ÉCOLOGIE & ENVIRONNEMENT

Découverte d’OpenRefine

Chloé MARTIN - UAR 2047 DoHNEE Données de recherche pour l'Histoire Naturelle, l’Écologie et l'Environnement
Stefan GAGET – UMR 8199 EGID European Genomic Institute for Diabetes

OpenRefine

Pour s’y retrouver …

- **Lien vers etherpad :**
  https://etherpad.in2p3.fr/p/OpenRefine_ET_SEE-Life

- **Vous y trouverez :**

  Lien de téléchargement vers

  - les supports de cours
  - Les TPs
  - OpenRefine et VIB BITS
  - Tous ce qui sera nécessaire à la formation

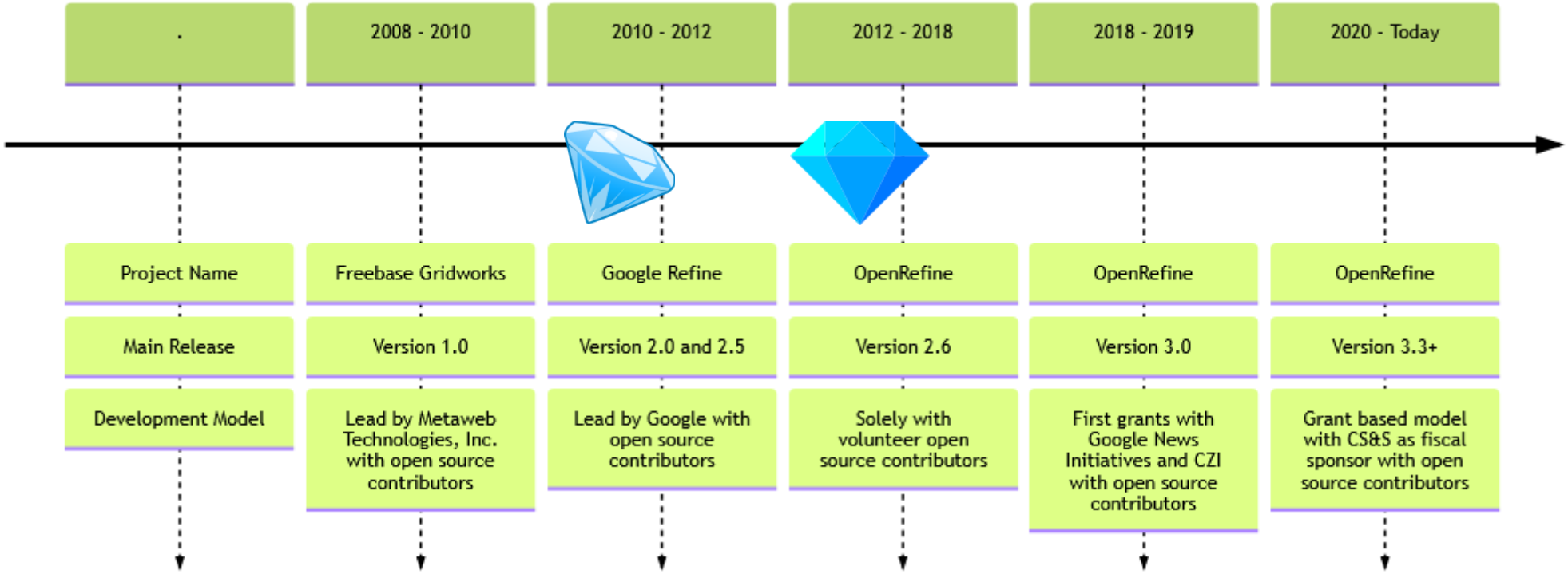
Programme …

1. **Historique/ Installation / Découverte des interfaces**

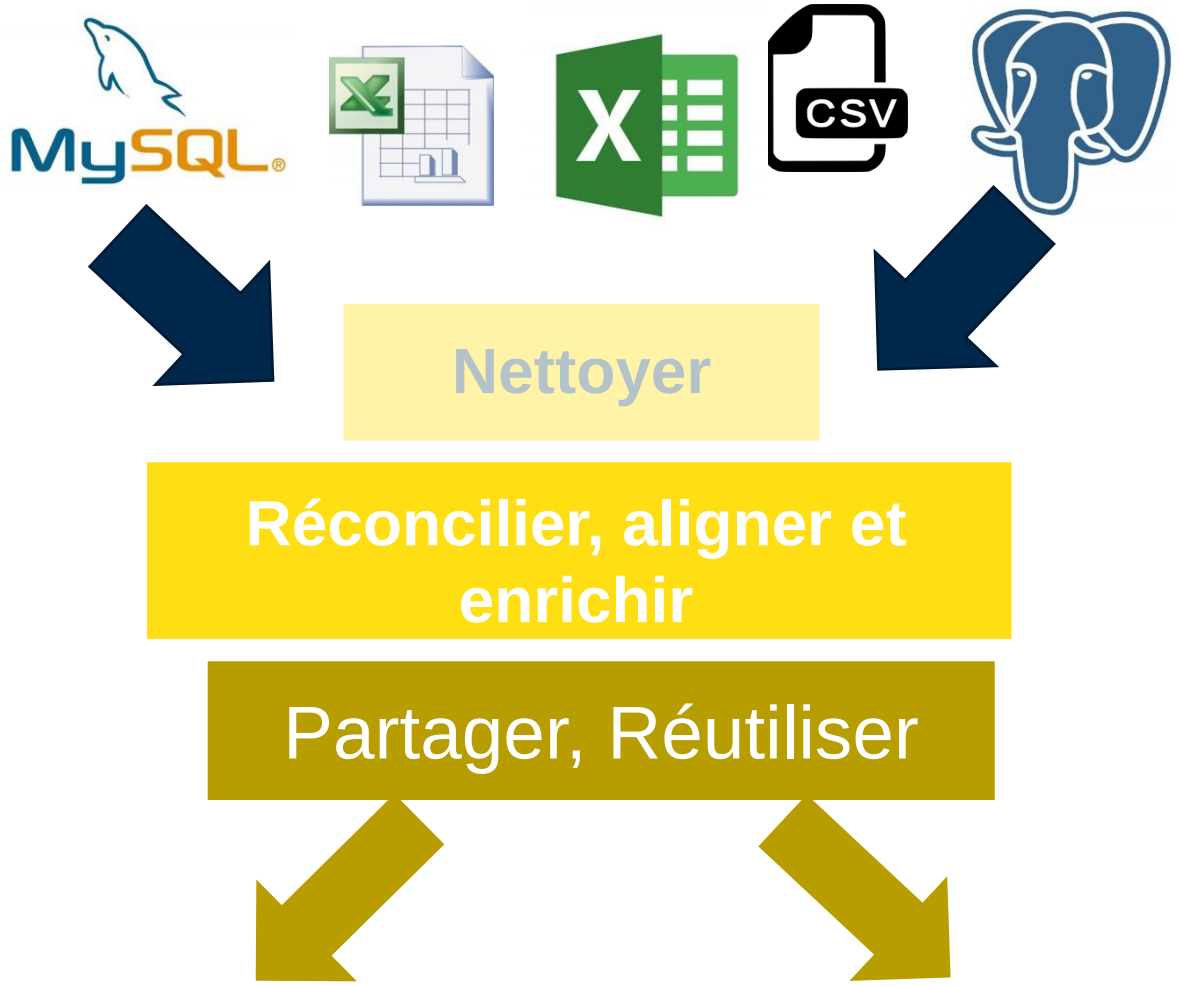
2. **Nettoyer des contenus (importation, modification, facettes, exportation …)**

