Standard data elements for classification

F. Christian Thompson

Classification and nomenclature are broadly defined to include those data elements useful not only for classification sensu stricto, but for making and documenting them. The data elements are clustered into three major groups: Characters, Classification, and Literature. Standards for data elements about biological associations are also treated here.

Classification Data Elements

Under this heading both nomenclatural and classification data are treated. For some databases, nomenclatural data are not necessary, but classification data are required for all databases as names form the "back-bone" of biological information. These data should conform to the minimal standards provided by the *International Code of Zoological Nomenclature*. Secondarily the standards used by the *Zoological Record* (BIOSIS) are followed.

Part I - Classification (Names)

Classifications are nothing more than lists of the correct names for taxonomic groups. To store and retrieve classifications, only TWO data elements are essential, the name and the name of the more inclusive group. For more formal classifications, the rank of the name may be desired. Some database models may represent classification data in a more rigid structure, defining separate structures for each formal level of the hierarchy, such as one for family, another for genus, others for species subfamilies, tribes, etc. However, by using modem database structures the classification a hierarchy of names can be collapsed into a single table with a self-relationship.

Classification data are inherently unstable. Classifications are really scientific theories hypotheses about relation ships among organisms. And there are different methodologies for translating such theories of relationships into hierarchical classifications. Therefore, there may be different views on the proper data for the following elements.

NAME	Essential. Name of the taxon. This is either a unique single word or a unique combination of two words (species).
RANK	Recommended. The category to which a valid name is assigned. Within each group of names, there may be two or more different hierarchical ranks (= categories, =levels). FAMILY group names may be of many different ranks (Superfamily, family, subfamily, supertribe, tribe, subtribe, etc.) GENUS group names may be of two different ranks (genus and subgenus). These are the only ranks recognized by the CODE. Systematists have, however, used additional "informal" levels in their classification (section, series, etc.). SPECIES group names may be of two different ranks species and subspecies. These are the only levels recognized by the CODE as part of a scientific name.
GROUP	Essential. The name of the taxon to which the name belongs. The precise placement of a taxon may not always be known. In this case, the incertae sedis convention should be used.
PHYLOGENETIC SEQUENCE	Optional. A number to allow names to be sorted by a phylogenetic sequence, instead of an alphabetic one.

Part II- Nomenclature (Name documentation)
The following data are fixed (static) in the sense they are determined by the CODE and the publication process. While not all the data may be available or agreed upon, proper use of the CODE (and Commission through its plenary powers) will eventually lead to permanent fixation of these data.

In zoological nomenclature, names are of three distinct groups, each having slightly different documentation requirements. These are: Family group; Genus group; and Species group.

Depending on the data model, the documentation for each group of names can be handled separately or all names treated together with a code used to indicate group of the name. Handling each group of names separately is probably the best approach as documentation requirements vary significantly between the groups.

Family Group Names:

Nomenclatural documentation for family group names is recommended.

SYNONYM Recommended. The family group name. Should be given in its original spelling. The use of the word synonym may be confusing. In some connotations, a synonym is viewed as the incorrect name. Synonym is used in a neutral sense of just a name. All correct taxonomic names have at least one synonym, which is their original form. Some taxonomic names may have two or more synonyms, in which case, the senior synonym is usually the correct taxonomic name and the junior synonyms are incorrect. Unique key; see Part III.

TAXONOMIC NAME Essential. Link to classification table.

TYPE Optional. Type of a family group name is a GENUS name. Optional as the family group name is formed from the name of the type genus, and a knowledgeable worker can determine this item from the name itself.

TYPE DOCUMENTATION
Optional.

None required. As the family group name is formed from the name of its type genus, there is no real need for documentation on typification. [However, it may be useful to give the stem from which the family groups names are formed.]

AUTHOR Recommended. Persons who is to be credited with the introduction of the name into scientific literature.

YEAR Recommended. Year in which the SOURCE (see below) was published. Ideally, this should be a year-month-day string. The CODE provides rules as how to fix the actual date of publication and given these rules precise dates can be generated for all names. [Uncertainty would be indicated by question-marks. So, when only the month and year are known, for example, the string would be 194404?? for April 1944. Given the ASCII collating sequence, this date will be greater than or sort after April 31 1944, etc.]

NB: The year (or publication date) should be a separate data element from author. Combining it with the author forces one to parse the

AUTHORITY field before doing logical operations (sort, comparisons, etc.). And the YEAR is a more important data element than is the author. For example, priority operates on the date, so one frequently wants lists ordered by date. Author is only part of a reference to the original source.

SOURCE Optional. Publication where the name was first noted in the sense of being "made available." Sub elements include title, serial source, volume, page, etc. In a complete database, this data element need only be a key pointer, etc. to the bibliographic citation.

