The Catalogue of Life Standard Dataset

Version 6.3, August 2012

The Catalogue of Life plans to deliver a standard set of data for every known species. This document presents a simple description of the standard dataset which is both the core knowledge set of which the Catalogue of Life is composed and around which processes and protocols are designed. This standard dataset is used in many contexts and includes minimum content of data transmitted between components of the programme, and the minimum content of data transmitted in public products. These data are drawn from an array of participating taxonomic databases: Global Species Databases (GSD) - databases containing worldwide coverage of all the species within one taxon - or Regional Species Databases (RSD). In this document we will use the name 'Source Database' for both GSDs and RSDs.

The Catalogue of Life has defined **13 field groups to be the standard set of data** for each species (or infraspecies).

1. **Accepted Scientific Name** linked to **Reference(s)** (obligatory)
2. **Synonym(s)** linked to **Reference(s)** (obligatory, where available)
3. **Common Name(s)** linked to **Reference(s)** (obligatory, where available)
4. **Classification above genus, and up to the highest taxon in the database** (obligatory, where available)
5. **Distribution** linked to **Reference(s)** (obligatory, where available)
6. **Life zone** (obligatory, where available)
7. **Additional Data** (optional)
8. **Latest taxonomic scrutiny** (obligatory)
9. **Reference(s)** (obligatory, where available)
10. **Taxon Globally Unique Identifier** (obligatory, where available)
11. **Name Globally Unique Identifier** (obligatory, where available)
12. **Source Database** (obligatory)
13. **Catalogue of Life LSID** (obligatory)

Some of the source databases additionally supply subspecies or varieties. The same dataset is used for each of these. Also, all information from field groups # 1, 2, 3, 5, 6, 7, 8, 9, 10, 11 & 12 from infraspecific taxa should be given for both the species and infraspecific taxa (i.e. the 'replicated' system of TDWG Plant Names Standard).

Additional information is available either within the appropriate Source Database, or through hyperlinks to other databases.

1. Accepted Scientific Name
(obligatory)

The Accepted, Valid or Correct scientific name (terminology for this name varies between the Codes of Nomenclature, in the Catalogue of Life we use the term 'Accepted') currently accepted for the species as a taxon. There should be exactly one per species. Two variants of NameStatus are possible in databases: 'Accepted name' or 'Provisionally accepted name'.

'**Accepted name**' is the name currently accepted for the species by the compiler or editor of dataset as a quality taxonomic opinion.

'**Provisionally accepted name**' is the name currently accepted for the species by the dataset compiler, but with some element of taxonomic or nomenclatural doubt.

Content:

a) Accepted Name of species

Genus | SubGenusName (where appropriate) | Species | AuthorString | Sp2000NameStatus | Reference(s) (obligatory)

b) Accepted Name of infraspecific taxon
only subspecies for taxa under ICZN; only subspecies, varieties and forms for taxa under ICBN:

Genus | SubGenusName | Species | AuthorString | InfraspeciesMarker (where appropriate)| InfraspeciesEpithet | InfraspeciesAuthorString | Sp2000NameStatus | Reference(s) (obligatory)

In the case of Virus Names (i) the Genus is placed in the Genus field, and (ii) the polynomial species name is placed in the species epithet field. Virus species names have no official author.

Where:

Genus	= Latin genus name.
SubGenusName	= Latin subgenus name.
Species	= second part of species name, Latin epithet
AuthorString	= name of author(s), who described this species or published current combination (Style of authorstring depends on nomenclatural practices under different Codes)
InfraspeciesMarker	= marker of infraspecific rank, where appropriate following Code regulations, for example, subsp., var., f. for plants. (Presence and style of infraspecific markers depends on nomenclatural and taxonomic practices under different Codes)
InfraspeciesEpithet	= third part of trinomial name, Latin epithet

InfraspeciesAuthorString	= name of author(s), who described this infraspecific taxon or published current combination (Style of authorstring depends on nomenclatural practices under different Codes; it could include year where appropriate)
Sp2000NameStatus	= the Catalogue of Life name status translated from source database: Accepted or Provisionally Accepted
Reference(s)	= just one reference that contains the original (validating) publication of taxon name or new name combination – Nomenclatural Reference, *or* one or more references that accept this species in the same taxonomic status, and with the same name – Taxonomic Acceptance Reference(s)

Example: `Acacia | sieberiana | DC. | Accepted name | ReferenceID`

(Accepted name record for Acacia sieberiana extracted from ILDIS database)

2. Synonym(s)
(obligatory, where available)

The list of Synonyms can include from 0 to many species or infraspecific names, which are given the Catalogue of Life synonymic status (Sp2000NameStatus). The three possibilities give the information sufficient for clear synonymic indexing, but do not give the full nomenclatural details, as these differ markedly in structure and context across different Codes. It is therefore necessary to 'translate' the very varied sorts of synonymic status in the source databases to create a uniform, accurate, but broad set of synonymic links for use in the Catalogue of Life.

