# Package 'Taxonstand'

November 2, 2017

**Type** Package

**Title** Taxonomic Standardization of Plant Species Names

**Version** 2.1

**Date** 2017-11-02

**Author** Luis Cayuela, Anke Stein, Jari Oksanen

**Maintainer** Luis Cayuela <luis.cayuela@urjc.es>

**Depends** pbapply (>= 1.3-2)

**Description** Automated standardization of taxonomic names and removal of orthographic errors in plant species names using 'The Plant List' website (www.theplantlist.org).

**License** GPL (>= 2)

**LazyLoad** yes

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2017-11-02 17:10:42 UTC

## R topics documented:



---

`bryophytes` *List of bryophytes*

---

### Description

`bryophytes` contain species names for 100 Mediterranean bryophytes

## Usage

```
data(bryophytes)
```

## Format

A data frame with 100 observations on the following 5 variables.

`Full.name` a character vector

`Genus` a character vector

`Species` a character vector

`Var` a character vector

`Intraspecific` a character vector

## Examples

```
data(bryophytes)
str(bryophytes)
```

---

# TPL *Standardize plant names according to The Plant List.*

---

## Description

Connects to The Plant List (TPL) website and validates the names of a vector of plant taxa, replacing synonyms by accepted names and removing orthographical errors in plant names.

## Usage

```
TPL(splist, genus = NULL, species = NULL, infrasp = NULL,
infra = TRUE, corr = TRUE, diffchar = 2, max.distance = 1,
version = "1.1", encoding = "UTF-8", author = TRUE,
drop.lower.level = FALSE, file = "", silent = TRUE, repeats = 6)
```

## Arguments

| | |
|---|---|
| `splist` | A character vector specifying the input taxa, each element including genus and specific epithet and, potentially, author name and infraspecific abbreviation and epithet. |
| `genus` | A character vector containing the genera of plant taxon names. |
| `species` | A character vector containing the specific epithets of plant taxon names. |
| `infrasp` | A character vector containing the infraspecific epithets of plant taxon names. |
| `infra` | Logical. If TRUE (default), infraspecific epithets are used to match taxon names in TPL. |
| `corr` | Logical. If TRUE (default), spelling errors are corrected (only) in the specific and infraspecific epithets prior to taxonomic standardization. |

`diffchar` A number indicating the maximum difference between the number of characters in corrected and original taxon names. Not used if corr = FALSE.

`max.distance` A number indicating the maximum distance allowed for a match in `agrep` when performing corrections of spelling errors in specific epithets. Not used if corr = FALSE.

`version` A character vector indicating whether to connect to the newest version of TPL (1.1) or to the older one (1.0). Defaults to "1.1".

`encoding` Encoding to be assumed for input strings from TPL website; defaults to "UTF-8" (see `read.csv` and `Encoding`).

`author` Logical. If TRUE (default), the function tries to extract author names from the input taxon (see **Details**).

`drop.lower.level` Logical. If TRUE, the variety is dropped from the input taxon if both subspecies and variety are given, and the forma is dropped from the input taxon if both subspecies or variety and forma are given. If specific and subspecific epithet are identical, the subspecies [variety] part is dropped instead, and variety [forma] is kept. Defaults to FALSE.

`file` Either a character string naming a file or a connection open for writing. "" (default) indicates output to the console.

`silent` Logical. If FALSE, the function prints the taxon name that is currently processed in the output. Defaults to TRUE.

`repeats` A number indicating how many times `TPLck` should be called if no connection to TPL website can be established (temporarily).

## Details

The procedure used for taxonomic standardization is based on function `TPLck`. A progress bar indicates the proportion of taxon names processed so far. In case the TPL website cannot be reached temporarily, the function returns an error but repeats trying to match the given taxon multiple times (see `repeats`). If standardization is still not successful, the input taxon is returned in field 'Taxon' with NA in all other fields.

