Standardizing taxa for the Switchboard example script

Roeland KINDT

2025-05-22

Contents

1 Packages needed

```
library(WorldFlora)
library(data.table)
library(readxl)
```

If you are not yet familiar with the WorldFlora package, check this publication.

Standardizations are done with a taxonomical backbone dataset created by World Flora Online.

2 Introduction

For this example, I use a subset of taxa included in World Economic Plants in GRIN-Global for the main economic importance of **Human food** and the subclass of **nut**. This is a subset of the data that was harvested on 8th January 2024 when compiling version 4 of the Switchboard.

Wiersema, J.H. and León, B., 1999. World economic plants: a standard reference. CRC press. https://npgsweb.ars-grin.gov/gringlobal/taxon/taxonomysearchwep

USDA, Agricultural Research Service, National Plant Germplasm System. 2024. Germplasm Resources Information Network (GRIN Taxonomy). National Germplasm Resources Laboratory, Beltsville, Maryland. URL: https://npgsweb.ars-grin.gov/gringlobal/.

Similar analysis pipelines were used for the different databases included in the Switchboard. These scripts can easily be adapted also to standardize other data sets with vascular plant names.

3 Input data

3.1 Provide the location of the Excel file (WorldFlora_data.xlsx)

First download the WorldFlora_data.xlsx file from the Zenodo archive.

```
input.file <- choose.files()
```

3.2 Read the data

```
input.data <- data.frame(read_excel(input.file, sheet="input_1", skip=5))
nrow(input.data)
```

```
## [1] 87
```

```
head(input.data)
```

```
##   SEQ taxonomy_species_id                          Taxonomy        Family
## 1   1                3060          Anacardium occidentale L. Anacardiaceae
## 2   2                7022 Bertholletia excelsa Humb. & Bonpl. Lecythidaceae
## 3   3                8064         Buchanania latifolia Roxb. Anacardiaceae
## 4   4                8815               Canarium indicum L.   Burseraceae
## 5   5                8819              Canarium ovatum Engl.   Burseraceae
## 6   6                9251            Carya cathayensis Sarg.  Juglandaceae
##   Economic.Use Usage.Type
## 1  Human food        nut
## 2  Human food        nut
## 3  Human food        nut
## 4  Human food        nut
## 5  Human food        nut
## 6  Human food        nut
```

4 Step 1 : Split taxonomic names from authorities

WorldFlora has a specific function to split naming authorities from the taxonomic names.

```
output1 <- WorldFlora::WFO.prepare(spec.data=input.data,
                                   spec.full="Taxonomy")
```

```
## Subpattern without punctuation mark (' ind') detected for: Canarium indicum

## Subpattern without punctuation mark (' ind') detected for: Castanopsis indica

## Punctuation mark detected for: Juglans ×bixbyi
## Punctuation mark detected for: Juglans ×bixbyi
```

```
head(output1[, c("spec.name", "Authorship")])
```

```
##                spec.name     Authorship
## 1 Anacardium occidentale             L.
## 2   Bertholletia excelsa Humb. & Bonpl.
## 3   Buchanania latifolia          Roxb.
## 4       Canarium indicum             L.
## 5        Canarium ovatum          Engl.
## 6      Carya cathayensis          Sarg.
```

After saving the file, a manual check was done.

```
file.save <- paste0(getwd(), "//working.txt")
fwrite(output1, file=file.save, sep="|", row.names=FALSE)
```

After the manual check, the checked data was put in a new sheet.

```
input2 <- data.frame(read_excel(input.file, sheet="input_2"))
nrow(input2)
```

```
## [1] 87
```

```
head(input2)
```

```
##   SEQ taxonomy_species_id                            Taxonomy        Family
## 1   1                3060           Anacardium occidentale L. Anacardiaceae
## 2   2                7022 Bertholletia excelsa Humb. & Bonpl. Lecythidaceae
## 3   3                8064          Buchanania latifolia Roxb. Anacardiaceae
## 4   4                8815                Canarium indicum L.   Burseraceae
## 5   5                8819               Canarium ovatum Engl.   Burseraceae
## 6   6                9251             Carya cathayensis Sarg.  Juglandaceae
##                spec.name     Authorship
## 1 Anacardium occidentale             L.
## 2   Bertholletia excelsa Humb. & Bonpl.
## 3   Buchanania latifolia          Roxb.
## 4       Canarium indicum             L.
## 5        Canarium ovatum          Engl.
## 6      Carya cathayensis          Sarg.
```

5 Step 2 : Standardize

Now the input data is ready to be standardized.

5.1 First modify the names of hybrid species in the downloaded taxonomic backbone

In this document, I use the same version of the World Flora Online (WFO, version 2023.12 downloaded from Zenodo) taxonomic backbone for the standardizations for the Switchboard.

As shown previously in Rpubs, I recommend to first use a **text editor** to replace instances of ' × ' by ' ×' for newer (2023 and later) versions of the taxonomic backbone.

