TerraFund Tree Species Data Cleaning and Connection to ICRAF data

Cleaning report level trees species data in TerraMatch using ICRAF support and utilizing ICRAF resources to attach tree species data to academic datasets on nativity, distribution, and uses

Contents

[Purpose 1](#_Toc170315413)

[Goals 2](#_Toc170315414)

[Timeline 2](#_Toc170315415)

[Resources 3](#_Toc170315416)

# Purpose

Data cleaning

Tree species data in TerraFund programs, both Top 100 and Landscapes (cohort 1 and 2, respectively), is collected in myriad questions and formats and at several stages of the application, selection, and reporting process. During application submission, we ask champions to detail which tree species they have grown historically and which tree species they intend to plant in their proposed project. At the contract negotiation stage after champions have been selected for funding, champions provide a finalized list of the tree species they have agreed, with our vettors approval, to plant. Lastly, we track the ongoing planting of these tree species throughout the project via 6-month reports, in which champions list all of the tree species they have planted to that point.

Each of these questions relies upon champions entering data in open text format. This presents several challenges for evaluating and analyzing tree species data:

1. Spelling mistakes are frequent.
2. Complete Latin taxonomic species names in *Genus species* format are not provided. Often, only a genus is provided.
3. Common names are often provided in place of or in addition to Latin species names, in French, English, and other languages.
4. Local names are often used.
5. Different species names are used to describe the same species, due to updates in taxonomy over time.

It is necessary to clean the tree species data within TerraFund using an updated, scientifically relevant, taxonomic backbone from the [World Flora Online](https://www.worldfloraonline.org/downloadData) to enable the team to perform high level data analysis and create visualizations using our tree species data. Furthermore, cleaning the tree species data will allow for connections to other databases used by ICRAF and other researchers in this space, including the [GlobalUsefulNativeTrees (GlobUNT)](https://patspo.shinyapps.io/GlobalUsefulTrees/), Botanic Gardens Conservation International [GlobalTreeSearch (GTS)](https://tools.bgci.org/global_tree_search.php), and the Royal Botanic Gardens, Kew’s [World Checklist of Useful Plant Species (WCUPS).](https://knb.ecoinformatics.org/view/doi:10.5063/F1CV4G34)

Connecting to other datasets

Once the tree species data is cleaned, we can connect these data using the tree species name to existing datasets that will provide information on the following topics:

1. Country distribution
2. Climate or observed environment range
3. Use

These data will largely come from the GlobUNT dataset, which contains data on 14,014 tree species. When not included in GlobUNT, we can access the WCUPS data for species use, GTS for country distribution, and [Tree Globally Observed Environmental Ranges (TreeGOER)](https://onlinelibrary.wiley.com/doi/full/10.1111/gcb.16914) for the environmental ranges, as these datasets feed into GlobUNT.

Eventually, this data will be directly imported to TerraMatch to help champions with their tree species selection.

# Goals

1. Cleaned species data in *Genus species* format for all tree species submitted by champions in Cohort 1 project reports on TerraMatch using the WFO backbone and the [RPub series](https://rpubs.com/Roeland-KINDT/1134151) provided by Roeland Kindt at ICRAF.
2. Connect the cleaned tree species data to the GlobUNT dataset provided by ICRAF for all African countries.
3. For any tree species not in GlobUNT, connect to the TreeGoer, GTS, and WCUPS datasets.
4. Import the cleaned tree species data to TerraMatch and the Unified Database Projects table.
5. Import the datasets to TerraMatch.
6. Improve the functionality of the TerraMatch application to allow for dropdown or automatic entry of tree species information, rather than open text boxes.

# Timeline

* **May 2024:** 
  + Collect tree species export from TerraMatch
  + Download WFO Taxonomic Backbone
  + Receive export from ICRAF of all African country data in GlobUNT
  + Set up code for cleaning tree species data
* **June-August 2024:**
  + Follow RPub series and use WFO backbone to clean TM tree species data using R Studio.
  + Verify the accuracy of the tree species matching using the backbone.
  + Edit the tree species data from TM that did not Fuzzy Match to the backbone
    - E.g., changing different spellings of avocado so it will match to *Persea americana*
* **August 2024:**
  + Join the cleaned data tree species data to the GlobUNT dataset to get data on tree species distribution, environmental range, and use
  + Import data to UDB and TerraMatch
* **September 2024**
  + Run analysis on data to see the portfolio of tree species in Restoration

# Resources

All R scripts and raw data will be shared on [GitHub](https://github.com/wri/restoration-tree-species-cleaning).

1. WFO (2024): World Flora Online. Version 2024.June. Published on the Internet; http://www.worldfloraonline.org. Accessed on:June 3, 2024.
2. 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): e11388.
3. Beech E, Rivers M, Oldfield S and Smith PP. 2017 GlobalTreeSearch: the first complete global database of tree species and country distributions Journal of Sustainable Forestry 36(5): 454-489.
4. BGCI. 2022. GlobalTreeSearch online database (version 1.6) Botanic Gardens Conservation International. Richmond, UK. DOI: 10.13140/RG.2.2.34206.61761 https://tools.bgci.org/global\_tree\_search.php
5. Diazgranados M, Allkin B, Black N, Cámara-Leret R., Canteiro C., Carretero J, Eastwood R, Hargreaves S, Hudson A, Milliken W, Nesbitt M, Ondo I, Patmore K, Pironon S, Turner R and Ulian T. 2020. World Checklist of Useful Plant Species Produced by the Royal Botanic Gardens, Kew. Knowledge Network for Biocomplexity.
6. Kindt R, Graudal L, Lillesø J-P, Pedercini F, Smith P and Jamnadass R. 2024. GlobalUsefulNativeTrees: country-specific selections of native tree species for forest landscape restoration and biodiverse plantings https://patspo.shinyapps.io/GlobalUsefulTrees/
7. Kindt R, Graudal L, Lillesø J-P, Pedercini F, Smith P and Jamnadass R. 2023. GlobalUsefulNativeTrees, a database documenting 14,014 tree species, supports synergies between biodiversity recovery and local livelihoods in landscape restoration Sci Rep 13, 12640 (2023). https://doi.org/10.1038/s41598-023-39552-1