## Value

A `data.frame` with the following components:

`$Taxon` Original taxon name as provided in input.

`$Genus` Original genus name as provided in input.

`$Hybrid.marker` Hybrid marker, if taxon is indicated as hybrid in the input.

`$Species` Original specific epithet as provided in input.

`$Abbrev` Original abbreviation other than infraspecific rank included in input taxon, including "cf.", "aff.", "agg.", "nom. cons.", "nom. cons. prop.", "nom. inval.", "s.l.", and "s.str." and their orthographic variants.

`$Infraspecific.rank` Original infraspecific rank abbreviation as provided in input, including "subsp.", "var.", "f.", and their orthographic variants.

`$Infraspecific` Original infraspecific epithet as provided in input. If infra = FALSE, this is not shown.

`$Authority` Original author of taxon name as provided in input.

`$ID` The Plant List record ID of the matched taxon before resolving synonyms.

`$Plant.Name.Index`
Logical. If TRUE the name is in TPL. If a taxon at infraspecific level is not in TPL, Plant.Name.Index equals FALSE, except for nominal infraspecies. Also compare `Higher.level`.

`$TPL.version` Version of TPL used.

`$Taxonomic.status`
Taxonomic status of the matched taxon in TPL, either 'Accepted', 'Synonym', 'Unresolved', or 'Misapplied'.

`$Family` Family name, extracted from TPL for the valid form of the taxon.

`$New.Genus` Genus name, extracted from TPL for the valid form of the taxon.

`$New.Hybrid.marker`
Hybrid marker, extracted from TPL for the valid form of the taxon.

`$New.Species` Specific epithet, extracted from TPL for the valid form of the taxon.

`$New.Infraspecific.rank`
Infraspecific rank abbreviation, extracted from TPL for the valid form of the taxon, including "subsp.", "var." and "f.".

`$New.Infraspecific`
Infraspecific epithet, extracted from TPL for the valid form of the taxon.

`$New.Authority` Author of taxon name, extracted from TPL for the valid form of the taxon.

`$New.ID` The Plant List record ID of the taxon, once synonyms have been replaced by valid names. For accepted and unresolved names, this field will be equivalent to `ID`.

`$New.Taxonomic.status`
Taxonomic status of the resolved taxon in TPL, once synonyms have been replaced by valid names. 'Accepted' or 'Unresolved'.

`$Typo` Logical. If TRUE there was a spelling error in the specific or infraspecific epithet that has been corrected.

`$WFormat` Logical. If TRUE, fields in TPL had the wrong format for information to be automatically extracted as they were not properly tabulated or, alternatively, there was not a unique solution.

`$Higher.level` Logical. If TRUE, the input taxon is at infraspecific level and does not occur in TPL, and the higher (species) level is provided in the output instead. Also see `Plant.Name.Index`.

`$Date` Current date according to `Sys.Date`.

## Author(s)

Luis Cayuela & Anke Stein

## References

Cayuela, L., Granzow-de la Cerda, I., Albuquerque, F.S. and Golicher, J.D. 2012. Taxonstand: An R package for species names standardization in vegetation databases. *Methods in Ecology and Evolution*, 3(6): 1078-1083.

Kalwij, J.M. 2012. Review of 'The Plant List, a working list of all plant species'. *Journal of Vegetation Science*, 23(5): 998-1002.

## See Also

TPLck.

