Desription of ANSA:
an automated workflow to standardise taxon names for South African alien species lists

Katelyn T. Faulkner[1,2]
[1]South African National Biodiversity Institute, Kirstenbosch Research Centre, Cape Town, South Africa
[2]Department of Zoology and Entomology, University of Pretoria, Pretoria, South Africa

3 December 2024

Contents

Introduction

The automated workflow Alien Names South Africa (ANSA) was developed to standardise the names of taxa for South African alien species lists. It was developed specifically using the taxonomic backbones implemented in South Africa. Although initially developed within the context of the species list (http://dx.doi.org/10.5281/zenodo.8217197) for the report "The Status of Biological Invasions and their Management in South Africa" (https://dx.doi.org/10.5281/zenodo.8217182), the workflow will be useful for South African alien species lists, more generally.

The workflow makes use of three taxonomic backbones: that of the Global Biodiversity Information Facility (GBIF); that of the Botanical Database of Southern Africa (BODATSA); and that of The World Checklist of Vascular Plants (WCVP), accessed via Plants of the World Online (POWO) (https://powo.science.kew.org/).

The taxonomic backbone used depends on the taxonomic group: 1. The GBIF backbone is used for non-plant taxa (e.g., animals, fungi, and chromista) 2. The BODATSA backbone is used for alien plants recorded outside of captivity (i.e., those included in BODATSA) 3. The WCVP backbone is used for alien plants not found outside of captivity (i.e., those not included in BODATSA)

In addition to standardising a list of taxon names based on the three taxonomic backbones, the workflow obtains canonical names and higher taxonomic information for the taxa from GBIF, and flags issues and uncertainties that arise during the taxonomic standardisation.

This document describes the workflow.

The workflow has been created in R software, version 4.4.0 (R Core Team 2024).

Requirements

The following are required to execute of the workflow:

1. Installed R software (version 4.4.0 or higher) and Rstudio
2. Installed R packages: "tidyr", "dplyr", "rgbif", "stringdist", "gtools", "stringr", "sjmisc", "rWCVP", "remotes", "rWCVPdata", "purrr"
3. A stable internet connection

The R environment

R (version 4.0.0 or higher) and Rstudio needs to be installed, which are freely available at: https://cran.r-project.org and https://www.rstudio.com/

Eleven R packages and their dependencies must be installed. Ten of these packages can be obtained through the R CRAN, and the remaining package "rWCVPdata" can be obtained from GitHub, using the package "remotes".

If executed the following code will load and, if required, install the packages.

Specify the packages required from R CRAN:

```
packages = c("tidyr", "dplyr",
             "rgbif", "stringdist", "gtools", "stringr", "sjmisc", "rWCVP", "purrr")
```

Install R CRAN packages (if required) and load:

```
package.check <- lapply(
  packages,
  FUN = function(x) {
    if (!require(x, character.only = TRUE)) {
      install.packages(x, dependencies = TRUE)
      library(x, character.only = TRUE)
    }
  }
)
```

Install and load “remote” package and install and load “rWCVPdata” package from GitHub

```
remotes.check<-"remotes" %in% rownames(installed.packages())
if(remotes.check == FALSE){
  install.packages("remotes")
  remotes::install_github(’matildabrown/rWCVPdata’)
}
require(’rWCVPdata’)
```

Scripts

The R script for this workflow is provided as a .R file and .Rmd file on a GitHub repository “https://github.com/KatelynFaulkner/alien-species-names”.

The repository can be downloaded onto your computer as a zip file. The downloaded zip file will contain all the required folders. The zip file will need to be extracted. The main folder contains the Rstudio project file (ansa-workflow.Rproj) of the workflow. This file should be opened to use the workflow. In the main folder is also this Description of the workflow as a R markdown (.Rmd file). The subfolder `R/` contains the script as .R and .Rmd files. The .Rmd file contains notes and examples to provide guidance on the code and outputs. The subfolder `data/` contains three subfolders named `inputs/`, `interim/`, and `outputs/`. The `inputs/` subfolder contains all files required to run the workflow. Outputs that are produced while executing the workflow, but that are not the final output are stored in the folder `interim/`. The final output of the workflow is stored in the `outputs/` folder.

