DELTA Newsletter

Number 5 April 1990

Note from the Editor – The DELTA NEWSLETTER is designed to promote communication among scientists developing and applying computer technology in the collection, storage, analysis, and presentation of taxonomic data for the production of descriptions, keys, interactive identification, and information retrieval. To achieve this goal the DN will be issued in April and October of each year. Contributions in the form of short comments or long discussions and explanations are encouraged from all developers and users of DELTA format and similar systems. Comments on methods of application, suggestions for improvements, project descriptions, or criticisms of current technology are encouraged. – Robert D. Webster, USDA/ARS/SBML, Bldg. 265, BARC - East, Beltsville, MD 20705, USA.

ALICE : the taxonomic and nomenclatural core module

By Bob Allkin and Peter Winfield

ALICE is a microcomputer system for biodiversity databases or checklists. It gives users great flexibility while saving them from learning database languages, designing databases or spending time coding data. Menu options are described in familiar terminology and selected by single keystrokes.

ALICE is taxon-based. For each species, subspecies or variety entered, information can be stored about synonymy, geography, uses, common names and habitats using dictionaries of terms built up by the user. Users may also define as many characters as they like to describe species morphology, ecology, etc. Any fact can be linked to a bibliographic reference (e.g. who published this name or who said that plant X is poisonous).

In the last two DELTA newsletters we described the ALICE family of programs and ALICE's relation to DELTA. In this issue we give more detail about the most important ALICE module: the checklist and describe the taxonomic and nomenclatural facilities currently available in ALICE Version 2. Later releases will expand upon these features.

At the heart of each ALICE database lies the taxonomic and nomenclatural information about the organisms involved. The system incorporates taxonomic and nomenclatural rules which it uses to deal correctly with binomials, authorities, homonyms, misapplied names etc. and to shield nontaxonomists wherever possible, from unnecessary complexity. To ask about species "X", for example, it isn't necessary to know the authorities or even to use the currently accepted scientific name. The software redirects those using synonyms or misapplied names to the right information via the "accepted" name.

Most users remain blissfully unaware of the existence of homonyms, but when these occur, ALICE detects them and prompts the user to select the appropriate usage. The user can explore each alternative and then base a decision on opinions recorded in the database and the publications cited.

DATA ELEMENTS

i) Taxonomy

Information can be stored for species, subspecies or varieties. For simplicity we will use "species" here to mean taxa of any of these ranks. Each species is assigned to a genus (implicitly through its scientific name) and to a taxon of a higher taxonomic rank (e.g. family or tribe). At present only minimal information can be stored explicitly for taxa above the specific rank. As described below, however, users can ask questions about genera or higher taxa, reorder data in reports or select data subsets using them.

ii) Nomenclature

There are two principal categories of name

A) ACCEPTED NAME: - the name considered by the database author to be the proper scientific name of that taxon. There are two subcategories

i) Fully accepted names: where the author doubts neither the authenticity of the species nor the usage of this name.

ii) Provisionally accepted names: where the author has doubts about the validity of the species itself or the correct application of the name. Suspect "taxa" for which data is available, can thus be stored in the database, at least temporarily, labelled "provisional". They need not necessarily appear in reports published from the database.

B) SYNONYMS: - any alternative scientific name applied to this taxon. ALICE Version 2 recognises the following classes of synonym : -

i) Synonyms: - taxonomic or nomenclatural synonyms

ii) Doubtful synonyms: - probable or potential synonyms that might best be omitted from reports to be published.

iii) Orthographic variants: - spelling variations of a name

iv) Misapplied names: - names that have been incorrectly applied in the literature to this taxon. A bibliographic reference should be cited.

LOGICAL AND VALIDATION RULES

ALICE incorporates many logical rules used to check data input and prevent the introduction of errors during editing. WE cannot list all of these rules and cross checks here but the following describes some of them: -

i) A taxon can have one (and only one) accepted or provisional name.

ii) A taxon may have any number of synonyms.

iii) A name can occur in an ALICE database once only (other than misapplied names which may occur any number of times).

iv) A name and its usage for a given taxon can be linked to any number of citations in the bibliography.

v) Any authority, higher taxon, genus, specific epithet, subspecific epithet will be stored only once in the database and thus will be spelt identically throughout all reports.

EDITING DATA

ALICE provides a variety of editing facilities subject to the same logical and validation rules.

dictionary editor

A single edit of a dictionary can replace for e.g. all occurrences of "Linnaeus" by "L." throughout the database.

nomenclatural editor

The nomenclatural editor allows you, for example,

- to add or delete synonyms
- to change the 'status' of a name (e.g. provisional to accepted)
- to make a synonym into the accepted name (ALICE forces you, first of all, to remove or demote the current accepted name).
- to add or delete citations for a name

When users change the status of a name, ALICE automatically creates a text note in the database recording the previous status of that name, the bibliographic citation(s) which had been attached and the date and time. Thus users can keep track of changes in nomenclatural opinion.