## Examples

```
## Not run:
data(bryophytes)

# Species names in full
r1 <- TPL(bryophytes$Full.name[1:20], corr = TRUE)
str(r1)

# A separate specification for genera, specific, and infraspecific
# epithets
r2 <- TPL(genus = bryophytes$Genus, species = bryophytes$Species,
infrasp = bryophytes$Intraspecific, corr = TRUE)
str(r2)

#----------------------------------
# An example using data from GBIF
#----------------------------------
# Download all records available in GBIF of all species within genus
# Liriodendron (GBIF table; note that a list of species can be also
# downloaded from GBIF for a defined geographical area)
require(dismo)
liriodendron <- gbif("Liriodendron", "*", geo = TRUE)

# Perform taxonomic standardization on plant names (TPL table)
sp.check <- TPL(unique(liriodendron$scientificName), infra = TRUE,
corr = TRUE)
head(sp.check)

# Join GBIF table with TPL table
require(dplyr)
liriodendron.check <- liriodendron %>%
left_join(., sp.check, by = c("scientificName" = "Taxon"))


## End(Not run)
```

TPLck *Standardize a plant name according to The Plant List.*

## Description

Connects to The Plant List (TPL) website and validates the name of a single plant taxon, replacing synonyms by accepted names and removing orthographical errors in plant names.

## Usage

```
TPLck(sp, infra = TRUE, corr = TRUE, diffchar = 2,
max.distance = 1, version = "1.1", encoding = "UTF-8",
author = TRUE, drop.lower.level = FALSE)
```

## Arguments

| | |
|---|---|
| sp | A character vector specifying the input taxon, i.e. genus and specific epithet and, potentially, author name and infraspecific abbreviation and epithet. |
| infra | Logical. If TRUE (default), infraspecific epithets are used to match taxon names in TPL. |
| corr | Logical. If TRUE (default), spelling errors are corrected (only) in the specific and infraspecific epithets prior to taxonomic standardization. |
| diffchar | A number indicating the maximum difference between the number of characters in corrected and original taxon names. Not used if corr = FALSE. |
| max.distance | A number indicating the maximum distance allowed for a match in agrep when performing corrections of spelling errors in specific epithets. Not used if corr = FALSE. |
| version | A character vector indicating whether to connect to the newest version of TPL (1.1) or to the older one (1.0). Defaults to "1.1". |
| encoding | Encoding to be assumed for input strings from TPL website; defaults to "UTF-8" (see read.csv and Encoding). |
| author | Logical. If TRUE (default), the function tries to extract author names from the input taxon (see **Details**). |
| drop.lower.level | Logical. If TRUE, the variety is dropped from the input taxon if both subspecies and variety are given, and the forma is dropped from the input taxon if both subspecies or variety and forma are given. If specific and subspecific epithet are identical, the subspecies [variety] part is dropped instead, and variety [forma] is kept. Defaults to FALSE. |

## Details

The function searches for a taxon name on The Plant List (TPL) website and provides its taxonomic status (http://www.theplantlist.org). If the status is either 'Accepted' or 'Unresolved' (i.e. names for which the contributing data sources did not contain sufficient evidence to decide whether

they were 'Accepted' or 'Synonyms'), the function returns the taxon name unchanged. In cases where the input taxon is recognised as a 'Synonym', the according valid taxon is provided in the output, i.e. the current accepted name or, in some cases, an unresolved name. Some data sets which contributed to TPL record not only how plant names should be used but also where in the published literature a given name may previously have been used inappropriately (to refer erroneously to another species). In those cases, the taxonomic status is 'Misapplied', and the function returns the name of the taxon to which this name has been previously and erroneously applied.

If the author of the taxon name is provided in the input, it will be used for taxon matching and distinction between homonyms. The best author match is based on `adist`, i.e. spelling variations are taken into account. Both the full authority and the current author only (i.e. dropping potential basionym author or author preceding the word 'ex') are compared (see **Examples**). Warnings are given in case of imperfect author match (except if differences are in spaces, dots, or '&' vs. 'et' only). The function distinguishes between author name and infraspecific epithet based on upper-case/lowercase only, so this procedure does not work properly if an infraspecific epithet is capitalized or if an author name is spelled with lowercase (but several lowercase author components like "auct", ex", "f./fil." etc. are accounted for). Note that if an input taxon at infraspecific level does not exist in TPL and the species level is given in the output (with Plant.Name.Index = FALSE and Higher.level = TRUE), the function so far does not use the species-level author (if given in the input) for picking the corresponding species.