Inputs

The following four inputs are required:

1. The list of taxon names that need to be standardised in csv format
2. Information on the format of the taxon names to be standardised
3. The BODATSA dataset in csv format. Downloaded from https://posa.sanbi.org/sanbi/Explore
4. The date that the BODATSA dataset was downloaded

Inputs 1 and 3 are provided by the user, who saves the files in the `inputs/` subfolder.

Input 2 is provided by the user, and entered in Step 7 of the workflow. There are two possible options: "canonical" and "scientific". "canonical" is stipulated when the species names do not include the authority. "scientific" is stipulated when the species names do include the authority.

Input 3 is provided by the user and can be downloaded from https://posa.sanbi.org/sanbi/Explore.

Input 4 is provided by the user, and entered in Step 4 of the workflow in DMY (Day, month, year) format. For example "05 December 2024".

Input file, names, formats, and column names

The execution of this workflow requires two input datasets in csv format. These datasets must have specific names; their columns must have specific names; and they must and contain specific data.

List of taxa you want to standardise This dataset must contain the list of taxon names that need to be standardised. The dataset must be named 'OriginalNames.csv'. The dataset must contain only one column, which will contain the names of the taxa. The name of this column must be 'verbatimScientificName'.

BODATSA data This dataset must be downloaded from https://posa.sanbi.org/sanbi/Explore. To do so, simply click download, indicated with using the symbology for Microsoft excel (a green 'X'). The standard dataset must be downloaded (i.e., no columns must be added or removed before download). The dataset will be downloaded as an excel file, but must be saved in the `inputs/` folder as a csv. The data must be used in exactly the format in which it is downloaded (i.e., it must not be changed in any way). The file must be named: BODATSA.csv.

The date the dataset is downloaded must be noted (see Step 4 below).

Executing the ANSA workflow

To execute the workflow the Rstudio project file `ansa-workflow.Rproj` found in the main folder for the workflow must be opened. This will set the working directory.

Below each step of the workflow is described.

Step 1: Install and load R CRAN packages

In this step the R environment is prepared by installing from the R CRAN the R packages that need to be installed, and by loading these packages.

Step 2: Install and load "remote" package from R CRAN, and install and load a package from GitHub

In this step the R environment is prepared by installing from the R CRAN the 'remote' package, and using this package to install from GitHub the 'rWCVPdata' package. The 'rWCVPdata' package is loaded.

Step 3: Load dataset from BODATSA

The BODATSA dataset is loaded and assigned the name 'BODATSA'

Step 4: Provide required information on BODATSA

Information on the date the BODATSA dataset was downloaded is provided in DMY (Day, month, year) format. For example "05 December 2024". This information is assigned the name 'BODATSADate'.

Step 5: Prepare BODATSA data

In this step the downloaded BODATSA data is prepared to be used for taxon name standardisation. There are many instances where data are missing in the dataset, these are assigned 'NA'. In the 'Rank1' column "ssp." is replaced with "subsp.", to align with the formatting of taxon names in the 'Accepted' column. In BODATSA, taxon names are split, with each part found in a different column. For example, there are seperate columns for genus ('Genus') and species ('Sp1'). The columns related to taxon names are joined so that there are single columns for canonical name (scientific name without authorities) and scientific name (scientific name with authorities). For synonyms, BODATSA provides the accepted species name in the column 'Accepted'. In some instances the name in the 'Accepted' column is in a different format from that in the 'scientificName' column, created using the columns of the dataset that are related to taxon names. This issue is addressed by identifying the incorrectly formatted species names in the 'Accepted' column, and correcting them using fuzzy matching techniques. The original data in the 'Accepted' column is saved in a column called 'AcceptedIssues' and records with this issue are flagged in a column called 'synAccIssue'. The prepared BODATSA dataset is written to the interim folder. The file is called 'BODATSAPrepared.csv'.

Step 6: Load list of names for standardisation

The names that need to be standardised are read in, and are named 'NamesDat'.

Step 7: Provide required information on list of names

Information on the format of the names that need to be standardised is provided. There are two options: 'canonical' (scientific name without authorities), and 'scientific' (scientific name with authorities). These details are given the name 'NameType'.

