Field Studies 8 (1992), 1–30

RANDOM-ACCESS IDENTIFICATION GUIDES FOR A MICROCOMPUTER

COLIN J. LEGG

*University of Edinburgh Institute of Ecology and Resource
Management, The King's Buildings, Mayfield Road
Edinburgh, EH9 3JU*

ABSTRACT

A random-access identification system is described which can be used for the identification of a group of organisms from a database prepared for either IBM-PC or BBC-B compatible computers. The programs are easy to use and are designed either for plant/animal identification for the amateur biologist, or for educational purposes. The KEY program enables the user to enter a few characters describing an unknown specimen in any order, and then interrogate the data base for a ranked list of the most similar species. Comparisons between species, and suggestions for further characters which would best discriminate between species are also provided. The ranking of species is based on a weighted similarity system which is robust to natural variability in the material and a small number of errors on the part of the user. Facilities are included for users to modify existing data bases or to construct their own keys to any reasonably homogeneous group of organisms. The programs are described in relation to, and distributed with, a "Guide to the Sedges of the British Isles" (Legg 1992). The programs and database are available on disc from FSC, Preston Montford, Montford Bridge, Shrewsbury, SY4 1HW, UK.

INTRODUCTION

The first part of this paper describes the essential information about a random-access key system for identifying organisms using a microcomputer. These basic instructions will provide most of the information required by the majority of users. This is followed by a more detailed explanation of how the system works and instructions for some additional features and the ancillary programs. This will be of interest to teachers and other advanced users of the system who wish to construct their own database.

The programs are designed to run on IBM-PC compatible machines using PC-DOS, MS-DOS (version 3 onwards) or MS-Windows. A different version is available for the BBC-B (32K) and upwardly-compatible machines using a random-access disc filing system. Either black and white or colour monitors are suitable. Note, however, that not all black and white monitors on IBM-compatible machines support highlighting of characters on the screen.

The capacity of the program is determined by both disc space and the amount of comment used in the data set, but 100K (40 track) discs with a 32K BBC-B will handle data for about 200 species with 300 character states and a few words of comment on each species; larger data sets are possible with extra memory and 80 track drives; the MS-DOS version is dimensioned for up to 1000 species and 1000 possible character states. Speed of operation and ease of use are hardly affected by the size of the database.

Instructions are also given for the design and construction of a new data base. New data files can be created with any text editor or standard word processor that can produce unformatted ASCII text files. Programs are included on disc which read and check these files and encode them into a form which can be used by the key program. Though written in BBC-Basic and Pascal respectively, the two sets of programs appear very similar on the two systems and use identical data files which can be transferred between systems with the appropriate communications software. The MS-DOS version is faster than the BBC and takes advantage of the 80 column screen and extra memory which, though a big advantage when creating new key data files, is not essential. A hard disc gives a slight increase in speed of access to the data on the MS-DOS version but is not essential. For users with access to relatively fast IBM-compatible computers with a hard disc, an additional program is available which enables data files to be created or edited with ease in the form of a spreadsheet.

Most of the instructions given here apply to the versions written for both the BBC-B and IBM-PC compatible computers, but, where these differ, *instructions specific to the MS-DOS version are written in italics,* ***while information specific to the BBC version is given in this bold italic typeface.*** The figures illustrate the 80 column IBM screen, but BBC users will have no difficulty in interpreting any differences in the 40 column BBC screen.

BASIC INSTRUCTIONS

Taking a Backup Copy

Before using your Key program it is wise to take a backup of the disc.

On IBM-compatible machines, first boot the computer with MS-DOS, then insert the program disc into drive A and a formatted disc in drive B if you have two disc drives. Type COPY A:*.* B: *and press return. On single drive machines you will be instructed to exchange the discs as appropriate. If using a hard disc, then the files can be copied into suitable directories on drive C with, for example* COPY A:*.* C:\KEYDIR\*.* *where KEYDIR is a directory on drive C.*

On the BBC this can be done using a double disc drive by placing a formatted disc in drive 0 and the program disc in drive 1 and typing *COPY 1 0 *.* and pressing return. If using a single disc drive then place the master copy in the drive and type *COPY 0 0 *.* and you will be instructed to exchange the discs for the transfer of information. Once the copy is complete place the new disc in the default disc drive and type *OPT4,3 in order to enable the SHIFT-BREAK start-up option using the !BOOT file.

Loading the Program and selecting the Data File

Switch on the computer and boot with PC-DOS or MS-DOS. If running the program from a floppy disc insert the program disc into the default disc drive and close the door. If running from the hard disc, then enter the appropriate directory with CD C:\KEYDIR *Enter (where KEYDIR is the name of the directory containing the KEY files). Type* KEY *and press Enter.*

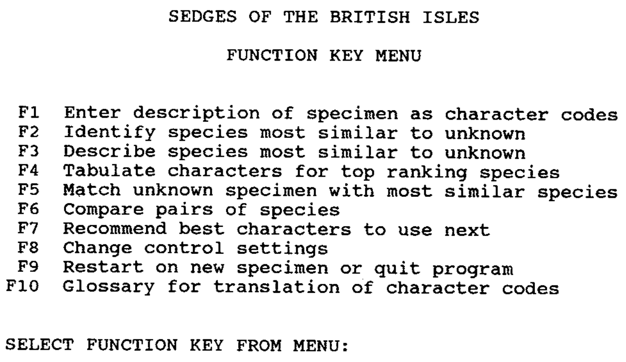
Place the program disc in the default disc drive (normally drive 0). If your BBC computer is normally configured to use cassette tapes or ADFS it may be necessary to type *DISC Return. To load and run the program either hold down the Shift key and press Break, or (with the computer in mode 6 or 7) type CHAIN "KEY" and press Return .

Instructions will now appear on the screen. Press the appropriate key for the identification guide you require (e.g. 1 for the “Guide to Sedges of the British Isles”. ***Newly acquired data files need to be installed on the disc for the BBC version. Refer to page 29 for details.*** *Data files for the MS-DOS version are assumed to be in the same directory as the key program. If not, then type X and enter the new path name or appropriate sub-directory.*

The Main Menu

Once the data file has been located, the “main menu” will be displayed on the screen (Fig. 1). This shows a list of ten options or “functions” which can be selected to instruct the program to perform particular operations such as displaying the characters of particular species. The options are selected by touching the function keys marked ***f0 to f9 on the BBC computer*** or *F1 to F10 on IBM keyboards.* Functions 2 to 5 cannot be used until you have entered some character data.

The menu can be re-displayed on the screen by pressing the **Escape** key at any time; the function keys, however, can also be used at any time and it is normally unnecessary to return to the menu once you are familiar with the key operations. The function keys can be labelled using a card similar to one of those given in Appendix 1. The following instructions give detailed explanations of each option. The first function you require will be number one: ‘Enter character data’.



```
                 SEDGES OF THE BRITISH ISLES

                      FUNCTION KEY MENU


 F1  Enter description of specimen as character codes
 F2  Identify species most similar to unknown
 F3  Describe species most similar to unknown
 F4  Tabulate characters for top ranking species
 F5  Match unknown specimen with most similar species
 F6  Compare pairs of species
 F7  Recommend best characters to use next
 F8  Change control settings
 F9  Restart on new specimen or quit program
F10  Glossary for translation of character codes


SELECT FUNCTION KEY FROM MENU:
```

FIG. 1 - The Main Menu

The screen display is obtained by pressing the ESCAPE key once the data have been loaded. The options are selected by pressing the function keys.

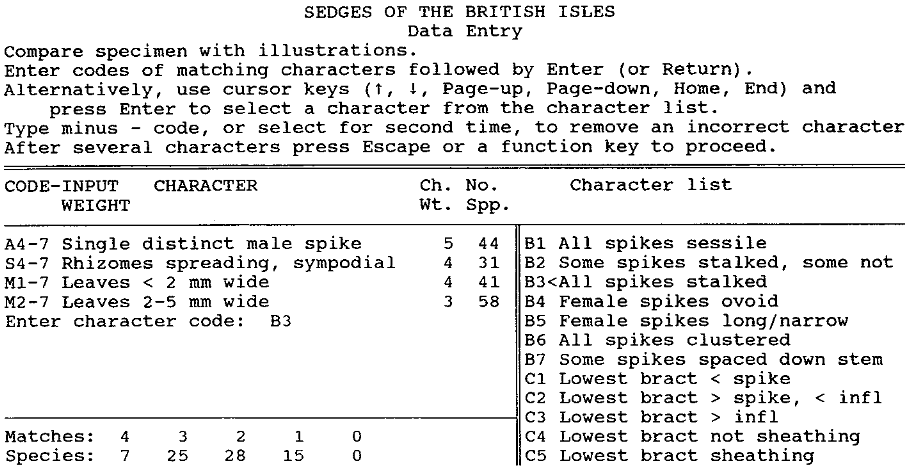
Function 1 - Enter Character Data

Select function key 1. Compare your specimen with the diagrams in the document of illustrated characters which is distributed with the database. Select character descriptions which accurately describe your specimen and type the character code into the computer and press return. The plant illustrated at the front of the *"Guide to Sedges of the British Isles"* (Legg, 1992), for instance, has the single male (terminal) spike of flowers appearing different from the female (lateral) spike, so type **A4 Return**). Make sure you use capital or lower case letters exactly as indicated in the illustrations. The computer will beep and an error message will result if you attempt to enter an invalid character.

For MS-DOS versions, a list of the character names is given on the right-hand side of the screen. You can scan through this list using the cursor-move keys (up or down arrow, page-up, page-down, home and end) or by pressing the initial letter of a character code. Once an appropriate character code is highlighted, this character can be selected by pressing Enter. The character will now appear highlighted in the character list.