If TPL includes multiple entries for a given input taxon, the function tries to match the correct taxon based on the author, infraspecific rank and hybrid marker, if provided. Otherwise, preference is given to an Accepted name over a Synonym, Misapplied, or Unresolved name (in this order). If the input taxon is at the infraspecific level and does not exist on the TPL website with the given infraspecific rank abbreviation, the taxon is matched to a different infraspecific rank, if possible (e.g. 'subsp.' versus 'var.'). In case of multiple synonyms, taxon names with a higher confidence level in TPL (regarding the status of name records; see website) are preferred. Otherwise, the first entry on the website that is not an Illegitimate or Invalid name (if possible) is selected, and a warning is given.

Orthographic errors can be corrected (only) in specific and infraspecific epithets. By increasing arguments 'diffchar' and 'max.distance', larger differences can be detected in typos, but this also increases false positives (i.e. replacement of some names for others that do not really match), so some caution is recommended here. Note that if the correction would result in multiple equally likely names (e.g. *Acacia macrocantha* could be corrected to *A. macracantha* or *A. microcantha*), no correction will take place, and Plant.Name.Index is set to FALSE. If a specific epithet cannot be matched after orthographic correction based on `agrep`, common endings are replaced by the equivalent masculine/feminine/neuter or genitive form (-a/-um/-us, -is/-e, -ii/-i) and the taxon is tried to match again.

If 'infra = FALSE', then infraspecific epithets are neither considered for species name validation in TPL, nor returned in the output.

The latest version of TPL, version 1.1, was released in September 2013. Version 1.1 replaces version 1.0, which still remains accessible. Version 1.1 includes new data sets, updated versions of the original data sets and improved algorithms to resolve logical conflicts between those data sets.

## Value

A `data.frame` with the following components:

| | |
|---|---|
| `$Taxon` | Original taxon name as provided in input. |

`$Genus` Original genus name as provided in input.

`$Hybrid.marker` Hybrid marker, if taxon is indicated as hybrid in the input.

`$Species` Original specific epithet as provided in input.

`$Abbrev` Original abbreviation other than infraspecific rank included in input taxon, including "cf.", "aff.", "agg.", "nom. cons.", "nom. cons. prop.", "nom. inval.", "s.l.", and "s.str." and their orthographic variants.

`$Infraspecific.rank`
Original infraspecific rank abbreviation as provided in input, including "subsp.", "var.", "f.", and their orthographic variants.

`$Infraspecific` Original infraspecific epithet as provided in input. If infra = FALSE, this is not shown.

`$Authority` Original author of taxon name as provided in input.

`$ID` The Plant List record ID of the matched taxon before resolving synonyms.

`$Plant.Name.Index`
Logical. If TRUE the name is in TPL. If a taxon at infraspecific level is not in TPL, Plant.Name.Index equals FALSE, except for nominal infraspecies. Also compare `Higher.level`.

`$TPL.version` Version of TPL used.

`$Taxonomic.status`
Taxonomic status of the matched taxon in TPL, either 'Accepted', 'Synonym', 'Unresolved', or 'Misapplied'.

`$Family` Family name, extracted from TPL for the valid form of the taxon.

`$New.Genus` Genus name, extracted from TPL for the valid form of the taxon.

`$New.Hybrid.marker`
Hybrid marker, extracted from TPL for the valid form of the taxon.

`$New.Species` Specific epithet, extracted from TPL for the valid form of the taxon.

`$New.Infraspecific.rank`
Infraspecific rank abbreviation, extracted from TPL for the valid form of the taxon, including "subsp.", "var." and "f.".

`$New.Infraspecific`
Infraspecific epithet, extracted from TPL for the valid form of the taxon.

`$New.Authority` Author of taxon name, extracted from TPL for the valid form of the taxon.

