating above the glumes (at least above the LI); disarticulating between the florets; with the rachilla prolonged apically. Glumes two; very unequal; decidedly shorter than the adjacent lemmas; awnless; not carinate (rounded on the back). Spikelets with incomplete florets. The incomplete florets both distal and proximal to the female-fertile florets. Proximal incomplete florets 1; male; awnless. Female-fertile florets (2–)3–10. Lemmas entire; pointed (acute to acuminate or aristulate); awnless, or awned (narrow-attenuate, muticous to aristulate). Awns (if lemmas aristulate) 1; apical; non-geniculate; much shorter than the body of the lemma. Lemmas 1–3 nerved. Palea present; conspicuous but

relatively short. Lodicules present; fleshy; ciliate, or glabrous. Stamens 3 (or two in the lower floret). Ovary glabrous. Stigmas 2; brown.

Fruit. Fruit small; smooth. Hilum short. Pericarp fused.

Photosynthetic pathway, leaf blade anatomy. C_3. XyMS+. Mesophyll with arm cells; without fusoids. Midrib conspicuous; with a conventional arc of bundles; without colourless tissue adaxially. All the vascular bundles accompanied by sclerenchyma.

Taxonomy. Arundinoideae; Arundineae.

Distribution. 3 species. Holarctic Kingdom, Paleotropical Kingdom, Neotropical Kingdom, Australian Kingdom, and Antarctic Kingdom.

Figure 18.3. The description in Figure 18.2 translated into natural language.

Identification Keys

In Figure 18.4 we have part of an identification key produced by first translating the data into an intermediate format, then passing them through our key generation program, KEY (Dallwitz 1974; Dallwitz and Paine 1986). Again, everything is completely automatic, including the typesetting. However, the user has a lot of control over the structure of the key, by changing parameter values.

1(0). Spikelets disarticulating above the glumes 2
Spikelets falling with the glumes 11
Spikelets not disarticulating 13
2(1). Female-fertile florets 1 3
Female-fertile florets 2 or more 8
3(2). Inflorescence of spike-like main branches; lodicules fleshy; C_4 4
Inflorescence paniculate; lodicules membranous; C_3 5
4(3). Glumes very unequal; lemmas awned; stigmas white; biochemical type PCK ..... **Chloris**
Glumes more or less equal; lemmas awnless; stigmas red pigmented; biochemical type NAD-ME ..... **Cynodon**
5(3). Ovary glabrous 6
Ovary hairy 7
6(5). Spikelets with female-fertile florets only; stamens 3; hilum short; mesophyll without arm cells; midrib with one bundle only ..... **Agrostis**
Spikelets with incomplete florets; stamens 5 to 6; hilum long-linear; mesophyll with arm cells; midrib with complex vascularization ..... **Oryza**

Figure 18.4. Part of a computer-generated key.

Foreign Languages

The character list can be translated into other natural languages. This is done manually, but then all the products (descriptions, keys, and interactive identification) are available automatically in that other natural language (e.g., French [Watson et al. 1986], Greek [Watson et al. 1988], Spanish and Portuguese [Webster et al. 1989], and Chinese [Xu Zhu et al. 1992]). Figure 18.5 shows part of a key in Greek, and Figure 18.6 shows a description in Chinese.

The programs themselves (directives, error messages, and manuals) have recently been translated into Spanish (Valdecasas et al. 1990) and Chinese (Xu Zhu et al. 1992). Future versions of the programs will be much more convenient to maintain in different languages, because all text will be in files separate from the program files.

1(0). Θηλυκά-γόνιμα σταχύδια αποκοπτόμενα πάνω από τα λέπυρα. 2
Θηλυκά-γόνιμα σταχύδια αποκοπτόμενα μαζί με τα λέπυρα. 96
Θηλυκά-γόνιμα σταχύδια μη αποκοπτόμενα. 141
2(1). Θηλυκά-γόνιμα σταχύδια με ατελή ανθίδια στη βάση. 3
Θηλυκά-γόνιμα σταχύδια χωρίς ατελή ανθίδια στη βάση. 8
3(2). Ταξιανθία με φύλλα ή μέσα σε σπάθη· ελάσματα φύλλων με εύκολα ορατές κάθετες νευρώσεις· μεσόφυλλο με ατρακτοειδή κύτταρα. . . . . **Arundinaria**
Ταξιανθία όχι με φύλλα ή μέσα σε σπάθη· ελάσματα φύλλων χωρίς εύκολα ορατές κάθετες νευρώσεις· μεσόφυλλο χωρίς ατρακτοειδή κύτταρα. 4
4(3). Καρπός με μικρή ουλή. 5
Καρπός με μία μακριά-γραμμική ουλή. 7
5(4). Γλωσσίδα μεβρανώδης χωρίς κροσσό από τρίχες· θηλυκά-γόνιμα σταχύδια χωρίς επιμηκυσμένη στην κορυφή ραχίλλα· εμβρυο μικρό· φύλλα χωρίς ωτία· ελάσματα φύλλων μη αποκοπτόμενα από τους κολεούς. 6
Γλωσσίδα κροσσωτή· θηλυκά-γόνιμα σταχύδια με ραχίλλα επιμηκυσμένη στην κορυφή· εμβρυο μεγάλο· φύλλα με ωτία· ελάσματα φύλλων τελικά αποκοπτόμενα από τους κολεούς. . . . . **Phragmites**