QUERYING THE DATABASE

Using the AQUERY program the database can be searched using the Genus or Higher taxon e.g. "List native Indian species belonging to the genus Citrus". Searches, are designed to be convenient for infrequent users that know nothing of the database contents or structure, and involve simple selection from lists. Alternatively information about individual taxa can be obtained in the following ways:-

i) explore the database using the program NVIEW. Select names from lists at all taxonomic ranks. Select "Vicia" from a generic list, for example, and a list of all the species epithets published under Vicia appears. Select one species epithet (e.g. "faba") from among them and a list of subspecies within V. faba appears. ALICE resolves complications due to synonymy and will access all the information stored for that species.

ii) enter any binomial directly (with or without authorities). ALICE will find missing authorities and resolve the problems of homonyms and synonymy, redirecting users to the data they sought.

iii) select from a list of all vernacular names - if more than one species share a common name then AQUERY lists the alternative scientific names.

WRITING REPORTS AND EXPORTING DATA TO OTHER PROGRAMS

ALICE provides various predefined report formats including:-

- accepted names with their synonyms
- species descriptions
- fully synonymised checklists
- bibliographic lists
- dictionary printouts (eg authorities)
- database statistics

The AWRITE program allows users to define their own report formats. Individual reports may contain any combination of data elements for taxa (geography, morphology, uses, habitats etc.) as well as their accepted names and synonyms. In setting up each new report format the user decides

- which data items are to be included (e.g. synonyms)
- in what order the data items are to appear

- whether or not generic names are abbreviated to a single letter (ie. "V.faba L." rather than "Vicia faba L.")
- what typeface should be used for each class of name
- what indentation should be used for each class of name
- whether bibliographic citations are attached to names.
- whether or not data headings like "Synonyms" should appear.

All reports are output to ASCII files. Reports or data subsets to be exported to other programs (e.g. DELTA or dBASE) can be prepared either for individual species, Genera, Families etc. or for sets of taxa sharing particular attributes e.g. the Legume timber trees of Africa.

SUMMARY

We clearly cannot describe here the full functionality of the checklist module but hope that this account gives you a feel for how the taxa and nomenclature are managed within ALICE databases. In the next newsletter we will describe the geographical module.

POSTSCRIPT

Readers may be interested to know that all of the ALICE satellite programs such as AWRITE and AQUERY are now available in Spanish and Portuguese. For further information please contact

Bob Allkin: Royal Botanic Gardens KEW, Richmond, TW9 3AB UK. Tel. +44-1-940-1171 ext.4715 Telex 296694

Miscellaneous Notes

by M. J. Dallwitz

Character ranges in DELTA items

As mentioned by Richard Pankhurst in the previous Newsletter, CONFOR accepts items containing attributes with a range of characters, for example, 5–7,2. This is not documented in the description of DELTA format in Chapter 2 of the DELTA User's Guide. The characters in the range must be of the same type, except that ordered and unordered multistate characters may be combined. The TRANSLATE INTO DELTA FORMAT directive of CONFOR produces attributes in this form only if a DATA COMPRESSION directive is used (DELTA User's Guide, p. 43). This feature is mainly intended for reducing the size of the data for publication in printed form. It is not recommended for general use, as it tends to make the data awkward to edit.

Implicit values in CONFOR

The IMPLICIT VALUES directive allows certain attributes or state values to be omitted from items. The omitted attributes or values are assigned default values, which are automatically inserted in all CONFOR output except natural language and DELTA format. The directive can be used to save time and space when entering data, by making the most common state of a character implicit, and entering only the attributes corresponding to the other states. Careful use of this feature can improve, as well as shorten, natural-language descriptions. For example, the presence of bright red pigmentation in the roots or rhizomes distinguishes the Haemodoraceae from other angiosperm families. The corresponding negative state can be made implicit. However, it may be worthwhile to explicitly code the negative state for the widespread families with which the Haemodoraceae are likely to be confused: the Anthericaceae, Liliaceae, and Iridaceae. The absence of pigmentation will then be stated explicitly in natural-language descriptions of these families.

Implicit values are not implemented in most of the PANKEY programs (see page 7). You can use the PANKEY DEDIT program, or the CONFOR directive INSERT IMPLICIT VALUES, to insert them explicitly. Note, however, that this procedure will destroy any distinction you have made between uncoded and coded implicit values (as described above). In this case, you must treat the output file as a temporary copy, and continue to use the original as your master data file. If you want to insert the implicit values for only *some* characters, you can do it by temporarily deleting the other characters from the IMPLICIT VALUES directive.