`$New.ID` The Plant List record ID of the taxon, once synonyms have been replaced by valid names. For accepted and unresolved names, this field will be equivalent to `ID`.

`$New.Taxonomic.status`
Taxonomic status of the resolved taxon in TPL, once synonyms have been replaced by valid names. 'Accepted' or 'Unresolved'.

`$Typo` Logical. If TRUE there was a spelling error in the specific or infraspecific epithet that has been corrected.

`$WFormat` Logical. If TRUE, fields in TPL had the wrong format for information to be automatically extracted as they were not properly tabulated or, alternatively, there was not a unique solution.

| | |
|---|---|
| `$Higher.level` | Logical. If TRUE, the input taxon is at infraspecific level and does not occur in TPL, and the higher (species) level is provided in the output instead. Also see `Plant.Name.Index`. |
| `$Date` | Current date according to `Sys.Date`. |

## Author(s)

Luis Cayuela, Anke Stein & Jari Oksanen

## References

Cayuela, L., Granzow-de la Cerda, I., Albuquerque, F.S. and Golicher, J.D. 2012. Taxonstand: An R package for species names standardisation in vegetation databases. *Methods in Ecology and Evolution*, 3(6): 1078-1083.

Kalwij, J.M. 2012. Review of 'The Plant List, a working list of all plant species'. *Journal of Vegetation Science*, 23(5): 998-1002.

## See Also

`TPL`.

## Examples

```
## Not run:
# An accepted name
(TPLck("Amblystegium serpens juratzkanum"))

# An unresolved name
(TPLck("Bryum capillare cenomanicum"))

# A synonym
(TPLck("Pottia acutidentata"))

# A misapplied name
(TPLck("Colutea istria"))

# A name that is not in TPL
(TPLck("Hypochoeris balbisii"))

# A spelling error in the specific epithet
(TPLck("Pohlia longicola", corr = TRUE))

# A spelling error that is not corrected ('max.distance' defaults to 1)
(TPLck("Microbryum curvicollum", corr = TRUE))

# If increasing 'max.distance', the spelling error is accounted for
(TPLck("Microbryum curvicollum", corr = TRUE, max.distance = 3))

# A spelling error where the ending is changed to the
# neuter/feminine form
(TPLck("Symphytum officinalis"))
```

```
(TPLck("Schinus terebinthifolium"))

# A spelling error that is not corrected because two different results
# are possible (see Details)
(TPLck("Acacia macrocantha"))

# A taxon matched through author name
(TPLck("Gladiolus communis L. subsp. byzantinus (Mill.) A.P.Ham."))

# If only the current author is provided (without the author of the
# basionym), the function still matches the correct taxon (even though
# adist returns a better overall match for the author of the homonym,
# Abies alba Mill.)
(TPLck("Abies alba Michx."))

# A difference between TPL versions 1.0 and 1.1
(TPLck("Fallopia japonica", version = "1.0"))
(TPLck("Fallopia japonica", version = "1.1"))

# Avoid illegitimate names when choosing between multiple synonyms
(TPLck("Anthemis altissima"))

# A nominal subspecies not in TPL (Higher.level == TRUE;
# Plant.Name.Index == TRUE)
(TPLck("Callitriche brutia Petagna subsp. brutia"))

# A variety not in TPL (Higher.level == TRUE; Plant.Name.Index == FALSE)
(TPLck("Asplenium ruta-muraria var. lanceolum"))

# A taxon matched through infraspecific rank abbreviation
(TPLck("Heliopsis helianthoides subsp. scabra"))

# Drop variety and keep subspecies
(TPLck("Vicia sativa L. subsp. nigra (L.) Ehrh. var. minor (Bertol.) Gaudin",
drop.lower.level = TRUE))

# Drop nominal subspecies and keep variety
(TPLck("Anagallis arvensis subsp. arvensis var. caerulea",
drop.lower.level = TRUE))

## End(Not run)
```

# Index