



A Comparison of Formats for Descriptive Data

15 July 2010

M. J. Dallwitz

Introduction

This document compares the capabilities of three data formats which were available in 1999 for recording descriptive data. The formats are:

- DELTA (Dallwitz 1980; Dallwitz, Paine and Zurcher 1993a). The table shows the capabilities of the 'current' version, as described in the previously cited publications and implemented in the DELTA System of the CSIRO Division of Entomology, and an 'extended' version, as described by Dallwitz (2000) and Dallwitz, Paine and Zurcher (1993b).
- Lucid (Lucidcentral 1999). The table shows the capabilities of Versions 1 and 2 of the Lucid Interchange Format (LIF), and of the Lucid Builder (which uses a binary data format).
- Nexus (Maddison et al. 1997). The table shows the capabilities of the 'general' format, as described in the previously cited publication, and the implementations in Paup (Maddison and Maddison 1992) and MacClade (Swofford 1991).

Also mentioned, but not included in the table, are some **other desirable features** that are not currently [1999] available in any of the data formats.

For an introduction to requirements for descriptive data, see Dallwitz (2000).

Comparative Table

The entries in the first column are linked to notes on the topics.

- Y The implementation of the feature is satisfactory.
y The implementation of the feature is unsatisfactory. See Notes for details.
blank The feature is not implemented.
? Not known.
(...) Application-related information that need not be part of the data format.
* See Notes.

	DELTA		Lucid		Nexus	
	Current	Extended	LIF	Builder	Paup & MacClade	General
Unordered multistate characters	Y	Y	Y	Y	Y	Y
Ordered multistate characters	Y	Y			y	y
Real numeric characters	Y	Y	Y	Y		
Integer numeric characters	Y	Y				
Molecular sequence characters					Y	Y
Text characters	Y	Y				
Cyclic characters		Y				
'List' characters		Y				
Character dependencies	Y	Y	Y	Y*		
Comments in characters	Y	Y			Y	Y
Alphanumeric character identifiers		Y				

	DELTA		Lucid		Nexus	
	Current	Extended	LIF	Builder	Paup & MacClade	General
Units for numeric characters	Y	Y		Y		
Unlimited number of states	Y	Y				Y
Unlimited character text	Y	Y	Y	Y		Y
Character notes	Y	Y		Y	Y*	Y
Character illustrations	Y	Y			Y*	Y
Character-state illustrations	y	y		Y		
Character keywords	y	y		Y		
Taxon names	Y	Y	Y	Y	Y	Y
Unlimited name text	Y	Y				Y
'or' in attributes	Y	Y	Y	Y	Y	Y
'to' in numeric attributes	Y	Y	Y	Y	Y	Y
'to' in multistate attributes	Y	Y				
'and' in attributes	Y	Y				
Unknown attributes	Y	Y	Y	Y	Y	Y
Inapplicable attributes	Y	Y				
Attribute comments	Y	Y				
Value comments	Y	Y				
Implicit values	Y	Y				
Indefinite values		Y				
Value is rare		Y	Y	Y		
Value is misapplied		Y	Y	Y		
Value is unknown		Y	Y	Y	y	y
Value probability		Y				Y
Value 'only'		Y				
Value 'not'		Y				
Value 'for'		Y				
Value is approximate		Y				
Value is guessed		Y				
Attribute weights		Y				
Attribute passed up		Y				
Attribute passed down		Y				
Taxon notes	Y	Y		y	Y*	Y
Taxon illustrations	Y	Y		Y	Y*	Y
Taxon keywords	y	y				
Typesetting markup	Y	Y			y	y
Alternative natural languages	Y	Y				
Alternative wordings		Y				
Private notes	Y	Y				
(Character weights)	Y	Y			Y	Y
(Description layout)	Y	Y				
(Character transformation types)					Y	Y
(Taxon trees)					Y	Y

Notes

Unordered multistate characters

Characters in which the states have no natural order. ‘State 1 to state 3’ does not imply the presence of state 2.

Ordered multistate characters

Characters in which the states have a natural order. ‘State 1 to state 3’ implies the presence of state 2.

Lucid does not distinguish between ordered and unordered multistate characters.

Nexus does not allow the connection of states by ‘to’, which provides the crucial distinction between ordered and unordered multistate characters in natural-language descriptions.