Step 8: Prepare the list of names for standardisation

The names for standardisation are prepared. This involves standardising the formatting of the spaces in the names and removing any double spaces.

Step 9: Match taxon names to the GBIF backbone and get canonical names and higher taxonomic information from GBIF

The GBIF API is used to standardise all the taxon names based on the GBIF taxonomic backbone. Canonical names, and higher taxonomic information (e.g., kingdom, family, class) are also obtained from GBIF. If the taxon name provided is a synonym according to the GBIF backbone, then the accepted scientific name is obtained. The output file will contain information on whether the taxon name provided was a synonym, the rank to which the taxon's name could be resolved, the status of the name provided, the type of match (e.g., fuzzy, exact, higherrank), and the taxon's unique GBIF 'usageKey'. This output is written to the interim folder as a csv named 'GBIFInterim.csv'.

Step 10: Separate dataset into plants and other organisms

The output from the GBIF standardisation is separated into two datasets, one for plants called 'PlantDat', and one for non-plant taxa called 'GBIFMatch'. All information obtained from GBIF besides the canonical name and higher taxonomic information are removed from the plant dataset ('PlantDat'). The taxon names of the plants will be standardised using other taxonomic backbones. The GBIFMatch dataset, which contains only non-plant taxa, will be used for further processing.

Step 11: Assess results for non-plants and flag issues

The results of the taxonomic standardisation for non-plants are evaluated and issues are flagged.

The prevalence of certain issues is also assessed. This is done by looking at how any of the taxa:

1. Fall into each status category (e.g., accepted, synonym etc)
2. Had a specific match type (e.g., exact, fuzzy etc)
3. Fall into different ranks (e.g., species, subspecies etc)

Taxonomic issues are flagged. These are:

1. Synonyms
2. Matches that were not exact (e.g., fuzzy matches)
3. Doubtful taxon names
4. Unrecognised taxa - those not recognised or only recognised at higher levels than the taxon name provided (i.e., genus or above level for a name provided at the species level, or a species level for a name provided at a sub-specific level).

In addition, all taxa that either had a non-exact match, a doubtful name, or were not recognised are flagged as having a possible error or issue.

Sub-specific entities are flagged

Duplicated names in either the returned accepted scientific name, canonical name, or the taxon name provided are also flagged.

Step 13: Post-processing

A column is added with the source of the scientific name and the date that source was consulted. The source contains a link to GBIF that is unique to the taxon, and which is stable over time. Note that the accepted name for the taxon could change over time as the GBIF taxonomic backbone is revised. Unnecessary columns are removed from the dataset, and the output is written to the 'interim' folder as a csv named 'NonPlantOrganisms_GBIF_standardisation.csv'.

Step 14: Match plant names with those in BODATSA

The taxon names in the 'PlantDat' dataset created in Step 10 are standardised, if possible, using BODATSA. This is done by merging the 'PlantDat' dataset with the BODATSA dataset to identify the taxon names that occur in both datasets. Depending on the format of the provided taxon names (stipulated in Step 7), this is done based on the canonical name (taxon name without authority) or the scientific name (taxon name with authority). If the taxon name provided is a synonym according to the BODATSA, then the accepted scientific name is obtained. Taxa for which there are multiple matches (i.e., the taxon name provided matches multiple names in BODATSA) are flagged. The output file will contain information on whether the taxon name provided had multiple matches in BODATSA, its status (e.g, accepted, synonym etc), whether the accepted name was corrected using fuzzy matching in Step 5, and some original columns from BODATSA. This output is written to the interim folder as a csv named 'BODATSAInterim.csv'.

Taxon names with multiple matches are resolved by taking the matching name that is an accepted scientific name. If none of the matching names is an accepted scientific name, then the taxon is flagged and the results are concatenated and separated with a pipe delimiter '|'.

Step 15: Assess results and flag issues

The results of the taxonomic standardisation for plants based on BODATSA are evaluated and issues are flagged.