Figure 18.5. Part of a computer-generated key, produced in Greek by translation of the character list.

9. 肥披碱草 Elymus excelsus Turcz.

茎140厘米高，基部具白色粉层．叶鞘基部光滑，或被微柔毛．叶扁平，20-30厘米长，10-16毫米宽，叶两面粗糙．穗状花序稠密，直立，15-22厘米长，绿色．穗轴被短硬毛，节间膨大，光滑．小穗4-5，12-25毫米长．颖狭披针形；10-13毫米长，等长，脉5-7，粗糙，背部光滑，芒7毫米长．外稃披针形，背部光滑，基部光滑，脉5，8-12毫米长，芒15-40毫米长，反折，粗糙．内稃短于外稃；截形，脊全部具毛，脊间无毛．

染色体基数，X= 42．地理分布 东北，华北，四川；山披，草地，路边．

饲用价值：优良牧草，适口性好，产草量高，具有较高的营养价值．

Figure 18.6. A computer-generated natural-language description in Chinese.

Typesetting

By default, CONFOR and KEY produce plain ASCII files, suitable for viewing on a computer screen or printing on an ordinary printer. However, they can be instructed to put typesetting marks in their output, which then may be processed by our typesetting program, TYPSET (Dallwitz and Zurcher 1988). The input data may also include typesetting marks (e.g., superscripts and subscripts, font changes). The programs normally pass these through to TYPSET, but they can be made to remove them, for example, to produce plain text for display on a screen (Dallwitz 1984). CONFOR and KEY were designed so that they would be easy to adapt to other typesetting or word-processing systems. A cruder way to convert to other typesetting systems would be to edit the typesetting marks in the intermediate files.

The Interactive Identification Program INTKEY

Introduction

Our interactive identification program, INTKEY, was developed from version 3 of Richard Pankhurst's ONLINE program (Pankhurst and Aitchison 1975), which we got in 1982, modified, and eventually completely rewrote as INTKEY. We are currently completely rewriting it again, to add new features suggested by experience with the earlier versions. The new version was released in October 1991. Pankhurst has also continued to enhance ONLINE, which is now in version 6.

INTKEY provides tools for identification and information retrieval. It does not provide a fixed sequence of actions: it lets you choose what the actions are to be, and you are free to follow a quite complex path through it. It is a complex system, but users can easily be instructed in the use of simple sequences of operations.

INTKEY has complete online help, and "?" can be entered at any prompt to get some information about the required response. The new version will be completely menu driven, although a command-line interface will still be there and will be preferred by experienced users.

*Best: display the best characters to separate the remaining taxa
*CHaracters: display names and numbers of characters
COmment: ignore text
DAta: read main data files
DEFine: define a keyword to represent a set of characters or taxa
*DELete: delete a previously used character
*DEScribe: display the description of a taxon
DIAgnose: generate diagnostic descriptions of taxa
*DIFferences: display the differences between taxa
DISplay: set screen display and prompts
EXAct: specify characters not subject to error
EXClude: exclude characters or taxa
FILes: menu for file input/output, display, and prompts
*FINish: exit from the program
FIX: retain the current character values when restarting
Help: display information about commands
INClude: include characters or taxa
INPut: read commands from a file
Keywords: display keywords
Log: send input and output to a file
MAtch: set criteria for matching of taxon descriptions
MEnu: return to main menu
OMit: omit inapplicable or unknown characters from descriptions
*OUtput: send output to a file
Parameters: menu for setting or displaying parameters
RELiabilities: set character reliabilities
REMark: copy text to the output file
*REStart: restart an identification
SAve: generate files for input to other programs
*SEParate: display the best characters to separate a taxon from the rest
*SET: set autobest, *autotaxa, rbase, stopbest, *tolerance, varywt
SHow: display text on the screen
Similarities: display the similarities between taxa
STatus: display parameter values
SUBset: generate files containing subsets of the data
SUMmary: display a summary of the data
*TAxa: display names and numbers of taxa
*Use: use a character to describe the specimen

Figure 18.7. The INTKEY commands, with short descriptions. The asterisks mark features that were present in ONLINE version 3.

