

# Utilisation d'OpenRefine

*Biodiversity Data Mobilization - Data Cleaning - OpenRefine Exercise (Français)*



## CONTENU

### [CONTENU](#)

#### [1. CONVENTIONS](#)

#### [2. UTILISATION BASIQUE](#)

##### [2.1. LE CHARGEMENT DE FICHIERS ET LES PROJETS](#)

[2.1.1. Avant de commencer](#)

[2.1.2. Exercice 1. Créez un projet](#)

##### [2.2. FACETTAGE](#)

[2.2.1. Avant de commencer](#)

[2.2.2. Exercice 2. Facettage et édition de masse](#)

[2.2.3. Exercice 3. Facettage et les espaces blancs I](#)

[2.2.4. Exercice 4. Facettage et espaces blancs II](#)

[2.2.5. Exercice 5. Facettage et doublons](#)

##### [2.3. FILTRAGE](#)

[2.3.1. Exercice 6. Filtre basique](#)

[2.3.2. Exercice 7. Filtrage avancée I](#)

[2.3.3. Exercice 8. Filtrage avancée II](#)

##### [2.4. REGROUPEMENT](#)

[2.4.1. Exercice 9. Regroupement basique](#)

#### [3. UTILISATION BASIQUE API](#)

[3.1. Avant de commencer](#)

[3.2. Exercice 1. Higher taxonomy](#)

#### [4. LIENS UTILES ET REFERENCES](#)

# 1. CONVENTIONS

*Formules (à copier-coller)*

Texte en bleu

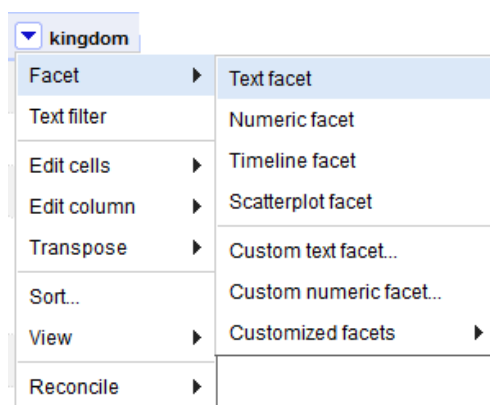
Par exemple...et coller l'expression `^[a-z]`



*Commands in Refine*

Texte en rouge

Par exemple...et suivez le chemin vers **Text facet** comme ci-dessous:



*Column names*

Texte en vert

Par exemple...et cliquez sur la colonne **Cat. Numb**

| Show as: rows records |            | Show: 5 10 25 50 rows |                         |            |  |
|-----------------------|------------|-----------------------|-------------------------|------------|--|
| All                   | Cat. Numb. | University            | Collector               |            |  |
| ☆                     | 7.         | UWP:157339            | University of Guatemala | Betancur J |  |
| ☆                     | 8.         | UWP:157339            | University of Guatemala | Betancur H |  |
| ☆                     | 224.       | UWP:122471            | University of Guatemala | Vargas P   |  |
| ☆                     | 225.       | UWP:122471            | University of Guatemala | Vargas I   |  |

*Hyperlinks*

[www.gbif.org](http://www.gbif.org)

*Column menu*



## 2. UTILISATION BASIQUE

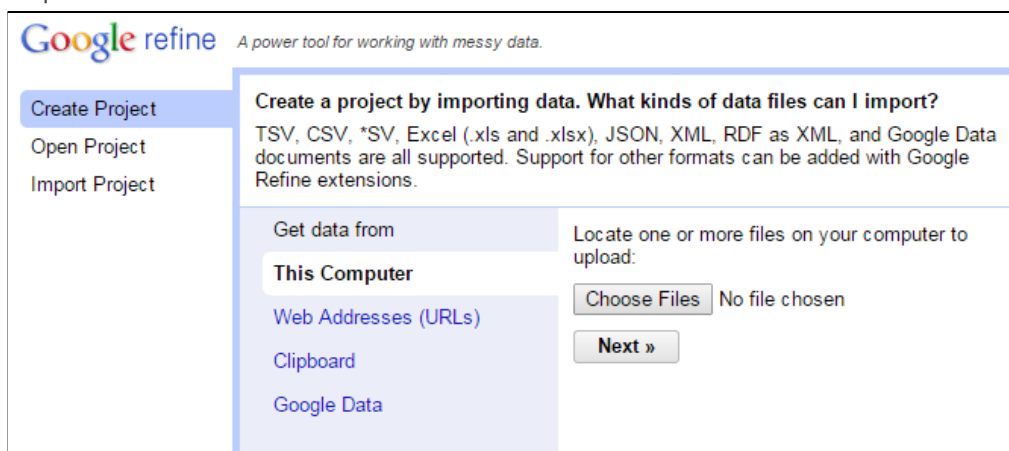
### 2.1. LE CHARGEMENT DE FICHIERS ET LES PROJETS

