# Guide explicatif de la stratégie d’harmonisation des espèces dans la base de données PostgreSQL

## Emplacement du projet R :

P:\F1272\ProjetCC\USDA-FIA\harmonisationEspeces

## Contenu du projet R :

Fichiers R :

* « .RData »  – Workspace du projet R
* « harmonisationEspeces.Rproj » -- Fichier du projet
* « script\_harmonisation.R » -- Script principal du projet contenant le code R
* « fonctions.R » -- Script secondaire contenant les fonctions utiles au projet

Fichier SQL :

* « especes\_tableau\_ref\_species.sql » -- Script contenant la requête SQL pour interroger la base de données sur la liste des espèces pertinentes

Fichier texte :

* « modifications\_codes\_ccbio.txt » -- Fichier texte expliquant les modifications apportées aux codes CCBIO existants
* « codeESSoriginal.csv » -- fichier contenant les codes d’espèces avant l’harmonisation
* « liste\_especes\_francais\_mod.csv » -- fichier dérivé de « codeESSoriginal.csv » où certains noms de codes et d’essences ont été modifiés tel qu’indiqué dans « modifications\_codes\_ccbio.txt »
* « doublons.csv » -- Fichier des doublons générés lors de la requête
* « table\_des\_equivalences\_especes.csv » -- Fichier final des équivalences entre les noms d’espèces

Autre :

* «Tableau d'équivalence des espèces et modifications apportées aux codes CCBIO existants.msg» -- Copie du courriel contenant les résultats
* « shape\_parcelles » -- Dossier contenant le projet QGIS et les shapefiles qui ont servi à sélectionner les états d’intérêt au projet avec une intersection

## Stratégie d’harmonisation :

Actions accomplies par le script :

* Connection à la base de données Postgres (package RPostgreSQL)
* Exécution de la requête et récupération du résultat en tant que data frame dans R
* Vérification des instances où les codes CCBIO générés pour les espèces américaines par la requête ont créé des doublons
* Vérification manuelle des doublons et modification de certains codes dans le data frame pour harmoniser avec les codes existants
* Modification de certains codes existants au besoin (ces modifications sont explicitées dans le fichier texte (voir section précédente)
* Fusion du data frame des espèces noms américains avec celui des espèces noms français basée sur le champ CODE\_CCBIO
* Écriture en CSV de la table résultante.

## Explication de la requête SQL pour obtenir la liste des espèces :

La requête PostgreSQL se divise en 5 étapes qui comprennent chacune un SELECT, les étapes vont du plus au centre de la requête vers l’extérieur :

1. SELECT DISTINCT sur le code unique d’espèces (spcd) et l’identifiant unique de la parcelle (plot) dans la table TREE pour sélectionner les essences et les parcelles qui les contiennent pour les états concernés dans le projet. La liste d’états provient de l’intersection fait dans QGIS (voir etats.xlsx dans le dossier « shape\_parcelles »)
2. JOINTURE des champs plot et lat (latitude des parcelles) de la table PLOT avec la sélection précédente basée sur l’identifiant unique plot. Un GROUP BY est effectué pour rassembler les parcelles qui contiennent la même espèce.
3. SELECT DISTINCT des spcd et calcul du nombre de parcelles (nb\_parcelles) et de la latitude maximale parmi celles des parcelles (lat\_max)
4. JOINTURE des champs de la table REF\_SPECIES avec la sélection précédente basée sur le champ spcd. Parmi les champs, subspecies, variety et species servent à éliminer des espèces non désirées avec des contraintes dans un WHERE.
5. SELECT des champs désirés dans la table finale ainsi que calcul du champ code\_ccbio à l’aide des champs genus et species.