Assuming that what you have typed represents a valid character, a translation will appear on the screen (Fig. 2), together with an indication of the total number of species which may at some time display that character. A 'character weight' is also given indicating (on an arbitrary scale from one to seven) the usefulness of that character in distinguishing between species within the group.

The bottom of the screen shows two lines of numbers which indicate the total number of species which may possess all of the characters entered so far, the number of species which match all but one of the characters entered, all but two, and so on. This gives some indication of the progress of the identification process.



```
                         SEDGES OF THE BRITISH ISLES
                                  Data Entry
Compare specimen with illustrations.
Enter codes of matching characters followed by Enter (or Return).
Alternatively, use cursor keys (↑, ↓, Page-up, Page-down, Home, End) and
    press Enter to select a character from the character list.
Type minus - code, or select for second time, to remove an incorrect character
After several characters press Escape or a function key to proceed.
================================================================================
CODE-INPUT   CHARACTER                     Ch. No.      Character list
     WEIGHT                                Wt. Spp.
--------------------------------------------------------------------------------
A4-7 Single distinct male spike             5   44 |B1 All spikes sessile
S4-7 Rhizomes spreading, sympodial          4   31 |B2 Some spikes stalked, some not
M1-7 Leaves < 2 mm wide                     4   41 |B3<All spikes stalked
M2-7 Leaves 2-5 mm wide                     3   58 |B4 Female spikes ovoid
Enter character code:   B3                         |B5 Female spikes long/narrow
                                                   |B6 All spikes clustered
                                                   |B7 Some spikes spaced down stem
                                                   |C1 Lowest bract < spike
                                                   |C2 Lowest bract > spike, < infl
---------------------------------------------------|C3 Lowest bract > infl
Matches:  4    3    2    1    0                    |C4 Lowest bract not sheathing
Species:  7   25   28   15    0                    |C5 Lowest bract sheathing
```

FIG. 2 - Data Entry - Function Key 1
The left-hand side of the screen shows the character codes and names of characters which have so far been entered as description of the unknown, together with the number of species which may match that character, and the character weight. The bottom of the screen shows that seven species match all four of the characters entered so far, 25 species match three characters, etc. The list of character names on the right-hand side of the screen is not present in the BBC version.

Though the characters are listed in the illustrations in some convenient order, it is better to start by scoring any distinctive or unusual features of your specimen (*e.g.* the distinct rhizome in our example, **S4**); the illustrated document will direct you to the best characters to score first if you are unfamiliar with the group.

Characters of which you are uncertain should be left at this stage. However, if your specimen is exactly intermediate between two classes of a series, it may be better to enter both classes since this at least distinguishes your specimen from species which are normally in a third class (e.g. the specimen at the front of the Sedge key has leaves exactly 2 mm wide but typing both **M1** (leaves < 2 mm) and **M2** (leaves 2-5 mm) distinguishes your specimen from species which are consistently **M3** or **M4**).

Type in five or six characters in the first instance. The accuracy of the identification increases rapidly as you enter more information, but the computer can suggest the most useful characters to look at next; this may save both time and errors. When you press **Escape** to return to the menu (or when you press any of the function keys) the character information which has been entered so far will be summarised on the top line of the screen. Note that the character codes are condensed so that **M1** and **M2** appear as **M12.**

Correcting a Wrong Entry

If you accidentally enter an invalid character code an error message will be displayed. If, however, you enter a code which is valid, but inappropriate for your specimen it can be "subtracted" by returning to the data-entry option (Function 1) and typing, for example, **-M2 Return** . Character M2 will then be removed from the description of your unknown specimen.

If characters are selected from the list on the right-hand side of the MS-DOS screen for a second time then they will be subtracted from the list of characters entered.

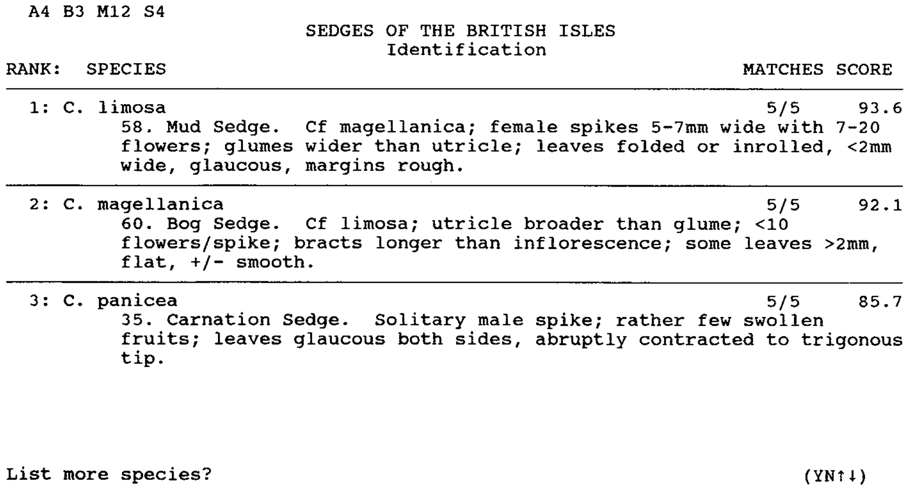
Function 2 - List of Most Similar Species

Function 2 displays a list of species which are similar to the unknown specimen with the most similar at the top (Fig. 3). The sorting process may take several seconds (up to thirty seconds for very large data sets on the BBC computer) but is only necessary once after each new set of character information has been entered with Function 1.

The list shows information for three or four species at a time, ranked in order of similarity to the unknown specimen. The display includes the rank number of each species, its name, the number of the characters typed in so far which match that species, and the 'score' which is an index of similarity (0 - no characters in common, 100 - exactly similar) to the unknown specimen based on weighted scores. The sorting and weighting systems and the ways in which they can be changed are described in more detail under Function 8. The species names are colour coded (or highlighted on monochrome monitors) to indicate remaining species which may be considered as being similar to the unknown specimen. The numbers of matches and scores are also highlighted where they are similar to those of the top ranking species. The criteria, used to determine which should be highlighted, are necessarily arbitrary and are intended as a rough guide only.

Each species name is also followed by short comments on diagnostic features, the names of similar species, or other useful information.

The question 'Want to list more species?' appears at the bottom of the screen (if there is insufficient space on the screen then the question will appear when you press any key). Typing **N** (or pressing the **Escape** key) will return you to the main menu. Responding by hitting Y or the down-arrow cursor key will display information about the next three or four species in rank order, while the up-arrow cursor key will move back up through the list re-displaying the previous screen of information.



```
  A4 B3 M12 S4
                         SEDGES OF THE BRITISH ISLES
                                Identification
RANK:  SPECIES                                            MATCHES SCORE
------------------------------------------------------------------------
  1: C. limosa                                              5/5     93.6
          58. Mud Sedge.  Cf magellanica; female spikes 5-7mm wide with 7-20
          flowers; glumes wider than utricle; leaves folded or inrolled, <2mm
          wide, glaucous, margins rough.
------------------------------------------------------------------------
  2: C. magellanica                                         5/5     92.1
          60. Bog Sedge.  Cf limosa; utricle broader than glume; <10
          flowers/spike; bracts longer than inflorescence; some leaves >2mm,
          flat, +/- smooth.
------------------------------------------------------------------------
  3: C. panicea                                             5/5     85.7
          35. Carnation Sedge.  Solitary male spike; rather few swollen
          fruits; leaves glaucous both sides, abruptly contracted to trigonous
          tip.



List more species?                                            (YN↑↓)
```

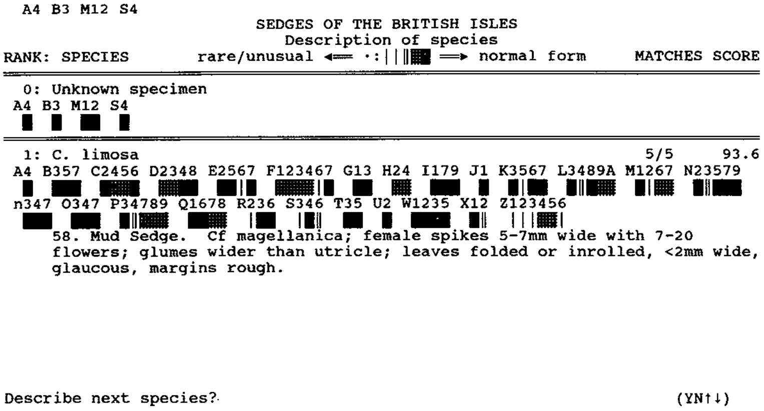
FIG. 3 – List of most similar species - Function Key 2
A list of the names of the species most similar to the unknown, together with the number of matches, weighted scores and comments for each species

Function 3 - Display Descriptions of Similiar Species

This option lists species, one at a time, in the same order as Functicn 2, but gives a detailed list of all the characters which may be displayed by each species. The top half of the screen display ***(or first screen display on the BBC)*** (Fig. 4) shows the characters typed in for your unknown specimen (note that the character codes are condensed so that M1 and M2 appear as M12). Each character has a square symbol on the line beneath it. This implies that the character has been scored 'reliably' (- different symbols will only appear if 'input weights' have been used, but these are not discussed until later).