#### 2.1.1. Avant de commencer

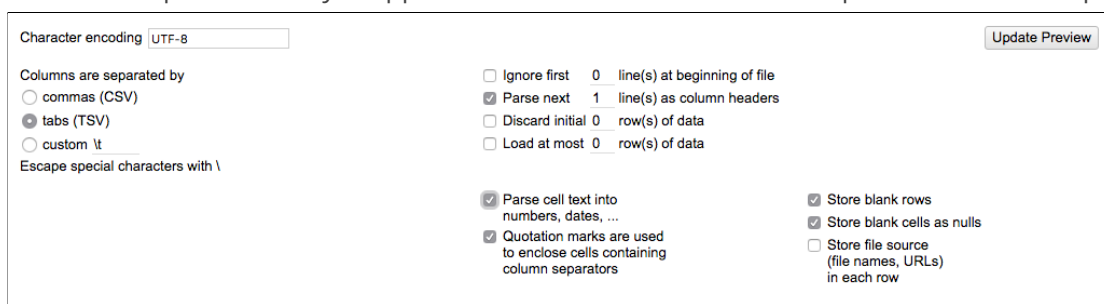
Le chargement des données peut se faire à partir de diverses sources de données: TSV, CSV, SV, Excel (.xls et .xlsx), JSON, RDF et des données XML comme des Google Docs. Le chargement des données comporte deux étapes: la première est de charger le fichier et la deuxième est de créer un projet.

#### 2.1.2. Exercice 1. Créez un projet

1. Chargez le fichier de base à partir du dossier indiqué.
2. Ouvrir *OpenRefine* (GoogleRefine), cliquez sur **Create Project**, et suivez le chemin **Get data from > This Computer**, puis cliquez sur **Choose Files**. Sélectionnez le fichier dans le dossier indiqué.
3. Cliquez sur **Next**.



4. Un menu d'options d'analyse apparaît. Assurez-vous de laisser les options comme indiqué dans l'image:



5. En haut à droite, vous pouvez renommer votre fichier et cliquez sur **Create Project** et vous serez prêt à travailler!

