Species Validation Tool

30 July 2021

Lauri Vesa, Javier Garcia-Perez, Elisée Tchana
National Forest Monitoring Team
FAO Forestry Division
Rome, Italy

Developed under the FAO Open Foris initiative (http://www.openforis.org/)

1. About this application

The purpose on this tool is to provide an easy application for the validation of tree or plant species taxonomic information. With the help of this tool, a botany expert can improve the quality of information for example in (national) forest inventories and other studies where plant data are collected.

In the Internet there are multiple global databases (e.g., Catalogue of Life, Integrated Taxonomic Information System, Tropicos) that encompass millions of species and most of these databases that can be freely accessed online or by downloading the database into the local machine (see e.g. García-Pérez 2017, Kindt 2020). Our method was to use R language, a set of R packages already specialized in retrieving and processing taxonomic data, and chain these packages to facilitate the validation of tree or plant species lists. In general, taxonomic uncertainty is one of the major sources of error across global plant ecological, biogeographical, and conservation studies and applications, and misspellings of scientific names can lead to failures to retrieve data from global databases that encompass millions of species (Boyle *et al*. 2013, Meyer *et al.* 2016).

The current version works only offline in a local machine. Later, FAO aims to add this application into the SEPAL Toolbox so that it can be accessible online.

To cite this R application and this manual in publications use:

```
Vesa, L., Garcia-Perez, J., Tchana, E. (2021). Species Validation Tool for R. NFM
Team, FAO Forestry Division. Rome, Italy.
```

2. How this tool works

This tool can be used to validate a list containing trees' or plants' **scientific names**. The data imported to the Species Validation Tool are subject to a series of iterative processes and checks. The following databases are used to validate the inputted species names:

1) **LCVP** - Leipzig Catalogue of Vascular Plants [offline],
2) **Tropicos** - Missouri Botanical Garden [online],
3) **Kew** - Plants of the World Online [online], *)
4) **NCBI** - National Center for Biotechnology Information, db="taxonomy" [online],
5) **WFO** - World Flora Online [offline],
6) **GBIF** - Global Biodiversity Information Facility [online].

[offline]: search database is loaded to the server/computer, and search is done locally,
[online]: search via online with an application programming interface (API).

The search algorithm first validates all inputted names against LCVP (Figure 1). After running this search, it drops out from the search list all names where ***Status*** is tagged as "accepted" or "synonym". Then the search continues with the next repository, Tropicos. From Tropicos up to GBIF, the script drops out all "accepted" cases, and continues to up to the end of this repository list, if there are any species names left.