The IMPLICIT VALUES directive can specify two different default states: one to be used when a whole *attribute* (e.g. 29,2) is omitted, and another to be used when only the *state value* is omitted. The second default value is useful if there are large numbers characters for recording simply the presence or absence of features. For example, suppose we have 50 characters of the form

```
#1. metabolite 1/
    1. present/
    2. absent/
```

and we set

```
*IMPLICIT VALUES 1–50,2
```

If only metabolites 3, 5, 6, 7, and 20 are present in Species 1, we can code it

```
# Species 1/ 3,1 5,1 6,1 7,1 20,1
```

However, if we set

```
*IMPLICIT VALUES 1–50,2:1
```

and also make use of character ranges as described above, the item can be coded

```
# Species 1/ 3 5–7 20
```

The natural-language descriptions could be improved by changing the wording of the characters to

#1. <metabolite 1>/
1. metabolite 1 <present>/
2. metabolite 1 absent/

Then, with an appropriate item subheading and linking, the coded description

Species 2/ 2 7 15,2 22

(where the 15,2 has been added for emphasis) would be rendered

Species 2
Metabolites present: Metabolite 2;
metabolite 7; metabolite 15 absent; metabolite 22.

OS/2 versions of CONFOR and KEY

OS/2 is an operating system which is available as an alternative to MS–DOS on many computers similar to the IBM AT, using 80286 or 80386 processors. Programs running under MS–DOS can directly access only 640kB of memory, even if more is present in the machine. OS/2 programs can directly access up to 16MB of memory, and a program can make use of almost the full 16MB even if that much is not physically present in the machine (this ability is called 'virtual memory'). However, OS/2 requires a minimum of about 4MB of physical memory to work efficiently.

OS/2 is a 'multitasking' operating system: that is, two or more programs can run simultaneously. This can be advantageous, as long program runs do not prevent the computer being used for other purposes.

Versions of CONFOR and KEY are now available to run under OS/2. They have about 50 times as much space available for data as the MS–DOS versions.

Data size limits

None of our DELTA programs has fixed limits on the numbers of items or characters. All storage is allocated from a single pool, whose size depends on the amount of memory (RAM) available in the computer. All of the programs require some storage proportional to the number of characters, and some proportional to the number of items (taxa). Some of the programs also require additional storage. KEY requires storage proportional to the number of items multiplied by the number of characters. CONFOR has a 'data buffer' which must be able to hold the longest item. The length of an item is increased by text material (text characters and comments), and decreased if some characters are not coded (for example, because they inapplicable to a particular item). The same storage is also used to hold individual characters from the character list, but this is rarely the limiting factor.

Leslie Watson's data for grass genera currently have 475 characters and 770 items. CONFOR and INTKEY can handle these data on an MS–DOS PC with 640k of memory. I am aware of only one set of data that is too big for CONFOR on MS–DOS. These data (on viruses) have 536 characters and 1508 items, and large amounts of text in some items. These data are converted to INTKEY format by CONFOR running under OS/2, and distributed to users in INTKEY format, for running under MS–DOS.

To further investigate the size limits, I made items files containing multiple copies of the items in the sample data (which are distributed with the programs). I found that under MS–DOS, CONFOR could handle 1750 items, but not 1800. Under OS/2, CONFOR ran successfully with 3500 items, and could certainly deal with many times this number. INTKEY was able to handle the 3500 items under MS–DOS.

We are currently working on a new version of INTKEY, and a new version of CONFOR is planned. Both will make much more efficient use of memory, and will be able to handle much larger data sets under MS–DOS. We hope to have the new INTKEY ready towards the end of the year.

The current version of CONFOR has two error messages which indicate that it does not have enough storage available. The first is

Not enough space in data buffer. The length is *n*.

This means that the item (or character) currently being read will not fit in the data buffer. The default length of the data buffer is 17 times the number of characters. This can be increased or decreased by means of the DATA BUFFER SIZE directive, which should be placed in the SPECS file immediately under the MAXIMUM NUMBER OF ITEMS directive.

The second message is

Not enough storage. *n* locations required, *m* available.

This means that the *total* amount of storage available is insufficient. The amount required must be reduced, or the amount available increased. A substantial reduction in storage requirements can be obtained by adding a SPECIAL STORAGE directive to the end of the SPECS file, so that the character list is stored on disk instead of in memory. Reducing the size of the data buffer (see above) will also help, but the scope for this is usually limited. However, any reduction is multiplied threefold, because there are three data buffers.

If these measures do not solve the problem, the next step is to maximize the amount of storage available. Under MS–DOS, our DELTA programs use all of the available 'conventional' memory (not extended or expanded memory). They cannot use memory being used by the operating system and other memory-resident programs such as device drivers and network software. The MS–DOS command CHKDSK will tell you the total amount of conventional memory, and how much is free. MS–DOS itself takes about 65kB (depending on the version), leaving about 575kB free in a 640kB system. If you have substan-

tially less than this, look at your CONFIG.SYS and AUTOEXEC.BAT files to see what other software is loaded, and remove any that you don't use regularly. Make backup copies of the original files, and, if necessary, obtain expert advice before changing them. Your system may become (temporarily) unusable if you delete important parts of these files. If there is still not enough memory, you can try making special versions of these files, containing only the software that is essential for the system to work. Rename your original files CONFIG.STD and AUTOEXEC.STD, and call the new ones CONFIG.DEL and AUTOEXEC.DEL. To set up the system for using DELTA, enter

COPY CONFIG.DEL *.SYS
COPY AUTOEXEC.DEL *.BAT

and reboot. Your standard configuration can be restored by a similar procedure using the .STD files. A simple batch file can make these operations easier and less error prone.