(Category A) List of "**Synonyms**" - names which point unambiguously at one species (synonyms, in the CoL sense, include also orthographic variants and published misspellings)

(Category B) List of "**Ambiguous synonyms**" - names which are ambiguous because they point at the current species and one or more others e.g. homonyms, pro-parte synonyms (in other words, names which appear more than in one place in the Catalogue).

(Category C) List of "**Misapplied names**" - names that have been wrongly applied to the current species, and may also be correctly applied to another species.

Some synonyms of species can be trinomials, and have taxonomic rank of subspecies (in zoology), or subspecies, variety and form (in botany).

Content: Genus | SubGenusName (where appropriate) | Species | AuthorString | Sp2000NameStatus | Reference(s) (obligatory)

or for trinomial synonyms (subspecies and varieties):

Genus | SubGenusName (where appropriate) | Species | AuthorString | InfraspecificMarker | Infraspecies | InfraspecificAuthorString | Sp2000NameStatus | Reference(s) (obligatory)

Where:

Genus	= as above
SubGenusName	= as above
Species	= as above
AuthorString	= as above
Sp2000NameStatus	= the Catalogue of Life synonym status translated from source database: Unambiguous Synonym, Ambiguous Synonym, Misapplied Name
Reference(s)	= as above

Examples:

```
Acacia | purpurascens | Vatke | Misapplied name | ReferenceID

Acacia | sieberiana | DC. | subsp. | vermoesenii | (De Wild.)Troupin | Unambiguous synonym | Refs.#,#,#

Acacia | abyssinica | sensu auct. | Misapplied name | ReferenceID
```

(Synonym records for Acacia sieberiana extracted from ILDIS database)

3. Common Name(s)

(obligatory, where available)

List of Common Names can include from 0 to many names.

Content: CommonName | TransliteratedName | Country | Area (optional, where appropriate) | Language | Reference(s)

Where:

Common name	= one-word or multi-word name in original script (if available; if name in original script is not available, go for transliterated name described in the next field)

Transliteration	= a single text string in roman characters free from any diacritics or other symbols other than numbers and some punctuation (ASCII): - EITHER: Transliteration of the Common Name (in the original Common Name field) into Roman alphabet without diacritics (into this field). - OR: Repeat entry of the Common Name itself, if already in Roman alphabet without diacritics. - OR: Directly supplied Transliteration (into this field), but *without* the original name in a non-roman script (in the original Common Name field).
Country	= country, where this name is in use
Area	= local geographical area within megadiverse country, where this name is in use
Language	= language of the common name
Reference(s)	= list of source references

Example: `Landlocked salmon | Canada | English | ReferenceID`

(Common name record for Salmo salar extracted from FishBase)

4. Classification above genus, and up to the highest taxon in the source database
(obligatory, where available)

The Catalogue of Life has decided to use a single taxonomic classification (also called a hierarchy) for management purposes – the management classification. The current classification in use is "The Catalogue of Life Taxonomic Classification, Edition 2, Part A". It is is regularly updated (http://www.catalogueoflife.org/testcol/info/hierarchy). This decision does not preclude future technical developments that would make other classifications available for linkage with the same species checklists.

The Catalogue of Life uses the current management classification **above** the node of attachment of each database. **Beneath** this node it uses the classification provided by the GSD. Where Global Species Databases, or GSD Sectors (that is Sectors rather than the whole) are used, each GSD or GSD Sector is linked at one node in the classification. The taxonomic rank of the highest taxon at this attachment node varies from one GSD to another (e.g. sector of Conifer Database is attached as phylum, sector of Cercopoidea Organised On Line is attached as superfamily, sector of ILDIS World Database of Legumes is attached as one family). The Catalogue of Life requires each GSD to indicate the highest taxon that is given in the GSD, and to provide the classification beneath it down to species level.

The Catalogue of Life management classification includes taxa of seven basic ranks only: **Kingdom** –**Phylum** – **Class** – **Order** – **Superfamily** – **Family** – **Genus**.

