# Guide to the Plant Taxon Name Cleaner

Nicole Kinlock

7 December 2022

## Program overview

- This program will edit plant taxon names in a way that best aligns with modern databases while adhering to international plant taxonomy standards.
  - The standards are the International Association for Plant Taxonomy (IAPT) *International Code of Nomenclature for Algae, Fungi, and Plants* for scientific names and the International Society for Horticultural Science (ISHS) *International Code for the Nomenclature for Cultivated Plants* for cultivar names.
  - This program was written because I was working with files listing plant taxon names in a variety of different ways, many of which were from older sources (published 1700-1950). Though I've found it useful to clean names from all kinds of sources using this program, it is especially useful when working with older sources.
  - This program will NOT standardize plant taxon names! But, it will reduce the chance that you get a false match during standardization.
- Briefly, this program can standardize spacing and capitalization in scientific names and authorities, split taxon names into their parts, replace common misspellings of genera, compare genera with lists of vascular plant, fungus, and bryophyte genera, replace authorities with standard forms according to Brummitt & Powell (1992) *Authors of Plant Names*, replace special characters, and remove extraneous information about taxonomic alterations/circumscriptions, among other things (see the "Details on what the program does" section).
- In addition to the output CSV file, the program will write an additional text file that describes the changes it made and that includes additional information that should be checked through manually (for example, genera that were not present in the World Checklist of Vascular Plants).

## Files included

- `runPlantTaxonNameCleaner.sh`: bash script used to execute python script and to write console output to file
- `cleanTaxonNames.py`: python script used to clean plant taxon names in a CSV file
- `AuthorAliases.csv`: CSV file listing the long forms (e.g., full surnames) and standard forms for over 300 commonly-cited authorities
- `AuthorAliases_OldText.csv`: CSV file listing the abbreviated forms (e.g., initials) and standard forms for over 350 commonly-cited authorities
- `GeneraAliases.csv`: CSV file listing over 1,400 commonly mispelled plant genera and the accepted spelling for each
- `WCVP_Genera_Jun2021.csv`: CSV file listing all genera (not just accepted ones) in Kew's World Checklist of Vascular Plants (WCVP) database as of June 2021 (version 5)
- `MOBOT_BryophyteGenera.csv`: CSV file listing all bryophyte genera from the Missouri Botanical Garden (MOBOT) database
- `GBIF_FungiGenera.csv`: CSV file listing all fungi genera from the Global Biodiversity Information Facility (GBIF)
- `Example1.csv` and `Example2.csv`: Two example files with plant taxon names, presented in different ways. Example 1 comes from Bailey (1924) *Manual of cultivated plants*, and Example 2 comes from the Northeastern Asia Biodiversity Institute (2021) *List of Korean Vascular Plants*.
- `PlantTaxonNameCleaner_Manual.pdf`: the current file

## Program requirements

- **Bash**: Bash is a Unix shell that is already available on your computer if you use macOS or Linux, and is available on Windows via Windows Subsystem for Linux (WSL)
  - *Why Bash?*: I highly recommend running R or Python scripts within a shell script because you can write the console output directly to file. This way you can systematically check through the console output. You can also refer back to past output in the future, which increases transparency. I find this to be a major improvement over scrolling through the (temporary) console output in RStudio or in Jupyter lab.
  - *What is WSL and how do I get it?*: Windows Subsystem for Linux is a useful and easy way to run a Linux distribution from within your Windows computer so that you can run Unix-based programs. To install, open PowerShell or Windows Command Prompt in administrator mode (right click and select "Run as administrator") and enter `wsl --install` (more details here, https://learn.microsoft.com/

windows/wsl/install). Then, I recommend installing Windows Terminal (https://learn.microsoft.com/windows/terminal/install), and using it to access WSL.

