Technique de nettoyage de données en utilisant le logiciel libre « Open refine »

Tsiky Rabetrano

Biodiversity Data Manager

REBIOMA

tsiky@rebioma.net

rebioma@rebioma.net

DLC Anosy, 15 Octobre 2015

C’est quoi OpenRefine?

- Outil très puissant qui peut considérablement accélérer le nettoyage des données.
- Données dans un tableur ou d’autres formats
  - TSV, CSV, *SV, Excel (.xls and .xlsx),
  - JSON,
  - XML,
  - RDF as XML,
  - Wiki markup, and
  - Google Data sont tous supportés.

  Logiciel anciennement connu sous le nom de «GoogleRefine»

Ressources pour OpenRefine

- Documentation

https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users

- Screencasts (via YouTube)

https://github.com/OpenRefine/OpenRefine/wiki/Screencasts

- Blog Posts

https://github.com/OpenRefine/OpenRefine/wiki/Recipes

http://openrefine.org/

Download OpenRefine

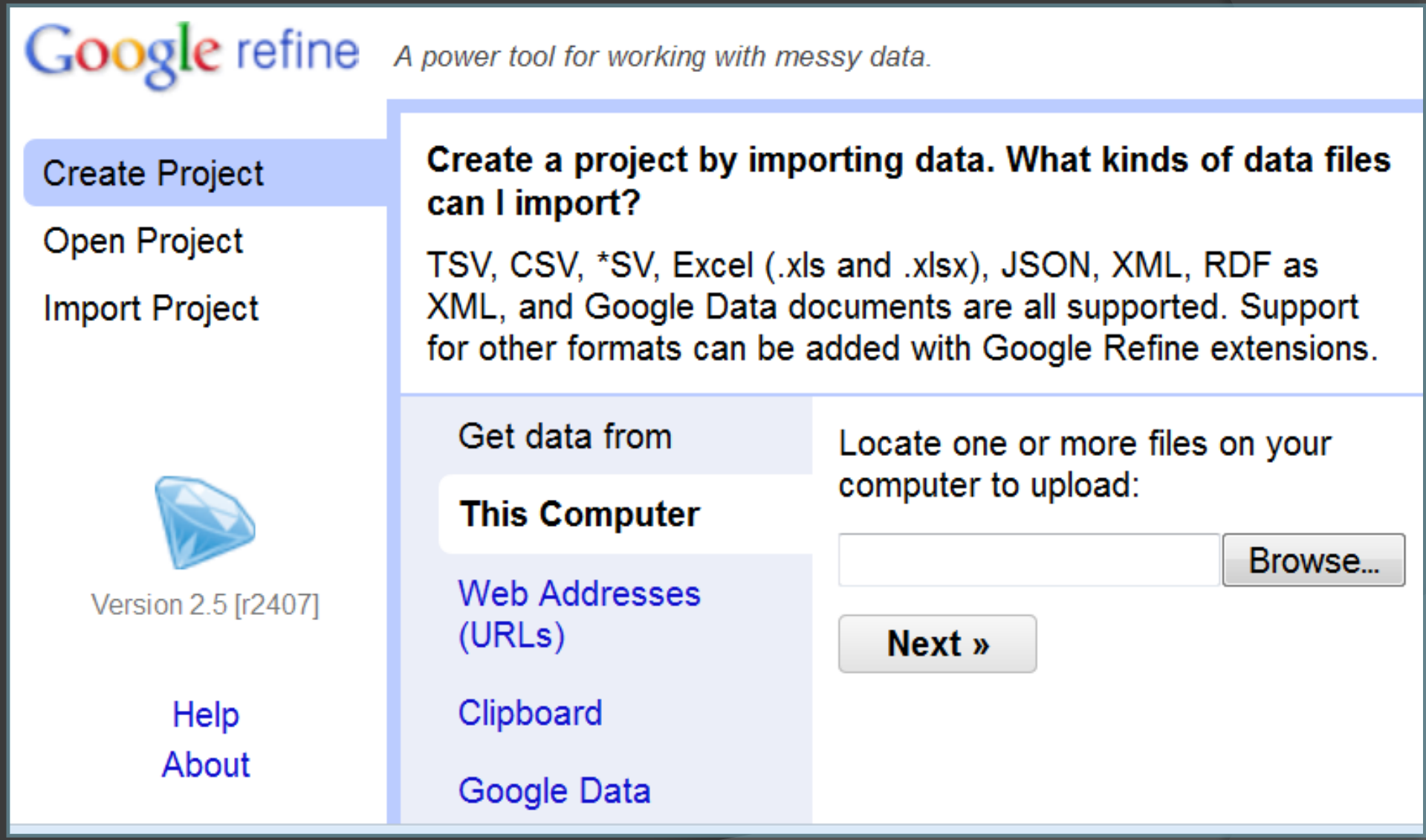
Currently the lastest stable version of OpenRefine is Google Refine 2.5. The next release (OpenRefine 2.6) will carry the new branding.