Database systems are now readily available off the shelf, and there is a growing tendency to think that it should be easy to use these for interactive identification. After all, is there anything more to it than finding which taxa have a certain value in a certain data field? We have spent many years enhancing Pankhurst's ONLINE (which was already quite a powerful program), adding features that we thought to be essential in the light of experience with nontrivial data sets (several hundred characters or taxa). Figure 18.7 is a current list of the program commands, taken from the menus. The asterisks mark features

that were present in ONLINE version 3, though all of these have been enhanced (with the possible exception of the FINISH command). Figure 18.8 is the Help for one of the commands, as a detailed illustration of the type of facility that is needed for a practical system.

MATCH options

where options is one or more of the letters
I – inapplicables
U – unknowns
S – subset
O – overlap
E – exact

This command specifies which character values are to be regarded as equal – i.e. 'match' – in the commands USE, DIFFERENCES, SIMILARITIES, or TAXA. MATCH I and MATCH U specify respectively that 'inapplicable' and 'unknown' match any value. MATCH S specifies that two sets of values match if one set (usually the values of the specimen) is a subset of the other. (E.g. 1/2 matches 1/2/4 but not 2/3; 2–5 matches 1–6 but not 4–10). MATCH O specifies that two sets of values match if they overlap, i.e. if they have any values in common (e.g. 1/2 matches 2/3; 2–5 matches 4–10). (S and O cannot be used together.) MATCH E or MATCH without parameters specifies that two sets of values match only if they are identical.

The default setting is MATCH O U I, which is usually the most appropriate for identification. For information retrieval, the most appropriate setting is usually MATCH O.

Figure 18.8. The Help text for the INTKEY command match.

Examples

Watson et al. (1989) and Dallwitz (1989a,b) give extensive examples of the use of the program. However, I may be able to give you some idea of the flexibility of the program by describing some of the possible courses of action once a tentative identification is made, that is, once the program has indicated that only one taxon matches the specimen description that you have entered. Actually, any of the commands below might be useful at any stage of the identification, and we feel strongly that programs should allow this kind of flexibility, rather than leading the user along predetermined pathways. This certainly means that some effort is required to learn to make the best use of the program, but this should be acceptable to professional users wanting to achieve professional results. (By "professional," I mean not just taxonomists, but everyone who needs identification or information retrieval as part of their job.)

DESCRIBE SPECIMEN
Recapitulate the specimen description that you have entered, so that you can check it.

DESCRIBE REMAINING
Display the full description of the "remaining" taxon. REMAINING is an example of an automatically denned "taxon keyword" representing a set of taxa. At this stage of the identification, it represents a single taxon, but at earlier stages it would represent several.

DESCRIBE REMAINING HABIT DISTRIBUTION ECOLOGY
Display the description of the remaining taxon in terms of its habit, distribution, and ecology. These are examples of user-defined "character keywords" representing sets of characters. They would generally have been defined by the person who prepared the data.

DIAGNOSE REMAINING
Generate and display a diagnostic description of the remaining taxon, in terms of characters not used in the identification. This description will distinguish the remaining taxon in at least one respect from all the other taxa, and so provides an independent check.

DIFFERENCES (SP 6)
Display the differences between the specimen description and taxon 6. (Maybe you thought your specimen was taxon 6. What is the evidence that it is not?)

MATCH EXACT

DIFFERENCES (SP REM)
Set exact matching (see Fig. 18.8) and display the differences between the specimen description and the remaining taxon. If the match setting were left as it was during the identification (normally Overlap, Unknown, Inapplicable), no differences would be shown, because the remaining taxon is, by definition, the one that matches the specimen. Setting match exact allows the difference command to pinpoint characters where the specimen and the remaining taxon differ because of variability or because the character is unknown or inapplicable for the remaining taxon.

SET TOLERANCE 1
Set the "tolerance" parameter to 1. This brings back as "remaining" taxa all those that differ in not more than one respect from the specimen description. You can then continue with the identification as before. This is particularly useful if you suspect or know that there has been an error; for example, if the number of taxa remaining is 0, or if the description of the remaining taxon does not fit the specimen.