Data for grass genera available on TAXACOM

Leslie Watson's data on the grass genera of the world are now available in INTKEY format on the TAXACOM bulletin board, as Issue No. 22 of Flora Online. The data include synonyms, morphology, anatomy, physiology, cytology, classification, pathogens, world and local distributions, and references. The TAXACOM number is 1–716–896–7581. See DELTA Newsletter 3 for more information about TAXACOM, or contact Richard H. Zander, Buffalo Museum of Science, 1020 Humboldt Pkwy, Buffalo, NY 14211, U.S.A.

Computer hardware requirements

Previous versions of our DELTA programs ran on any MS–DOS computer equipped with: (1) at least 640kB of RAM; (2) two floppy disk drives of at least 360kB each, or a hard disk. However, recent changes to CONFOR have made it too large to fit on a 360kB disk: it now requires a hard disk, or dual floppies of at least 720kB. INTKEY will still run on dual 360kB floppies or a single 720kB floppy, so it is usable on most laptop machines. Future versions of the programs will require strict IBM compatibility, because they will have more elaborate screen displays. If you are buying a new computer, I would suggest the following.

1. Computer with 80286 or 80386 processor and at least 640kB of RAM, compatible with IBM AT. The processor should be as fast as you can afford – some operations can be quite time-consuming with large data sets. More RAM could be useful for some programs. Even if you do not get a large amount of RAM at the start, ensure that there are plenty of sockets for more on the motherboard, so that it can be added later without buying more cards.

80386 machines, as well as usually being faster than 80286 machines, are more efficient at 'multitasking' (the running of two or more programs simultaneously), and some multitasking systems (e.g. VM386) work only on 80386 processors. If plan to get a multitasking system, make sure that it will run on your computer.

If you think you may eventually need OS/2 (see above), make sure that the manufacturer of the computer is licensed to supply it (most small manufacturers are not), or that they guarantee that their machine will run IBM's version of OS/2. OS/2 has an 'DOS compatibility box' which can run some MS–DOS software, but it has limitations, particularly in the amount of memory available. Until you have OS/2 versions of most of your major software, you will probably want to run true MS–DOS most of the time, and boot OS/2 only when necessary. Thus, it is very convenient to be able to boot either system from the hard disk. This requires extra software with OS/2 versions 1.1 and above (e.g. MultiBoot from Bolt Systems, Inc., 4340 East-West Hwy, Bethesda, MD 20814, U.S.A. 1–301–656–7133). Version 2.0, which has not yet been released, is reported to have a much better implementation of the DOS compatibility box, but will require an 80386 processor.

2. Hard disk of at least 40MB (preferably 70MB or more), average access time 28ms or less, 1:1 interleave.

3. 1.2MB, 5.25" floppy drive, able to read and write 360kB disks. Make sure that the drive will write 360kB disks that can be reliably read on a standard 360kB drive. When testing the drives, use floppy disks that have already been written on a 360kB drive. This makes the test harder, but it is within the capabilities of most modern drives. It is also useful to have a 720kB/1.44MB, 3.5" floppy drive.

4. Two serial ports and one parallel port (for modem, serial mouse, and printer).

5. VGA card and colour monitor. A new version of INTKEY, currently under development, will be able to display illustrations. I think that a resolution of at least 640x480 is necessary for satisfactory display of many biological illustrations, and 256 colours are needed for colour or grey-scale pictures. A standard VGA card has display modes 640x480x16 (horizontal resolution x vertical resolution x number of colours) and 320x200x256. This is adequate for line drawings, but not for colour and grey-scale. Remember that, even if you don't intend to use illustrations in your own work, you may want to use illustrated data provided by others. It is best to get an 'enhanced' VGA card, capable of 640x480x256, 800x600x256, and 1024x768x16. These have 512kB of RAM. Make sure that you get a monitor capable of correctly displaying all the resolutions you are interested in. In particular, check that the combination of card and monitor maintains the size and proportions of the display, without manual adjustment, when switching to resolutions of 800x600 and higher. The card should be 'register compatible' with the IBM VGA. If you want to

use the higher resolutions with software such as Windows 286 or 386, a word processor, or a spreadsheet, make sure that the card is supplied with 'drivers' for this software, or that the software can be configured to use the required modes of your card. To configure a program for a card, it is usually necessary to know the register values which set the different modes, so check that this information is included in the documentation of the card.

6. Two or three free expansion slots. You may need these later for such additions as a CD drive or a scanner.

Mike Dallwitz, CSIRO Division of Entomology, G.P.O. Box 1700, Canberra, A.C.T. 2601, Australia. Telephone 61-6-246-4911. Facsimile 61-6-246-4000. Telex AA 62309.