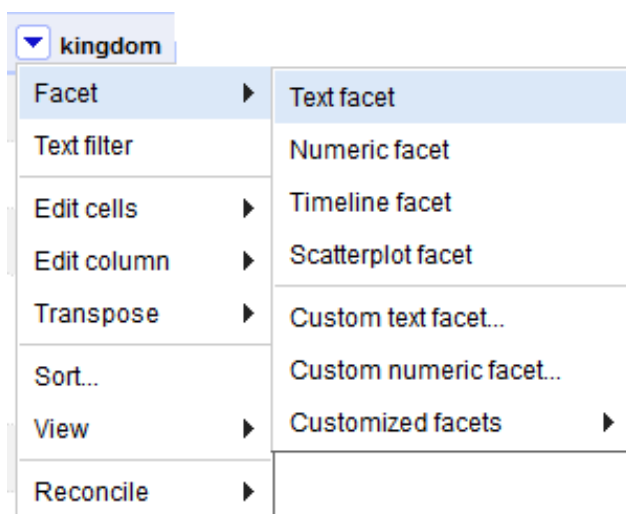
## 2.2. FACETTAGE

### 2.2.1. Avant de commencer

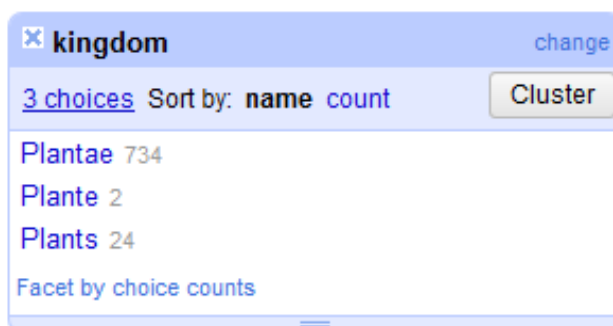
Facettage est une fonctionnalité qui va nous permettre d'obtenir un aperçu général des données, et de filtrer les enregistrements que nous voulons changer ou voir. Il facilite l'utilisation et l'analyse des données et peut être fait avec des cellules contenant tout type de texte, des chiffres et des dates ...

### 2.2.2. EXERCICE 2. Facettage et édition de masse

1. Aller sur la colonne **kingdom**, et puis cliquez sur le menu colonne  et suivre le menu jusqu'au **Text facet** comme montré ci-dessous :

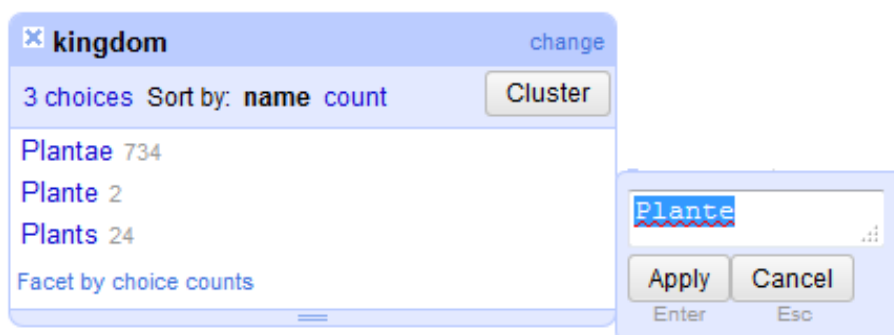


2. Sur la gauche, une fenêtre avec le nom de la colonne apparaît, ceci est la facette:



Cliquer sur **count** pour trier par le nombre, puis cliquer sur **name** pour trier alphabétiquement.

3. Corriger les fautes d'orthographe. Placez le curseur sur le texte dans la fenêtre, puis cliquez sur **edit**, puis corriger les erreurs dans la boîte de dialogue, et sauvegarder en cliquant sur **apply**.



Toutes les valeurs seront corrigées automatiquement.

### 2.2.3. EXERCICE 3. Facettage et les espaces blancs I

1. Aller sur **Country col.** Et cliquer sur le menu colonne  et effectuer une **Text Facet**.



Sur une vue rapide, le pays semble être correctement orthographié, mais la facette montre trois valeurs différentes en raison des espaces supplémentaires à la fin du texte.

2. Corriger les erreurs à partir de la colonne menu sur **Country col.**, continuer sur **Edit Cells > Common transforms > Trim leading and trailing whitespace**. Vous allez avoir ce message de notification:

**Text transform on 38 cells in column Country col.: value.trim() Undo**

3. Maintenant vérifier la fenêtre facette, seulement une seule valeur restera.

## 2.2.4. EXERCISE 4. Facettage et espaces blancs II

1. Aller sur la colonne **Full name** et cliquer sur  puis aller sur **Text facet**. Cliquez sur **count**. La facette va afficher:









| Full name                              |               | change |
|--|---------------|--------|
| 253 choices                            | Sort by: name | count  |
| <input type="button" value="Cluster"/> |               |        |
| Guzmania lingulata                     | 25            |        |
| Aechmea veitchii                       | 24            |        |
| Guzmania coriostachya                  | 22            |        |
| Guzmania lingulata                     | 20            |        |
| Aechmea tillandsioides                 | 17            |        |
| Aechmea penduliflora                   | 15            |        |
| Aechmea servitensis                    | 14            |        |
| Guzmania angustifolia                  | 13            |        |
| Aechmea dactylina                      | 12            |        |
| Catopsis sessiliflora                  | 12            |        |
| Aechmea Angustifolia                   | 11            |        |
| Aechmea nubescens                      | 10            |        |

Comme montré ci-dessus, *Guzmania lingulata* est le premier enregistrement sur la liste avec 25 spécimens, mais il est aussi présent à la 4ème place avec 20 spécimens.