For the lower half of the screen display ***(or second screen display on the BBC obtained by pressing Y in response to the question at the bottom of the screen)*** you are shown the rank, name, number of matches and weighted score for the top ranking species as indicated for Function 2. This is followed by a complete list of the characters for that species, each marked with a symbol (in green or dark grey) on the line below. The symbols indicate the constancy or reliability with which that character may be scored for that species. *(On MS-DOS versions the symbols are highlighted except for codes which have already been entered as characters for the unknown.)* As indicated at the top of the screen, the heavier the symbol the more reliable the character will be in describing the species. Thus the heaviest symbol (a solid rectangle) marks constant characters; the next two symbols *(stippled rectangles on MS-DOS)* ***(hash and double bar on the BBC)*** mark characters which are variable or intermediate, but are not uncommonly found in that species; the vertical bar symbols represent somewhat unusual features in that species; and the lightest symbols (colon and period) mark characters which may be regarded as only being present in very rare forms. Characters which are not likely to be scored for that species do not appear in the list.

The character list is again followed by comments pertinent to that species. Typing **Y** or the down-arrow cursor key will display a full description of the next species in rank order of similarity to the unknown; **N** or **Escape** will return you to the menu.



```
 A4 B3 M12 S4
                          SEDGES OF THE BRITISH ISLES
                             Description of species
RANK: SPECIES        rare/unusual ◄══ ·:||‖▓ ══► normal form      MATCHES SCORE
═══════════════════════════════════════════════════════════════════════════════
 0: Unknown specimen
A4 B3 M12 S4
█  █  ██  █
═══════════════════════════════════════════════════════════════════════════════
 1: C. limosa                                                           5/5     93.6
A4 B357 C2456 D2348 E2567 F123467 G13 H24 I179 J1 K3567 L3489A M1267 N23579
█  ███  ▓▓▓▓  ▓▓██  ██|█  ▓▓▓▓|█  ██  ▓▓  ███  █  █|██  █‖▓▓▓  █|▓▓  █‖███
n347 O347 P34789 Q1678 R236 S346 T35 U2 W1235 X12 Z123456
███  ███  █‖▓▓▓  ██▓▓  |██  |█‖  ██  █  ████  █‖  ||‖▓▓|
     58. Mud Sedge.  Cf magellanica; female spikes 5-7mm wide with 7-20
     flowers; glumes wider than utricle; leaves folded or inrolled, <2mm wide,
     glaucous, margins rough.




Describe next species?                                        (YN↑↓)
```

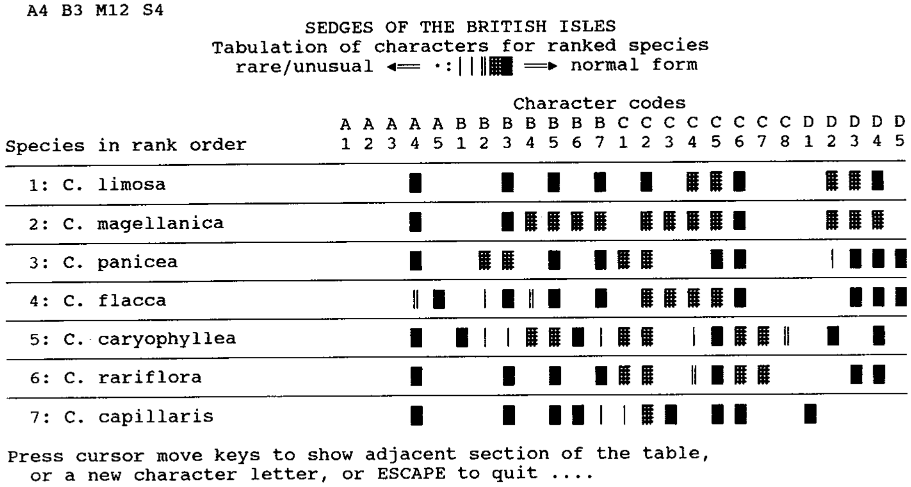
FIG. 4 – Description of the top-ranking species - Function Key 3
A display showing all of the possible character states for the top-ranking species. Heavy symbols indicate characters which are reliably present in that species. Other symbols indicate rare or unusual forms. The BBC version shows the characters for the unknown and the top ranking species on separate screens obtained by pressing the cursor-arrow keys.

Function 4 - Tabulate Characters

The data stored on disc can be summarised in the form of a table, each row representing a species and each column a character state. A section of this table can be viewed with Function key 4 ***(followed on the BBC by the initial letter of the character codes to be viewed).*** The species will be arranged in order of similarity to the unknown specimen (Fig. 5). Columns which appear in a different colour (or shade of grey) represent character states which have already been entered as part of the description of the unknown specimen. The symbols indicate the constancy, or reliability, of the character as described under Function 3.

Only a small portion of the table is displayed at any one time but the adjacent sections can be viewed by pressing the arrow cursor movement keys (left, right, up and down arrows). Alternatively, a particular set of characters can be viewed by pressing the initial letter of the character code (the computer will beep if an invalid code is pressed).

This table is a 'lateral key' which enables the comparison of several species to be made at a glance. However, unlike printed tabular keys, comparisons are made much easier through arrangement of the species of particular interest at the head of the table.



A4 B3 M12 S4

SEDGES OF THE BRITISH ISLES
Tabulation of characters for ranked species
rare/unusual ◄══ ·:||‖▦█ ══► normal form

	Character codes																								
	A	A	A	A	A	B	B	B	B	B	B	B	C	C	C	C	C	C	C	C	D	D	D	D	D
Species in rank order	1	2	3	4	5	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	1	2	3	4	5
1: C. limosa				█				█		█		█		█		▦	▦	█				▦	▦	█	
2: C. magellanica				█				█	▦	▦	▦	▦		▦	▦	▦	▦	█				▦	▦	▦	
3: C. panicea				█			▦	▦		█		█	▦	▦			█	█				|	█	█	█
4: C. flacca				‖	█		|	█	‖	█		█		▦	▦	▦	▦	█					█	█	█
5: C. caryophyllea				█		█	|	|	▦	▦	█	|	▦	▦		|	█	▦	▦	‖		█		█	
6: C. rariflora				█				█		█		█	▦	▦		‖	█	▦	▦				█	█	
7: C. capillaris				█				█		█	█	|	|	▦	█		█	█			█				

Press cursor move keys to show adjacent section of the table,
or a new character letter, or ESCAPE to quit ....

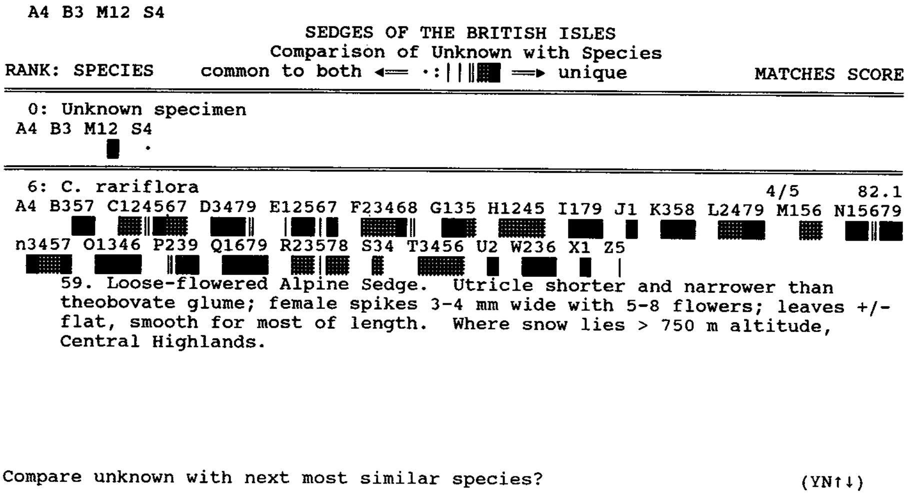
FIG. 5 - A table of characters - Function Key 4
A table of characters for the species most similar to the unknown. Heavy symbols indicate reliable characters. Adjacent sections of the table can be displayed using the cursor arrow keys, or by pressing the initial letter of the character code.

Function 5 - Compare Unknown with Top Species

Though using a screen display superficially similar to Function 3, the characters of both the unknown specimen and a named species are displayed on the screen simultaneously (Fig. 6). The meaning of the symbols (in yellow or near white) however, is quite different as indicated at the top of the screen display. The symbols now indicate the differences between the two descriptions. Solid rectangles beneath characters for the unknown specimen mark recorded characters which never occur in the named species; other symbols represent characters which, though sometimes present in the named species, are not constant characters. The heavier the symbol the less likely it is that the character would have been scored for the named species - thus a period marks a character which is common, but variable in the named species, but the second heaviest symbol marks a character recorded for your specimen which only rarely occurs in the named specimen.

Similarly, the squares beneath the codes for the named species represent reliable characters which have not yet been recorded for the unknown specimen; other symbols for characters not yet recorded have the same meaning as in Function 3. Characters with no symbols *(colour coded on the MS-DOS version)* may be present in that species, but have already been scored for the unknown and therefore represent matches.