Microcomputer-assisted telephone identification of houseplants using ONLIN6

by Deborah Metsger

The identification of unknown plant materials which have been eaten pose problems for staff of poison information centres the world over. Though well equipped to assess toxicity once the plant materials are identified, they are often unable to make identifications themselves. As a result they are forced to rely on the botanists and botanical resources available to them. Staff of the Botany Department of the Royal Ontario Museum regularly respond to telephone requests for plant identification that are referred to us by the regional poison control centres. To assist us in reponding to these requests, we have developed a preliminary database of 103 houseplant taxa in DELTA for use with ONLIN6 in answering poison plant calls.

Characters in the database are described in natural language so as to be understood by the lay person. At present the characters used describe only vegetative structures since most houseplant specimens usually lack reproductive structures. Moreover, untrained lay people often are unable to describe such structures clearly. As more taxa are added to the database vegetative characters alone may prove insufficeient and we will add reproductive features.

The current database comprises 103 houseplant taxa representing 42 families and 92 genera. 38 taxa represent whole genera or groups of genera, and 65 are individual species. 25 of these taxa are poisonous and 34 can cause dermatitis.

Data are recorded using dBase III Plus, and then converted to DELTA format. Though I have cursed the KCONP program while revising the database and trying to get ONLIN6 to run, once it is running, I find it very useful to know that all the taxa can be distinguished - especially in this application where unequivocal identification is vital. Several features of ONLIN6 have also proved very useful, since the operator almost always cannot see the plant being identified. The BEST feature, which orders the remaining characters according to their separation potentials, can guide the operator in asking questions about the plant. The DESCribe and DIAGnostic feature allow the operator to test their impressions or guesses against the specimen. Finally, the variability limit LIMI allows the operator to take account of unreliable information.

The database is currently being evaluated, and trials to date have indicated a greater than 75% success rate for identifications. We are beginning to compile graphics files to be used with the database. This feature will be vital if the system is to be used by non-botanists in hospital emergency centers. Future plans include expansion of the houseplant database and compilation of databases for outdoor cultivated and native plants.

Deborah Metsger, Department of Botany, Royal Ontario Museum, Toronto, Canada, M5S 2C6

MPI System Upgrades for 1990

by Laurence C. Hatch

The Micromputer Plant Identification (MPI) System developed by Taxonomic Computer Research has been expanded and upgraded for 1990. All MPI volumes will utilize the Version 1.2 Master Menu when processing single access or dichotomous style keys.

Version 1.2 of the Master Menu implements the new MPI L programming language, a dBASE-style query language. Over 32 commands and subcommands are supported. These include such functions as count taxa, count v (vars.), count f (f.), count ' (cvs.) color (change colors), char [taxon] (list characters), append note, print (several printer commands), trans (several textfile transcript commands), outline (list subkeys), author (display author information), back (several backtrack commands), define count (count number of dictionary terms), define auto (show list of automatically displayed definitions in keyfile), define [term] (lookup term), dir (directory), jump [taxon or term] (move to node with string), lit (display literature references cited in keyfile), test (debug key format), scroll, menu, quit, list (display all lines with term), undo, refer (reference all traces to given node #), load (new key), reload (reinitialize key after TSR edit), session (display key use statistics), and several others.

An on-line dictionary of several hundred phytographic terms is available during keying sessions. It may be accessed using the F7 keyword function MPI L define command, or displayed automatically when selecting a node. The AUTOMATIC TERM DISPLAY FUNCTION pops up a term definition whenever the user selects the appropriate node. Key authors may program a keyfile to anticipate the user's need to understand a new or variously

defined term. The phytographic dictionary in ASCII textile format is easily modified by key developers from within any wordprocessor. Terms and their definitions may be modified, deleted, or appended as desired.

Other new functions include optional feature such as password access, automatic instruction screen, automatic transcript file, automatic start- up key, and return to host key.

All volumes now include both BASICA source code and compiled versions. Orders for any MPI volume after December 1, 1989 may receive a free upgrade to Master Menu Version 1.2 for payment of $3.00 to cover postage and disk cost. Orders before December 1, 1989 may upgrade from Version 1.1 to 1.2 for $29.00 postpaid.

The basic strengths of the MPI have been strengthened by the revised Master Menu: 1) Key composition in wordprocessor of user's choice, 2) available library of over 125 different keys, 3) 100% disclosure in easy to modify BASICA code, 4) compact code and file size for operation in the field on portable PC with 360K drive, and 5) 6 alternative formats for displaying key data on screen.

A free copy of the MPI Brochure is available from Taxonomic Computer Research, PO Box 12011, Raleigh, NC 27605 USA.

Differences between PANKEY and CONFOR, revisited

by R. Pankhurst

Since this note was first published in DELTA Newsletter 3, Mike Dallwitz has pointed out to me some further differences. For the convenience of readers, I am putting them all together in one place. The new information is enclosed with ** .... **.