Content: Kingdom | Phylum | Class | Order | Superfamily | Family | Genus

**Incertae sedis* or *not assigned* taxa are also allowed in ranks of phylum, class, order, superfamily and family, but not in ranks of kingdom and genus.

Plus, Catalogue of Life Taxon LSID with every taxon in the classification.

Where:

Kingdom	= Latin scientific name of the kingdom that includes the specified phyla
Phylum	= Latin scientific name of the division or phylum that includes the specified classes
Class	= Latin scientific name of the class that includes the specified orders
Order	= Latin scientific name of the order that includes the specified families or superfamilies for insects
Superfamily	= Latin scientific name of the superfamily that includes the specified families (for insect groups only)
Family	= Latin scientific name of the family that includes this species. If the taxon is not known then this must be stated (e.g. family labelled *incertae sedis* (not assigned) in taxonomic treatments) and the next higher taxon must be given with its rank.
CoL LSID	= CoL Taxon Matcher software issues permanent CoL Global Unique Identifiers at the stage of optimisation of CoL database for every taxon recognised in the Catalogue of Life using the Life Science Identifier (LSID) system (http://sourceforge.net/projects/lsids).

Example:

```
Plantae (kingdom) | Rhodophyta (phylum) |
Rhodophyceae (class) | Bangiales (order) | Bangiaceae
(family) | Phyllona carnea (species)

Plantae LSID
urn:lsid:catalogueoflife.org:taxon:d755b8fe-29c1-
102b-9a4a-00304854f820:col20121017
```

(Example extracted from AlgaeBase)

5. Distribution
(obligatory, where available)

Field Group contains three fields: i) Area system or systems used, ii) For each area system used, List of zero to many Areas of Occurrence, and for each Area of Occurrence, iii) Status in that Area

Content: DistributionElement | StandardInUse | DistributionStatus

DistributionElement = 3-letter code or Name of an Area using one of the agreed Area systems in use:
- for the land areas of the world: Updated TDWG Level 4 Areas (preferred), or ISO 3-letter country codes.
- for the sea areas of the world: Intersect of IHO's and EEZ areas (see: VLIZ (2010), Intersect of IHO Sea Areas and Exclusive Economic Zones (v5, 2009). Available online at http://www.vliz.be/vmdcdata/vlimar/downloads.php)

AreaStandard Short name for each Area System in use (from a dictionary provided). Area systems in use should be limited to an agreed set, *e.g.* TDWG Level 4 code, TDWG Level 3 name, FAO_ISO Code, or Text when providing free text description

DistributionStatus = multi-state descriptor code. Score for multiple states where more than one applies. Proposed codes (fixed, non-extensible)
N = Native
D = Domesticated
A = Alien
U = Uncertain

Example: `Botswana | TDWGL4 | Native`

(Distribution record of Acacia sieberiana extracted from ILDIS database)

n.b.
i) The proposed Catalogue of Life Standard does *not* include a source reference for this data. The source is effectively the source database.
ii) However, it is recommended to GSDs, as part of the 'best practice', that the GSD *does* record the source reference linked to each Area of Occurrence record.
iii) Where the source GSD is accessible on the web, this source reference data, which may be quite extensive, is available to CoL users by clicking

6. Life zone
(obligatory, where available)

A single multi-state descriptor field, for which multiple scores can be recorded. Scores are recorded by the Source Database custodian using expert knowledge *without* recording source or reference in

the Catalogue of Life. Descriptor states are fixed and non-extensible. The descriptor states are: **Marine**, **Brackish**, **Freshwater**, **Terrestrial**, **Unknown** (n.b. These field states follow the GISIN standard titled as 'Realm' by GISIN.). Scoring Rules (but these can be changed if needed):

i) Species or Infraspecies occupying more than one state are recorded for several states.

ii) Species occupying parasitic, epiphytic or other conditions dependent on another organism are scored as the state(s) appropriate to that other organism.

iii) There will need to be a set of guide notes on what to score under each state, and of how to deal with difficult cases, such as organisms in ground water, mosquito larvae in water pools up trees in epiphytes etc...

Content:	LifeZone	
Where:	LifeZone	A single multi-state descriptor according to the fixed and non-extensible following coding (interoperable with the GISIN standard titled as 'Realm' by GISIN.): • Marine • Brackish • Freshwater • Terrestrial • Unknown
Example:	`Freshwater, Terrestrial`	

7. Additional Data

(optional)

This field can contain free text up to 255 characters. It can contain information from one or several data fields from the source database (for example, type specimen, taxonomic comments, common name of the family, habit/life form, detailed ecology, host, etc.) as decided by the custodian of the source database. Unlike all other field groups, there is no intention to make these data compatible across taxa. It can therefore be distinctive or particular to the species supplied by one database.