Stepping through the species list by responding with **Y** or the down-arrow cursor key, to the question at the bottom of the screen will show an increasing number of differences between your specimen and the named species. If the codes you have entered do not give an exact match to even the top ranking species, this option allows you to find out where the differences lie, and hence which characters to check on the original specimen.



```
  A4 B3 M12 S4
                              SEDGES OF THE BRITISH ISLES
                           Comparison of Unknown with Species
RANK: SPECIES     common to both ◄══ ·:||‖▓█ ══► unique             MATCHES SCORE
══════════════════════════════════════════════════════════════════════════════
  0: Unknown specimen
 A4 B3 M12 S4
       █  ·
══════════════════════════════════════════════════════════════════════════════
  6: C. rariflora                                                    4/5    82.1
 A4 B357 C124567 D3479 E12567 F23468 G135 H1245 I179 J1 K358 L2479 M156 N15679
 n3457 O1346 P239 Q1679 R23578 S34 T3456 U2 W236 X1 Z5
    59. Loose-flowered Alpine Sedge.  Utricle shorter and narrower than
    theobovate glume; female spikes 3-4 mm wide with 5-8 flowers; leaves +/-
    flat, smooth for most of length.  Where snow lies > 750 m altitude,
    Central Highlands.


Compare unknown with next most similar species?                     (YN↑↓)
```

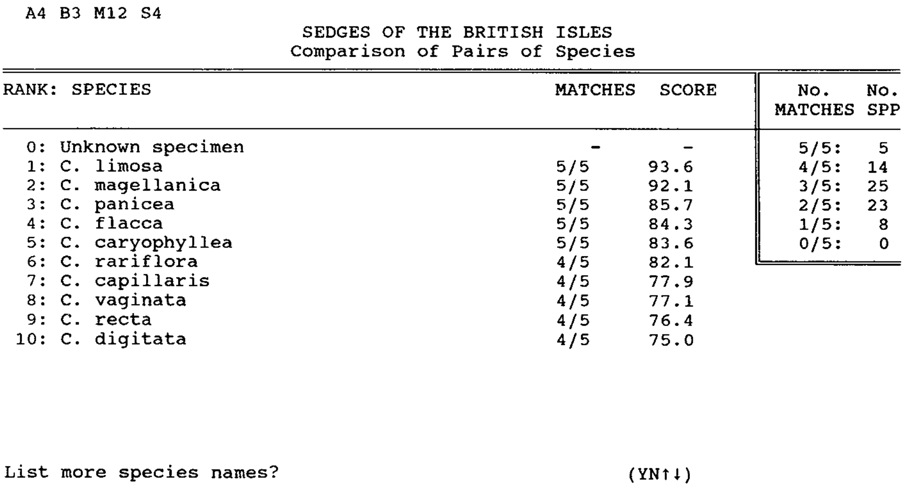
FIG. 6 - Compare unknown with top species - Function Key 5
The vertical bar beneath M2 for the unknown species indicates that this is a somewhat unusual character in *Carex rariflora.*

Function 6 - Compare Given Pair of Species

This is an extension of Function 5 which allows you to compare any two named species and highlight the differences between them. The first screen display (Fig. 7) is a summary table of the top ranking species giving rank, name, number of matches and weighted score for each; these are colour coded or highlighted as described under Function 2. Select the species you wish to compare and note the rank numbers. If the species of interest do not appear, then type **Y** or a cursor key to step through the species list.

After typing **N** you will be invited to type in the rank numbers of the two species to be compared (pressing **Return** each time). A display of yellow (or near white) symbols somewhat similar to that given for Function 5 will then be presented for each of the two species (Fig. 8). As in the previous option the solid rectangle indicates a character which is normally present in that species but is never present in the other. This symbol indicates the main features which reliably distinguish between the two species. Other symbols are of less use for discriminating between them, but the heavier symbols mark characters which are more likely to occur in one species than the other. The hash symbol, or stippled rectangle, for instance, could either mean that the one species always displays the character but the other species may also do so on rare occasions, or it may indicate a variable character which never occurs in the other species. If the former is the case, the character code will be listed for the other species without any symbol, but if the latter is the case the character will not be listed for the second species. Unsigned character codes indicate a character which may occur in that species, but with equal or less constancy than in the other species.

Because of limitations on space, the comments sections for the two species are not listed on this screen display on the BBC version. They may be viewed, however, by pressing Y in response to the question at the bottom of the screen. Typing **N** will return you to the menu, or, of course, function key 6 will re-display the species list for you to choose a second pair of species. The up- and down-arrow cursor keys in this instance will display a comparison of the first named species at the top of the screen with the species preceding or succeeding in rank order that in the bottom half of the screen.



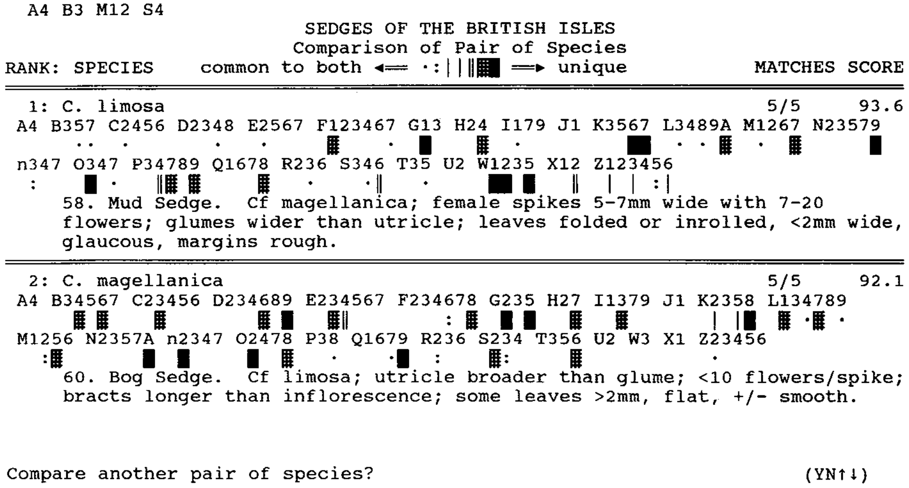
```
 A4 B3 M12 S4
                     SEDGES OF THE BRITISH ISLES
                    Comparison of Pairs of Species
==========================================================================
RANK: SPECIES                                    MATCHES   SCORE  ║  No.    No.
                                                                  ║ MATCHES SPP
------------------------------------------------------------------║-----------
  0: Unknown specimen                               -        -    ║  5/5:   5
  1: C. limosa                                     5/5     93.6   ║  4/5:  14
  2: C. magellanica                                5/5     92.1   ║  3/5:  25
  3: C. panicea                                    5/5     85.7   ║  2/5:  23
  4: C. flacca                                     5/5     84.3   ║  1/5:   8
  5: C. caryophyllea                               5/5     83.6   ║  0/5:   0
  6: C. rariflora                                  4/5     82.1   ╚═══════════
  7: C. capillaris                                 4/5     77.9
  8: C. vaginata                                   4/5     77.1
  9: C. recta                                      4/5     76.4
 10: C. digitata                                   4/5     75.0



List more species names?                             (YN↑↓)
```

FIG. 7 - Summary table - Function Key 6

The first screen display for Function 6 is a summary table showing the numbers of matches and scores for the 10 top-ranking species. Typing N enables you to select two species from the list for comparison.



```
 A4 B3 M12 S4
                       SEDGES OF THE BRITISH ISLES
                      Comparison of Pair of Species
RANK: SPECIES    common to both ◄══ ·:||‖▓ ══► unique            MATCHES SCORE
=============================================================================
 1: C. limosa                                                    5/5    93.6
 A4 B357 C2456 D2348 E2567 F123467 G13 H24 I179 J1 K3567 L3489A M1267 N23579
      ··  ·       ·   ·     ▓    ·  █    ▓  ·         ██  · ·  ▓  ·  ▓      █
 n347 O347 P34789 Q1678 R236 S346 T35 U2 W1235 X12 Z123456
  :    █ ·   ‖▓ ▓     ▓   ·    ·‖   ·     ██ █   ‖  | | :|
     58. Mud Sedge.  Cf magellanica; female spikes 5-7mm wide with 7-20
     flowers; glumes wider than utricle; leaves folded or inrolled, <2mm wide,
     glaucous, margins rough.
=============================================================================
 2: C. magellanica                                               5/5    92.1
 A4 B34567 C23456 D234689 E234567 F234678 G235 H27 I1379 J1 K2358 L134789
      ▓ ▓    ▓        ▓ █   ▓‖        : ▓  █ █   ▓   ▓        | |█  ▓ ·▓ ·
 M1256 N2357A n2347 O2478 P38 Q1679 R236 S234 T356 U2 W3 X1 Z23456
   :▓       █  █     █  ▓   ·    ·█  :    ▓:     ▓            ·
     60. Bog Sedge.  Cf limosa; utricle broader than glume; <10 flowers/spike;
     bracts longer than inflorescence; some leaves >2mm, flat, +/- smooth.


Compare another pair of species?                                   (YN↑↓)
```

FIG. 8 - Comparison of pair of species - Function Key 6

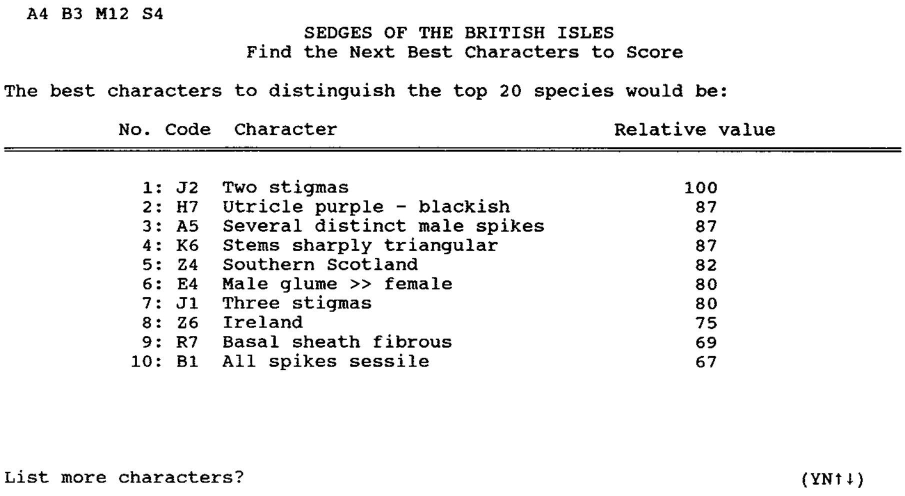
The symbols indicate the characters for the two selected species which best discriminate between them. Solid rectangles are characters reliably found in that species but not recorded in the other. Other symbols are characters more frequent in the one species than the other.