The PANKEY programs still use Version 2 of DELTA. The DELTA editor DEDIT reads DELTA 3 but outputs DELTA 2. The differences are not great in practice, and since DELTA 2 is upward compatible with DELTA 3, you can simply use DELTA 2 for both PANKEY and CONFOR. There are also a number of differences of practice between PANKEY and CONFOR which were not intentional, but there has nevertheless been some divergence.

FORMAT differences

The following differences have been observed:

1) Running together character numbers in directives. In DELTA 2 you need to write out CHARACTER TYPES and NUMBERS OF STATES separately for each character e.g., 23,RN 24,RN 25,RN and not 23–25,RN and similarly 6,3 7,3 8,3 and not 6–8,3.

2) IMPLICIT VALUES are not implemented in PANKEY, except in DEDIT and in KCONP (for ONLIN5 and 6 and KCONI) "See also 13".

3) Qualitative character states may not be put into ranges in ** PANKEY **, so that if character 5 has states 1,2 or 3, you must put 5,1/2/3 and not 5,1–3

4) PANKEY expects the directive CHARACTER DESCRIPTIONS instead of CHARACTER LIST.

5) DEPENDENT CHARACTERS are programmed as a type 5 directive in PANKEY, so it appears after the CHARACTER LIST and before the ITEM DESCRIPTIONS.

6) Quantitative KEY STATES in PANKEY are not programmed to accept truncated ranges at the beginning and end of a sequence i.e. the forms ~t and t~ (DELTA manual p 54) are not accepted. Also, PANKEY does not yet accept KEY STATE data for qualitative ** (UM,OM) characters, only for quantitative, (IN, RN). **

7) The feature of V.3 where quantitative characters in descriptions can have bracketted parts of their ranges is not in version 2. e.g 5,2–9.5 instead of 5,(2-)4–7.5(-9.5)

8) PANKEY only computes (where applicable) with CHARACTER WEIGHTS which are integers.

9) The '+' feature for ITEM DESCRIPTIONS is not in Version 2.

10) PANKEY expects every line of the DELTA file to have a blank in column 1 in order to show that there are no sequence numbers. Sequence numbers, if present, are ignored.

11) PANKEY expects the *HEADING directive to end with a /

12) '&' in item descriptions is treated as an error (see newsletter 2).

**13) IMPLICIT VALUES. The only PANKEY programs which accept these are KCONP and DEDIT. Both these programs expand the implicit values according to the statements on p.47 of the DELTA 3 manual. CONFOR does not insert the implicits in DELTA or natural language output. **

**14) PANKEY requires discontinuous integers (character type IN) to be in ascending order in descriptions

e.g. 3,1/3 not 3,3/1

This is to make it easier to compare descriptions. DEDIT would accept 3/1 on input but would turn it into 1/3 on output. These remarks do NOT apply to qualitative characters in the same format**

**15) PANKEY treats a missing *END as an error. DEDIT just gives a warning message. **

**16) DEDIT and PANKEY can correctly handle only a single comment at the end of a feature description, state description, or attribute. On output, all the comments are collected and placed at the end, which can sometimes alter the meaning. Here are some examples from the sample data supplied with CONFOR: 'not <consistently> in distinct long-and-short combinations <implicit>/' becomes 'not in distinct long-and-short combinations <consistentlyimplicit>/'; '45,1<usually>/3' becomes

'45,1/3<usually>'; '65,1<5>/2<14>/3<4>' becomes '65,1/2/3<5144>'.

Differences of practice

These differences are in the way the programs work, which affect the data.

1) DELTA files. In PANKEY the normal practice is to keep all the DELTA data in one file, and not in several files as for CONFOR.

2) Dependent characters. PANKEY is stricter about dependent characters than CONFOR is. If the DEPENDENT CHARACTER rules imply that a character state should be inapplicable, then PANKEY checks that the state is inapplicable (if scored), or records it automatically as inapplicable if it is not scored. If the controlling character is variable, then the dependent characters might be either inapplicable or scored. PANKEY automatically allows for this, and it is never necessary to write characters such as 5,-/2 if character 5 depends on 4 (say) which is variable. This means that you can simply leave out inapplicable characters and there is never any need to score them. PANKEY checks for the consistent use of inapplicables and issues error messages accordingly, whereas CONFOR does not – specifically, PANKEY does not allow the use of the '-' coding unless it is implied by a dependency.

** Example CONFOR and DEDIT will accept dependency rules such as

1,1:15 1,2:15

whereas PANKEY programs will insist on having 1,1/2:15 i.e. each pair of characters e.g. 1 and 15 can only be connected by one entry. DEDIT will recombine rules in the former form on input to the latter on output ***

3) MAXIMUM NUMBER OF ITEMS. PANKEY treats this number as the actual number of items to be expected, and gives an error message if the actual number of items encountered is not correct.