Real numeric characters

Characters that take real values, e.g. ‘length of leaves’.

In Paup and MacClade, true numeric characters are not implemented – numerical data must be recorded in multistate characters.

Integer numeric characters

Characters that take integer values, e.g. ‘number of veins’.

In Lucid, integer numeric characters are not distinguished from real numeric characters, i.e., all numeric characters can take non-integer values.

In Paup and MacClade, true numeric characters are not implemented - numerical data must be recorded in multistate characters.

Molecular sequence characters

Characters to record molecular sequence data.

Text characters

‘Characters’ to record arbitrary text, which may be separately searched, and intercalated among the other characters in natural-language descriptions.

Cyclic characters

Characters to record cyclic data, e.g. ‘month of flowering’.

‘List’ characters

The same as multistate characters, except that the states are contained in an external list, which may be used in several characters (and also in other contexts). For example, the list might consist of names of countries, and these could be used in characters to record the actual distribution and the native distribution of the taxa.

Character dependencies

Character dependencies specify sets of characters — the ‘dependent’ characters — that are inapplicable when certain other characters — the ‘controlling’ characters — take certain values. For example, the attribute ‘leaves absent’ implies that characters describing the nature of the leaves (e.g. length, shape) are inapplicable, and must not be recorded. If any of the recorded values of a given controlling character do not make its dependent characters inapplicable, then the dependent characters may be recorded. For example, ‘leaves present or absent’ allows other leaf characters to be recorded. If a given dependent character is dependent on more than one controlling character, then the dependent character can be recorded only if allowed by all of its controlling attributes.

In the current DELTA format (and in Lucid), the controlling characters must be multistate characters.

It is important that any program for maintaining data should check the data for consistency with the dependencies; otherwise, large numbers of errors are almost inevitable. The Lucid builder does not carry out such checks.

Comments in characters

Comments in characters may be included or omitted depending on context. This can make it possible to use a single character list for different applications, such as keys and natural-language descriptions. A more powerful mechanism is provided by alternative wordings.

Alphanumeric character identifiers

Character numbers are satisfactory identifiers for characters within a single data set. However, they may change when characters are added, deleted, or reordered. Identifiers that are fixed within a given context (e.g. within an organization, or worldwide) would be useful for

purposes such as merging data sets and checking the consistency of wording.

Units for numeric characters

Units for numeric characters, e.g. 'mm'.

Unlimited number of states

No limit on the number of states for multistate characters. In Lucid, the maximum number of states is 15.

Unlimited character text

No limit on the length of text for character feature and state descriptions. Paup and MacClade have a limit of 32 symbols for each feature and state.

Character notes

Notes on the interpretation of characters.

Implemented in MacClade but not in Paup.

Character illustrations

Illustrations of characters. Programs that allow state selection from the illustrations need to be able to incorporate other information, such as text overlays and active regions.

Implemented in MacClade but not in Paup.

Character-state illustrations

Illustrations of individual character states.

In DELTA, the association of illustrations with states is by means of human-readable 'subjects' only.

Character keywords

Named subsets of characters, not necessarily hierarchical.

Keywords are implemented in Intkey, but not in the current DELTA format.

Taxon names

Labels for taxa.

Unlimited name text

No limits on the length of taxon names.

'or' in attributes

An 'attribute' consists of the values of a particular character that are recorded for a particular taxon, and any associated information such as comments and probabilities. The values may be connected in various ways, the most

basic of which is 'or', i.e. the taxon may exhibit any or all of the specified values.

'to' in numeric attributes

Connecting numerical values by 'to', indicating a range of values.

'to' in multistate attributes)

Connecting multistate values by 'to', indicating that the taxon is intermediate between the specified values. Intermediate states (if any) are included for ordered multistate characters, but not for unordered multistate characters.

'and' in attributes

Connecting values by 'and', indicating that the values are simultaneously present in the taxon.

Unknown attributes

An indication that a character is unknown for the taxon. In DELTA, the attribute may simply be omitted, but may also be explicitly coded 'unknown'.

Inapplicable attributes

An indication that a taxon is inapplicable for the taxon, for cases not covered by character dependencies.

Attribute comments

Free-text information associated with attributes.

Value comments

Free-text information associated with particular values in attributes.

Implicit values

Missing attributes that take a default value, rather than being interpreted as 'unknown'.