Function 7 -Suggest The Best Characters To Use Next

If characters are scored for the unknown specimen in a random fashion, the identification will not be very ‘efficient’; a proportion of the characters chosen may contribute very little information towards distinguishing the most similar species. Function 7 scans the data for up to ten of the top ranking species *(twenty on the MS-DOS version)* and compares the value of each remaining character for distinguishing between them.

Characters are listed on the screen about ten at a time (Fig. 9). The number on the right-hand side gives an indication of the discriminating value of that character as a percentage of the most useful character. Details of how many species are to be scanned and how the discriminating value of the characters are calculated are given later. It is best to select from this list those characters which are easy to score for the unknown specimen without error, and then return to the data entry phase (Function 1) to add in these characters.

Pressing **Y** or the down-arrow cursor key in response to the question at the bottom of the screen will list the next ten most useful characters and so on until there remain no characters which provide reliable distinction between any of the top ranking species.



```
  A4 B3 M12 S4
                         SEDGES OF THE BRITISH ISLES
                      Find the Next Best Characters to Score

The best characters to distinguish the top 20 species would be:

          No. Code  Character                        Relative value
===============================================================================

           1: J2   Two stigmas                             100
           2: H7   Utricle purple - blackish                87
           3: A5   Several distinct male spikes             87
           4: K6   Stems sharply triangular                 87
           5: Z4   Southern Scotland                        82
           6: E4   Male glume >> female                     80
           7: J1   Three stigmas                            80
           8: Z6   Ireland                                  75
           9: R7   Basal sheath fibrous                     69
          10: B1   All spikes sessile                       67


List more characters?                                               (YN↑↓)
```

FIG. 9 - Next Best Characters - Function Key 7
The list shows the characters which, if scored next, would give the most information for discriminating between the top-ranking species.

Function 8 - Reset Control Options

There are three control settings which can be changed with this function. These concern the method of ranking the species and the use of the weighting systems in the calculation of the weighted similarity scores. When starting on a new specimen these will be given the default settings, and it will not normally be necessary for these to be changed. However, to provide the user with extra control, it is possible to give ranking priority to the weighted scores, and to modify the way in which the weighted scores are calculated. The relative merits of the different settings are discussed later but the default settings are recommended for most purposes.

Function 9 - Start on New Unknown or Quit Program

When you have completed the identification of a specimen and want to start afresh with a new set of characters, select Function 9. This will delete all the character codes entered so far and reset the default ranking and weighting systems. To avoid accidental loss of the data it is necessary to confirm this option by typing **Y** in response to the question which appears at the bottom of the menu. If you press the **Escape** key then no action is taken.

Responding N to the first question but Y to the second enables you to return to the very beginning of the program and select a different data base for identification of a different group of organisms. Answering N to the second question will quit the program, clear all the function keys, close the data files and return control to BBC Basic or MS-DOS.

Function 10 (MS-DOS) or Function 0 (BBC)- Translate Character Code

Function key 0 on the BBC or *Function 10 on MS-DOS* replaces the top line of the screen display with a glossary which provides a translation of any valid character code. You are first invited to enter the two characters of the code required; pressing **Return** will then display the name of that character.

The number at the right-hand margin of the screen is the character weight as described above under Function 1. Pressing any key will return you to whichever part of the program you left when selecting this option. This means that the glossary may be accessed by the function key at any time. If the glossary is accessed during the data entry phase (see above) then the screen is cleared on leaving the glossary and the character codes entered so far are listed on the top line of the screen.

Break Key - Start Again on New Data File

The Break key on the BBC computer will also interrupt the program. The effect is the same as answering Y to the second question in Function 9 and the program will re-start from the beginning. However, there is no pause for confirmation of the option, so all current data will be lost. It is possible to disable the Break key on some Models of the BBC. It is better to leave the program using Function 9.

Procedure for Identification

The normal procedure for identifying a specimen will be to enter five or six characters, choosing those characters which are easy to score without error. The most efficient identification will derive from characters which are distinctive or unusual for that group of organisms. The illustrated guide will probably indicate a suitable set of characters with which to start. If this results in several species with high similarity scores (Function 2) then select Function 7 to suggest which characters should be examined next and add in some of these with Function 1. If the top ranking species do not match all of the characters you have entered, then select Function 5 to identify where the differences occur; these characters should then be checked carefully on your specimen.

When the list of species is reduced to two or three possible contenders, Function 6 can be used to find the specific differences between pairs of species. Alternatively, if you have reason to believe that the specimen belongs to some other species, the characters of that other species can be compared with the unknown or the top ranking species. Continue adding information until the top ranking species has a similarity score much higher than the second ranking species and agrees with your specimen on several more matches. Knowing when to stop adding more characters and to accept the identification requires a subjective decision. The top-ranking species from this key should not be considered the correct identification 'beyond reasonable doubt' unless it is identical to your specimen on all characters, and the next nearest species differs on several characters; the weighted similarity score should then be high (around 90%).

The specimen should always be checked against an illustration or full description in a conventional text. **This system should be considered as a guide which is complementary to conventional texts, and not a substitute for printed material.** If your specimen differs from the top ranking species, or the second species is very similar, then it would be wise to check two or three of the top species against illustrations. Remember that your specimen may belong to a species or taxon (*e.g.* a hybrid) which is not included in the key!

ADVANTAGES OF MULTI - ACCESS IDENTIFICATION SYSTEMS

Random-access (or multi-access) keys are a powerful tool for the identification of biological material. The simplest form of random-access key is a table showing the presence or absence of a range of characters in every species of a group of organisms. An unknown specimen is scored for a number of characters and compared with each of the species in the table. The specimen is most likely to be a member of the species which has the most characters in common. This type of key, though valuable for distinguishing between members of a small group of similar species, soon becomes unwieldy to use on paper when more than twenty or thirty species or characters are to be considered. Microcomputers, however, enable large tables of data to be searched with an ease and speed which greatly enhances the power of random-access keys. The advantages and disadvantages of different types of identification key are discussed by Tilling (1984), and random-access keys are described by Pankhurst (1978).

This program is designed to be simple to use, and yet provide the user with many facilities which are not readily available from conventional dichotomous keys. These include:

1. Free choice as to which, and how many, characters are used, and in what order. Damaged or incomplete specimens can therefore be identified, and characters which are difficult to score can be ignored.

LABEL CARDS FOR FUNCTION KEYS

F1 Enter data	**F2** List species
F3 Describe species	**F4** Tabulate characters
F5 Compare unknown	**F6** Compare species
F7 Best characters	**F8** Change options
F9 Restart or quit	**F10** Glossary

F1 Enter data	**F2** List species	**F3** Describe species	**F4** Tabulate characters

F5 Compare unknown	**F6** Compare species	**F7** Best character	**F8** Change options

F9 Restart or quit	**F10** Glossary

f0 Glossary	**f1** Enter data	**f2** List species	**f3** Describe species	**f4** Tabulate characters	**f5** Compare unknown	**f6** Compare species	**f7** Best character	**f8** Change options	**f9** Restart or quit	**BREAK** Restart program

FIG. 10 Label cards for Function keys

2. Species names are listed in order of decreasing similarity to the unknown specimen so that a short-list of likely species can be obtained. This means that an untypical specimen (or even where mistakes are made) will still give a high rank to the correct species, and a hybrid will probably give high rank to both its parents. This feature makes the identification 'robust' in that it is not dependent on a perfect specimen.
3. Additional comments about diagnostic features, or about similar species are given for each of the species listed.
4 The user has a choice of system for ranking species: either by the number of characters which each species may have in common with the unknown, or by a flexible weighting system which indicates the 'most likely' species.
5. A full coded description of each of the most likely species can be displayed on the screen. A marker is placed against each character showing how constant it is, or how reliably it may be scored in that species.
6. Direct comparisons may be made between the characters of any pair of species, or between the characters typed in for the unknown species and any named species. Each character is then marked to show its value for distinguishing that pair of species.
7. The user is guided as to the most useful characters which could be scored next in order to distinguish between those species most similar to the unknown.
8. The characters are described and defined in a separate, illustrated document where technical jargon can be avoided or explained by simple diagrams.
9. The character information is typed into the computer in a simple coded form. This is quick, and a lot of information can be displayed on the screen at one time. An on-screen glossary is available to translate the codes if necessary.

DEFINITION OF TERMS

The data base on which the key program operates is considered to be a table in which each row represents a taxon (or species) and each column represents a character state (such as "Leaves hairy above"). Each character state is given a particular code (e.g. N8) with related character states grouped together for convenience with the same character letter (N) (*e.g.* "Leaves not hairy above" is N7). However, the key program does not recognise such groupings, so the two character states N7 and N8 are treated quite independently (unlike some other computer keys where the character, which can exist in only one of a limited set of states, is clearly defined). These character states will henceforth be referred to simply as 'characters'.

Identification is by comparing a list of characters which describe the unknown specimen with each row of the table in turn. The computer calculates a similarity score for each species in the data base by matching the characters entered with those recorded in the character table. Species are ranked according to this similarity score.