- **Python 3**: Python is a programming language that is available for a variety of operating systems. You may already have Python 3 installed if you use macOS, Linux, or WSL. You can check this by typing `python3` into a command line program (Terminal in macOS, Console in Linux, Terminal in Windows running WSL or WSL2). Otherwise, you can install python at https://www.python.org/downloads/.
  - *Why Python?*: I wrote this program in Python rather than R because its purpose is to manipulate text and Python has better tools for text manipulation, in my opinion, particularly regular expression (regexp) matching and substitution.
- **Python modules**: Python modules are analogous to R packages. `sys`, `os`, `datetime`, `re`, `random`, `ast`, and `csv` come pre-installed with Python 3. `numpy` and `pandas` must be installed. This can be done entering the following into a command line program:

```
python3 -m pip install numpy
```

```
python3 -m pip install pandas
```

## How to run the program

Make sure your computer meets all of the requirements listed in the previous section.

Download the files, and place them in your desired folder.

Open your command line program (Terminal or Console). The program is run by entering the following in the command line (but including the paths to the folders where you put the files):

```
PATHTOSCRIPT/runPlantTaxonNameCleaner.sh PATHTOFILE/FileToEdit.csv
```

For example, for me to run the program on my Linux machine with Example 1:

```
/home/nlkinlock/PlantTaxonNameCleaner/runPlantTaxonNameCleaner.sh
/home/nlkinlock/PlantTaxonNameCleaner/Example1.csv -oc
```

And to run the program with Example 2:

```
/home/nlkinlock/PlantTaxonNameCleaner/runPlantTaxonNameCleaner.sh
/home/nlkinlock/PlantTaxonNameCleaner/Example2.csv -ac
```

The output files will be saved in the same folder as the taxon names file. Note the flags included (`-oc` for Example 1 and `-ac` for Example 2) and see the following section for an explanation of these flags.

### Program options

The program will run differently depending on what flags you add to the command line call.

- `-c` (CV): This flag tells the program that you want it to correct the format of cultivar names, to resemble 'Cultivar Name'. International standards for naming cultivars (the International Code of Nomenclature for Cultivated Plants) were not developed until 1953 (so pre-1953 sources intrinsically do not follow international standard), and many modern sources do not always follow these conventions. I recommend always including this flag unless you have complex, already-formatted cultivar names in a separate 'Cultivars' column. If your file doesn't include cultivars at all, this flag will be ignored.
- `-a` (AUTHORSPLIT): This flag tells the program that your taxon names include authorities. Note that this flag only matters if you want the program to split your taxon names into parts. If you have already split the taxon names into parts, this flag will be ignored.
- `-o` (OLDTEXT): This flag tells the program that your file is an old source, i.e., from approximately 1900 or earlier (definitely do not include this flag if your source is from 1950 or later). This flag is very important because standards for recording taxon names used to be much less uniform. In particular, authority abbreviations used to be very different, e.g. W. for Willdenow (today Willd.), Lin. for Linneaus (today L.), or W. & K. for Waldstein and Kitaibel (today Waldst. & Kit.). By including this flag, the program will replace these older abbreviations with modern ones. However, on a newer source, this process is likely to fail, as the pool of potential authorities is much larger today than it used to be. Also, in old sources, authors did not distinguish cultivars from infraspecific taxa. Sometimes, names that clearly referred to cultivars were included as infraspecific names, e.g., 'flore pleno marginalis'. Including these long cultivar names as scientific names in a taxonomic name standardization program could lead to errors (or, at the very least, such "infraspecific names" will be removed during standardization). If you include this flag, the program will reassign these types of infraspecific names to cultivars as follows: (1) infraspecific names with periods (e.g., fl. pl. or fol. margin.) will

always be assigned to be cultivars, and (2) infraspecific names that begin with flore, fructu, fructo, or foliis will be assigned to be cultivars as long as they do not have an infraspecific rank (e.g., 'var. flore pleno' will not be assigned to be a cultivar).

To include these flags, enter them into the command line, for example:

```
PATHTOSCRIPT/runPlantTaxonNameCleaner.sh PATHTOFILE/FileToEdit.csv -cao
PATHTOSCRIPT/runPlantTaxonNameCleaner.sh PATHTOFILE/FileToEdit.csv -o
-c
PATHTOSCRIPT/runPlantTaxonNameCleaner.sh PATHTOFILE/FileToEdit.csv -a
```

If you ever forget the commands or the options, enter the shell script alone in the command line without a file or any flags, for example:

```
PATHTOSCRIPT/runPlantTaxonNameCleaner.sh
```

Then, usage details, reminding you of what to write on the command line and the meaning of the flags, will be printed to the terminal.