If anybody knows of any further problems, please let me know.

Bug in MACINTOSH versions of PANKEY programs

Early versions of MacPankey (prior to February, 1990) have suffered from an intermittent bug in the program KCONP. This program will sometimes give erroneous error messages about missing or inapplicable character states. The error has something to do with reallocation of RAM during execution time. The same copy of KCONP with a correct data file e.g. JUR2D.DAT will sometimes run correctly and sometimes not. The error comes and goes on different machines and on different days. The problem has been referred to the manufacturers of the FORTRAN compiler.

The MacPankey programs have been developed in my own time and at my own expense, and enquiries should be addressed to RJP at 203, Sheen Lane, London SW14 8LE, England.

DELTA editor Version 0.2

DEDIT is an editing program for DELTA format. It reads DELTA versions 2 or 3 and outputs DELTA in version 2 (for PANKEY programs). It also outputs binary files for the ONLIN6 and KCONI interactive programs, and enables data capture and editing for DELTA descriptions. It has menus, windows and help screens, and is FREE.

My thanks to those patient users who tested DEDIT version 0.1 for me and drew attention to a number of bugs. Version 0.2 is now available, with all known bugs corrected.

Progress of book

The sequel to "Biological Identification" by R. J. Pankhurst (1978, E. Arnold) is now complete in draft. It should be published later this year under a title such as "Complete Methods for Taxonomists". It includes databases, classification, (phenetic & cladistic) and expert systems as well as identification.

General purpose descriptors

I agreed at the last TADWG meeting to organize a group to discuss general purpose descriptors. Such a descriptor may be defined as a character which is of general application to higher plants. The character of 'habit' (herb or shrub or tree) is an example. There are perhaps about 20 such characters, and they tend to be characters which figure prominently in the classification of higher plants, or in keys to families. The characters in the following list are intended to serve in order to stimulate debate. I therefore invite readers to

1) comment on which characters should be added (or removed)

2) volunteer themselves to join the group

Aerial stem: absent or present in what form
Breeding system: in-or out-breeding
Carpels free or fused
Chromosome number (s)
Class: pteridophyte, gymnosperm or angiosperm (or the characters which these imply)
Cotyledons: monocot or dicot (or the characters implied)
Flowers (suitable defined): absent or present
Fruit type
Habitat: aquatic, terrestrial, epiphytic
Indumentum: plant entirely glabrous or with indumentum of some kind
Inflorescence type
Leaves: absent or present

Life: annual, biennia, perennial
Nutrition: autotrophic, partly autotrophic, heterotrophic (and for non- autotrophic plants, whether saprophytic, parasitic etc.)
Ovary: inferior or superior
Reproduction: asexual, sexual, or both
Roots: absent or present
Sexuality: hermaphrodite, monoecious, dioecious etc.
Sympetaly: petals free or combined

PANDORA database for monographs aand Floras

For some time it has been obvious that DELTA ought to be available as a database, and that a DELTA database ought to also handle other data at the same time, such as nomenclature, specimens, bibliography and geographical distribution. This is a progress report on the development of PANDORA, which is a database application for monographic revision or writing Floras or Faunas.

My paper in Taxon 37(3): 733-746, (1988) 'Database design for monographs and floras' was a first attempt. This database did not allow for DELTA data however, and was written in dBASE3, which seems to me to be inadequate for handling DELTA. The recent paper by Skov, F. (1989). 'HyperTaxonomy - a new computer tool for revisional work' in Taxon 38: 582-590, is recommended reading. It takes the idea further by combining morphological (DELTA) data with many other relevant types of data. It is implemented in HyperCard on Macintosh. Skov tells me that he will not be able to develop his system any further. HyperCard has the advantage that its user interface uses graphics, and that it can handle graphic data (if the hard disc is large enough). It has the disadvantage of not being a relational database, and I am not aware whether it is or will be available on PCs.

PANDORA has been developed from the original dBASE3 version, but has had DELTA input and output and editing features added. The DELTA features of PANDORA are roughly the same as those in the current version of DEDIT. PANDORA also handles nomenclature, types and other specimens, loans, bibliography and geographical distribution. It generates reports on nomenclature, distribution and the management of loans. It is being written as an application of Advanced Revelation, and is currently running under AREV version 1.1. For those who may not be familiar with it, AREV is a highly sophisticated relational database system. It has a program generator (4GL) and is closely related to the PICK database system, as used in TROPICOS at Missouri Botanic Garden. Another important advantage is that AREV works entirely with variable length fields and records. AREV is also extensively used for biological records in the USA and the UK, and is being used for the European plant database at the BM.

PANDORA is already being used for my own work on the Rosaceae for Floras of Central America, but is still undergoing rapid expansion. A definitive version should be expected by the end of 1990. In the mean time, I would be glad to hear from anyone who would like to help with beta-testing.