3. **Comparer, réconcilier et partager ses données**

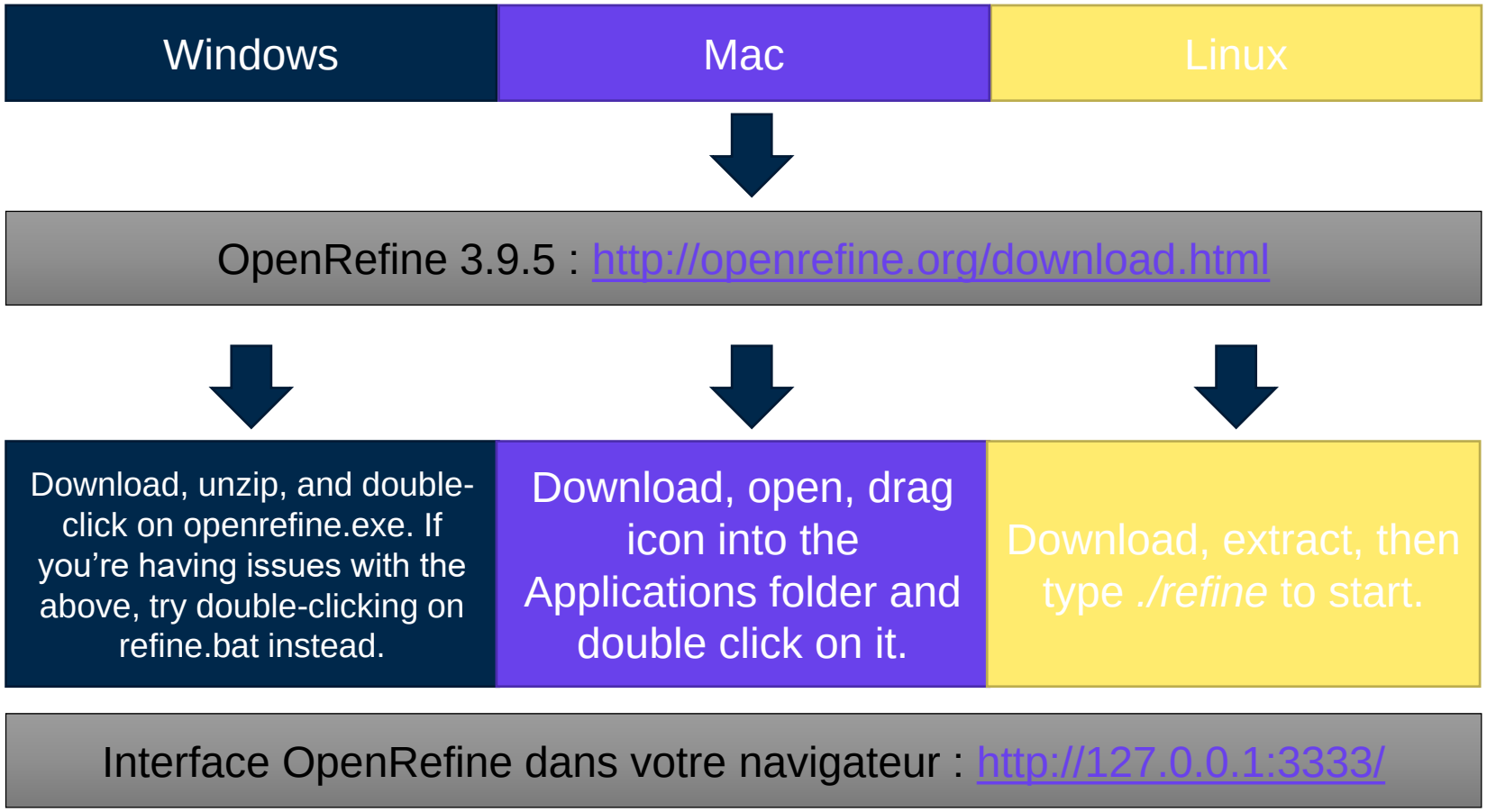