## How to prepare a file with taxon names

- The program will accept a comma-separated values (CSV) file with columns including: `Genus`, `SpecificEpithet`, `InfraspecificRank`, `InfraspecificName`, `Authority`, `InfraspecificAuthority`, and `Cultivar`. Any of the previous columns may be included as long as there is at least a `Genus` and `SpecificEpithet` column. In `Example1.csv`, taxon names are split into these parts. A few examples of plant taxon names and their parts follow:
  - '*Cornus officinalis* Siebold & Zucc.' Genus: *Cornus*, Specific epithet: *officinalis*, Authority: Siebold & Zucc.
  - '*Aralia cordata* Thunb. var. *continentalis* (Kitag.) Y.C.Chu' Genus: *Aralia*, Specific epithet: *cordata*, Infraspecific rank: var., Infraspecific name: *continentalis*, Authority: Thunb., Infraspecific authority: (Kitag.) Y.C.Chu
  - "*Thuja occidentalis* L. 'Smaragd Variegated'" Genus: *Thuja*, Specific epithet: *occidentalis*, Authority: L., Cultivar: 'Smaragd Variegated'
  - '*Sorghum vulgare* var. *drummondii* (Steud.) Hitchc.' Genus: *Sorghum*, specific epithet: *vulgare*, Infraspecific rank: var., Infraspecific name: *drummondii*, Infraspecific authority: (Steud.) Hitchc.

- Alternatively, you can include the entire taxon name in a `TaxonName` column. Then, the program will split the name into its parts. In `Example2.csv`, taxon names are not split into parts.
  - Splitting names should work correctly, but this depends somewhat on the complexity of the names. Potential pitfalls include: infraspecific names without ranks (especially if authorities are included), cultivars that not specified in a standard way (e.g., something other than 'Cultivar Name' or cv. Cultivar Name), or errors in capitalization. Always read the console output to troubleshoot these issues.
- Including `TaxonName` and `Authority` columns separately will also work fine. A number of suitable combinations of columns exist, but **always check the console output carefully for errors**. The program is designed to fail before it manipulates a name incorrectly, but errors can always occur.

## Details on what the program does

### First steps

- If you have already split taxon names into their parts, the program will figure out which parts are present and create an `OriginalTaxonName` column with these parts arranged. Alternatively, if you have entire taxon names, the `TaxonName` column will become the `OriginalTaxonName` column.
- If you supplied names in the `TaxonName` column, they will be split (see the "Split taxon names" section for more details).
- For all of the following steps, fairly extensive output will be reported in the console output file. Be sure to check through this carefully.

### Split taxon names

- Again, this is only done if you provided names in a `TaxonName` column and did not split names into their parts.
- The program will first check whether the taxon is a genus hybrid (e.g., ×*Cupressocyparis*) or a graft hybrid (e.g., +*Crataegomespilus*).
- Then, the program will check whether a genus is included, by looking for a capitalized word with at least two letters. Failure at this stage could be because a taxon is not capitalized, or because the first letter is a special character (not allowed according to the IAPT code).