ILLUSTRATE TAXA REM
Display illustrations of the remaining taxon. This is not available in the current version (1990), but is implemented in the new one (1991).

Conclusion

We are aiming to produce practical tools, not just to develop methods: we want to put the methods in the hands of a wide variety of users. We support the programs, and they evolve through feedback from the users. We aim to avoid manual manipulation of data wherever possible, so we provide pathways from one program to another. We want the programs and the data to have depth and flexibility, without ad hoc restrictions built in, so that people can use them in ways we did not anticipate. We want the programs to be able to benefit both the compiler of the data and end user. Perhaps these aims are rather ambitious, but I think we are succeeding to some extent.

Acknowledgments

I trained in physics and mathematics and joined the CSIRO Division of Entomology to do ecological modeling, but my work evolved into general computing, and I developed a particular interest in taxonomy.

In about 1971 I started writing a program for generating identification keys (Dallwitz 1974), and a few years later I met Leslie Watson of the Australian National University, who was using Richard Pankhurst's KEYGEN program for the same purpose (Pankhurst 1970). In collaboration with Peter Milne, of the CSIRO Division of Computing Research, Watson had devised a more flexible format for preparing data for this program (Watson and Milne 1972). Leslie and I talked about it and decided that the idea could be improved, and this was the origin of the DELTA format. I have worked closely with Leslie ever since, and the development of the programs would not have been possible without his collaboration.

References

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

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

Dallwitz, M.J. 1984. Automatic typesetting of computer-generated keys and descriptions. In 'Databases in systematics', Systematics Association Special Volume No. 26, pp. 279–90. (Eds R. Allkin and F.A. Bisby.) Academic Press, London.

Dallwitz, M.J., and Paine, T.A. 1986. User's guide to the DELTA system: a general system for processing taxonomic descriptions. 3rd edition. CSIRO Aust. Div. Entomol. Rep. No. 13, 106pp. A later version is available at http://delta-intkey.com.

Dallwitz, M.J., and Zurcher, E.J. 1988. User's guide to TYPSET: a computer typesetting program. 2nd edition. CSIRO Aust. Div. Entomol. Rep. No. 18, 25pp.

Dallwitz, M.J. 1989a. Diagnostic descriptions from INTKEY and CONFOR. DELTA Newsletter 3: 8–13. Also available at http://delta-intkey.com.

Dallwitz, M.J. 1989b. Diagnostic descriptions for groups of taxa. DELTA Newsletter 4: 8–13. Also available at http://delta-intkey.com.

Pankhurst, R.J. 1970. A computer program for generating diagnostic keys. Comput. J. 13, 145–51.

Pankhurst, R.J., and Aitchison, R.R. 1975. An on-line identification program. In 'Biological Identification with Computers', pp. 181–5. (Ed. R.J. Pankhurst.) Academic Press, London.

Partridge, T.R., Dallwitz, M.J., and Watson, L. 1988. A primer for the DELTA System on MS–DOS and VMS. 2nd edition. CSIRO Aust. Div. Entomol. Rep. No. 38, 17pp.

Rohlf, F.J., Kishpaugh, J., and Kirk, D. 1981. NT-SYS – Numerical taxonomy system of multivariate statistical programs. The State University of New York, Stony Brook, N.Y.

Swofford, D.L. 1984. Phylogenetic analysis using parsimony. Version 2.2. Illinois Natural History Survey, Champaign.

Valdecasas, A.G.-, Elvira, J.R., Becerra, J.M., and Bello, E. 1990. Dallwitz's DELTA: CONFOR, KEY, DIST, TRANSNTE. Museo Nacional de Ciencias Naturales, Madrid. (Spanish User's Guide for the DELTA System.)

Watson, L., and Dallwitz, M.J. 1981. An automated data bank for grass genera. Taxon 30: 424–429 + 2 microfiche.

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.

Watson, L., Dallwitz, M.J. and Johnston, C.R. 1986. Grass genera of the world: 728 detailed descriptions from an automated database. Aust. J. Bot. 34: 223–230.

Watson, L., Damanakis, M., and Dallwitz, M.J. 1988. The grass genera of Greece: descriptions, classification, keys. In Greek. University of Crete, Heraklion.

Webster, R.D., Kirkbride, J.H., and Reyna, J.V. 1989. New World genera of the Paniceae (Poaceae: Panicoideae). SIDA 13: 393–417 + microfiche.

Xu Zhu, Dallwitz, M.J., and Watson, L. 1992. Chinese and English botanical keys generated by computer. Grassland of China 12(1): 53–57.