2. Corriger les erreurs à partir du menu de la colonne **Full name**, **Edit Cells** > **Common transforms** > **Collapse consecutive whitespaces**.
3. Une fois les espaces blancs supprimés, *Guzmania lingulata* devrait seulement apparaitre une fois dans la liste avec 45 enregistrements.

## 2.2.5. EXERCISE 5. Facettage et doublons

1. Aller sur le catalogue de la colonne **Cat. Numb**, et suivre le menu **Facet** > **Customized facets** > **Duplicates facet**. La facette va montrer 4 doublons
2. Cliquer sur **true, et** et vous allez voir les valeurs sur la fenêtre principale :


| Show as: rows   |            | records    |                         |            |  | Show: 5 10 25 50 rows |  |
|---|------------|------------|-------------------------|------------|--|-----------------------|--|
| All   | Cat. Numb. | University | Collector               |            |  |                       |  |
|   | 49.        | UWP:122471 | University of Guatemala | Vargas P   |  |                       |  |
|   | 50.        | UWP:122471 | University of Guatemala | Vargas I   |  |                       |  |
|   | 117.       | UWP:157339 | University of Guatemala | Betancur J |  |                       |  |
|   | 118.       | UWP:157339 | University of Guatemala | Betancur H |  |                       |  |

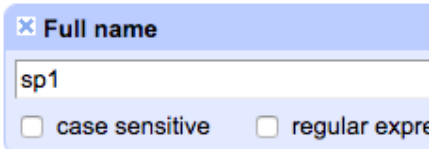

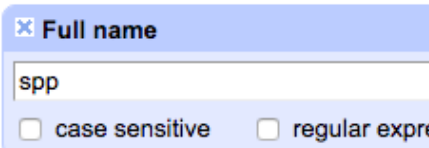

Après vérification avec les étiquettes des spécimens, corriger les erreurs en cliquant directement sur les cellules avec le vrai numéro de catalogue:

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

## 2.3. FILTRAGE

### 2.3.1. EXERCISE 6. Filtre basique

1. Aller encore sur le menu de la colonne **Full name** et puis **Text facet** pour visualiser les valeurs, puis aller encore sur  et cliquer sur **Text filter**, effectuer les filtres suivants et les corriger comme indiqué ci-dessous:

| Filtre   | Comment corriger  | Valeur correcte  |
|--|---|--|
|   | Modifier directement dans la cellule  | Cyperus  |
|  | Modifier directement dans la cellule, vérifier la majuscules et minuscules  | Cyperus  |
|  | <ol style="list-style-type: none"> <li>1. Aller sur  <b>Full name</b>, puis cliquer <b>Edit cells &gt; Transform...</b></li> <li>2. Dans la zone de texte coller la formule <code>value.replace(" spp.", "")</code></li> <li>3. Cliquer sur <b>OK</b></li> </ol> | Aechmea<br>Chusquea<br>Eleocharis<br>Greigia<br>Navia<br>Neurolepis<br>Rhynchospora<br>Tillandsia<br>Xyris |

### 2.3.2. EXERCISE 7. Filtrage avancée I

1. Aller sur la colonne **genus** et effectuer un **Text filter**.
2. Cocher **regular expression** et **case sensitive**, puis coller l'expression `^[a-z]`




Cette expression régulière filtre les chaînes dont la première lettre est en minuscule.

3. Effectuer une correction si le genre doit être en majuscule.

Note: Si vous voulez en savoir plus sur les expressions régulières cliquer [ici](#).

### 2.3.3. EXERCISE 8. Filtrage avancée II

1. Aller sur la colonne **Full name** et effectuer un **Text filter**.
2. Cocher **regular expression** et **case sensitive**, puis coller l'expression `^[A-Z].*\s[A-Z]`



Cette expression régulière filtre les chaînes de caractères qui commencent par une majuscule suivie par n'importe quel caractère, puis un espace, puis une majuscule.

3. Effectuer une correction afin que le deuxième mot du nom soit en minuscule.

Note: Si vous voulez en savoir plus sur les expressions régulières cliquer [ici](#).