- The program will next check whether the taxon name includes an infraspecific name or cultivar. It will use this to narrow the search range in which to check for (1) a specific epithet (must be a word beginning with a lowercase letter) and (2) a species-level hybrid (e.g., 'Genus × hybrid epithet' or 'Genus specific epithet 1 × specific epithet 2'). Including taxa that are not species should cause the program to fail at this stage (as is desired), for example, trying to split the name '*Aster* sect. *Alpigeni*' will cause an error. In contrast, taxa with undefined species names, e.g., *Lolium* sp., *Zingiber* spp., *Forsythia* 'Bronxensis', or *Rosa* 'Soleil d'Or' will not cause an error, but they will be removed at a later stage (see the 'Remove undefined species' section).
- Next, the program will pull out the cultivar name as the part of the name surrounded in single or double quotes or the part following 'cv.'. Cultivar names are allowed to include a wide variety of characters and numbers (as according to the ISHS code), and cultivars are not followed by authorities.
- Then, infraspecific names will be pulled out by looking for lowercase words following infraspecific ranks. Ranks include: subsp., var., convar., subvar., f. subf., or Greek lowercase letters (α, β, etc., this is an older way to indicate varieties). If taxon names include multiple infraspecific epithets, only the finest taxonomic resolution is kept. For example, *Saxifraga aizoon* subf. *surculosa* may also be referred to as *Saxifraga aizoon* var. *aizoon* subvar. *brevifolia* f. *multicaulis* subf. *surculosa*. Both names are perfectly legal according to the IAPT code, so the simplified form is preferred in order to maximize cross-dataset harmonization potential.
- Infraspecific names without ranks are difficult to parse. If the third word of the name is lowercase with no special characters, this will be assigned as the infraspecific name. Otherwise, if you have indicated that you did not include authorities in the `TaxonName` column (by not including the `-a` flag), the remaining part of the name will be assigned as an infraspecific name.
- If your `TaxonName` does include authorities (indicated by including the `-a` flag), the remainder of the name is assigned to be the authority. Because this is a bit of a catch-all category, look through the `Authority` and `InfraspecificAuthority` columns in the output particularly carefully.

### Remove undefined species

- This program only returns taxon names defined to the species-level or below. So, taxon names that are missing genera or specific epithets (e.g., *Lolium* sp. or *Zingiber* spp.) will be removed from the file. Taxon names that just list genus and cultivar are also removed, as these are likewise not defined to the species level (e.g., *Forsythia* 'Bronxensis' or *Rosa* 'Soleil d'Or').

- Taxon names involving open nomenclature are removed because they are not defined to the species level. Open names can include: *sp. nov.* (*species nova*), in which the taxon is a previously undescribed (and thus unpublished) species; *aff.* (*species affinis*), in which the taxon is assumed to be closely related to the indicated name; or *cf.* (*confer*), in which the taxon is believed to be the indicated name.

### General cleaning steps

- Extra whitespace is removed from taxon names, genera are put in sentence case, and specific and infraspecific epithets are put in lowercase.
- Ligatures are converted to their respective separate letters
- Alternative quotation characters are converted to ' and "
- Alternative whitespace characters are converted to unicode spaces and linebreak characters are removed
- Question marks are removed
- Accented characters (diacritical signs) in genera, specific epithets, and infraspecific epithets are transcribed according to the IAPT Shenzhen Code Article 60.7, "ä, ö, ü become, respectively, ae, oe, ue; é, è, ê become e; ñ becomes n; ø becomes oe; å becomes ao."
- All non-ASCII characters in genera, specific epithets, and infraspecific epithets are identified and reported in the console output.
- All numbers are identified and reported in the console output. Numbers in genera, specific epithets, and authorities will trigger an error, and you will need to check these and resolve them.
- Special characters are allowed in authorities and cultivars. However, they may be indicative of an error in file conversion. For this reason, all unusual characters in these columns are identified and reported in the output for you to review.
- Spacing around parentheses is standardized
- Synonyms in names are removed, and are identified as (= Synonym)
- Hybrid symbols and spacing are standardized. The hybrid symbol, ×, is used, and is followed by one space.

### Genera

- Genera are checked for common mispellings and corrected to the accepted spelling.
  - **More about genus spelling**: Errors in genus spelling are extremely common for a number of reasons: (1) classical Greek and Latin letters do not correspond with the

