DELTA Newsletter

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.

Distribution records and habitat data in ALICE databases

by Bob Allkin and Peter Winfield

   ALICE is a suite of menu driven database programs for biologists with no prior knowledge of computers who wish to design, build or distribute biodiversity databases. For each taxon in an ALICE database users can enter or retrieve information about taxonomy, synonymy, geography, uses, common names, habitats or the associated literature. Users may further define any number of DELTA- like descriptive characters to describe particular morphological, ecological or other features relevant to their group of animals or plants. They may also store any amount of free text information per taxon.

   In previous newsletters we described the nomenclatural part of the ALICE system and how it links with DELTA. To the taxonomic and nomenclatural framework at the heart of any ALICE database can be added any amount of other information. This month we describe how the current version of ALICE handles geographic distribution and habitat data and mention how these features are being expanded on in the next version.

DATA ELEMENTS

The gazetteer

   Each database project may use its own three-tier geographic hierarchy. A database for the world's legume species, for example, has gazetteer entries for "continents", "countries" and "states" while a Mexican ornithologist might enter "major regions", "states" and "municipalities". ALICE databases for National Parks are built around hierarchies of much smaller unit size. ALICE Version 3.0 will allow users any number of levels in their hierarchy.

   Users decide the name and, if desired, abbreviation of each geographical unit within the database. They also define the hierarchy i.e. which states belong to which countries, which countries belong to which continents etc. As standard gazetteers such as that developed by the Taxonomic Databases Working Group, become available, they will be optionally provided with ALICE. Users may define their gazetteer either at the outset of building their database or 'on the fly' as they enter species data.

The data

   For each taxon (or putative taxon) in the database, users can enter distribution records at all levels in the gazetteer. Each record of species X in place Y may have any number of source references attached.

Habitat data

The user may enter habitat data into an ALICE database in a number of ways.

1) by defining a multistate unordered habitat descriptor with one "state" for each habitat category required.

2) by entering free text habitat descriptions.

3) by defining a suite of habitat categories linked to species distribution. In this report we explore only the latter alternative.

Users define a different habitat classification for each of the places at the highest level of the geographical hierarchy. Thus it is possible, for example, to use a completely different habitat classification for each continent. Data records for the presence of species X in habitat type Y in place Z may have any number of source references attached.

Introduction Status and Conservation

Users may record whether a species is "native" or "introduced" or of "unknown" status in each place. This set of alternatives is currently fixed by ALICE but users will be free to define their own categories using ALICE Version 3. Conservation status is entered as a DELTA like descriptor and users have complete control over the categories used. Once more source references may be included.

Geographically related descriptors

A significant advance in ALICE Version 3 will be the ability to define any number of "geographically related descriptors". It will be possible to associate user defined descriptors with one or more of the levels defined within the geographical hierarchy. Thus it will be possible to use a conservation descriptor at international, national and local levels. It will be possible to record geographically related ecologically or morphologically related distribution patterns.

LOGICAL AND VALIDATION RULES

ALICE uses many rules to check data as it is entered and to prevent logical errors through subsequent editing. The rules governing entry and editing of geographical data include:-

- species recorded as being present in "India" cannot simultaneously be recorded as absent from "India".
- species recorded as present in "India" cannot simultaneously be recorded as absent from "Asia".
- places cannot be removed from the gazetteer while species are still recorded from that place.

EDITING THE DATA

The gazetteer and dictionary editor

Allows users to create new places or delete them, correct spellings, convert habitats or place names to foreign languages or to edit the geographical hierarchy (so that existing countries can be moved to some newly defined sub- continent, for example). Users also have complete control over the habitat classification used.

Species records

Users can delete data records, amend or delete introduction, habitat or conservation records or delete any of the attached references.