Indefinite values

An indefinitely large or small value, as typically expressed in natural-language as 'many' or 'up to'.

Value is rare

A flag indicating that a value in an attribute is rare.

Value is misapplied

A flag indicating that a value in an attribute is not actually present in the taxon, but is likely to be thought to be present in error. Effectively a

special case of ‘value ‘only’’ (see below), as the purpose of this flag is to allow the value to be used in identification, to allow for common errors.

Value is unknown

A flag indicating that a value in an attribute is unknown, i.e. the value may possibly be present in the taxon or it may be absent.

In Nexus this is called ‘uncertainty’, and is indicated by enclosing the values in braces { } rather than parentheses (). However, it is not possible to indicate that some states are definitely present while others are uncertain.

Value probability

The probability that the value will be shown by a specimen of the taxon, i.e. the frequency of the value in the taxon.

Value ‘only’

Specifies that a character value is for use in specified applications only. It would be omitted from other applications. See Dallwitz (1999) for examples.

Value ‘not’

Specifies that a character value is not for use in specified applications. It would be omitted from these applications. See Dallwitz (1999) for examples.

Value ‘for’

Only the values so marked would be used in specified applications. Equivalent to ‘not’ on the other values. See Dallwitz (1999) for examples.

Value is approximate

The numerical value is approximate. Rendered as ‘about’ in natural language.

Value is guessed

A guessed value.

Attribute weights

Weights for individual attributes, to modify the overall character weights.

Attribute passed up

The attribute has not been entered directly, but has been generated by combining data entered for lower-level taxa (or specimens).

Attribute passed down

The attribute has not been entered directly, but has been inherited from data entered for a higher-level taxon.

Taxon notes

Free-text information associated with a taxon.

In Lucid, this can only be in the form of a separate file.

Implemented in MacClade but not in Paup.

Taxon illustrations

Association of illustrations with taxa.

Implemented in MacClade but not in Paup.

Taxon keywords

Named subsets of taxa, not necessarily hierarchical.

Keywords are implemented in Intkey, but not in the current DELTA format.

Typesetting markup

Markup to represent ‘italics’, ‘bold’, etc.

Nexus has limited capabilities, implemented in MacClade only.

Alternative natural languages

Alternative wordings for characters and comments in different natural languages (e.g. English, Spanish).

Alternative wordings

Alternative wordings for characters and comments, to meet the requirements of different applications (e.g. natural-language descriptions and keys).

Private notes

Comments to be seen only by the author of a data set.

(Character weights)

Character weights (or ‘reliabilities’) to be used by various applications, e.g. generation of conventional keys.

(Description layout)

Specification of the formatting of natural-language descriptions, e.g. headings, new paragraphs.

(Character transformation types)

The specification of the costs and rules imposed on specific state-to-state changes in parsimony or phenetic analysis (see Maddison et al. 1997, p. 618).

(Taxon trees)

Placement of taxa on the terminal nodes of trees.

Other Desirable Features**Structured names**

Information about the structure of taxon names, e.g. 'genus', 'species', 'variety'.

Character hierarchy

A tree structure with named nodes (headings), for organizing characters.

Taxon hierarchy

A taxonomic hierarchy, i.e. higher taxa that contain lower taxa.

Acknowledgements

Thanks to Susan B. Farmer for information about Paup and MacClade.

References

- Dallwitz, M.J. 1980. A general system for coding taxonomic descriptions. *Taxon* 29: 41–6.
- Dallwitz, M.J. 2000 onwards. Data requirements for natural-language descriptions and identification. <http://delta-intkey.com/www/descdata.htm>
- Dallwitz, M.J., Paine, T.A., and Zurcher, E.J. 1993a 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. 1993b onwards. New features for the DELTA System. <http://delta-intkey.com>
- Lucidcentral. 1999 onwards. Lucid home page. <http://lucidcentral.org>
- Maddison, W.P., and Maddison, D.R. 1992. MacClade: Analysis of phylogeny and character evolution. Version 3. 398pp. (Sinauer Associates: Sunderland, Massachusetts.)
- Maddison, D.R., Swofford, D.L., and Maddison, W.P. 1997. NEXUS: an extensible file format for systematic information. *Syst. Biol.* 46: 590–621.
- Swofford, D.L. 1991. PAUP: phylogenetic analysis using parsimony. Version 3.1. (Illinois Natural History Survey: Champaign.)