English alphabet (u/v, i/j, and eu/ev are interchangeable), (2) an author of a popular book included a misspelling that was then copied by their contemporaries, and (3) typos are incredibly common. Many of these errors, particularly those stemming from older source, have been recognized in the International Plant Names Index (IPNI) as orthographic variants (orth. var.) or misprints (*sphalm.*, *sphalma typographicum*). However, these known errors are often not linked to the correct spelling of the genus. I've gathered over 1,400 of these common mispellings and added them to the `GeneraAliases.csv` file. I have found that errors in genus spelling are much more difficult to resolve when using taxonomic name standardization programs relative to errors in specific epithets. All taxonomic name standardization programs that I am familiar with use only fuzzy matching based on Levenshtein distances to resolve misspelling. These distances are not informed at all by taxonomic information, and so matches between unrelated taxa occur (*Tulipa* and *Sulipa*, *Hemerocallis* and *Hesperocallis*, *Xeranthemum* and *Eranthemum*, *Mahonia* and *Magonia*, etc.). Pre-cleaning the genera and forcing genus matches to be perfect during standardization can avoid this.
  - The `GeneraAliases.csv` file can, of course, be added to or subsetted as desired.

- The following steps describe checks that are returned only as text in the console output file, these checks do not alter the taxon name output.

  - Genera are checked against all genera listed in the World Checklist of Vascular Plants (which includes genera that are not accepted). If a genus on your list is not found in the WCVP, then it probably means that either the genus is misspelled or the taxon is not a vascular plant. Genera that are not found in the WCVP are listed in the console output file for you to check through manually.
  - Genera are checked against all fungus genera listed on GBIF. If a genus on your list matches a fungus genus from GBIF, then it is listed in the console output file for you to check through manually. Note that some fungus genus names are also vascular plant genus names. You may wish to remove fungi from your list of taxa before, for example, proceeding with taxonomic standardization.
  - Genera are checked against all bryophyte genera listed on the MOBOT Bryophyte Names Authority List, http://www.mobot.org/mobot/tropicos/most/bryolist.shtml. If a genus on your list matches a bryophyte genus from MOBOT, then it is listed in the console output file for you to check through manually. Note that some bryophyte genus names are also vascular plant genus names.

## Authorities

- The spacing for initials in authority names is standardized, e.g., M.Bieb. instead of M. Bieb., following the guidelines in Brummitt & Powell (1992).
- When two authors are connected by 'and' or 'et', this is replaced with an ampersand. Ampersands are edited to have a single space on either side. Both 'et' and '&' are allowed according to Brummitt & Powell, but '&' seems to be more common in modern databases.
- 'ex' in authorities is converted to lowercase, with no period, and surrounded by a single space ('Authority A ex Authority B' means that Authority A did not validly publish the name, but Authority B did).
- 'in' and following content in authorities is removed because it is not a required element of authorities. See the IAPT Shenzhen Code Article 46.2 Note 2, 'When authorship of a name differs from authorship of the publication in which it was validly published, both are sometimes cited, connected by the word "in". In such a case, "in" and what follows are part of a bibliographic citation and are better omitted unless the place of publication is being cited.'
- Some other features occasionally included in authorities are actually part of the citation of the publication, but not the authority specifically. These include *pro. syn.* (published as a synonym), *nomen nudum* (published without a description), *non* before authorities of homonyms (this and following authority are removed, otherwise "*sensu* G.L.Nesom, *non* Lam." may match "Lam.", which would not be desirable), *auct. non* (e.g., in *auct. non Name A*, Name A is a misapplied name, and so Name A and the following authority are removed), *nom. cons.* and *orth. cons.* (conserved name/spelling), *nom. rej.* (rejected name). Taxonomic alterations are also sometimes included (e.g., *emend*, *p. p.*, *s. l.*, etc.), but are not required in authorities according to the IAPT Shenzhen Code Article 47.1 "An alteration of the diagnostic characters or of the circumscription of a taxon without the exclusion of the type does not warrant a change of the author citation of the name of the taxon. When an alteration as mentioned in Art. 47 has been considerable, the nature of the change may be indicated by adding such words, abbreviated where suitable."
- Sometimes errors in authorities can be identified by looking for words that start with lowercase letters. These are identified (with the exception of "fil." and "hort.", which are common and are not errors) and printed in the console output for you to check through.
- Some sources tend to include the full surnames of authorities rather than using standard abbreviated forms. This impedes taxonomic harmonization. Authorities are checked for long forms and converted to standard forms using the the `AuthorAliases.csv` file. You can add to or remove from this file as needed.
- For old texts, authorities are checked for alternative abbreviations and corrected to

standard abbreviations using the `AuthorAliases_OldText.csv` file. This should not be used for modern sources (more recent than c. 1950), because the larger pool of authorities included in modern sources could increase the number of errors in replacement. Again, you can add to or remove from the `AuthorAliases_OldText.csv` file as needed.

