Desirable attributes for interactive identification programs
7 April 2007

M.J. Dallwitz

Contents

Introduction

This is an updated and expanded version of a posting to the Taxacom mailing list (Dallwitz 1994).

Dallwitz (2000) gives a detailed comparison of several programs, in terms of these attributes.

For more discussion of the attributes, and other aspects of interactive keys, see Dallwitz, Paine and Zurcher (1998, 2000).

Advantages over conventional keys

Unrestricted character use. No restrictions on the order in which characters can be used (apart from restrictions imposed by character dependencies — see below).

Character deletion/changing. Removing characters used in an identification, or changing their values. (Undesirable limitations: removal only in the reverse order of use.)

Error tolerance. The ability to reach a correct identification after errors have been made, or if there are errors in the data.

Locating errors. When the error tolerance is non-zero, a taxon that differs from the specimen can remain in contention in an identification. The differences may be due to errors by the user, errors in the data, or both. The program should be able to display these differences.

Expressing uncertainty. The user can specify uncertainty by entering more than one state value, or a range of numeric values.

Numeric characters. Using numeric characters directly (without converting to multistate by dividing the values into ranges).

Guidance in character selection

Best characters. Advice on the most suitable characters for use at any stage of an identification. (Undesirable limitations: a lack of flexibility in this area. Examples: Inability to handle numeric characters. Recommendations built into the data, as in a conventional key or a rule-based expert system.)

Separating a taxon. Ranking characters according to how well they separate a given taxon from the rest.

Best routes. Paths, similar to conventional keys, embedded in the interactive key. A path may be followed from the start of an identification; after it is left, ordinary interactive identification is resumed. This provides some guidance in the choice of characters. The method is inherently much less flexible than 'best characters' (see above), and this limits its usefulness.

Differentiating attributes. Attributes that are exhibited by only a small number of the remaining taxa. This is not a desirable feature for identification, but it is included here because it is

implemented in some programs. These characters are the worst to use in identification, apart from those which do not differentiate the taxa at all or which differentiate only parts of taxa.

Removing redundant characters. Removing from the list of available characters those that cannot separate the remaining taxa in an identification.

Removing redundant character states. This is a way of preventing the selection of character states that are not exhibited by any of the remaining taxa in an identification. It can be done either by removing the states entirely from the display, or by greying them (which is preferable, as the user is then aware of the situation). Numeric characters can be treated similarly, by not allowing the entry of values not exhibited by the remaining taxa. This is not a desirable feature for identification, but it is included here because it is implemented in some programs. The feature encourages the user to enter values consistent with the values of previously used characters, thereby hindering the detection and correction of errors.

Character reliabilities. The 'reliability' of a character is a subjective measure, usually supplied by the author, of the character's accuracy and/or ease of use. It should be taken into account when choosing 'best' characters. (Undesirable limitations: higher reliabilities always implying 'better' characters, regardless of other considerations.)

Attribute reliabilities. An 'attribute' is the value or values of a character for a particular taxon. The 'reliability' of an attribute is a subjective measure, supplied by the author, of the attribute's accuracy and/or ease of use. It is recorded as a increase or decrease in the overall or average reliability of the character. Attribute reliabilities allow better performance of the 'best' algorithm.

Searching the character list. Finding text strings in the character list.

Recording and matching character values

Retaining unknowns. Taxa for which a character is not recorded are retained when that character is used (with any value) in an identification. This is essential for correct identification.

Character dependencies. Relationships specifying that some characters are inapplicable when other characters take certain values.

Automatic controlling characters. Automatically setting controlling characters to the appropriate value(s) when dependent characters are used. The author must have the option of overriding the automatic setting in the (rare) cases when it would lead to incorrect or inconvenient results. In these cases, the user should be required to use the controlling character before the dependent character.

Gaps for integer numeric characters. The possibility of gaps in recorded values for integer numeric characters, e.g. '5 or 10' distinguishable from '5 to 10'.

Text characters. Storing and searching free-text information about taxa.

Special values for keys. Flagging values in the data for use only in keys. This would be done for values which are not strictly exhibited by a taxon, but which a user might be likely to assign erroneously to a specimen belonging to the taxon. (Undesirable limitations: the use of these values not being under the control of the user of the key.)