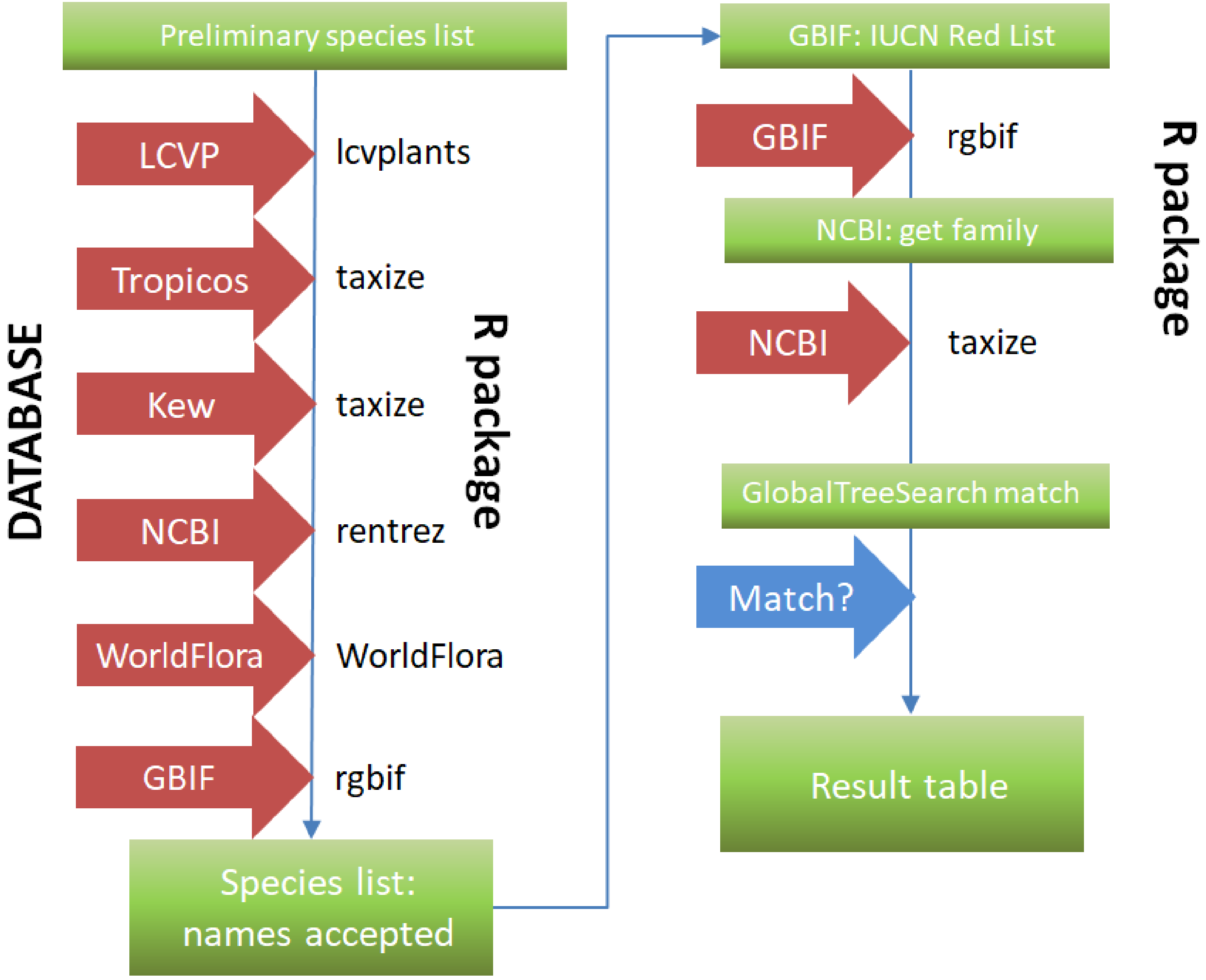


Figure 1. Data processing chain.

The reason for running the search in this order is purely pragmatic. We start with a complete offline-type database, LCVP, because an online search with a long species list is always a bit risky: connection may be cut or refused by the remote server, or the server does not offer a bulk query mode. It is efficient to start searching the hits from locally installed database first, and then use the online servers.

Note:
GBIF cannot be considered as a trustful source for validation of the species names, because this repository contains information from different sources, and in some cases the stored names in GBIF may be outdated. So please be critical to these GBIF hits and validate the hits from other sources too!

The following repositories are not used for validation of the names, but for other purposes as follows:

7) GBIF search to look for the International Union for Conservation of Nature (IUCN) **Red List status** [online],
8) **GTS - Global Tree Search** [offline]. This repository is used to check name occurrences in this list.

Note that the results provided through data processing may not be completely true or accurate due to absences of the name in the selected global databases or due to the unreliability of the source. In addition, notice that in the World Flora Online (WFO) search we use fuzzy matching algorithm that returns just one match, although a more efficient method would be to return several fuzzy alternatives.

GBIF and IUCN Red List categories

The search algorithm currently uses GBIF as the source for checking the IUCN Red List status. GBIF is free and fast source for querying. However, the real source of information should be the IUCN database, but it contains limitations how access it programmatically. This part of the search may be changed soon. The second fact to note is that GBIF may not contain/give Red List status results for all species names.

The Red List categories are as follows:

- NE: Not evaluated,
- DD: Data deficient,
- LC: Least Concern,
- NT: Near Threatened,
- VU: Vulnerable,
- EN: Endangered,
- CR: Critically Endangered,
- EW: Extinct in the wild,
- EX: Extinct.

The additional category '*No Red List Data*' (ND) tells that this species name exists in GBIF, but its IUCN Red List category is not given.

3. How to access the application and offline databases

Currently, this tool is available as R scripts that need to be downloaded and run locally. The following two software need to be installed: R (version 3.6 or newer) and RStudio.

This documentation and R scripts are available through Open Foris website:
http://www.openforis.org/materials.html

and from GitHub: https://github.com/openforis/shiny-species_validation

The following R packages are required (install all them except *lcvplants* from CRAN):

- shiny
- tidyverse
- leaflet.extras
- rvest
- parallel
- utils
- stringr
- data.table
- taxize
- rentrez
- rgbif
- plyr
- WorldFlora
- textclean
- foreach
- doParallel
- lcvplants

```
library(devtools)
devtools::install_github("idiv-
biodiversity/lcvplants")
```

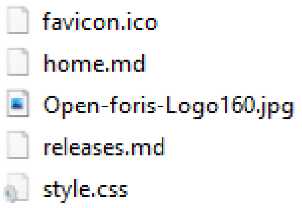
- if Linux OS: doMC

The application consists of the following files that all need to be in the same folder:

1) *server.R,*
2) *ui.R,*
3) *Species_search_30_July_2021.R*[1]

and three databases (LCVP, WFO, GTS; see instructions below).

For getting files for Shiny user-interface, get copy of the folder **/www**
It contains the following files:



favicon.ico
home.md
Open-foris-Logo160.jpg
releases.md
style.css

Three local (offline) databases need to be downloaded:

1) WFO: classification.txt (unzip the downloaded file into the application's main folder)
   This file is originated from here (ver. 2019.05, May 17, 2019). Copy this file to the application's folder.