There are three different weighting systems available to the user which determine how the similarity score is calculated; two of these (character weighting and species weighting) are built into the data base; they are the default options and will be in operation at the start of every new identification run unless disabled by Function 8. The third (input weighting) can be applied by the user during data entry. The three different systems can be used in any combination. The similarity score for each species is the sum of the combined weights for every character in common with the unknown specimen, and is expressed on the screen as a percentage of the maximum score possible for a species which matches all characters perfectly. Calculations of the actual species scores are illustrated with a worked example of data taken from the "*Random-access Guide to British Sedges using a Microcomputer*" (Legg, 1992) which is distributed with the program.

Character Weights

Each character state (column) is assigned a character weight on an arbitrary scale between 1 and 7. A character with weight 4 contributes as much to the similarity scores as two characters with a weight of 2. The weights are arbitrarily defined by the author of the database and represent the assumed 'taxonomic value' of that character. The scale range from 0 to 7 was chosen for convenience when compressing the information into byte form in the data files. High values will be given to reliable characters which distinguish the major taxonomic divisions within the group while low weights will be assigned to characters which are unreliable or difficult to ascertain with certainty.

If the character weighting is turned off, using Function 8, during the identification then all characters entered subsequently will be given a character weight of 7. The letter c will appear in the top left-hand corner of the screen as a reminder. Note that the default character weighting is resumed at the start of every new unknown specimen (*i.e.* after Function 9).

If an incorrect character code is entered then the same weighting system must be used when it is "subtracted" or removed. It is expected that most people wishing to change the weighting system will do so at the beginning of each new specimen and retain the same system throughout.

Species Weight

The second weighting system applies an individual weight to every species/character combination (*i.e.* every cell in the data table) and expresses the constancy of the character in that species. The weights range from 7, (constant and reliable character for that species), through 5 or 6 (variable but commonly expressed characters), 3 or 4 (somewhat unusual characters), to 1 or 2 (rarely expressed). Characters which are never found in that species effectively have a species weight of zero.

The values of the species weights appear both in the descriptions of named species and the table of character data (Function 3 and Function 4) as the square symbol (weight 7) through to a period (weight 1). However, the symbols appearing in the comparisons of species (Function 6) represent the numerical difference between the weights for the two species where that is positive (*i.e.* the symbols only appear against the species with the higher species weight for each character). For the purposes of comparison of the unknown and named species (Function 5), all characters for the unknown are considered to have "species weights" of 7.

Species weighting may again be turned off or on with Function 8 affecting all subsequent characters entered; the letter s will then appear in the top left-hand corner of the screen. If erroneous characters are entered then the same weighting system must be in use when they are removed.

If both character weighting and species weighting are turned off then the weighted score will be equal to the ratio of character matches to characters entered expressed as a percentage.

Input Weights

Input weighting enables the user to express his own confidence in the characters at the time of data entry by modifying the current character weights. By default, all characters have an input weight of 7, which is multiplied by the character and species weights. If you are uncertain of a character, however, you may wish to reduce its importance to below that indicated by the character weight. This can be done at data entry by appending an input weight on an integer scale from 1 to 7 separated from the character code by a minus sign. The number displayed at the right-hand end of the line after data entry is then the product of the character and input weights divided by seven. Thus if character A4 has a character weight of 5, but is entered with an input weight of 3 by typing **A4-3**, the display will show the combined weight of 5 x 3 / 7 = 2.1.

The input weights appear on the screen for the description of your specimen (Function 3) as the usual symbols representing 7 (square symbol) through to 1 (.). Note, however, that in the comparison of your specimen with named species (Function 5) each character for the unknown is considered accurately described and the display reverts to the square symbol (equivalent to a 'species weight' of 7).

If the character weighting is turned off to give all characters the default value of 7, then the input weighting can be used as a substitute to give complete control to the user in defining the relative importance of each character.

TABLE 1. *The calculation of weighted scores (with example taken from the "Guide to British Sedges" (Legg, 1992)). The character and species weights are embedded in the data file. The default input-weight of 7 is assumed.*

	Weights					
Code entered:	A4	B3	M1	M2	S4	
Char. weights:	5	3	4	3	4	
Species weights for *Carex flacca:*	5	7	4	7	7	
Default weighting Total Scores (CxSxI):	5x5x7	3x7x7	4x4x7	3x7x7	4x7x7	756
Maximum possible:	5x7x7	3x7x7	4x7x7	3x7x7	4x7x7	931
Similarity Score	756/931*100 = 84.3					

Calculation of the Weighted Scores

The number of matches for species j is calculated as follows:

$$Mj = \sum_{i=1}^{n} a_{ij}$$

Where: n = number of characters entered so far k
a_{ij} = 1 if character i may be expressed by species j
or = 0 if character i is never expressed by species j

The weighted scores are calculated as:

$$Wj = \sum_{i=1}^{n} \frac{(aij \times Ci \times Sij \times Ii)}{(Ci \times 7 \times Ii)}$$

Where: C_i = Character weight for character i
or = 7 if character weighting is turned off

S_{ij} = Species weight for character i, species j,
or = 7 for all positive characters if species weighting is turned off

I_i = Input weight for character i
or = 7 if no input weight specified

Examples of the calculations using various combinations of weighting systems are shown in Table 1.

Ranking Methods

When first used, the program will rank species according to the numbers of character matches with the unknown specimen. Species with a similar number of matches are then ranked within the list according to the weighted scores. Selecting the alternative system by typing Y in response to the first question in function 8 will cause species to be ranked primarily by the weighted similarity scores; species with identical scores will then be ranked by matches. The letter w will appear in the top left-hand corner of the screen while this ranking system is in use.

Next Best Character

The ranking of characters in Function 7 is based on both the character weights and the qualitative discriminating value of the characters. The number of species to be compared is determined as the number of species which are within two matches of the top-ranking species and have a similarity score within 80% of that of the top ranking species. The maximum number of species considered is 10 for the BBC or 20 for the MS-DOS version.

The computer scans each of the species to be considered and every remaining character which has not yet been entered with Function 1. A score is then calculated for each character as the product of the character weight, the number of species with species-scores greater than 5, and the number of species which never display that character (species score 0). These scores are then expressed as a percentage of the highest, and the characters are listed in ranked order. Characters with low character weights will appear low in the list, while those which divide the remaining species into two more-or-less equal groups with reliable characters (*i.e.* with species weights of 0 and 6/7) will be ranked high in the list.

If Function 7 is used before any data have been entered with Function 1 then the ranking of characters is based entirely on the character weights.

ADDITIONAL FEATURES OF THE KEY PROGRAM

The following additional facilities provided in the main KEY program are not essential for the running of the program and have not therefore been included in the Introductory section but may be of use to the more experienced user.

Changing Directories

The MS-DOS version assumes by default that both the KEY.EXE program and the .KEY data files are in the current directory. Data files available in this directory are then listed on the opening screen menu. The directory can be changed either within the program by using the X option provided, or by appending the new path name to the command line when loading the program. Thus typing **KEY \MONOCOTS** *from MS-DOS will run the KEY program from the current directory which will then display a menu of .KEY data files in directory \MONOCOTS. If MONOCOTS contains a key data file called CAREX.KEY then this can be opened direct from the command line with* **KEY \MONOCOTS\CAREX.** *This can also be used to attempt to run the program on a data file with a file extension other than .KEY, for example* **KEY B:CAREX.TMP** *would open a file CAREX.TMP on the B: drive as though it were a .KEY file. The complete file name can also be given at the instruction 'Enter new path name'.*

The ancillary programs TABLE.EXE and ENCODE.EXE described below use a similar procedure for locating the .CHA, .SPE and .PCC files though these filenames cannot be specified on the command line.

Printing the Screen Dislay

It is not normally necessary to print the information displayed as part of the identification process and a separate program (TABLE) is provided for printing the full information stored in the data base. It is possible, however, to send a copy of the screen display to a printer by typing **CTRL-P** ***on the BBC*** or *using the print-screen key* ***Shift-Print Screen*** *on the IBM keyboard when using the MS-DOS versions.* This will send an approximate copy of the screen to the default printer. **Note that if the printer is not connected or not in the 'on line' condition the program will hang.** If this happens you should either plug in and switch on a printer, or interrupt the program with ***the BREAK key on the BBC*** or *CTRL-ALT-DEL on MS-DOS. The MS-DOS version assumes that the printer has the full IBM character set; if not, the symbols and boxes on the screen will be printed as letters. Most dot-matrix and laser printers can be set to IBM mode using the appropriate combination of DIP-switches. It may be necessary to eject the paper manually after printing.*

Function 1. Species Elimination with Exclusive Characters

Several other computer-based identification systems available operate on an exclusion basis - all species which do not show the character entered are eliminated from the key. This can lead to very "efficient" identification in that very few characters may be needed to leave a single species remaining. On the other hand, these systems are very susceptible to errors either on the part of the user who incorrectly scores a character (or has an atypical specimen), or on the part of the author who writes the database. The primary objective in the design of this program was to provide a robust system resistant to such occasional errors. However, for characters which are known to be taxonomically reliable and are easily scored without error, it is useful to be able to eliminate those species which can never show that character.

If an exclamation mark is entered during data entry (function 1) in the place of a character code, then all species which have a score of zero (*i.e.* no matches to any of the characters entered so far) are effectively eliminated from the key until function 9 is used. Further, the scores of all remaining species are set to zero. This enables the user to start afresh on a reduced key including only those species which matched at least one of the previously entered codes.