OpenRefine timeline



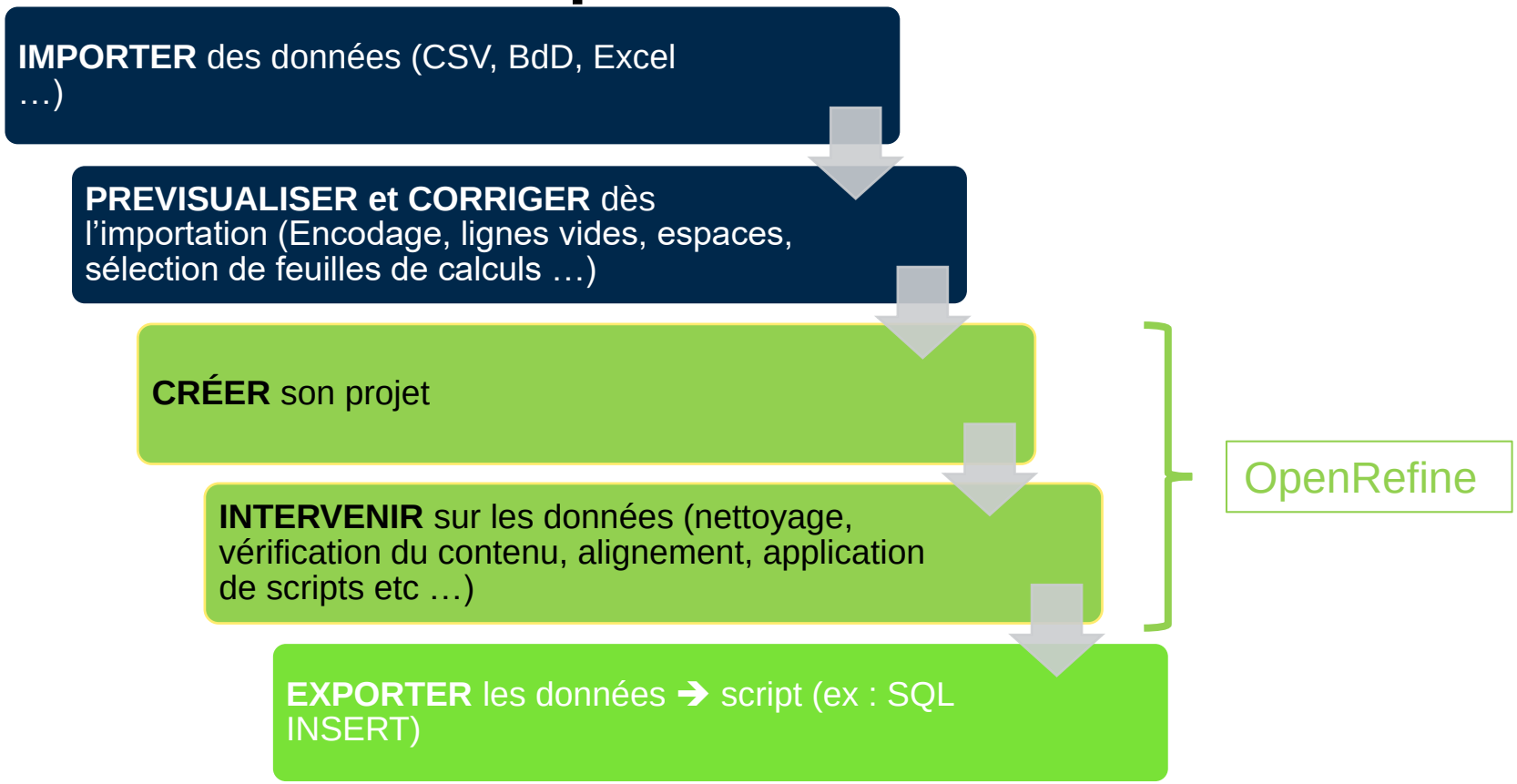
Pourquoi OpenRefine ?