- **For Windows**, Download, unzip, and double-click on *google-refine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **For Mac**, Download, open, drag icon into the Applications folder and double click on it.
- **For Linux**, Download, extract, then type *./refine* to start.

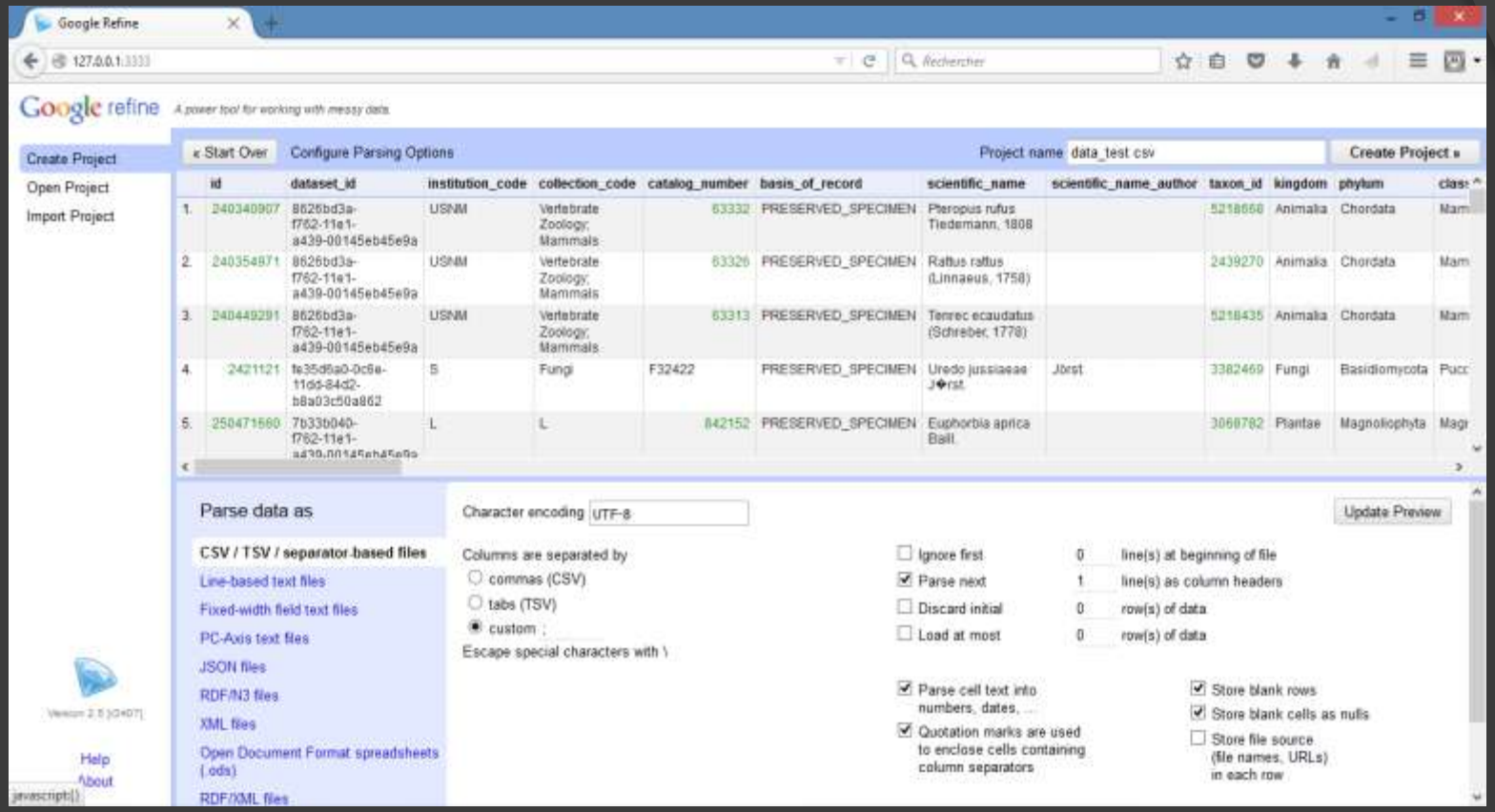
Pour commencer

1. Ouvrir Open refine
   - Double click sur « *google-refine.exe* »
1. Click Create Project
2. Importer les données
   - Choose files
   - Next
3. Vérifier le paramètre des données
4. Choisir l'encodage
5. Click Create project
6. Affichez « 50 row » (pour voir plus de données)