QUERYING THE DATABASE

As for any datatype in ALICE databases, AQUERY provides simple mechanisms to search for species showing particular combinations of features. Compound searches are built up using the logical operators "OR", "AND", and "NOT". You might explore, for example, geographical distribution patterns for any combination of places - those birds in Mexico AND Guatemala but NOT in the US. Those plants in Africa but NOT in Kenya that are found in high montane forest. The following screens illustrate a simple example search.

STAGE 1: Select a level in the geographical hierarchy and then a place.

```
        AREA                COUNTRIES
        County              Gabon
        Country             Galapagos
        Continent           Gambia The
                            Gambier Is
                            Germany-F.E.
                            Ghana
                            Gibraltar
                            Gilbert Is
                            Great Britain
                            Greece
                            Greenland  ....
```

STAGE 2: Formulate the first part of the query

```
      Build a new query
        Country               equals Gabon
                                                        Press any key to continue...
```

STAGE 3: Then add additional elements to the search - one at at time

```
      Build a new query
        Country               equals Gabon
        Habitat               equals Guineo-Congolian forest
                                              Press any key to continue...
```

STAGE 4: Do the search once the list of criteria is complete

```
  EXECUTE QUERY
  Database: ILDIS Phase 1 Database:  Version 1.07.02 August 1990
  Date: 10/05/90  Time: 21:21:18          Total no. of Taxa:  15861
        Country        equals Gabon                          375     and
        Habitat        equals Guineo-Congolian forest        267     and
        Habit          equals Tree                           200     and
        Use            equals Wood          RESULT  =         64
```

Thus for this database there are 375 taxa in Gabon of which 267 are found in Guineo-Congolian forest. Of these 267, 200 are tree species of which 64 are used for wood.

STAGE 5: We can now list those 64 taxa satisfying the search criteria

```
LIST NAMES
   ...
   Tetraberlinia bifoliolata (Harms)Hauman
   Tetraberlinia moreliana Aubrev.
 CAESALPINIEAE
   Erythrophleum ivorense A.Chev.
   Pachyelasma tessmannii (Harms)Harms
 CASSIEAE
   Cassia mannii Oliver
   Dialium pachyphyllum Harms
   Distemonanthus benthamianus Baillon
 DALBERGIEAE
   Pterocarpus osun Craib
   Pterocarpus soyauxii Taubert
 DETARIEAE
   Afzelia bipindensis Harms
   ....
```

WRITING REPORTS AND EXPORTING DATA TO DELTA AND OTHER PROGRAMS

AWRITE is used to generate reports from ALICE databases. Users decide whether geographical or habitat data is to appear in a particular report, where it appears in relation to the nomenclature, descriptive data, free text or other types of information. Source references may be included or not for either datatype, and each may appear in a different typeface or indented to a different degree. Country names may appear as full names OR in a 2-letter abbreviated form.

ALEX exports chosen subsets of your data into the format of your choice. We currently support SDF, dBASE, Fixed field output and XDF as well as DELTA. ASLICE creates small ALICE databases from subsets of your data.

Even distribution data can be exported to DELTA, with or without other data. Users thus benefit from the ease of data capture, verification, logical checks and flexible reports that ALICE provides while still being able to use other DELTA programs. ALEX defines a separate multistate character to represent the uppermost level within the geographical hierarchy: e.g. a DELTA character named "Continent" with one state for each of the continents defined by the user: "Africa, Asia, etc.". ALEX also creates a series of characters dependent on the first - one for each continent: "Countries of Africa", "Countries of Asia" etc. The states of each being those countries in the gazetteer for that continent. Finally, ALEX defines a further set of dependent characters to describe species distributions at the lowest level in the hierarchy (e.g. "States of India", "States of Pakistan" etc.) again using gazetteer entries as character states.

When users opt to export habitat data to DELTA, characters are defined in a similar way. Bibliographic source data, of course, cannot be exported to DELTA since DELTA offers no facility for such data.