Probabilistic identification. Using probabilities of state values in the taxa, and probabilities of user errors, to calculate the probabilities that the specimen belongs to a given taxon.

Inapplicable versus unknown. Distinguishing inapplicable values, including those not resulting from character dependencies, from unknown values.

Expanded ranges for numeric characters. Expanding the range of values recorded for a numeric character that has been poorly sampled in a taxon. (Undesirable limitations: the transformation not being under the control of the user of the key (as in the ABSOLUTE/PERCENTAGE ERROR mechanisms in Confor/Intkey).)

Unknown state values. Recording individual state values of multistate characters (as opposed to the character as a whole) as unknown.

Exact characters. Specifying characters whose values are assumed not be subject to error.

Fixing character values. Specifying character values that are not to be cleared when a new identification is started. This is convenient when identifying several specimens that are known (or thought) to share some attributes, usually the place of origin or belonging to a higher taxon.

Subsets

Named subsets of characters. A mechanism for naming subsets of the characters. (Undesirable limitations: subsets being built into the identification package, and not definable by the user.)

Global subsets of characters. Specifying subsets of characters to which all subsequent operations will be restricted.

Local subsets of characters. Specifying subsets of characters for a single operation.

Named subsets of taxa. A mechanism for naming subsets of the taxa. (Undesirable limitations: subsets being built into the identification package, and not definable by the user.)

Global subsets of taxa. Specifying subsets of taxa to which all subsequent operations will be restricted.

Local subsets of taxa. Specifying subsets of taxa for a single operation.

Character interpretation

Character notes. Extensive text to aid interpretation of characters conveniently available within the system.

Glossaries. Linking definitions to words in the database.

Character illustrations. Displaying illustrations of characters.

State selection from character illustrations. Selecting character state values from illustration screens during identification.

Images and sounds

Taxon illustrations. Displaying illustrations of taxa.

Taxon illustrations by subject. Selecting illustrations of taxa by subject (for example, flowers, habitat, distribution map).

Flexible display of illustrations. Illustrations of any size can be scaled, scrolled, repositioned, displayed simultaneously, and tiled or cascaded. All images can be closed with a single operation.

Text with illustrations. Text can be associated with illustrations (instead of being built into the illustrations). (Legible text after scaling, possibility of multiple languages.)

Sounds. Sounds can be attached to characters and taxa.

Videos. Videos can be attached to characters and taxa.

Running without illustrations. A package normally containing illustrations can run well without them.

Linked keys

Integral hierarchical keys. Having keys for different taxonomic levels within a single data matrix.

Separate hierarchical keys. Linking taxa to other keys (for example, a genus linked to a key to the species of the genus).

Information retrieval

Searching for taxon names. Finding text strings in the taxon names and synonyms. (Undesirable

limitation: separate searching for correct taxon names and synonyms.)

Descriptions from the data. Displaying descriptions generated directly from the data used in identification.

Differences between taxa. Finding the differences between members of a set of taxa. (Undesirable limitations: restrictions on the size of the set of taxa.)

Similarities between taxa. Finding the differences between members of a set of taxa.

Diagnostic descriptions. Finding diagnostic descriptions. A diagnostic description for a taxon matches that taxon, but differs from all the other taxa in at least one respect. (Undesirable limitations: inability to distinguish between taxon and specimen diagnostic descriptions; inability to restrict the choice of characters to those not used in the current identification; inability to set the strength of the descriptions (e.g. DiagLevel in Intkey).)

Taxon retrieval by attributes. Finding all taxa that have certain attributes or combinations of attributes. It must be possible to eliminate taxa for which the attribute is unknown or inapplicable.

Control of value matching. Control by the user over which values or pseudovalues are deemed to match other values: exact, subset, overlap, unknown, inapplicable. See Dallwitz, Paine and Zurcher (1995) or online Intkey documentation for details.

Character-value distributions. Displaying the distribution of character values within a set of taxa.

Most similar taxa. Ranking taxa by their similarity to a given taxon.

Text files attached to taxa. Text files containing information about taxa.

Data sharing

Importing DELTA format. Using DELTA-format data to generate the interactive identification system.

Exporting DELTA format. Exporting DELTA-format data from the interactive key, or from the system that generates the key.