If any data are given, it is recommended that at least the PAGE where the name first appeared be given. If the name appeared on more than one page in the original source, then the page where the most complete documentation is given should be cited. For example, a new name may appear in the table of contents, in a key, at the head of a description, in figure legends, and in the index. In this case, the page on which the description starts is recommended.

STATUS Recommended. Status of name. This may be simply:

AVAILABLE: Available for taxonomic use without qualification.
NOT ... : Not available or with special qualifications.

While there are many minor divisions, essentially names are either:

VALID, the correct name to be used for a taxon;
AVAILABLE, but not currently considered valid, that is, a name, given a different classification, could be valid, and
UNAVAILABLE, not a scientific name under the CODE, such as anincorrect spelling, etc.
HYBRID, a name ruled by ICZN as "unavailable for priority, but available for homonymy" Also, there are names which have "modified precedence."

Or a more informative code system could be used. At the Systematic Entomology Laboratory, we use a two digit code for status so that the various subclasses of status junior homonyms, incorrect original spellings, unjustified emendations, etc. can be recognized. These different subclasses are frequently treated differently typographically in printed catalogs.

1- = Available, valid
10 = Available, valid, not as below
12 = Available, valid, Not RECOGNIZED (nomen dubium)
15 = Available, valid, new status
16 = Available, valid, new combination
17 = Available, valid, new [replacement] name
18 = Available, valid, replacement name
2- = Available, invalid
20 = Available, invalid, junior synonym
22 = Available, invalid, dubious synonym
26 = Available, invalid, new (junior) synonym
27 = Available, invalid, unjustified new name
30 = Available, invalid, junior homonym
44 = Available, invalid, justified emendation
46 = Available, invalid, unjustified emendation
5- = Unavailable

50 = Unavailable, unspecified
55 = Unavailable, nomen nudum
56 = Unavailable, incorrect original spelling
57 = Unavailable, improper formation
58 = Unavailable, published in synonymy, not subsequently validated
60 = Unavailable, misspelling
70 = Unavailable, misidentification
80 = Unavailable, subsequent usage
etc.

Genus group names:

Nomenclatural documentation for genus group names is essential.

SYNONYM Essential. The genus group name. Should be given in its original spelling. Unique key; see Part III.

TAXONOMIC NAME Essential. Link to classification table.

TYPE Essential. Type of a genus group name is a SPECIES group name. Should be given in its original combination.

TYPE DOCUMENTATION

Essential. For genus group names documentation about typification is CRITICAL. The data elements that are needed are:

KIND of DESIGNATION -- two letter code is sufficient

[by original designation]
Original designation OD
Automatic AU

[by indication]
Typicus method TM
Monotypy MO
Tautonymy TT
Linnaean tautonymy TL

[by subsequent designation]
Subsequent designation SD
Subsequent monotypy SM

SOURCE of designation: For subsequent designations data are needed on when YEAR, where PUBLICATION SOURCE including PAGE and by whom AUTHOR. The YEAR, AUTHOR and PAGE elements are recommended, but the PUBLICATION SOURCE may be a key pointer, etc. to a bibliographic record or citation.

This arrangement reflects the current CODE. So, for a working database it is probably useful. However, it could be reduced to merely "fixed originally or subsequently," as the details of which kind of designation are only of interested to specialists.

AUTHOR See under Family group name.

YEAR See under Family group name.

SOURCE See under Family group name

ORIGINAL RANK Recommended. Whether the name was first used as a subgenus or not. This may be merely a logical field with the default condition being originally used as a genus. In those rare cases where two names were published simultaneously, the CODE states that the name which was used as a genus has priority over the one used as a subgenus.

STATUS See under family group name.

Species group names:

Nomenclatural documentation for species group names is essential.

SYNONYM Essential. The species group name. Should be given in its original spelling. Unique Key; see Part III

TAXONOMIC NAME Essential. Link to classification table.

ORIGINAL GENUS Recommended. The genus group name that was used with the species group name.

TYPE Recommended. Type of a species group name is a specimens or in special cases an interrelated group of specimens hapantotype. See below under type description.

TYPE DOCUMENTATION

Recommended. For species group names documentation about typification is desired [the present CODE does not require typification for species group names, but does provide rules for their typification]. The data elements that are needed are;

KIND of DESIGNATION - two letter code is sufficient

[by original designation]
HOLOTYPE **HT**
SYNTYPES **ST**

[by subsequent designation]
LECTOTYPE **LT**
NEOTYPE **NT**

[NO designation]
SYNTYPES **ST**

SOURCE of designation: For subsequent designations data are needed on when (YEAR), where (PUBLICATION SOURCE including PAGE) and by whom (AUTHOR).