Créer un projet

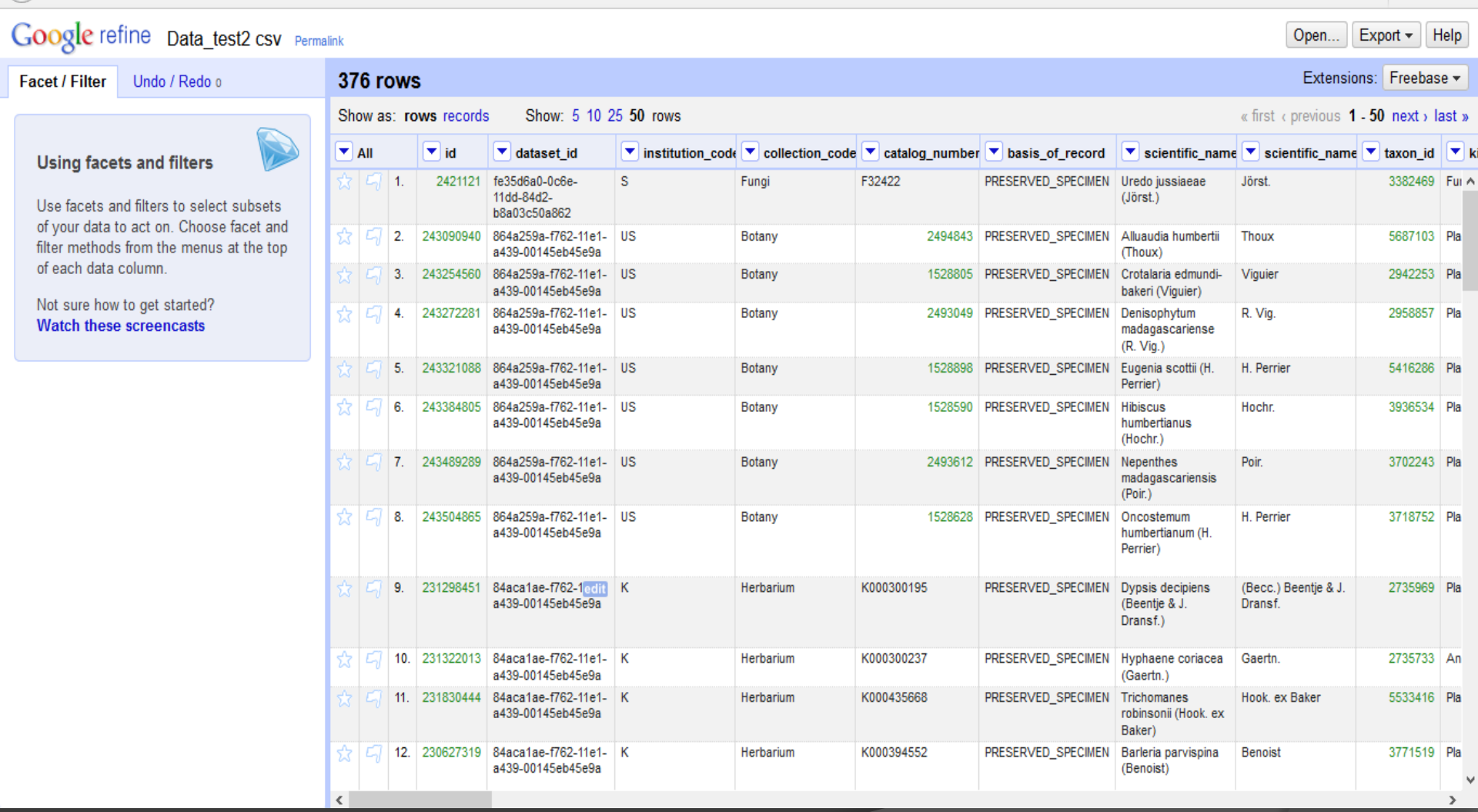


Configurer les options analyse syntaxique puis cliquez sur « Create project »