[1] This file name is subject to change.

2) Data of the LCVP package, which you can install as follows:

```
library(devtools)
devtools::install_github("idiv-biodiversity/LCVP")
```

3) GTS database[2]. Get *global_tree_search_trees_1_5.csv* from here. Copy this file to the application's folder.

4. Setting API keys

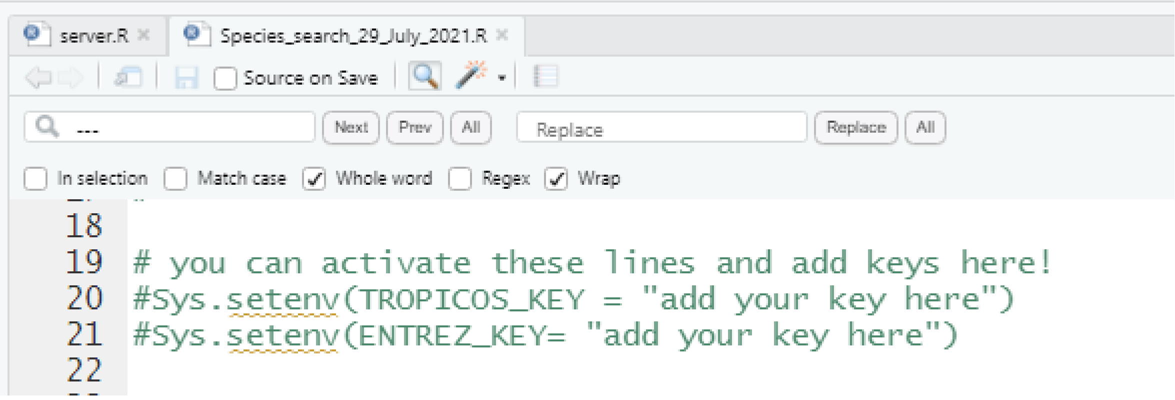
Tropicos and NCBI servers require that you have got the API keys. This key is a code that gets passed in by computer applications and it is used to identify the user. The keys can be accessed here (as in July 2021):

https://services.tropicos.org/help?requestkey

https://ncbiinsights.ncbi.nlm.nih.gov/2017/11/02/new-api-keys-for-the-e-utilities/

Read here more how to set an API key into with R. The keys can be added into the codes as follows:

```
Sys.setenv(TROPICOS_KEY = "your_Tropicos_API_key")

Sys.setenv(ENTREZ_KEY= "your_NCBI_API_key")
```



☐ Source on Save

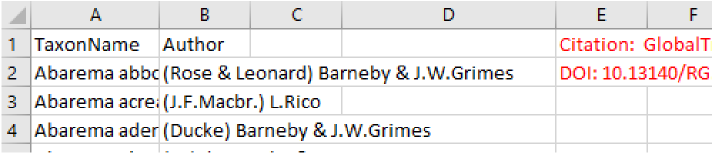
--- ☐ In selection ☐ Match case ☑ Whole word ☐ Regex ☑ Wrap

```

# you can activate these lines and add keys here!
#Sys.setenv(TROPICOS_KEY = "add your key here")
#Sys.setenv(ENTREZ_KEY= "add your key here")

```

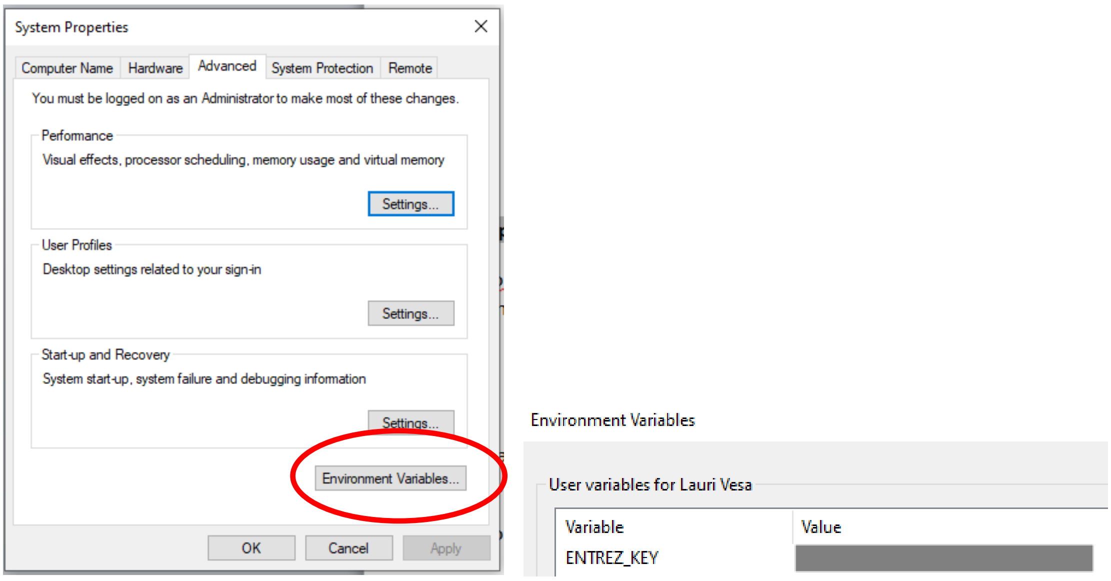
Figure 2. API keys in the code.