New versions of ONLINE identification program

While working at the Plant Resources Information Laboratory (PRIL) of A. Gomez-Pompa at the University of California, Riverside during February and March this year, RJP prepared two new versions of the ONLIN6 interactive identification program. Both programs allow the use of video images for characters and items (taxa). The TAXA command, which normally gives a list of current taxa, has been extended to the form TAXA n, where n is the number of an item. The effect is to produce image(s) on a screen corresponding to that item. There may be one or more images for each item, and you can scroll through them by hitting any key. The CHAR i command, which gives you the states for character i, behaves in the same way.

Two versions operate on different hardware available at Riverside.

a) using the TARGA16 frame-grabbing hardware and special monitor. TARGA16 works with a TV camera or other video input device to store images on DOS hard disc which can then be displayed on either a video monitor or on the PC digital display (VGA). There is a graphics drawing package included so that text and other graphics can be added to the image.

b) using a video disc player. This is connected to the RS232 channel of the PC and is driven by simple commands sent down the line. It has its own monitor and displays single or multiple video images as required. The images have to be previously written to a video disc.

Tests of both these programs were carried out using a DELTA dataset for seed ferns of Mexico and stored images from living plants and herbarium specimens that were produced at PRIL by A. Vovides.

CHARACTER ANALYSIS IN DELTA

Since taxonomic data is commonly qualitative rather than quantitative, some measure of correlation is needed which can be applied to qualitative data. This can be done using the information statistic. The method for doing this was first published by Estabrook (1967), who described a program called CHARANAL.

The information content of characters is represented pictorially in the diagrams (on page 10). Each complete and separate character is represented by a circle, which

may or may not overlap with others. Let the information contained by character "a" be written H(a). In diagram I the two characters do not overlap at all, and the circles are separate. These two characters are entirely independent of each other. In diagram II, character b is completely contained in character "a", so that "a" contains information that is not in "b", but "b" just duplicates what is in "a". The general situation is shown in diagram III, where the characters partly overlap. The conditional information of character "a" on character "b" is the information held exclusively by "a" when all the information held by "b" is removed, and is written H(a/b). Similarly, the information held by "a" without the contribution of "b" is H(b/a). Finally, the information held by both "a" and "b" is expressed as H(a.b). The areas for H(a/b), H(a.b) and H(b/a) are shown from left to right in diagram III.

The amount of information held by "a" but not by "b" is found by the usual formula but with character "b" held constant, in state 1, say.

Then

$$H(a/b_1) = - \sum_{\text{states in a}} p(a/b_1) \log p(a/b_1)$$

This quantity is then summed over all the states of "b" so that

$$H(a/b) = \sum_i H(a/b_i)\ p(b_i)$$

with a similar definition for H(b/a). H(a.b) is then defined just by subtraction so that

$$H(a.b) = H(a) - H(a/b) = H(b) - H(b/a)$$

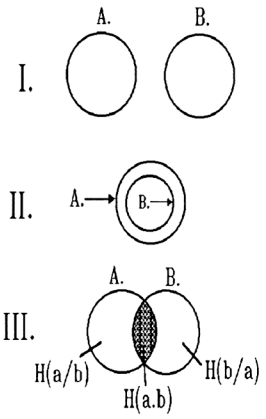
The proportion S of information shared by two characters as opposed to the total information they hold can be seen in diagram III as the ratio of the unshaded area to the total area

$$S = \frac{H(a.b)}{H(a.b) + H(a/b) + H(b/a)} = \frac{H(a) - H(a/b)}{H(b) + H(a/b)}$$

This value S may be regarded as a measure of correlation or similarity between two characters. In order to calculate the information values for actual data sets, allowance has to be made for missing, inapplicable and variable characters. Quantitative characters have to be converted into (pseudo)qualitative characters by dividing them up into ranges with *KEY STATES.

The highest possible value for S is 1.0 when two characters have an identical distribution of states i.e. they are exactly correlated. By analogy with the representation of information by circles, these would appear as two circles drawn on top of one another. In cladistic analysis this correlation might be recognised as a synapomorphy or a symplesiomorphy. If the S value is zero, this would indicate two characters which are completely independent of each other. This would mean that of two such characters, either one would distinguish any given pair of taxa or the other would, but not both together. If a character pair shows value zero for H(b/a), which is the situation of diagram II, this would be because the information in one character is already included in the other. Similarly if H(a/b) is zero. Pairs of characters which have high values of S can indicate meaningful character correlations. For example, with the genus *Epilobium* which is used in my book on identification it might be said that species with erect habit tend to have hairy stems, and that if there are glandular hairs on the stem, they tend to occur on the calyx as well.

A program called CHANAL, one of the PANKEY programs, has been written to calculate values of S and the other coefficients. CHANAL reads DELTA version 3 data files. CHANAL is available FREE of charge from RJP at the usual addresses.



ESTABROOK, G.F. (1967). An information theory model for character analysis. Taxon, 16: 86-96.

R. Pankhurst, Dept. of Botany, Natural History Museum, Cromwell Road, London SW7 5BD, England.