- Codage des caractères - sélectionner "ISO 8859-1" ou "UTF-8" - ce qui garantit l'affichage correct des caractères spéciaux ou des signes diacritiques
- Pour terminer l'importation des données, clic sur "Create Project"

Visualisation de données



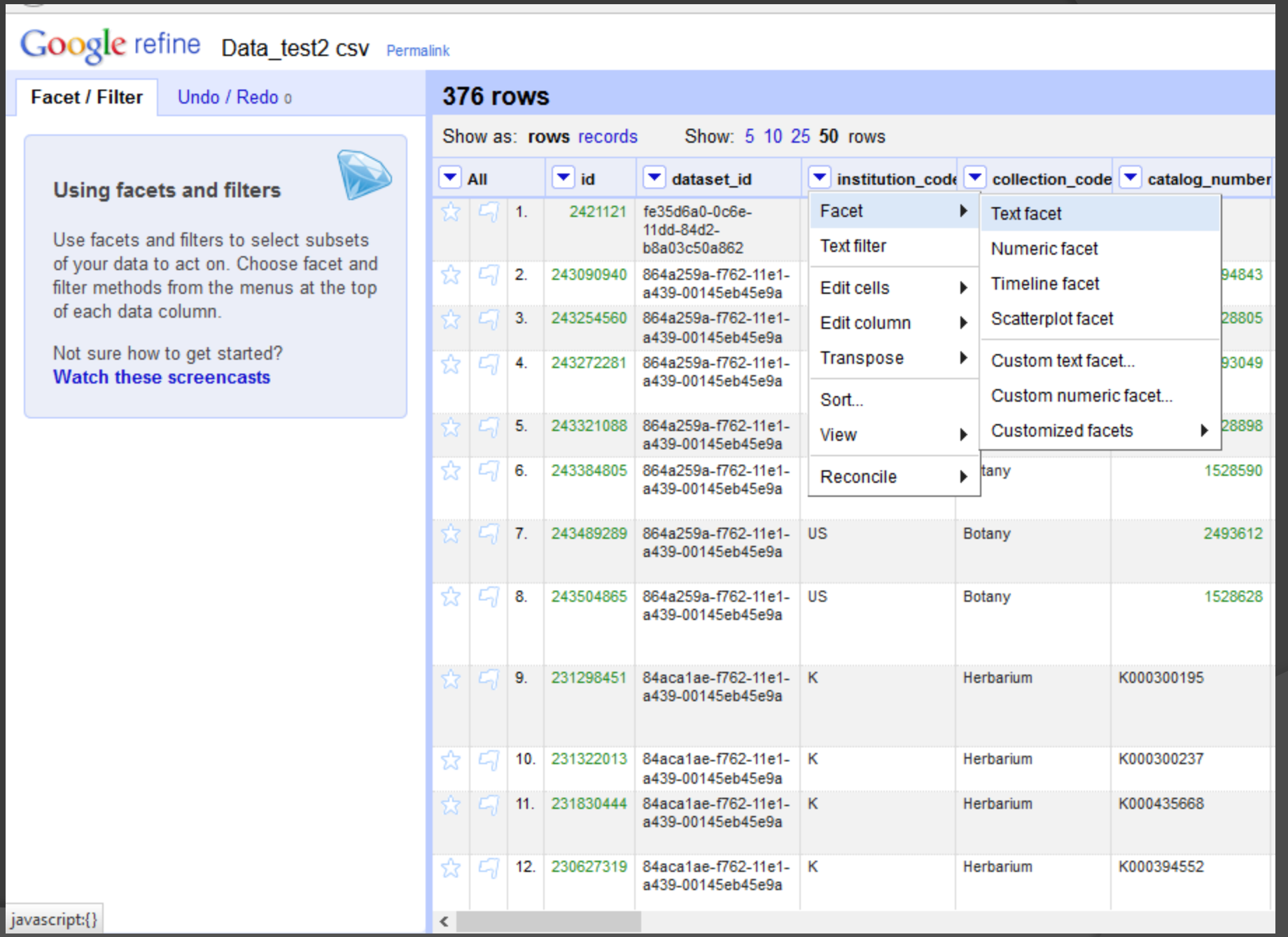
Se familiariser avec les facettes et les filtres