Another method to set in your API key in Windows machine is to use environment variables (and to add variables *TROPICOS_KEY* and *ENTREZ_KEY*).

[2] You can also download the GTS database from https://tools.bgci.org/global_tree_search.php but you need to edit it: delete extra items in cells E1:E2.



	A	B	C	D	E	F
1	TaxonName	Author			Citation: GlobalT	
2	Abarema abbc	(Rose & Leonard) Barneby & J.W.Grimes			DOI: 10.13140/RG	
3	Abarema acre	(J.F.Macbr.) L.Rico				
4	Abarema ader	(Ducke) Barneby & J.W.Grimes				



System Properties

You must be logged on as an Administrator to make most of these changes.

Performance

Visual effects, processor scheduling, memory usage and virtual memory

User Profiles

Desktop settings related to your sign-in

Start-up and Recovery

System start-up, system failure and debugging information

Environment Variables

User variables for Lauri Vesa

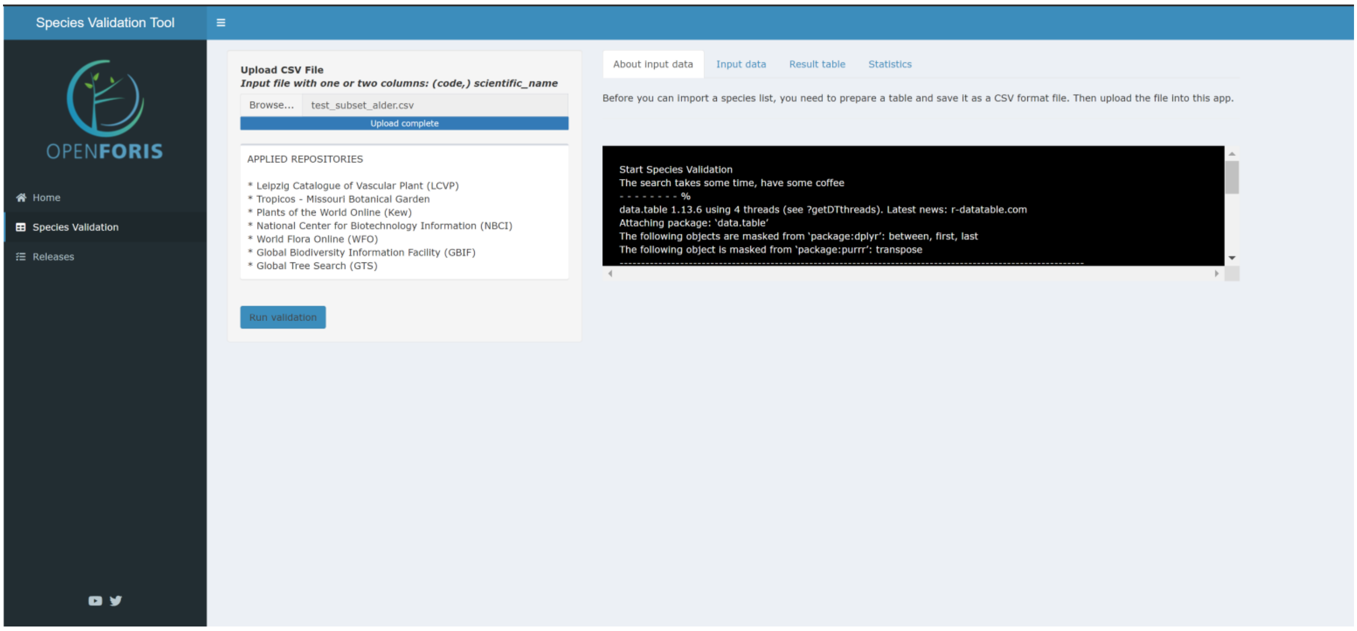
Variable	Value
ENTREZ_KEY	[redacted]

Figure 3. Setting an environmental variable into Windows OS.

5. User interface

The left side panel provides access to three main pages:

1) Home. This page contains general information about the application, and the links to the global taxonomic repositories.
2) Species Validation. This page contains the application (Figure 4 and 5).
3) Releases. This page contains information about the version.