Data output. Output of program results in forms suitable for input to other programs.

Links with description writing. Generating publication-quality descriptions from the same data that are used to construct the identification system.

Links with key generation. Generating conventional keys from the same data that are used to construct the identification system.

Links with classification. Carrying out cladistic and phenetic analyses from the same data that are used to construct the identification system.

Usability

Online help. Complete, context-sensitive, built-in help.

Command files or macros. A mechanism for storing and repeating a series of operations.

User-definable toolbar. The author or user can define toolbar buttons for easy access to commonly used operations or groups of operations.

External program text. The program text (commands, help, messages, etc.) is external to the program, allowing easy creation and use of different language versions.

Log files. Creating a file showing the history (input and output) of a session.

Unlimited data size. The numbers of taxa, characters, and states are unlimited.

Unlimited field lengths. The lengths of fields (for example, taxon names, text of characters, character notes) are unlimited.

No special memory requirements. The program will run with the minimum amount of memory normally needed to run the operating system (including dependence on data size, if applicable).

Execution speed. Execution times of representative operations on a reasonably large data set (e.g. 200 characters, 400 taxa).

Internet capability. The program can access data and images over the Internet.

Installation unnecessary. The program can be run directly from a CD-ROM, without installation.

Simple user interface. The user interface is simple, efficient, and consistent, and provides an easy transition from basic features to advanced features (if any).

References

Dallwitz, M.J. 1974. A flexible computer program for generating identification keys. *Syst. Zool.* 23: 50–7.

Dallwitz, M.J. 1980. A general system for coding taxonomic descriptions. *Taxon* 29: 41–6.

Dallwitz, M.J. 1989. Diagnostic descriptions from INTKEY and CONFOR. *DELTA Newsletter* 3: 8–13.

Dallwitz, M.J. 1989. Diagnostic descriptions for groups of taxa. *DELTA Newsletter* 4: 8–13.

Dallwitz, M.J. 1992. A comparison of matrix-based taxonomic identification systems with rule-based systems. In 'Proceedings of IFAC workshop on expert systems in agriculture', pp. 215–8. (Ed. F. L. Xiong.) (International Academic Publishers: Beijing.) Also available at http://delta-intkey.com

Dallwitz, M.J. 1993. DELTA and INTKEY. In 'Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision', pp. 287–296. (Ed. R. Fortuner.) (The Johns Hopkins University Press: Baltimore, Maryland.)

Dallwitz, M.J. 1994. Desirable attributes for interactive identification programs. http://usobi.org/archives/taxacom.html (December)

Dallwitz, M.J. 1996 onwards. Programs for interactive identification and information retrieval. http://delta-intkey.com

Dallwitz, M.J. 1999 onwards. A comparison of formats for descriptive data. http://delta-intkey.com

Dallwitz, M.J. 2000 onwards. A comparison of interactive identification programs. http://delta-intkey.com

Dallwitz, M.J., and Paine, T.A. 1993 onwards. Definition of the DELTA format. http://delta-intkey.com

Dallwitz, M.J., Paine, T.A., and Zurcher, E.J. 1993 onwards. User's guide to the DELTA system: a general system for processing taxonomic descriptions. 4th edition. http://delta-intkey.com

Dallwitz, M.J., Paine, T.A., and Zurcher, E.J. 1993. Preliminary suggestions for new features for the DELTA system. *DELTA Newsletter* 9: 2–13.

Dallwitz, M.J., Paine, T.A. and Zurcher, E.J. 1993 onwards. New features for the DELTA system. http://delta-intkey.com

Dallwitz, M.J., Paine, T.A., and Zurcher, E.J. 1995 onwards. User's guide to Intkey: a program for interactive identification and information retrieval. http://delta-intkey.com

Dallwitz, M.J., Paine, T.A., and Zurcher, E.J. 1998. Interactive keys. In 'Information technology, plant pathology and biodiversity', pp. 201–212. (Eds P. Bridge, P. Jeffries, D.R. Morse, and P.R. Scott.) (CAB International: Wallingford.) Dallwitz, Paine and Zurcher (2000) is an updated version.

Dallwitz, M.J., Paine, T.A., and Zurcher, E.J. 2000 onwards. Principles of interactive keys. http://delta-intkey.com