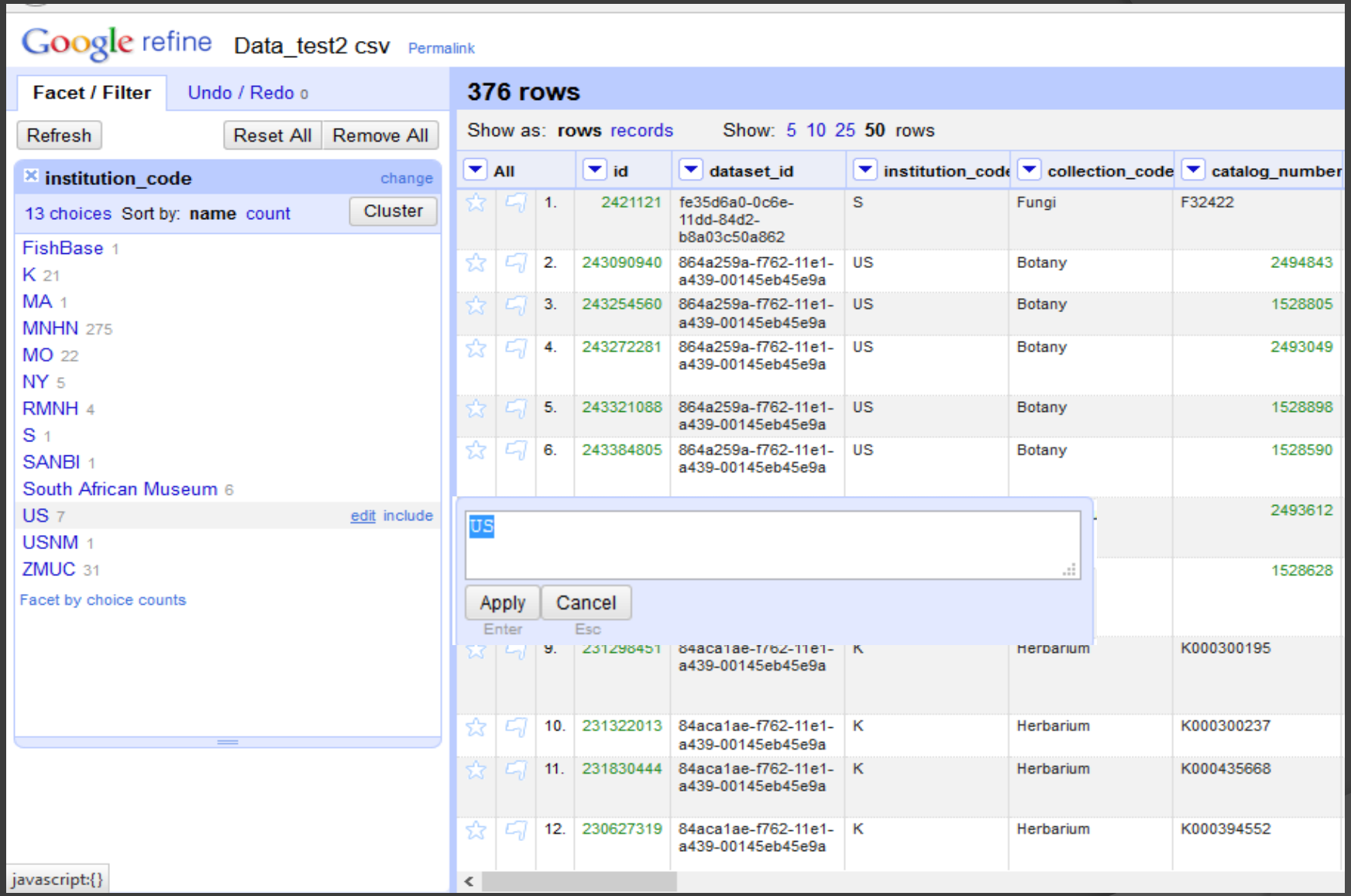
Vous utiliserez les facettes et des filtres très souvent dans Refine. Essayez de créer une facette de texte pour comprendre son fonctionnement :

- Cliquez sur la flèche déroulante de **l'en-tête de colonne> Facets > Text facets >**
- Vous verrez apparaître un encadré qui regroupe tous les contenus de cellules identiques et fournit un décompte du nombre de fois où l'occurrence apparait dans cette colonne

Ceci est utile pour plusieurs raisons :

- Repérer les fautes de frappe - Par exemple
- Repérer les colonnes vides (Vous pouvez ensuite supprimer cette colonne en cliquant sur la colonne, *Menu déroulant* > ***Edit Column > Remove this Column***)