## 2.4. REGROUPEMENT

### 2.4.1. EXERCISE 9. Regroupement basique

1. Aller sur **County**, puis sur le menu de la colonne cliquer **Text facet**.



County
change

10 choices
Sort by: name count
Cluster

- Andressan 1
- Flores 175
- La Libertad 50
- Libertad La 1
- Melchor de Mencos 67
- Mencos de Melchor 1
- San Anders 1
- San Andres 357
- San Jose 106
- SanAndres 1

Facet by choice counts

Les noms corrects des villes sont :

- Flores
- La Libertad
- Melchor de Mencos
- San Andres
- San Jose

2. En haut à droite de la fenêtre de la facette cliquer sur **Cluster**, une nouvelle fenêtre apparaîtra:

**Cluster & Edit column "County"**

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "GÄ[del]" and "Godel" probably refer to the same person. [Find out more ...](#)

Method key collision    Keying Function fingerprint    2 clusters found

| Cluster Size | Row Count | Values in Cluster   | Merge?                   | New Cell Value    | # Rows in Cluster |
|--------------|-----------|---|--------------------------|-------------------|-------------------|
| 2            | 68        | <ul style="list-style-type: none"> <li>Melchor de Mencos (67 rows)</li> <li>Mencos de Melchor (1 rows)</li> </ul> | <input type="checkbox"/> | Melchor de Mencos | 51 — 68           |
| 2            | 51        | <ul style="list-style-type: none"> <li>La Libertad (50 rows)</li> <li>Libertad La (1 rows)</li> </ul>             | <input type="checkbox"/> | La Libertad       | 11 — 17           |

Select All   Unselect All
Merge Selected & Re-Cluster   Merge Selected & Close   Close

3. Maintenant vous pouvez voir les informations sur les regroupements :
  - **Taille du regroupement:** Malgré le nombre de versions, les algorithmes de regroupement demeurent les mêmes
  - **Nombre de lignes :** le nombre d'enregistrements avec l'une des valeurs du regroupement.
  - **Valeurs du regroupement:** Les valeurs considérées comme étant identiques par l'algorithme. Il y a aussi le nombre d'enregistrements avec chaque valeur particulière, et la possibilité de parcourir le contenu du cluster dans un autre onglet.
  - **Fusionner?:** Vérifier si les valeurs doivent être fusionnées en une valeur standard unique.
  - **Nouvelle valeur de la cellule:** la valeur à appliquer à chaque enregistrement du cluster. Par défaut, c'est la valeur de la plupart des enregistrements. Vous pouvez également cliquer sur une valeur à appliquer à la **Nouvelle valeur de la cellule**.

Note: Si vous voulez en savoir plus sur le regroupement, cliquer [ici](#).

4. Cliquer sur **Select All** et puis sur **Merge Selected & close**, vous allez voir un message de notification:

**Mass edit 119 cells in column County Undo**

5. Pour corriger les reste des villes, aller encore sur **Cluster** dans la fenêtre facette de **County**.
6. Dans le regroupement et la fenêtre de modification, aller sur **Keying Function**, puis sélectionner **ngram-fingerprint**, et mettre **1** comme valeur dans le **Ngram Size**. Appuyer sur le bouton entrer.
7. Cliquer sur **Select All** puis aller sur **Merge Selected & close**, vous allez avoir une notification comme suit:

**Mass edit 360 cells in column County Undo**

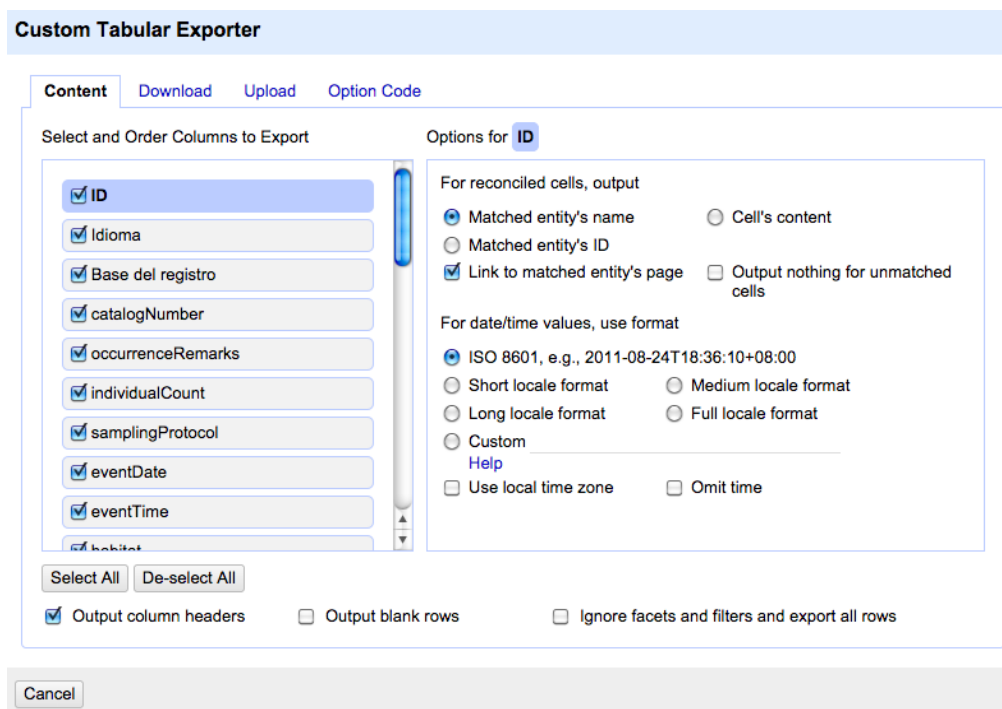
8. Vos villes sont maintenant corrigées et votre fenêtre devrait ressembler à l'image ci-dessous :



## 2.5. EXPORTATION

Vous avez plusieurs options pour exporter vos données nettoyées, mais l'option suivante est utile dans la plupart des cas.

1. Sur le coin supérieur droit, cliquez sur **Export** et sélectionnez **Custom tabular exporter...**
2. Vous verrez la fenêtre d'exportation:



**Custom Tabular Exporter**

Content Download Upload Option Code

Select and Order Columns to Export

- ID
- Idioma
- Base del registro
- catalogNumber
- occurrenceRemarks
- individualCount
- samplingProtocol
- eventDate
- eventTime
- habit

Options for ID

For reconciled cells, output

- Matched entity's name
- Matched entity's ID
- Link to matched entity's page
- Cell's content
- Output nothing for unmatched cells

For date/time values, use format

- ISO 8601, e.g., 2011-08-24T18:36:10+08:00
- Short locale format
- Long locale format
- Custom
- Medium locale format
- Full locale format

[Help](#)

- Use local time zone
- Omit time

Select All De-select All

- Output column headers
- Output blank rows
- Ignore facets and filters and export all rows

Cancel

3. Sur l'onglet **content** vous pouvez choisir les colonnes que vous souhaitez exporter. Si vous sélectionnez **Ignore facets and filters and export all rows** toutes les facettes et filtres seront ignorés, ce qui est utile si vous oubliez de les effacer avant d'exporter.
4. Allez à l'onglet **Download** et sélectionnez le séparateur que vous préférez. Ne modifiez pas les autres options, sauf en cas de besoin.

Vous pouvez également exporter l'ensemble du projet pour l'ouvrir dans OpenRefine sur un autre ordinateur en suivant le chemin **Export > Export project**. Dans ce cas, vous ne téléchargez pas un fichier de données pour ouvrir dans un tableur. À la place, vous aurez un fichier GZIP accessible uniquement en l'ouvrant avec OpenRefine.

### 3. UTILISATION BASIQUE API

#### 3.1. Avant de commencer

La Réconciliation fait correspondre une information contenue dans l'une de vos colonnes à une base de données externe. Cela est particulièrement utile pour la validation des noms, car il prouve que le nom que vous avez existe quelque part ailleurs. Ce service est très utile, mais peut prendre du temps. Dans ce cas, nous allons utiliser un processus avec seulement trois enregistrements en utilisant l'API de GBIF. Une connexion Internet est nécessaire.

#### 3.2. Exercice 1. Higher taxonomy

1. Aller sur [Collector](#), puis effectuer un **Text facet**. Sélectionner le collecteur Elsa P



2. Sous **Full name**, cliquer sur le menu de la colonne **Edit column** > **Add column by fetching URLs...**, nommer la nouvelle colonne **Api\_name**
3. Changer le **Throttle Delay** à **250** et coller l'expression suivante:

```
"http://api.gbif.org/v1/species/match?verbose=true&name="+escape(value, 'url')
```

### Add column by fetching URLs based on column Full name

New column name  Throttle delay  milliseconds

On error  set to blank  store error

**Formulate the URLs to fetch:**

Expression  Language Google Refine Expression Language (GREL) ▾

No syntax error.