Content:	AdditionalData *(Free text, which might be structured with headings)*	
Where:	AdditionalData	= Free text up to 255 characters
Example:	`Type strain: strain ATCC 33244 = CFBP 3612 = CIP 105207 = ICPB EA175 = LMG 2665 = NCPPB 1846 = PDDCC 1850, "Pantoea ananatis corrig. (Serrano 1928) Mergaert et al. 1993, comb. nov."` *(Additional Data record for Erwinia ananatis extracted from BIOS database)*	

8. Latest taxonomic scrutiny

(obligatory)

This field group should contain only one record of the Latest Taxonomic Scrutiny (LTS) of the species or infraspecific taxon in the source database. LTS includes (a) name(s) of the taxonomic expert or editor, who is responsible for the taxonomic concept accepted in the source database and (b) date when the expert or editor assessed the record. If the source database has multiple records, just the most recent should be passed to the Catalogue of Life. If the source database has no latest taxonomic scrutiny records, but is the work of one specialist or a small team, then for the whole database this field should show the scrutiny by that specialist or small team. Users of the Catalogue of Life content are obliged to cite the name of LTS specialist with each species taken from the Catalogue.

Content:	LTSSpecialist | LTSDate	
Where:	LTSSpecialist	= surname of taxonomic editor, initial(s)
	LTSDate	= date of record scrutiny (revision) in the source database; style specified by the custodian of Source Database; 'Year' is obligatory; 'Month' & 'Day' might be applied where available.
Example:	`Rico, M. L. | 1994`	
	(Scrutiny record for Acacia sieberiana extracted from ILDIS database)	

9. Reference(s)
(obligatory, where available)

References should be linked with Accepted Scientific Names, Synonyms and Common names.

Content:	Author(s) | Year | Title | Details | Reference Type	
Where:	Author(s)	= Author (or many) of publication
	Year	= Year (or some) of publication
	Title	= Title of paper or book
	Details	= Title of periodicals, volume number, and other common bibliographic details

	Reference Type	= Taxonomic status of reference: Nomenclatural Reference NomRef (just one reference which contains the original (validating) publication of taxon name or new name combination *or* Taxonomic Acceptance Reference(s) TaxAccRef (one or more bibliographic references, where the name is mentioned in the same taxonomic status (i.e. as a species or as a synonym) *or* Common Name Reference(s) ComNameRef (one or more bibliographic references that contain common names)

Example: `Ross, J.H. | 1979 | A conspectus of African Acacia | Mem. Bot. Surv. S. Afr. 44: 1-150 | TaxAccRef`

(Reference record extracted from ILDIS database)

10. Taxon Globally Unique Identifier (obligatory, where available)

This field is provisional in the Catalogue of Life; it is intended for future use to support an interoperability between Source Databases, CoL and its corporate users.

Content: GSDTaxonGUID

Where:	GSDTaxonGUID	A single text field containing the single Globally Unique Identifier supplied for every taxon at species or infraspecies rank by the Source Database supplying this taxon as a concept. Taxon GUID suppose to reflect changes in the accepted concept.

Examples:

Notes

i) This is for future use, once a pattern of Source Databases supplying Taxon Globally Unique Identifiers are established.
ii) At present this leaves open the possibility that the Identifiers may be LSIDs or some other form of unique identifiers (for example, if Source Database does not provides Taxon GUIDs, the Catalogue of Life may populate unique identifiers which are in Source Database use)..
iii) It is assumed that the unique identifier received does already contain some provenance indication, as is the case with LSIDs.
iv) This field is intended for unique identification and provenance metadata tracking and possible use in the Catalogue of Life web-services. It is not necessarily for display in the public interface.

11. Name Globally Unique Identifier (obligatory, where available)

Content: GSDNameGUID

GSDNameGUID	One or more text strings containing one or more Unique Identifiers supplied for a single name at species or infraspecific rank by the Source Database supplying this name.