Google refine Data_test2 csv Permalink
Facet / Filter
Undo / Redo 0
Using facets and filters
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.
Not sure how to get started?
Watch these screencasts
376 rows
Show as: rows records
Show: 5 10 25 50 rows
All
id
dataset_id
institution_code
collection_code
catalog_number
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
Text facet
Numeric facet
Timeline facet
Scatterplot facet
Custom text facet...
Custom numeric facet...
Customized facets
1. 2421121 fe35d6a0-0c6e-11dd-84d2-b8a03c50a862
2. 243090940 864a259a-f762-11e1-a439-00145eb45e9a 94843
3. 243254560 864a259a-f762-11e1-a439-00145eb45e9a 28805
4. 243272281 864a259a-f762-11e1-a439-00145eb45e9a 93049
5. 243321088 864a259a-f762-11e1-a439-00145eb45e9a 28898
6. 243384805 864a259a-f762-11e1-a439-00145eb45e9a tany 1528590
7. 243489289 864a259a-f762-11e1-a439-00145eb45e9a US Botany 2493612
8. 243504865 864a259a-f762-11e1-a439-00145eb45e9a US Botany 1528628
9. 231298451 84aca1ae-f762-11e1-a439-00145eb45e9a K Herbarium K000300195
10. 231322013 84aca1ae-f762-11e1-a439-00145eb45e9a K Herbarium K000300237
11. 231830444 84aca1ae-f762-11e1-a439-00145eb45e9a K Herbarium K000435668
12. 230627319 84aca1ae-f762-11e1-a439-00145eb45e9a K Herbarium K000394552
javascript:{}



- **Si vous souhaitez corriger un nom particulier, vous pouvez cliquer sur « Edit » qui apparaît à côté de chaque nom si vous déplacez votre souris dans la zone à droite de chaque nom**

Cluster

- Cliquez sur la flèche déroulante de **l'en-tête de colonne> Facets > Text facets >**
- Après avoir cliqué sur "cluster" vous pouvez ou ne pouvez pas obtenir un résultat avec les algorithmes de défaut
- Essayer toutes les options possible
- Cocher et renommer si nécessaire les propositions qui vous sembles juste
- Il suffit de sélectionner sur « **Merge**" si la conjecture est correcte

Cluster & Edit column "scientific_name_author"

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 "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method key collision Keying Function fingerprint 5 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	7	• (Drake) Drake (4 rows) • Drake (3 rows)	☐	(Drake) Drake
2	3	• (Parr, 1928) (2 rows) • Parr, 1928 (1 rows)	☐	(Parr, 1928)
2	6	• DC (4 rows) • DC. (2 rows)	☐	DC
2	2	• Boeck (1 rows) • Boeck. (1 rows)	☐	Boeck
2	4	• Renauld & Paris (3 rows) • Paris & Renauld (1 rows)	☐	Renauld & Paris