For example, British members of the genus *Carex* may have either two or three stigmas and this is a reliable taxonomic character. Entering **J1** return (stigmas three), followed by ! return will eliminate the species which only ever have two stigmas from all subsequent listings. If you then wish to further restrict the key to species known to occur in England and Wales, the characters **Z1** , **Z2** and **Z3** could be entered in turn followed by the ! . This will eliminate those species which are found only in Scotland and Ireland. Note that by using the ! function twice the key has been restricted to those species with both three stigmas and a southern distribution. But that by entering Z1, Z2 and Z3 before the second ! we have included those species which occur in either Southern England (Z1) or Northern England (Z2) or Wales (Z3) or any combination. All species scores are re-set to zero so it is now possible to re-enter Z1, for instance; this will make use of the species scores for this character giving higher weights to species which are commoner in Southern England in subsequent listings.

The basis used for species elimination will appear on the top line of the screen, in this case as J1! Z123! followed by the list of subsequent characters.

Function 6. Absolute Numbers

The species for comparison are normally selected from the ranked list of species using their rank numbers. It is often useful to be able to compare, say, the top ranking species with a particular known species; however, the rank numbers change as more characters are entered into the computer. As an alternative, the species may be referenced by their 'absolute numbers' which are indexed in the illustrated document. These numbers are prefixed by the letter A and represent the order in which the data are stored on disc. If Function 6 is selected before any characters have been entered, then the rank numbers are the same as the absolute numbers. Species number zero selects the unknown specimen.

To obtain a simple description of a particular named species (rather than a comparison) select Function 6 and enter either the rank or absolute number of the chosen species as the 'first species', and then give -1 in place of the second species number.

CTRL - Function 7 (BBC version). Re-display Suggested Characters

Calculation of the best characters to use next involves accessing large amounts of data on the disc which may be rather slow on the BBC computer for large data sets (up to about 30 seconds). It is however useful to be able to switch rapidly from f1 (data entry) to the list of next most useful characters and back again. This can be done on the BBC by pressing CTRL-f7 (holding down the control key while pressing f7). This will re-display the last version of the list without calculating any changes due to additional characters entered since the list was last displayed. After entering three or four new characters, the next-best list should be recalculated with f7. *This facility is not necessary on the MS-DOS version where the speed of calculation makes the re-calculated list the preferred option.*

CTRL - Function 7 (MS-DOS version). Distinguish Top n Species

The computer uses an arbitrary formula for estimating the number of species which should be considered in function 7. Pressing CTRL-F7 enables you to fix the number of species which will be compared. The display will then show the characters which best discriminate the chosen number of species at the top of the ranking list. Note that this function may take some time if you choose more than about 20 species. This facility is not available on the BBC version.

PREPARATION OF NEW DATA FILES

Format of the Data Files

The key program is designed to operate on a 32 K BBC-B microcomputer and yet access a large database containing information on up to several hundred species and several hundred characters. The data are therefore compressed into a form which occupies as little disc space as possible and which gives rapid disc access to both rows and columns of the data matrix. It is therefore necessary to use ancillary programs to prepare the data files and also to read and print out the contents of the files in an intelligible form. It should not be necessary for the user to gain access to the main data file except through the ancillary programs provided, so no details are given here.

Because of the condensed and rigid format of the main data files, the creation of a new database is either from text files, or through the program EDITKEY described below for MS-DOS machines. EDITKEY is designed for use on relatively fast machines with a hard disc but is the best way of editing files already in existence. Owners of older IBM-compatible machines may prefer to create databases using the text files described here, but use EDITKEY for subsequent corrections and additions.

The text files can be created by any word processing software or text editor which can produce an unformatted ASCII text file such as ***WORDWISE or VIEW on the BBC,*** *EDIT on MS-DOS or NOTEPAD* in MS-WINDOWS. EDLIN could be used on MS-DOS with difficulty, though any word processing package would be an improvement. These files are read by ancillary programs which convert the text information into the form suitable for direct use by the KEY program. The success of any key, however, depends on the choice of characters and definition of the character states.

Choice of Characters

The characters should be chosen not so much for "efficiency" of identification, but for ease and reliability of scoring. Though the random-access approach is particularly robust to a few errors, it also provides much greater scope for selecting characters on the basis of ease-of-use, including many which would not be appropriate in a conventional, dichotomous, key. These may include characters such as distribution and habitat which, though not reliable as taxonomic characters, are easy to score and do provide useful information; the relative importance of the information can be incorporated into the weighting system.

Detailed discussion of the criteria used in the selection of characters and definition of character states can be found in Tilling (1984) and Pankhurst (1978).

Character Codes

Character states are grouped for convenience into 'letters' though any standard keyboard character could be used other than -, !, comma, vertical bar and spacebar. The £ $ and # keys are not standard to all keyboards and should not be used. Upper case and lower case letters are treated as different though it may be useful to use them for related characters. For instance, upper case N and lower case n are used for the upper and lower surfaces of leaves respectively in the Sedges key. There are 31 characters which can be used for the second part of the character code. These are numbers 1 to 9 and capital letters A to V. These must be used in the correct order without omission.

Format of the Raw Text Files for Data Entry

There are two raw data files:

- the character-description file containing the names of the character states and all basic information about the structure of the data set - the number of letters, the number of states per letter, the character weights, etc. *(In MS-DOS these files are given the filename extension .CHA.)*

- and the species-description file containing the actual data for each species. *(Filename extension .SPE.)*

The MS-DOS version has the capacity to operate on all species in the key simultaneously and only one species file is required. However, the BBC programs are limited in the amount of data they can hold. For the 32 K machines it is recommended that the species-descriptions should be compiled into separate files of about 20 - 30 species only (though more may be incorporated in the BBC-B+ and Master computers depending on the amount of memory available, the number of character states and the amount of comment used). Each of these is then compressed into a separate KEY file; several of these KEY files can finally be merged into a single main data-file using the MERGE program. The same character-description file can be used for all of the species-description files.

Contents of the Character-Description File

The data are in free format as a text file. The data items are separated by commas or new lines. Continuation lines are marked with a backslash (\) at the end of the preceding line. Blank lines and leading spaces are ignored. Comments may be included anywhere in the file using the angle brackets < and >. Commas or angle brackets within textual information can be overridden by enclosing the item in double quote marks. If the file is prepared with a word processing package, line justification should be turned off and care taken not to include any control information such as tab characters, underline or bold commands, ruler lines, conditional page-throws, etc. Most word processing packages include the facility to save the unformatted document as an ASCII text file without control characters. On the MS-DOS version the raw data files should be given the filename extension **.CHA** to distinguish them from other files.

The information is presented in the following sequence as illustrated by the data for the Sedge Key:

"Character data for British Sedges - version dated 10/8/91"
<A title for the character file enclosed in double quotes.>

ABCDEFGHIJKLMNnOPQRSTUWXZ
<The letters used to define groups of character states>

A5, B7, C8, D8 <A list of the character letters and number of>
E8..., Z6 <states used in that letter.>

A1, 7, "Single, unbranched spike"
A2, 5, "Several similar spikes"
A3, 7, ... <A list of the codes, character weights and names of the character states. The state names should be enclosed in double quotes if they include punctuation and should not be longer than 30 characters including spaces. Characters must be in the correct order.>
Z6, 6, "Ireland"

Contents of the Species-Description File

The species-description file ***(containing data for no more than 20 - 30 species for the BBC version)*** should be in the following sequence using free format as above with backslash for line continuation and double quotes around textual information.

"SEDGES OF THE BRITISH ISLES#19/08/91 Colin Legg, University of Edinburgh"
<The title of the Key. Information following the # sign is a subtitle which is only displayed when the data file is first accessed Enclose in double quotes.>

75 <Number of species in this file.>

"C. paniculata" <Species name. Enclose in double quotes if the name includes punctuation.>

A24|37, B27|36|46|57|66|76, C16|26|47|76|86
D ..., Z17|27|37|46|54|66

<The list of data for all of the character letters. Each item contains pairs of digits for each of the positive character states within that character letter. The first digit of each pair is the number (1 - 9) or capital letter (A - V) designating the character state within that character-letter, and the second is the species weight. Thus character A2 has a species weight of 4 and A3 has a weight of 7 (A1, A4 and A5 do not occur in this species). The vertical bar (|) is optional and is ignored by the computer program but may be useful to clarify the layout of state numbers and species weights.>

"1. Greater Tussock Sedge. \
Cf. appropinquata; utricle \
greenish to blackish-brown."

<Comment concerning that species. Note the use of backslash (\) as continuation markers at the ends of lines, and the double quotes. The comment must not exceed 255 characters including spaces>

"C. appropinquata"
A2437, ..., etc.

All the data files are checked for consistency by the programs which encode the data; interpretive error messages are given where possible. To facilitate error checking it is important that all of the data items are in the correct order.

Once the raw data files have been created they are checked for errors and translated by the program ENCODE. The running of this program should be self explanatory. Type ***CHAIN "ENCODE" for the BBC version*** or *just* **ENCODE** *for the MS-DOS* version and press return. The program will then prompt you for the names of the raw character and species data files, and for the name of the main key file which is to be created. The MS-DOS version gives a list of files in the current directory with appropriate filename extensions, or permits you to change the directory as explained above. ENCODE then scans the two data files for syntax errors and reports all that it is able to detect. If no errors are found then the main KEY file is created. Explanatory error messages can be listed to the screen or to a printer for later corrections. If errors are detected then the file must be corrected with the word processor before the main key file can be created. Note that this program may take up to an hour to process large data files on the BBC and on older MS-DOS machines using floppy discs.

Though the MS-DOS and BBC versions detect the same errors, the former runs rather faster, can take larger data files and gives slightly more information about the nature and location of any errors found. *The MS-DOS version displays information in three 'windows' on the screen. The first gives a direct copy of the text files as they are being read, together with the line number. The second window gives the 'interpretation' of the information which shows if the data have been misunderstood by the program, and the third gives a message when an error is detected. These error messages can be sent to a printer for later checking.* After detecting an error the program attempts to recover and continue with the rest of the file. Note that one error may generate several different error messages due to the mis-interpretation of subsequent information; missing quote marks or comment markers may cause particular problems.