TYPE LOCALITY: While it is not part of typification, the type locality provides useful data for systematists and therefore should be captured. Again this arrangement reflects the current CODE and different arrangements are possible. A simpler arrangement for species group names would merely to state kind of type (Hapanto-, Holo-, Lecto-, Neo-Syn-, etc.).

AUTHOR Recommended. As under family group names.

YEAR Recommended. As under family group names.

SOURCE Recommended. As under family group names.

ORIGINAL RANK	Recommended. Whether the name was first used as a subspecies or not. This may be merely a logical field with the default condition being originally used as a species. In those rare cases where two names were published simultaneously, the CODE states that the name which was used as a species has priority over the one used as a subspecies. Also, whether a name was used as a "variety," "form," "morph," etc., should be recorded as this datum may be used to determine whether the name is available that is, whether it is a scientific name in the sense of the CODE.
STATUS	As under family group name.

Part III- Data Structures.

UNIQUE data elements (KEYs)

Different data structures are possible for these nomenclatural data. These data structures, in part, depend upon what assumptions one makes about stability of the data and inter-relationships among the data elements. However, whatever data structure is used, given the complexity of data (in the sense of being a combination of FIXED and VARYING data) keys must be used to link the different data groups (tables, files, etc.). For efficiency, KEYs must be unique. **UNIQUENESS is guaranteed for correct names** (or those that may potentially be correct names [=available names]) by the *CODE*. However, synonyms, unavailable names, etc. may be homonymous. So, to link nomenclatural data, homonyms needed to be made unique.

Uniqueness:

For family group names, the name itself must be UNIQUE.

[As family group names may take different endings depending on the hierarchical level one assigns them to, the unique key for a family group name should be made using a standard family level ending -idae. For example, the subtribal name Xylotina was first introduced for a tribe (based on the genus *Xylota*) and has been used as a subfamily name (Xylotinae). However, the unique key to nomenclatural data about this name would be Xylotidae. This is critical not only because the level category and hence the ending of the name may vary according to ones classification, but the ends for some hierarchical levels subtribe may generate a name identical to a genus group name (=their key). The subtribal form for *Xylota*, Xylotina, is identical to the genus group name *Xylotina*.]

For genus group names, the name itself must be UNIQUE.

For species group names, the valid combination, as well as the original combination, must be UNIQUE. For subspecies, the combination of the genus and subspecies names must be unique. So, the maximal number of words for a taxonomic key is two. The longest taxonomic name known to me is 44 characters and the longest potential taxonomic name would be 68 characters that is, the longest known genus group name (31 characters) plus the longest known species group name (37 characters) (see Thompson, 1986, Antenna 10: 6-7.)

To make homonyms unique, I recommend that YEAR or publication date be appended to the junior homonyms. Hence, the maximal number of digits that need to be added to a junior homonym is 7, but the senior homonym and the available and/or VALID remains unchanged. Also, digits are easily stripped from the junior homonyms to reveal the actual name. So, for example:

Unique KEY	
Noctua	for *Noctua Linnaeus* 1758 of insects
Noctual1771	for *Noctua Gmelin* 1771 of birds
Musca heraclei	for *Euleia heraclei* Linnaeus 1758
Musca heraclei1795	for *Musca heraclei* Fabricius 1795 now known as *Tephritis poslica* Loew 1884

The problem with "unique identifying numbers," such as BIOSIS use of TRPNUM, is that one needs a central organization to do the assigning, etc., or else one has chaos. And such requirements bring along many additional problems or at least perceptions of problems [control, etc.]. Also, numbers are not "user friendly." Why should users be burdened with a number for *Noctua* when the name itself is a UNIQUE combination that a computer system can use as well as a number [all are currently translated into binary representation anyway!]. The real beauty of this is that users DO NOT need numbers for available names for the name itself is it KEY!

Literature Data Elements

Literature data elements are of two functional groups: Citations and Bibliographic References. Citations are the linkage between bibliographic references and lots, specimens, and/or names. Bibliographic reference is all the data necessary to describe a publication and allow for its retrieval. Many standards exist for bibliographic data references, and a number are approved ISO/ANSI standards. These library and abstracting journal BIOSIS standards should be used, rather than generating new ones. Only the critical minimum data elements necessary to find references are given below.

Citation:

AUTHOR **DATE** **SOURCE**	The above three data elements should be included or any unique link to the bibliographic reference can be used instead of them.
PAGE	Page or specific location within the publication
CONTENTS	Nature of
CITATION	A unique key to identify the citation.
TAXONOMIC NAME	Taxonomic name or any unique link to classification, lots or association. Two names may be used if a full database is built. One name would be the current correct name which links the citation to classification and is always required. The second name would be the name used in the publication, which may be an incorrect synonym, misidentification, etc., and would link the citation to nomenclature.
GEOGRAPHY	Location data or any unique link to geography

Reference:

AUTHOR **DATE** **TITLE** **SOURCE** **COLLATION** **ANNOTATIONS**	
[Key]	A unique identifier to provide linkage to other files. This key could be built from the AUTHOR, DATE and SOURCE elements.