Upload CSV File

Input file with one or two columns: (code,) scientific_name

test_subset_alder.csv

Upload complete

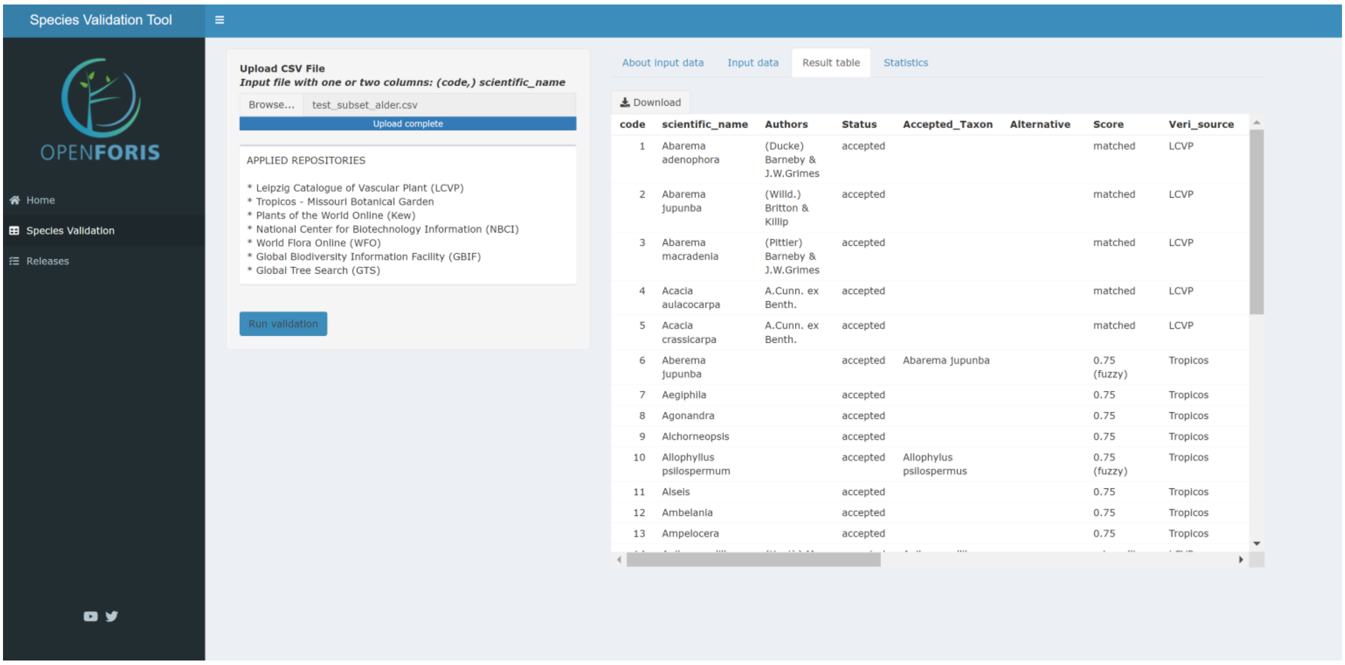
APPLIED REPOSITORIES

* Leipzig Catalogue of Vascular Plant (LCVP)
* Tropicos - Missouri Botanical Garden
* Plants of the World Online (Kew)
* National Center for Biotechnology Information (NBCI)
* World Flora Online (WFO)
* Global Biodiversity Information Facility (GBIF)
* Global Tree Search (GTS)

Before you can import a species list, you need to prepare a table and save it as a CSV format file. Then upload the file into this app.

```
Start Species Validation
The search takes some time, have some coffee
- - - - - - - - %
data.table 1.13.6 using 4 threads (see ?getDTthreads). Latest news: r-datatable.com
Attaching package: 'data.table'
The following objects are masked from 'package:dplyr': between, first, last
The following object is masked from 'package:purrr': transpose
```

Figure 4. The “Species Validation” window.



Upload CSV File
Input file with one or two columns: (code,) scientific_name

test_subset_alder.csv

APPLIED REPOSITORIES

- Leipzig Catalogue of Vascular Plant (LCVP)
- Tropicos - Missouri Botanical Garden
- Plants of the World Online (Kew)
- National Center for Biotechnology Information (NBCI)
- World Flora Online (WFO)
- Global Biodiversity Information Facility (GBIF)
- Global Tree Search (GTS)

code	scientific_name	Authors	Status	Accepted_Taxon	Alternative	Score	Veri_source
1	Abarema adenophora	(Ducke) Barneby & J.W.Grimes	accepted			matched	LCVP
2	Abarema jupunba	(Willd.) Britton & Killip	accepted			matched	LCVP
3	Abarema macradenia	(Pittier) Barneby & J.W.Grimes	accepted			matched	LCVP
4	Acacia aulacocarpa	A.Cunn. ex Benth.	accepted			matched	LCVP
5	Acacia crassicarpa	A.Cunn. ex Benth.	accepted			matched	LCVP
6	Aberema jupunba		accepted	Abarema jupunba		0.75 (fuzzy)	Tropicos
7	Aegiphila		accepted			0.75	Tropicos
8	Agonandra		accepted			0.75	Tropicos
9	Alchorneopsis		accepted			0.75	Tropicos
10	Allophyllus psilospermum		accepted	Allophylus psilospermus		0.75 (fuzzy)	Tropicos
11	Alseis		accepted			0.75	Tropicos
12	Ambelania		accepted			0.75	Tropicos
13	Ampelocera		accepted			0.75	Tropicos