 # Rows in Cluster

2 — 7

Average Length of Choices

2 — 15

Length Variance of Choices

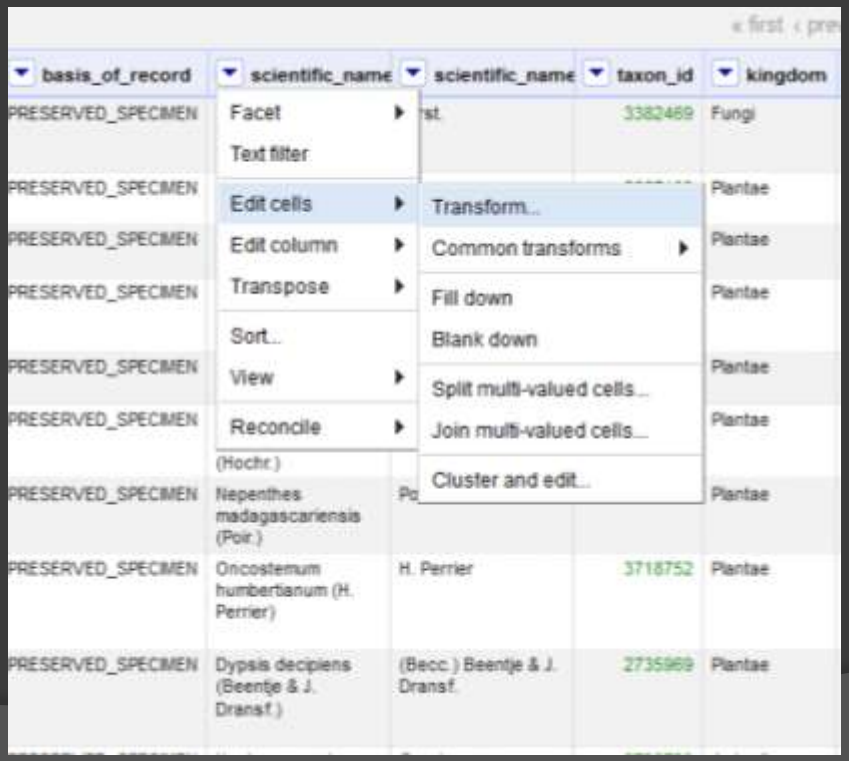
0 — 4

Select All Deselect All **Merge Selected & Re-Cluster** Merge Selected & Close Close

Suppression des textes en parentheses

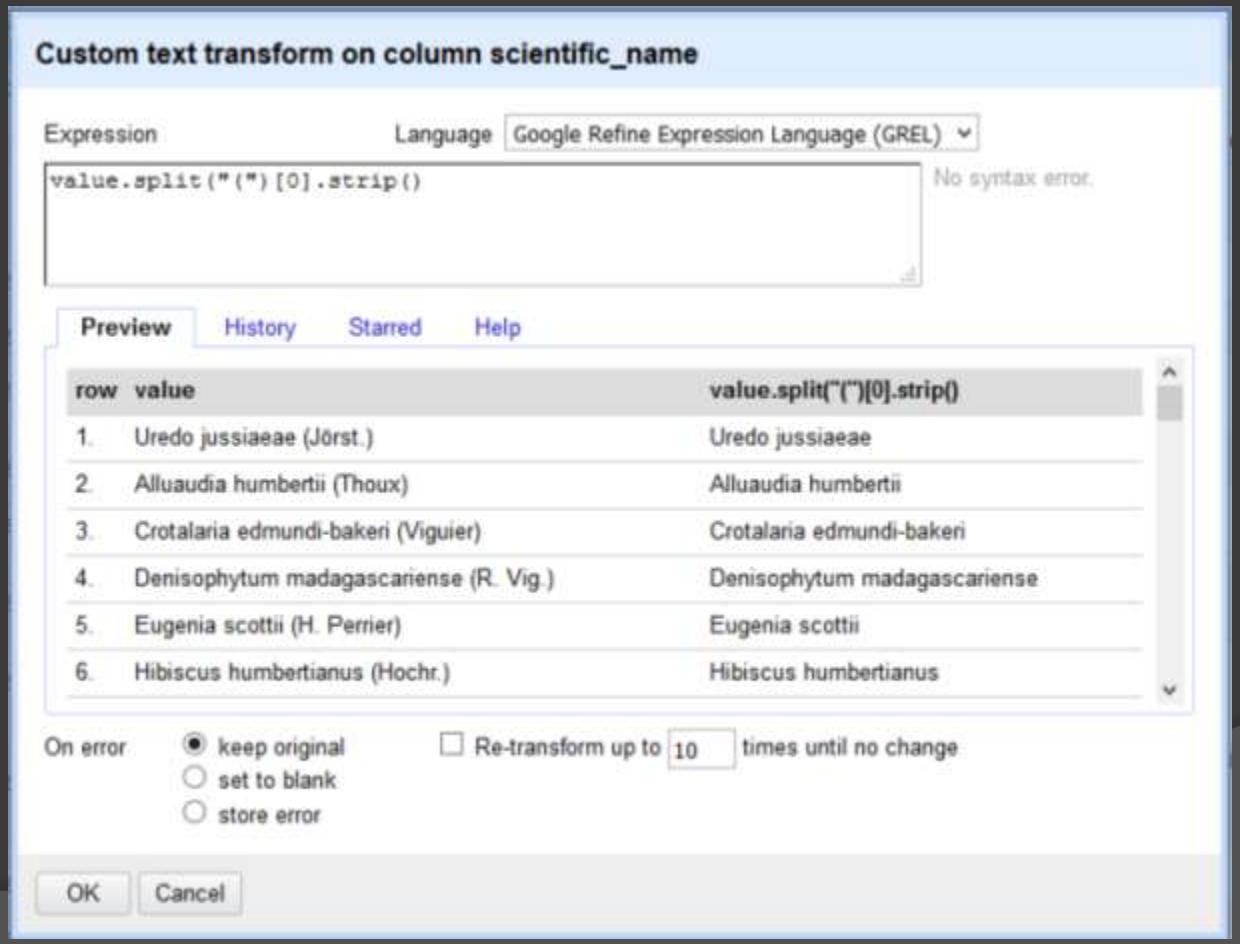
Vous remarquerez que dans a colonne Scientific_name, les noms de l'auteur de l'espèces apparaissent en entre parenthèses après le nom scientifique. vous pouvez les supprimer afin d'améliorer la lisibilité.