5.2 Load the taxonomic backbone data file of World Flora Online

I used the latest available version of World Flora Online of v.2023.12. The download was done earlier, followed by providing the location of the file to **WFO.download** via its argument 'WFO.file'.

(You could use **WFO.remember(WFO.file=file.choose())** to be certain that the right version of the WFO backbone is used.)

In the taxonomic backbone, World Flora Online lists about half a million current species names.

```
WFO.remember()
```

```
## Data sourced from: E:\Roeland\WorldFloraOnline\2023\wfo_202312_RK.csv (Mon Sep  9 14:10:58 2024)
```

```
## Reading WFO data
```

```
## The WFO data is now available from WFO.data
```

```
nrow(WFO.data)
```

```
## [1] 1576062
```

```
nrow(WFO.data[WFO.data$taxonRank == "species", ])
```

```
## [1] 1164296
```

```
nrow(WFO.data[WFO.data$taxonRank == "species" & WFO.data$acceptedNameUsageID == "", ])
```

```
## [1] 517755
```

5.3 Use WFO.match.fuzzyjoin

Everything is in place now to start the name matching pipelines.

To avoid a crash of the **WFO.match.fuzzyjoin()** function with large input datasets, however, the data needs to be split. This can be done relatively easily via the **cut()** function. For the example dataset, this is not required, but for larger datasets and especially those with many names that can not be directly matched, the **breaks** parameter needs to be increased to 10 or 20.

```
cuts <- cut(c(1:nrow(input2)), breaks=5, labels=FALSE)
cut.i <- sort(unique(cuts))

start.time <- Sys.time()

for (i in 1:length(cut.i)) {

cat(paste("Cut: ", i, "\n"))

input.i <- WFO.one(WFO.match.fuzzyjoin(spec.data=input2[cuts==cut.i[i], ],
                                       WFO.data=WFO.data,
                                       spec.name="spec.name",
                                       Authorship="Authorship",
                                       fuzzydist.max=3),
                   Auth.dist = "Auth.dist",
                   Old.author.dist="Old.author.dist",
                   verbose=FALSE)

if (i==1) {
  input.WFO <- input.i
}else{
  input.WFO <- rbind(input.WFO, input.i)
}

}
```

```
## Cut:  1

##
## Checking new accepted IDs

## Cut:  2

##
## Checking new accepted IDs

## Cut:  3

##
## Checking new accepted IDs

## Cut:  4

##
## Checking new accepted IDs

## Warning in min(as.numeric(WFO.case[WFO.case$New.accepted == TRUE,
## Old.author.dist]), : no non-missing arguments to min; returning Inf

## Cut:  5
```

```
##
## Checking new accepted IDs
```

```
## Warning in min(as.numeric(WFO.case[WFO.case$New.accepted == FALSE,
## Auth.dist]), : no non-missing arguments to min; returning Inf
```

```
end.time <- Sys.time()
end.time - start.time
```

```
## Time difference of 15.32039 secs
```

For datasets where there is no information on the name authority, arguments for **Authorship**, **Auth.dist** and **Old.author.dist** should not be included.

To make checking of fuzzy matches easier, the following function flags acceptable matches

```
accept.var <- WFO.acceptable.match(input.WFO,
                                   spec.name="spec.name",
                                   no.vowels=TRUE)

input.WFO <- data.frame(input.WFO,
                        acceptable=accept.var)
```

```
save.file <- paste0(getwd(), "//WFO matches.txt")

fwrite(input.WFO, file=save.file, sep="|", row.names=FALSE)
```

6 Step 3 : Manual checks

Outputs from standardizations by **WorldFlora** were examined afterwards, especially checking fuzzy matches and records with a large distance between the submitted and matched authorities.

One of the changes that was implemented after manual checks was to modify the match of **Prunus dulcis (Mill.) D. A. Webb** to **Prunus dulcis D.A. Webb** with taxonID wfo-0001005398.

7 Session Information

```
sessionInfo()
```

```
## R version 4.2.1 (2022-06-23 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.utf8
## [2] LC_CTYPE=English_United Kingdom.utf8
```

```
## [3] LC_MONETARY=English_United Kingdom.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] readxl_1.4.1      data.table_1.14.2 WorldFlora_1.14-5
##
## loaded via a namespace (and not attached):
##  [1] rstudioapi_0.14  knitr_1.40       magrittr_2.0.3   tidyselect_1.2.0
##  [5] R6_2.5.1         rlang_1.1.1      fastmap_1.1.1    fansi_1.0.3
##  [9] stringr_1.4.1    dplyr_1.1.2      tools_4.2.1      xfun_0.33
## [13] utf8_1.2.2       cli_3.4.1        htmltools_0.5.6  yaml_2.3.5
## [17] digest_0.6.29    tibble_3.2.1     lifecycle_1.0.3  vctrs_0.6.3
## [21] glue_1.6.2       evaluate_0.16    rmarkdown_2.16   stringi_1.7.8
## [25] compiler_4.2.1   pillar_1.9.0     cellranger_1.1.0 generics_0.1.3
## [29] pkgconfig_2.0.3
```