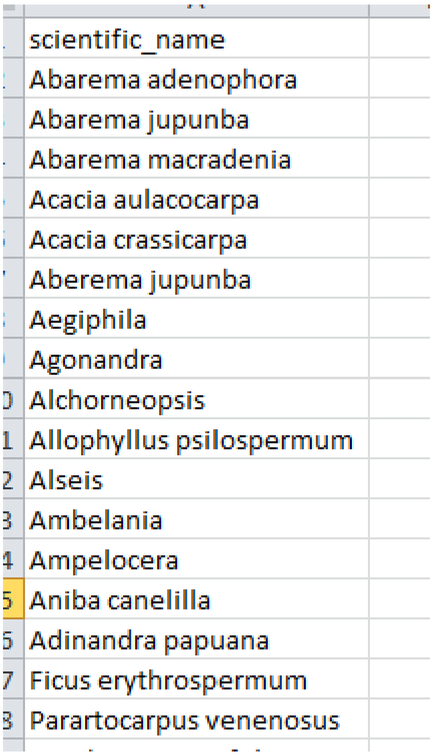
Figure 5. The result table shown in the application.

6. How to use this tool

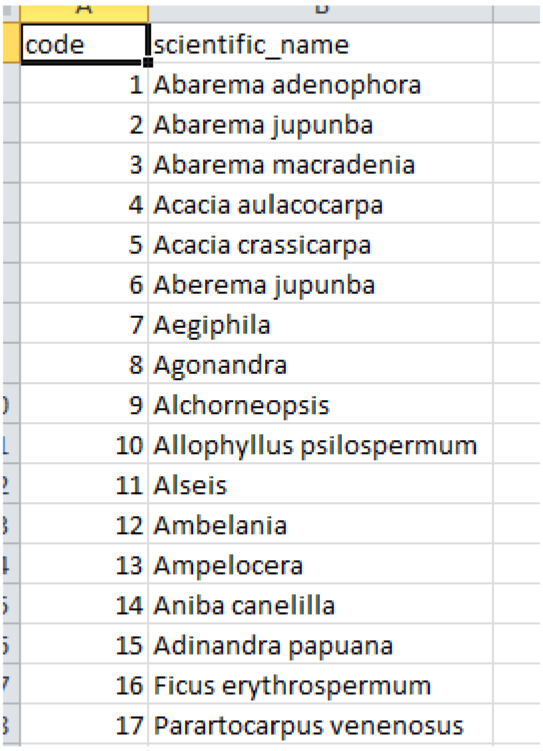
Download all application files and folder '*/www*' into your computer, see Chapter 3.

Start the application by opening file *server.R* in RStudio, then run the application by clicking ▶ Run App button. In RStudio browser, click Open in Browser and the application gets open in your default web browser.

First create a list of scientific species names are same it as a comma-separated text file (CSV). The input data can contain one or two columns, as follows:



scientific_name
Abarema adenophora
Abarema jupunba
Abarema macradenia
Acacia aulacocarpa
Acacia crassicarpa
Aberema jupunba
Aegiphila
Agonandra
Alchorneopsis
Allophyllus psilospermum
Alseis
Ambelania
Ampelocera
Aniba canelilla
Adinandra papuana
Ficus erythrospermum
Parartocarpus venenosus