### Infraspecific ranks and names

- Alternative spellings for infraspecific ranks are corrected, for example, ssp. is corrected to subsp., v. is corrected to var., and lowercase Greek letters (α, β, etc.) are corrected to var.
- Infraspecific names are checked for infraspecific ranks, and these are moved to the infraspecific ranks column. So, if you did not separate these ahead of time, the program will do that.
- Infraspecific names are checked for cultivars, e.g., cv. Flore Pleno or 'Flore Pleno', and these are moved to the cultivar column.

### Cultivars

- If the `-c` flag is included, cultivar names are standardized according to the ISHS code, meaning that cultivar names are surrounded by single quotation marks, and each letter of the epithet starts with a capital letter (depending on linguistic convention, e.g., 'Gloire de Lorraine' is standard).
- Note that other writing systems (for example, Chinese or Japanese characters, Cyrillic script) are technically allowed in cultivar names according to the ISHS code. However, I have not encountered these types of names yet, and thus they will likely cause errors in the program. The ISHS code recommends converting these types of names to Latin script using the ALA-LC Romanization Tables or the ISO transliteration standards.

### Final steps

- For specific and infraspecific epithets with multiple words, spaces are replaced by hyphens.
  - This is according to IAPT Shenzhen Code Article 20.3, "The name of a genus may not consist of two words, unless these words are joined by a hyphen" and Article 23.1, "The name of a species is a binary combination consisting of the name of the genus followed by a single specific epithet in the form of an adjective, a noun in

the genitive, or a word in apposition. If an epithet consisted originally of two or more words, these are to be united or hyphenated."

- Periods in specific and infraspecific epithets are removed.
- Infraspecific ranks with no corresponding name are removed, e.g., "*Globularia cordifolia* var." is edited to "*Globularia cordifolia*".
- Full names are generated in several formats
  - `TaxonNameCV`: Scientific name with cultivar name and no authority
  - `TaxonName`: Scientific name with no cultivar name and no authority
  - `TaxonNameAuthorCV`: Scientific name with cultivar name and authority
  - `TaxonNameAuthor`: Scientific name with authority, but no cultivar name
  - The authority used in these full names is the authority of the finest taxonomic level
- Last, duplicate rows are removed (all columns are considered when removing duplicates)
- The columns are arranged as: `index` (row index of original file, note that Python begins with an index of 0), `Genus`, `SpecificEpithet`, `InfraspecificRank`, `InfraspecificName`, `Authority`, `InfraspecificAuthority`, `AuthorityCorrected`, `Cultivar`, `OriginalName`, `TaxonNameAuthorCV`, `TaxonNameAuthor`, `TaxonNameCV`, `TaxonName`. All of the original additional columns included in your file are added after these columns.
- The output is saved as a `.csv` file in the same folder and with the same file base as the input file, plus `_TextCleaned_YYYYMMDD.csv`
- The console output file is saved as a `.dat` file in the same folder and with the same file base as the input file, plus `Console_YYYYMMDD.dat`

## Problems, requested changes, suggestions?

- Contact me! (nlkinlock@gmail.com)
- Feel free to edit the files yourself (locally) if you want. But if you have an idea, please do consider letting me know.

## Related links

International Association for Plant Taxonomy (IAPT) International Code of Nomenclature for algae, fungi, and plants, Shenzhen Code (2018) lists the complete rules for plant scientific name nomenclature

International Society for Horticultural Science (ISHS) International Code of Nomenclature for Cultivated Plants, Ninth Edition (2016) lists the complete rules for cultivated plant nomenclature

International Plant Names Index (IPNI) provides a list of authorities standardized according to the rules in Brummitt & Powell (1992) *Authors of Plant Names*