It is important to ensure that there is sufficient spare space on the disc for the KEY file being created before running the ENCODE program. If using a single disc-drive machine, delete all unnecessary files on the disc before running the program. The program itself can be on a different disc which is replaced by the data disc as soon as the program is loaded. ***For the BBC version use the *COMPACT command to remove all 'dead space' on the disc.*** If using double disc-drives or double-sided discs, the data files can be in different drives as indicated in the full file names. For example ***the BBC filename ":1.D.CAREXC" indicates that the file CAREXC is in directory D on drive 1.*** *For the IBM version, the filename B:\KEYS\CAREX.KEY indicates that the file CAREX.KEY is to be created in directory KEYS on drive B:. The MS-DOS ENCODE program will run faster if some or all of the data files are on a hard disc or virtual disc, but this is not essential.*

Merging Data Files

Where several small data files have been created for the BBC, as part of a single database, these are then merged using the MERGE program. The use of this program should be self explanatory but you must again ensure that there is plenty of space available on the destination disc (especially where a single disc drive is being used). Again, the program disc may be replaced by the data disc after typing CHAIN "MERGE".

It should not be necessary to merge files on MS-DOS where the whole data set can be processed at once.

The Edit Key Program

This program for MS-DOS reads a .KEY data file and displays a portion of the data on the screen in the form of a table. The data can be edited on screen with ease. It is possible to insert new characters or new species, or to change the character codes and change the order of character states. The amended data can then be stored on disc either as a new .KEY file ready for use by the KEY program, or as text files which can be read by the ENCODE program, further edited by a word processing package, or transferred to the BBC computer.

Answer **Y** to the first question to edit a .KEY file already on disc, or **N** to start a new database from scratch. The number of new species and characters to be added should be the maximum number to be added during this editing session.

The .KEY files are compacted to minimise the space used on disc. It is not possible to insert new information direct to the .KEY file. EDITKEY therefore creates a temporary file which can be added to or edited. New .KEY, .CHA and .SPE text files are then created at the end of the session. The temporary file is quite large (perhaps 3 x the size of the .KEY file) and should be put onto a hard disc if possible. The default filename (\TEMPKEY.TMP) will be placed in the root directory of the current drive unless the path name is changed. If the temporary file is put onto a floppy disc it is important that this disc should not be changed while running the program. Ensure that there is sufficient space to save the final .KEY file either on the same disc, or on a second drive. It will take several minutes for the temporary file to be created on slow machines.

If starting a new database from scratch you will be presented with a screen headed by the function key labels. Start by pressing F1 to insert a title and subtitle to the database (see below), followed by F3 and F4 to build the rows (species) and columns (character states) of the data table. The species codes can then be entered by moving the cursor around the table so formed.

If editing a data file already in existence, you will be presented with a table showing the 'species weights' for the first five species and the first 25 characters. Move around the table using the cursor move keys (left, right, up, down arrow keys, page-up, page-down, home, end), or by pressing the initial letter of a particular character code, or by using function F2 to move to a particular species number. Change the species weight indicated by the cursor position by pressing the appropriate number (1-7), or 0 to indicate absence of the character in that species. The amended codes will appear in red (not highlighted) until the screen is re-drawn.

If using a fast machine then the character name, species name and comment will be displayed for the whole time. For slow machines, however, this information can be displayed by pressing the spacebar.

Other information about the database is amended using the function keys as follows (note that the Escape key can be used to abandon any of the operations up until the final response):

F1 - Edit Title

Permits changes to be made to the title and subtitle of the database. Note that the title must not be more than 30 characters long. The computer will beep if you exceed this limit. Full edit functions are available except that you must not press return at the ends of lines when editing the subtitle. There is no 'word-wrap' facility and text is continuous. Press Enter only when the full subtitle is complete.

F2 - Goto Species

Moves the cursor position to a particular species number.

F3 - Insert Species

Permits insertion of a new species either at (or immediately above) the cursor position, or at the end of the file. You will be prompted for the species name and the associated comment. Do not press Enter at the ends of lines in the comment section. Press Enter only when the comment is complete.

F4 - Insert Character

Permits the insertion of a new character state (column) at the cursor position or at the right-hand end of the table. You will be prompted for the character name (maximum 30 characters), the character weight (1-7), and the initial letter of the character code. The initial letter must be either the same as that of the adjacent columns, or it must be an unique letter not used elsewhere in the database. The second part of the character code (number 1-9 or letter A-V) will be assigned automatically. When inserting a new character state in the middle of a letter group note that all subsequent character codes will be adjusted accordingly.

F5 and F6 - Delete Species and Delete Character State

These function keys will delete the species and character at the cursor position respectively. Warning: It is very easy to delete data, but very tedious to have to replace it!

F7 - Edit Species Information

This displays the current species name and the comment section for correction.

F8 - Edit Character Information

Character name and character weights can be amended with F8. You are then asked if the column should be moved to a new position in the table. This is done simply by moving the cursor to the new position and pressing Enter. You will then be invited to enter the new initial letter of the character code which will be constrained as described under F4 above.

F10 - Quit

This leaves the data-editing phase and enables you to save the new information as either a .KEY file or as text files (.CHA and .SPE). It is then possible either to return to do more editing, or to delete the temporary file and quit the program.

Note that this program is dangerous! It is very easy to change data and delete whole species and characters. You are advised to take a backup copy of the data file, and make a hard copy using the TABLE program, before running EDITKEY. Make the changes a few at a time and store the data at intervals using F10. The last version saved can then be restored if serious mistakes are made.

Note that the disc containing the temporary file must not be removed from its drive until the program has finished running. The temporary file cannot be used again and should be deleted at the end of the program.

Printing the Data Table

The TABLE program can be used to list the full data table to a printer. Both BBC and MS-DOS versions send information to the system (default) printer port using standard ASCII characters. Both programs enable the user to determine the table size, column widths, etc. ***The BBC program includes and option for EPSON-FX compatible printers with capacity for 130 columns of condensed type; this gives maximum information across the page. The default setting can be used for most other printers but continuous-feed paper is required for both.***

The MS-DOS version includes the facility to set machine-dependent printer control codes and store these on disc for future use. Files of codes for several common machines are included on the disc with the program and are given the filename extension .PCC. The printer control codes are entered as decimal numbers and are separated by spaces. Any text information after the codes is ignored and can be used as comment. The control codes can be easily modified by referring to the printer manual. After setting the printer control codes and selecting the file to be printed the table will be sent to the default printer port.

Installing Data Files

The start-up menu, displayed at the very start of the KEY program, shows a list of the data files which have been installed on the current disc. The MS-DOS version recognises any file in the current directory with the filename extension .KEY as a potential key data file. ***The BBC computer obtains this list from a special file called "KEYDATA" which should be present on every data disc. The KEYDATA file is created on BBC data discs by running the program called "INSTALL" which is present on the program disc (type CHAIN "INSTALL" and press return). This program scans the disc for files which can be recognised as key data and makes them available for the KEY program. If a new data file is acquired, or one is deleted from a disc it will be necessary to re-run this program.***

Problems

The program should be free of 'bugs'. The most likely problems concern damage to discs and you are advised to take a back-up copy. The following problems may occur:

1. Program restarts automatically. ***If you accidentally press the BREAK key on the BBC computer, the program restarts automatically and the data are lost. It is possible on some machines to disable the break key to prevent this happening. You should always leave the program with Function 9.***

2. All goes dead. ***(Note that sorting species into order and Function 7 may up to 30 seconds for large data sets on the BBC.)*** If the disc drive is whirring, check that the door is closed, the disc properly inserted ***and the drive set on 40 or 80 tracks as appropriate.*** If you have accidentally given the print-screen command ***(CTRL-P on the BBC*** or **Shift-Print screen** *on the IBM keyboard)* then ensure that the printer is plugged in and on-line. If this is not possible then you must ***press BREAK on the BBC*** or **Alt-Ctrl-Del** *on the IBM keyboard and re-start the program.*

3. Unable to find data file. If the data file cannot be found on the current disc, this may be because the ***files on that disc have not been "installed" on the BBC or*** *that the data files are in a different directory on MS-DOS - see section on changing directories above.*

4. ***If this happens. The program does not start running when pressing Shift-Break on the BBC ensure that this command is enabled on the disc by typing *OPT4,3 and pressing return .***

5. Insufficient memory. ***This may be a problem with larger data files on the BBC B with only 32K of memory. If this occurs with larger memory machines, check that the full amount of memory is available to BBC Basic; this may require that you run additional software supplied with the machine.*** *The MS-DOS programs should run comfortably on 256K machines. If problems occur, check that there are no large memory-resident programs loaded by the AUTOEXEC.BAT file when the machine is booted; in particular, WINDOWS can be very expensive on memory. If problems arise, run the programs from DOS.*
6. Errors when printing may occur if the printer runs out of paper or is switched off-line during printing. Resolve the problem and run the program again.

If any other errors occur please make a note of what you were doing at the time and of any error messages which appear on the screen (with a screen dump if possible) and inform the author.

REFERENCES

LEGG, C. J. (1992). A Random-access Guide to Sedges of the British Isles using a Microcomputer. *Field Studies*, **8**, 31-57.

PANKHURST, R. J. (1978). *Biological Identification: the Principle and Practice of Identification Methods in Biology*. London, Edward Arnold.

TILLING, S. (1984). Keys to biological identification: their role and construction. *Journal of Biological Education*, **18**, 293-304.