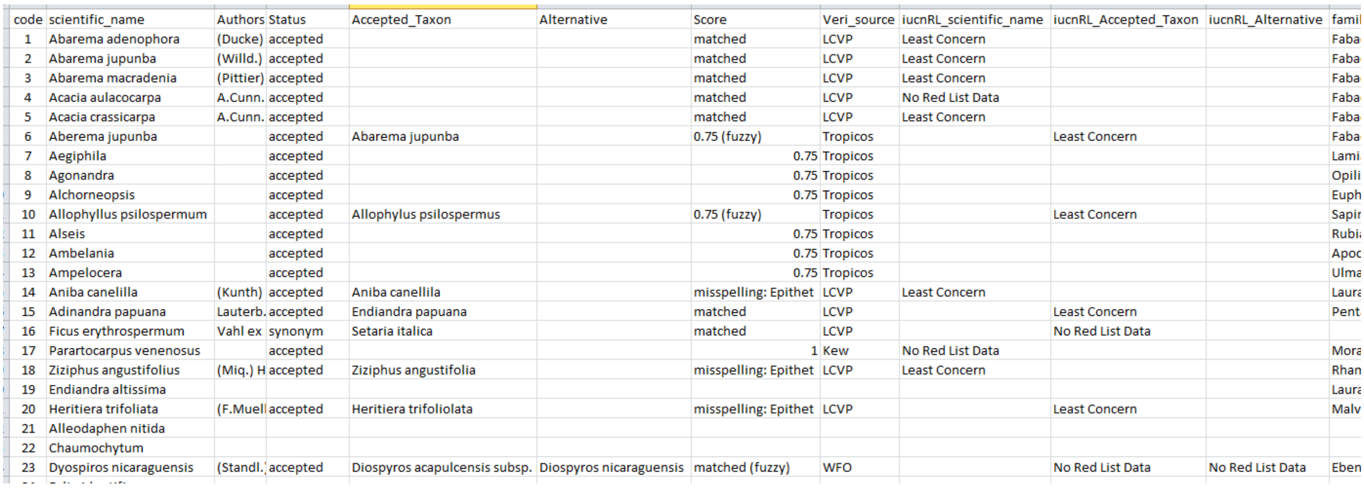
code	scientific_name
1	Abarema adenophora
2	Abarema jupunba
3	Abarema macradenia
4	Acacia aulacocarpa
5	Acacia crassicarpa
6	Aberema jupunba
7	Aegiphila
8	Agonandra
9	Alchorneopsis
10	Allophyllus psilospermum
11	Alseis
12	Ambelania
13	Ampelocera
14	Aniba canelilla
15	Adinandra papuana
16	Ficus erythrospermum
17	Parartocarpus venenosus

If a species code is given, it will be also written into the result table. The list can contain species names, or genera only. If the input data is a genus name, its Red List status nor match in the Global Tree Search database are not examined.

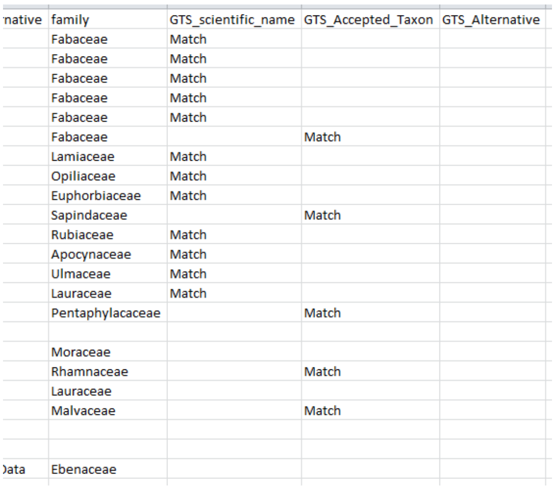
Upload the data into the application and click "Run validation".

Notice: **the current script is very slow, so let it run and take a cup of coffee (or two) but do not close your web browser!**

Once when the processing is completed, the results are shown in the tab sheet "Result table". The result table looks as follows (split into two images):



code	scientific_name	Authors	Status	Accepted_Taxon	Alternative	Score	Veri_source	iucnRL_scientific_name	iucnRL_Accepted_Taxon	iucnRL_Alternative	famil
1	Abarema adenophora	(Ducke)	accepted			matched	LCVP	Least Concern			Faba
2	Abarema jupunba	(Willd.)	accepted			matched	LCVP	Least Concern			Faba
3	Abarema macradenia	(Pittier)	accepted			matched	LCVP	Least Concern			Faba
4	Acacia aulacocarpa	A.Cunn.	accepted			matched	LCVP	No Red List Data			Faba
5	Acacia crassicarpa	A.Cunn.	accepted			matched	LCVP	Least Concern			Faba
6	Aberema jupunba		accepted	Abarema jupunba		0.75 (fuzzy)	Tropicos		Least Concern		Faba
7	Aegiphila		accepted			0.75	Tropicos				Lami
8	Agonandra		accepted			0.75	Tropicos				Opili
9	Alchorneopsis		accepted			0.75	Tropicos				Euph
10	Allophyllus psilospermum		accepted	Allophylus psilospermus		0.75 (fuzzy)	Tropicos		Least Concern		Sapir
11	Alseis		accepted			0.75	Tropicos				Rubi
12	Ambelania		accepted			0.75	Tropicos				Apoc
13	Ampelocera		accepted			0.75	Tropicos				Ulma
14	Aniba canelilla	(Kunth)	accepted	Aniba canellila		misspelling: Epithet	LCVP	Least Concern			Laura
15	Adinandra papuana	Lauterb.	accepted	Endiandra papuana		matched	LCVP		Least Concern		Pent
16	Ficus erythrospermum	Vahl ex	synonym	Setaria italica		matched	LCVP		No Red List Data		
17	Parartocarpus venenosus		accepted			1	Kew	No Red List Data			Mora
18	Ziziphus angustifolius	(Miq.) H	accepted	Ziziphus angustifolia		misspelling: Epithet	LCVP	Least Concern			Rhan
19	Endiandra altissima										Laura
20	Heritiera trifoliata	(F.Muell	accepted	Heritiera trifoliolata		misspelling: Epithet	LCVP		Least Concern		Malv
21	Alleodaphen nitida										
22	Chaumochytum										
23	Dyospiros nicaraguensis	(Standl.)	accepted	Diospyros acapulcensis subsp.	Diospyros nicaraguensis	matched (fuzzy)	WFO		No Red List Data	No Red List Data	Eben