SUMMARY

We have given an overview of the functionality of the geographic module but realize that there is no substitute for practical experience using the program. We remind you that demonstration copies of the program are available for a small fee to cover distribution costs!! In the next newsletter we plan to describe the common knowledge module including "Use" data and "Vernacular names" - unless readers suggest otherwise.

POSTSCRIPT

Since the last newsletter we have released a new version of the ALICE database maintenance program. Now called AMIE (for our french friends), this program is distributed free with ALICE and undertakes a host of housekeeping tasks associated with data maintenance, logical integrity, data redundancy and data compression. If you have a fully registered copy of ALICE and have not yet received your copy of AMIE please be sure to let us know.

Bob Allkin: Royal Botanic Gardens KEW, Richmond, TW9 3AB UK. Tel. +44-1-940-1171 ext.4715 Telex 296694 KEW-GAR G Fax. +44-1-948-1197. email email BTGOLD 81:bio023

Notes from Richard Pankhurst

by Richard Pankhurst

Crisis at the Natural History Museum, London

As a result of a management reorganization coupled with a financial crisis, the section of taxonomic computing, and many of the other computing projects at the museum are being closed down. As a result, Richard Pankhurst will be made redundant, and expects to have to leave the Museum by April 1991. At the time of writing, his future is uncertain, but he will remain committed to the PANKEY software and his various database projects.

PANDORA

The PANDORA database for taxonomic monographs and floristic projects, described in the DELTA Newsletter issue 5, has been considerably expanded. It will shortly be converted from Advanced Revelation version 1 to version 2. While not all its features have yet been implemented, PANDORA is in a usable state. It can be made available to users of Advanced Revelation, free of charge, in an alpha-test version. This means that I shall expect to hear from alpha-users about bugs and any other constructive criticisms as a condition for not having paid for it!

New version of ONLINE program

By the time this appears in print, a new version of the expert identification program, ONLIN7, will be available. The previous version, ONLIN6, is a full screen interactive program with graphic images (monochrome or colour EGA) to illustrate characters, and was reported in the DELTA Newsletter issue   .

ONLIN7 is a development of ONLIN6, but has windows and help screens for general background, commands and characters, and has optional menus for commands. Multiple images are available for both characters and taxa. ONLIN7 will require an EGA or VGA (colour or monochrome) monitor.

Improvements to DEDIT

The currently distributed version 0.2 is able to enter or correct item descriptions (as did the COND4 program) and is very useful for creating data for DELTA data files. DEDIT is indispensable for the preparation of binary files for use in the newer (C program) versions of ONLIN6 and KCONI (Interactive key construction).

By the time this appears in print, DEDIT version 1.0 will be available, and to this will have been added code for editing characters and state descriptions and taxon names.

ARTISYST

Richard Pankhurst and Mike Dallwitz both attended the workshop on Artificial Intelligence and Systematics organized by Renaud Fortuner of the Department of Nematology, University of California Davis on behalf of the National Science Foundation of the USA. The meeting was held at Napa, California, from the 9th -14th September, and its purpose was to bring together both computer scientists and systematists in order to report to the NSF on what kinds of research require support in these areas. A separate conference report will be published by the University of California Press. Most of the participants were of course citizens of the USA, but there were also representatives from Canada, the UK, Australia, New Zealand, and France. The four topics considered were cladistics, morphometrics, identification and databases, of which the last two contained the most interest for us.

There will be many recommendations in the final report to the NSF which cannot all be covered here, but here are some of the suggestions which were put forward:
- to the cladists, recommended that they use a standard format for descriptions e.g. DELTA
- to the computer scientists, that the main area for development of AI systems is for non rule-based and non-probabilistic datasets i.e. data matrices
- for taxonomic software, an overall major improvement in user interfaces is needed
- for databases, development of video image data.

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 was referred to the manufacturers of the FORTRAN compiler, and has now been corrected. Would users of older versions of MacPankey please return their master discs to me for an update.

R. Pankhurst, Dept. of Botany, British Museum (NH), Cromell Road, London SW7 5BD, England.