[Preview](#) [History](#) [Starred](#) [Help](#)

| row  | value                    | "http://api.gbif.org/v1/species/match?verbose=true&name="+value                 |
|------|--------------------------|---|
| 29.  | Tillandsia adpressiflora | http://api.gbif.org/v1/species/match?verbose=true&name=Tillandsia adpressiflora |
| 688. | Paspalum decumbens       | http://api.gbif.org/v1/species/match?verbose=true&name=Paspalum decumbens       |
| 753. | Guacamaya superba        | http://api.gbif.org/v1/species/match?verbose=true&name=Guacamaya superba        |

4. Cliquer sur **ok** et attendre, ceci peut prendre du temps suivant votre connexion internet et le nombre de taxa.
5. Aller sur **Api\_name**, cliquer sur le menu de la colonne, puis sur **Edit column > Add column based on this column...**  
Nommer la nouvelle colonne **Rank** et coller les expressions suivantes:

```
value.parseJson().get("kingdom")+
", "+value.parseJson().get("phylum")+
", "+value.parseJson().get("class")+
", "+value.parseJson().get("order")+
", "+value.parseJson().get("family")
```

Vous aller voir Kingdom, Phylum, Class, Order et family pour chaque taxon.

8. Sous **Rank** suivre le menu **Edit column > Split into several columns...**, laisser les mêmes paramètres que dans l'image ci-dessous :

### Split column Rank into several columns

**How to Split Column**

by separator  
 Separator   regular expression  
 Split into  columns at most (leave blank for no limit)

by field lengths  
  
 List of integers separated by commas, e.g., 5, 7, 15

**After Splitting**

Guess cell type  
 Remove this column

OK Cancel

9. Maintenant vous savez comment obtenir la taxonomie pour un taxon si elle est disponible via l'API GBIF. Les noms des colonnes peuvent être modifiés sur [Edit column > Rename this column](#).
10. Dans le cadre de cet atelier, les colonnes créées dans cet exercice devraient être supprimées. Sous **All**, qui est la première colonne, Aller sur [Edit columns > Re-order / remove columns...](#)
11. Glisser les colonnes comme montrés ci-dessous et cliquer **OK**:

### Re-order / Remove Columns

Drag columns to re-order

- coordinateUncertaintyInMeters
- identifiedBy
- typeStatus
- kingdom
- phylum
- class
- order
- family
- genus
- specificEpithet
- infraspecificEpithet
- Full name
- taxonRank
- Authorship
- Other name

Drop columns here to remove

- Api\_name
- Rank 1
- Rank 2
- Rank 3
- Rank 4
- Rank 5

OK Cancel

## 4. LIENS UTILES ET REFERENCES

- Tutoriel sur la validation des noms:  
[https://docs.google.com/document/d/1tkDRXIYhmassYAk5T4v5oac5prF0jAiSMr\\_JEGTvhRo/edit](https://docs.google.com/document/d/1tkDRXIYhmassYAk5T4v5oac5prF0jAiSMr_JEGTvhRo/edit)
  - Tutoriel sur la taxonomie supérieure:  
[https://docs.google.com/document/d/1XZ\\_pM9gldQzHzl8wfUCVea-52yub5T\\_3tc-snBgPRa0/edit](https://docs.google.com/document/d/1XZ_pM9gldQzHzl8wfUCVea-52yub5T_3tc-snBgPRa0/edit)
  - Documentation  
<https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users>
  - Liste des ressources pour OpenRefine:  
<https://github.com/OpenRefine/OpenRefine/wiki/External-Resources>
- 

Exercise concept and content developed by Néstor Beltrán.

Updated: 03 July 2019. Sophie Pamerlon

Updated: 11 July 2019. Laura Russell, Sophie Pamerlon.

Traduit en Français à partir de la version en Anglais par Andry Jean Marc RAKOTOMANJAKA, Mélianie Raymond et Maheva Bagard Laursen.