native	family	GTS_scientific_name	GTS_Accepted_Taxon	GTS_Alternative
	Fabaceae	Match		
	Fabaceae	Match		
	Fabaceae	Match		
	Fabaceae	Match		
	Fabaceae	Match		
	Fabaceae		Match	
	Lamiaceae	Match		
	Opiliaceae	Match		
	Euphorbiaceae	Match		
	Sapindaceae		Match	
	Rubiaceae	Match		
	Apocynaceae	Match		
	Ulmaceae	Match		
	Lauraceae	Match		
	Pentaphylacaceae		Match	
	Moraceae			
	Rhamnaceae		Match	
	Lauraceae			
	Malvaceae		Match	
)ata	Ebenaceae			

The result table's columns are explained in the next table.

Column header	Explanation
code	Species code (input data)
scientific_name	Species name (input data)
Authors	Authors' names (if given)
Status	*Accepted*: name found (may still contain typo, see the next column) *Synonym*. *Blank*: no hit founds in the validation search.
Accepted_Taxon	May contain result for a fuzzy search or a synonym.
Alternative	LCVP: match according to The Plant List (if exists).
Score	Score as returned by the repository (see the next column).
Veri_source	Source of verification (or hit).
iucnRL_scientific_name	GBIF: IUCN Red List status for 'scientific_name'
iucnRL_Accepted_Taxon	GBIF: IUCN Red List status for 'Accepted_Taxon'
iucnRL_Alternative	GBIF: IUCN Red List status for 'Alternative'
family	Family name. This data may be missing (as in case of *Ficus* sp.)
GTS_scientific_name	GTS hit for 'scientific_name'
GTS_Accepted_Taxon	GTS hit for 'Accepted_Taxon'
GTS_Alternative	GTS hit for 'Alternative'

Under the tab "Statistics" there are two tables. The first result table contains number of species names validated against the listed repository:

repo_name	species_count	
LCVP	28	28 inputted names (all) into the LCVP search. 10 hits (=28-18).
Tropicos	18	18 names inputted, 8 hits.
Kew	10	10 names inputted, 1 hit.
NCBI	9	9 names inputted, no hits.
WFO	9	9 names inputted, 3 hits.
GBIF	6	6 names inputted, no hits.
no hits	6	6 names totally unknown (i.e., no hits).

The second result table shows the counts of hits against the repositories, aggregated by the combined 'Status' and 'Score' fields.

Status_Score	LCVP	Tropicos	Kew	WFO	nodata
-	0	0	0	0	6
accepted - 0.75	0	6	0	0	0
accepted - 0.75 (fuzzy)	0	2	0	0	0
accepted - 1	0	0	1	0	0
accepted - matched	6	0	0	0	0
accepted - matched (fuzzy)	0	0	0	3	0
accepted - misspelling: Epithet	3	0	0	0	0
synonym - matched	1	0	0	0	0

References

Boyle, B., N. Hopkins, Z. Lu, J. A. R. Garay, D. Mozzherin, T. Rees, N. Matasci, *et al.* (2013). The taxonomic name resolution service: An online tool for automated standardization of plant names. BMC Bioinformatics 14: 16.

Freiberg, M., Winter, M., Gentile, A. *et al.* (2020). LCVP, The Leipzig catalogue of vascular plants, a new taxonomic reference list for all known vascular plants. Sci Data 7, 416. https://doi.org/10.1038/s41597-020-00702-z

García-Pérez, J. (2017). The use of global biodiversity databases to increase taxonomic quality in forest inventories. Inaugural Global Forest Biodiversity Initiative Conference & GFBI-FECS Joint Symposium. "Forest Research in the Big Data Era", September 6-9, 2017 Beijing, China. http://docs.wixstatic.com/ugd/07a4b9_0bc71281e26e44d59b762a1acf7f1110.pdf

Kindt, R. (2020). WorldFlora: An R package for exact and fuzzy matching of plant names against the World Flora Online taxonomic backbone data. Applications in Plant Sciences 8(9): https://doi.org/10.1002/aps3.11388

Meyer, C., Weigelt, P. and H. Kreft. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecology Letters 19: 992–1006.