Associate Data Elements

In biology, there are many types of associations between species, such as one species eating another (host-parasite, predator-prey, etc.) All these associations can be viewed as one to one relationships see figure, and can be reduced to three basic data elements (the two actors and what they do together!).

SUBJECT NAME Taxonomic Name; Link to classification

ASSOCIATE NAME Taxonomic Name; Link to classification

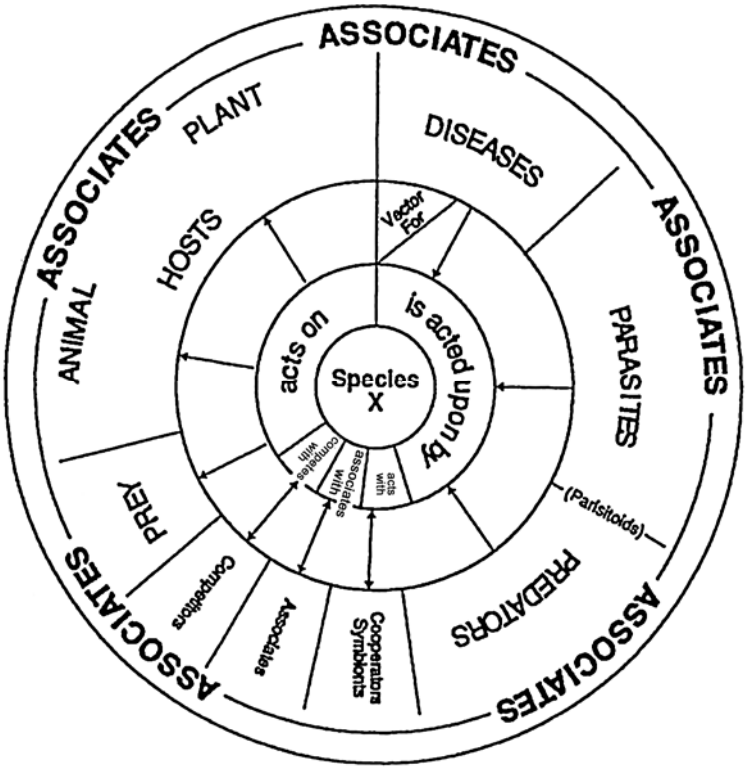
Two sets of names may be required, if nomenclature data is maintained. One set would be the correct valid names which link to classification, and the other set being the actual names used on the specimens, in the citation, etc., which may be incorrect synonyms.

LOCALITY Link to geography (SITENO)

CITATION Link to bibliography, if based on literature citation

LOT NUMBER Link to lots, if based on specimens

RELATIONSHIP What is the relationship between the subject and associate expressed in terms of the SUBJECT. That is, for entomologist working on fruit fly, the subject (a fly), the relationship with a plant (associate) would be that of HOST.



MODE OF ACTION What the subject is actually doing to the associate. For example, for the fruit fly this may be mining within the leaves of the plant, forming a gall in the flower, etc.

PART OR STAGE AFFECTED As the associate may be a complex organism, this data element more precisely defines the part or stage acted upon by the subject. For the fruit fly, this may be the leaves or flowers.

MODE OF COLLECTION How was the association discovered, that is, how was the association collected. For the fruit fly, this may be rearing of the larvae to the adult stage.

RELIABILITY	Assessment of the reliability of the identification ofboth the subject and associate should be recorded.
NOTES	Spaces for textual discussion of the nature of the association and/or mode of action. A standard vocabulary should be used for the data elements above (RELATIONSHIP, MODE OF ACTION, MODE OF COLLECTION, PART OR STAGE AFFECTED), whereas free style text should be permitted at the end of the record.

Character data elements

While the actual data elements for characters are few, there are many different approaches to encoding characters as the storage requirements and how the characters are analysed and used vary according to one of the data elements TYPE. Standards for character data, such as DELTA, do exist and should be carefully studied before new standards are developed.

Characters:

CHARACTER	Description of character
STATE	Description of the state of the character. Not always necessary as some types of character may have implied states numerical types.
TYPE	Type of character binary, ordered & unordered multistate, discrete and continuous numerical.

Operational Taxonomic Unit (OUT):

VALUE	Value of the character state.
SPECIMEN NUMBER	A unique Key
LOTNO	Link to GEOGRAPHY, ECOLOGY, etc.
TAXONOMIC NAME	Link to classification