The prevalence of certain issues is also assessed. This is done by looking at how any of the taxa:

1. Fall into each status category (e.g., accepted, synonym etc)
2. Were assigned accepted scientific names that had been corrected through fuzzy matching (Step 4)
3. Had multiple matches
4. Had multiple matches that could not be resolved

Taxonomic issues are flagged. These are:

1. Synonyms
2. Matches that were not exact (e.g., unresolved multiple matches and accepted names corrected through fuzzy matching)
3. Doubtful taxon names (uncertain, missapplied names)
4. Unrecognised taxa - those with no match in BODATSA

In addition, all taxa that either had a non-exact match, a doubtful name, or were not recognised are flagged as having a possible error or issue.

Sub-specific entities are flagged

Duplicated names in either the returned accepted scientific name, canonical name, or the taxon name provided are also flagged.

Step 16: Post-processing

A column is added with the source of the scientific name and the date that source was consulted (as provided in Step 4). The source contains the general link BODATSA; as a stable, unique link for each taxon is not available. Note that the accepted name for the taxon could change over time as BODATSA is revised. Unnecessary columns are removed from the dataset, and the output is written to the 'interim' folder as a csv named 'Plants_BODATSA_standardisation.csv'.

Step 17: Match plant names unrecognised by BODATSA with those in WCVP

The taxon names of plants that had no match in BODATSA (were not recognised) are standardised using the WCVP taxonomic backbone. Depending on the format of the provided taxon names (stipulated in Step 7), this is done based on the taxon name as provided (if provided without the authority) or the canonical name obtained from GBIF (if taxon name was provided with the authority). If the taxon name provided is a synonym according to the WCVP backbone, then the accepted scientific name is obtained. The output file will contain information on whether the taxon name provided had multiple matches in WCVP, the rank of the taxon's name, the status of the name provided, the type of match (e.g., fuzzy), and some other information obtained from WVCP, inclung the and the taxon's unique identification number on POWO 'powo_id'. This output is written to the interim folder as a csv named 'WCVPInterim.csv'.

Taxon names with multiple matches are resolved if possible by taking the matching name with the same authority as that in the provided taxon name (if taxon name was provided with the authority), or failing that, by taking the matching name that is an accepted scientific name. If neither of these processes resolves the issue, then the taxon is flagged and the results are concatenated and separated with a pipe delimiter '|'.

Step 18: Assess results and flag issues

The results of the taxonomic standardisation for plants based on WCVP are evaluated and issues are flagged.

The prevalence of certain issues is also assessed. This is done by looking at how any of the taxa:

1. Fall into each status category (e.g., accepted, synonym etc)
2. Had a specific match type (e.g., fuzzy etc)
3. Fall into different ranks (e.g., species, subspecies etc)
4. Had multiple matches
5. Had multiple matches that could not be resolved

Taxonomic issues are flagged. These are:

1. Synonyms
2. Matches that were not exact (e.g., not exact, unresolved multiple matches and not at species level)
3. Doubtful taxon names (invalid, illegitimate, missapplied)
4. Unrecognised taxa - those with no match in WCVP

In addition, all taxa that either had a non-exact match, a doubtful name, or were not recognised are flagged as having a possible error or issue.

Sub-specific entities are flagged

Duplicated names in either the returned accepted scientific name, canonical name, or the taxon name provided are also flagged.

Step 19: Post-processing

A column is added with the source of the scientific name and the date that source was consulted. The source contains a link to POWO that is unique to the taxon, and which is stable over time. Note that the accepted name for the taxon could change over time as the WCVP taxonomic backbone is revised. Unnecessary columns are removed from the dataset, and the output is written to the 'interim' folder as a csv named 'Plants_WCVP_standardisation.csv'.

Step 20: Create final output

The formats of the interim outputs from the taxon name standardisation processes using the three backbones are standardised. The outputs for plants from the two related processes (BODATSA and WCVP) are aligned (duplicated names are identified). The three interim output datasets are merged and unnecessary columns removed. The number of taxa included in the output is compared to that in the original input file (the number should be equal), and the data in the final output are re-ordered, so that the taxa appear in the same order as in the original input file. The output is written to the 'output' folder as a csv named 'Taxon_names_standardised.csv'.