Examples: `urn:lsid:ipni.org:names:30000959-2`

(A plant scientific name from IPNI)

`urn:lsid:indexfungorum.org:Names:213649`

(A scientific name of a fungi from Index Fungorum)

Notes

i) This is for future use, once a pattern of Source Databases supplying Name Globally Unique Identifiers are established.
ii) This field is intended for unique identification and provenance metadata tracking and possible use in the Catalogue of Life web-services. It is not necessarily for display in the public interface.
iii) There is no single authority for issuing unique identifiers for names of all organisms, although there is a perception that this task will be undertaken by existing the Nomenclator organisations. It is possible that we may receive more than one unique identifier per name, and may wish to store all of those supplied.

12. Source Database

(obligatory)

This information (metadata) will be supplied for each source database only once, but it will be shown as a part of every record in the Catalogue of Life.

Content: TaxonomicCoverage | DatabaseShortName | DatabaseFullName | DatabaseVersion | ReleaseDate | HomeURL | LogoFileName | ContactPerson | AuthorsEditors | Organisation | GroupNameInEnglish | Abstract | Coverage | Completeness | Confidence

Where:	TaxonomicCoverage	= Higher taxon(s), which represent taxonomic sector(s) covered by the database

DatabaseShortName	= Abbreviated or shortened memorable name of Source Database which intended for easy use in day-to-day communications; as supplied by the custodian
DatabaseFullName	= Full title of Source Database; as supplied by the custodian
DatabaseVersion	= Database version (number or code, plus date, where Month and Year are obligatory) provided by the custodian; style specified by the custodian of Source Database
ReleaseDate	= Original date (Year-Month-Date) of issue of the version for the Catalogue of Life.
HomeURL	= Uniform Resource Locator (Internet address) of Source Database home page
LogoFileName	= Name of the file containing the Source Database logotype. Technical requirements include minimum image size as 10x10 mm with 300 dpi
ContactPerson	= Name of contact person for the Source Database and email address; as specified by the custodian; for internal use by the CoL Secretariat only; this information will not be released to the public
AuthorsEditors	= Name(s) of Source Database editor or author; as specified by the custodian
Organisation	= Name of the Organisation that hosts the Source Database
GroupNameInEnglish	= English Name of the taxon covered by the Source Database
Abstract	= Standardised short database description (text of 50-70 words) for use in the Catalogue of Life supporting materials, such as the booklet published with the Annual Checklist on DVD

Coverage	= geographic coverage of the Source Database for the taxon – Global for a worldwide coverage, and – Regional for a geographically restricted coverage. If Regional, the region should be specified in brackets
Completeness	= percentage of completeness of species list of the taxon provided by the Source Database
Confidence	= quality of taxonomic checklist with values 1 to 5; quality is stated by the custodian. Confidence indicators are as follows: 1 - This database does not contain scrutinised taxonomic checklist. However, it is taken temporary to the Catalogue of Life to fill major gaps as only available source in the time. 2 - Taxonomic database is at an early stage of its development. 3 - Sector specialist database of high quality. It is taken to the Catalogue of Life to fill gaps on low classification levels (e.g. species, genera). 4 - Taxonomic database with good quality of expertise at the stage of its development. 5 - Taxonomic database with high quality of expertise with frequent updates covers nearly all known diversity worldwide.

Example: `Cercopoidea` | COOL | Cercopoidea Organised On Line | 4, Sep 2011 | 2011-09-07 | http://rameau.snv.jussieu.fr/cool/ | `http://www.catalogueoflife.org/images/databases/COOL.png` | Soulier-Perkins A. | Soulier-Perkins A. | Muséum National d'Histoire Naturelle, Paris, France | Froghoppers | COOL is a systematic...| `Global | 94 | 4`

(Source Database record extracted from COOL database)

13. Catalogue of Life Taxon LSID (obligatory)

This field is reserved for CoL Taxon Life Science Identifiers. CoL Taxon Matcher software issues

permanent CoL Global Unique Identifiers for every taxon recognised in the Catalogue of Life at the stage of optimisation of CoL database using the Life Science Identifier (LSID) system (http://sourceforge.net/projects/lsids). Life Science Identifiers are persistent, location independent and unique identifiers for specific data in the web. They should serve to trace changes in the CoL records.

Content:	CoL Taxon LSID	
Where:	CoLTaxonLSID	A single text field containing the single Globally Unique Identifier supplied for every taxon at species or infraspecies rank by the Source Database supplying this taxon as a concept. Taxon GUID suppose to reflect changes in the accepted concept.
Examples:	`Species Zorka angelinae` CoL Taxon LSID: urn:lsid:catalogueoflife.org:taxon:653bc0d9-f89c-11e0-af7a-6c3dedecd876:col20121017	