Installation



Utiliser OpenRefine

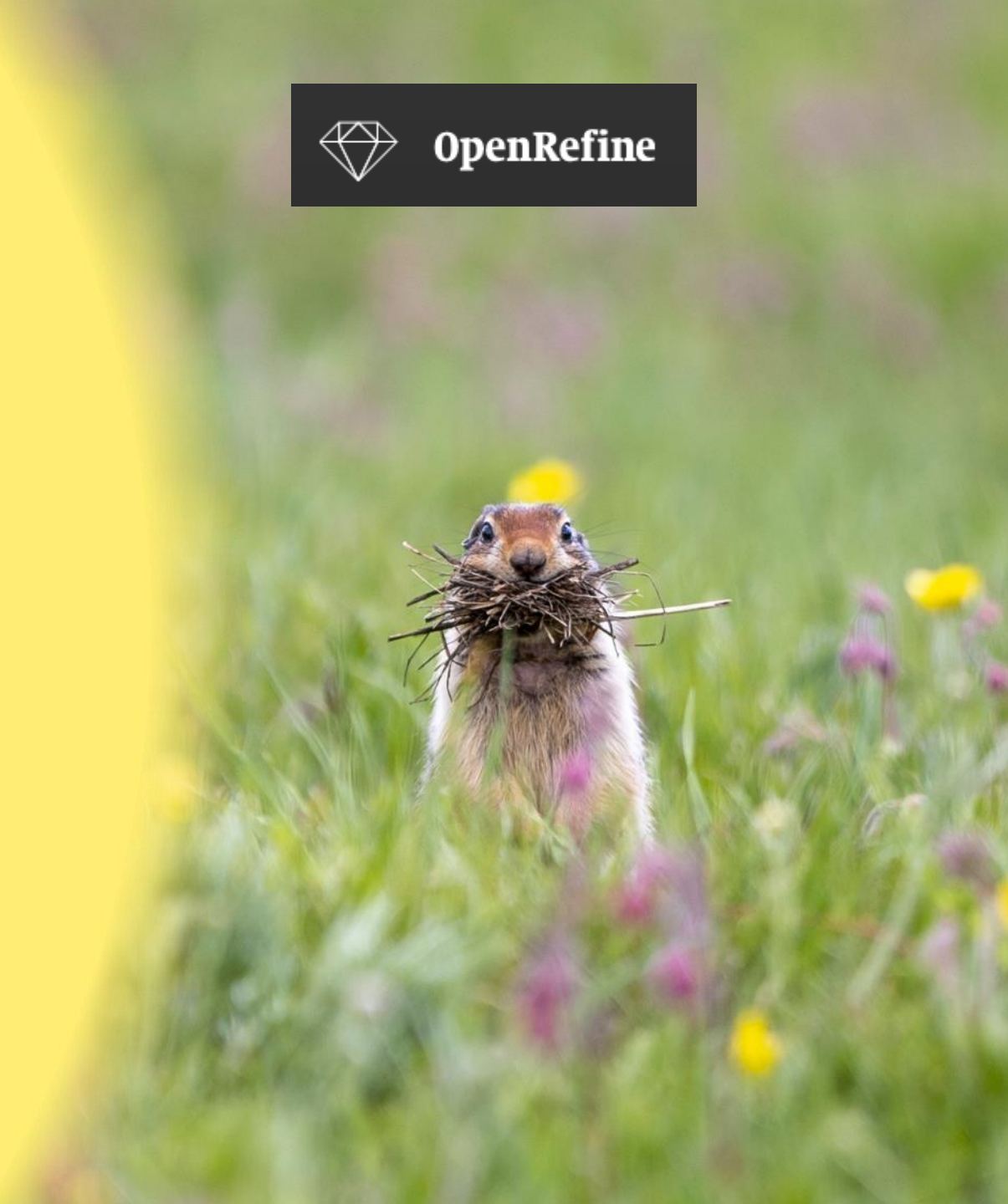


Documentation

- Documentation officielle : https://docs.openrefine.org/
- Mathieu SABY : https://msaby.gitlab.io/tutoriel-openrefine/
- Stefan GAGET : https://mate-shs.cnrs.fr/actions/tutomate/tuto18-gaget-openrefine/
- Vidéos / tutos : https://www.canal-u.tv/chaines/rbdd/tes-premiers-pas-avec-openrefine-0

cnrs ÉCOLOGIE & ENVIRONNEMENT

OpenRefine



OpenRefine
TP

Chargement de fichier

- Exercice 1 : Créez un projet à partir du fichier **openrefine_file.csv**

  - choisir les options au moment de l'importation

  - nommer le projet

Facettage et édition de masse

Exercice 2 :

- Explorer la colonne **kingdom** avec une facette (trier par nom, trier par nombre de resultats)

- Corriger les fautes d'orthographe à partir de la facette

Exercice 3 :

- Corriger les erreurs de la colonne **Country col**.

- Corriger les espaces blancs consécutifs de la colonne **Full name**

Facettage et doublons

- Exercice 4 :

- Rechercher les doublons de la colonne **Cat.Numb** et en fonction du contenu des autres colonnes (étiquettes des specimens) modifier les « vrais » doublons

UWP:122470	Vargas P
UWP:122471	Vargas I
UWP:157351	Betancur H
UWP:157339	Betancur J

Filtrage

- Exercice 5

- A partir d'une facette de la colonne **Full name**, utiliser les filtres pour corriger les noms de taxons :

Nous voulons

- Enlever les sp1, SP2, spp etc … et les remplacer par juste le nom du genre

- Vous pouvez utiliser les fonctions disponibles dans les menus de colonne ou utiliser la fonction replace en passant par Edit cells puis Transform

value.replace(«valeur à rechercher», «valeur qui remplace celle recherchée»)

Filtrage avancé

- Exercice 6

Utiliser une expression régulière pour filtrer les données et rechercher tous les cas où le nom genre + espece est indiqué avec des majuscules (ex : Aechmea Angustifolia à la place de Aechmea angustifolia)

L'expression **^[A-Z].*\s[A-Z]** permet d'afficher tous les résultats pour lesquels il y a un 1er mot commençant par une majuscule, suivi d'une 2eme mot commençant par une majuscule

Attention il faut indiquer au filtre qu'il s'agit d'une expression régulière et qu'il doit être sensible à la casse

- Lien vers doc sur les expressions régulières : https://openrefine.org/docs/manual/expressions#regular-expressions

ou l'outil : https://regexr.com/ qui vous permet de tester vos expressions régulières

Regroupement

- Exercice 7 :

A partir de la colonne County et en utilisant les fonctionnalités de groupement/cluster, corriger les noms des villes :

- Noms corrects :
- Flores
- La Libertad
- Melchor de Mencos
- San Andres
- San Jose

N’hésitez pas à tester plusieurs types de fonctions de groupement !

Exportation

- A partir de l’exportation tabulaire personnalisée, exporter vos données en prenant soin de ne pas prendre en compte les facettes et les filtres.

Licence

https://creativecommons.org/licenses/by/4.0

Vous êtes autorisé à :

- **Partager** — copier, distribuer et communiquer le matériel par tous moyens et sous tous formats
- **Adapter** — remixer, transformer et créer à partir du matériel pour toute utilisation, y compris commerciale.

Selon les conditions suivantes :

Attribution — Vous devez créditer l'Œuvre, intégrer un lien vers la licence et indiquer si des modifications ont été effectuées à l' Œuvre. Vous devez indiquer ces informations par tous les moyens raisonnables, sans toutefois suggérer que l'Offrant vous soutient ou soutient la façon dont vous avez utilisé son Œuvre.

Pas de restrictions complémentaires — Vous n'êtes pas autorisé à appliquer des conditions légales ou des mesures techniques qui restreindraient légalement autrui à utiliser l' Œuvre dans les conditions décrites par la licence.

Attribution : C. Martin & S. Gaget – Ecole thématique Data SEE-Life - 2025