Menu déroulant > Edit cells > Transform



Vous arrivez sur un écran qui vous invite à saisir une fonction dans la zone d'édition de texte

Expression:
value.split("(")[0].strip()

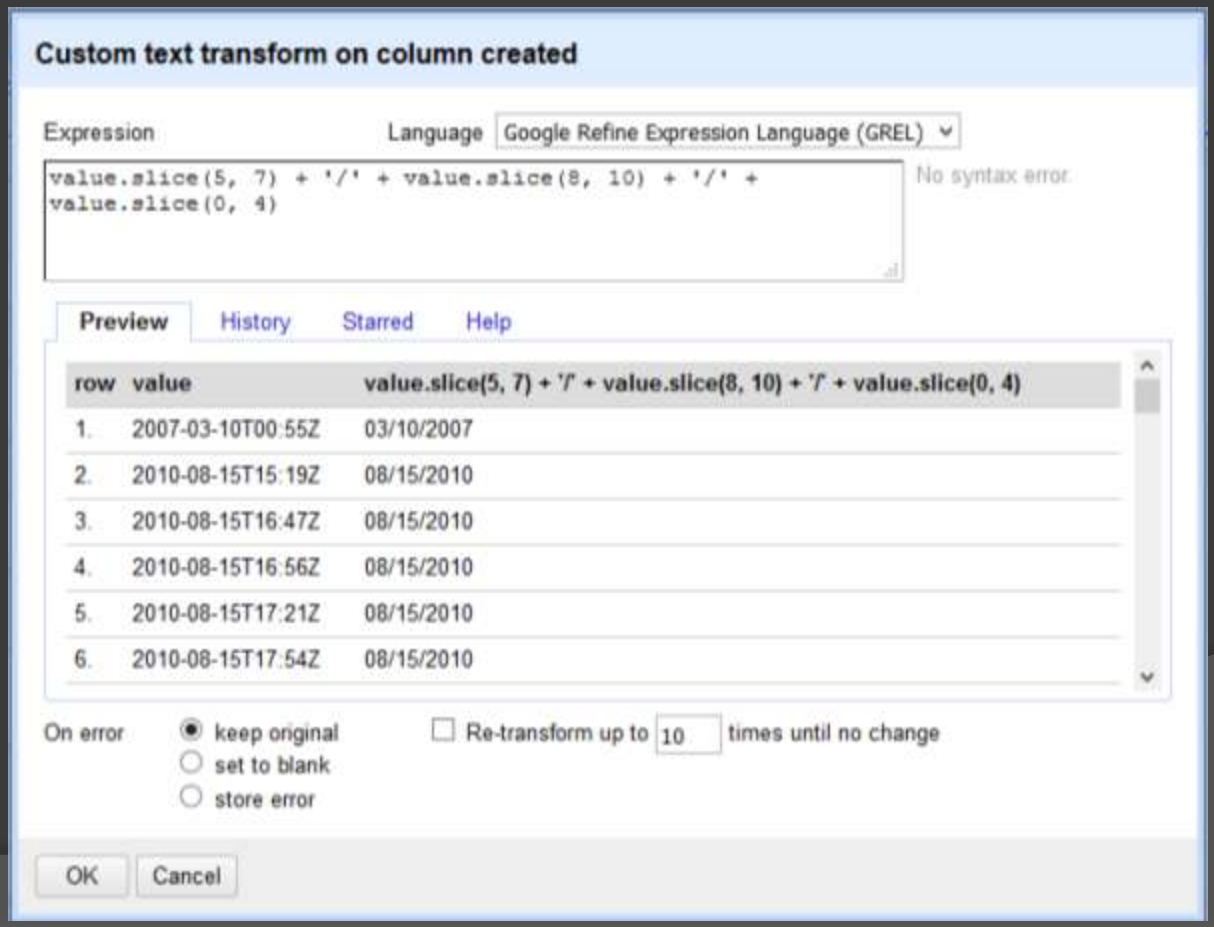


Modifier le format de la date

Exemple: Change "20100531T01:10:0Z" to "05/31/2010"

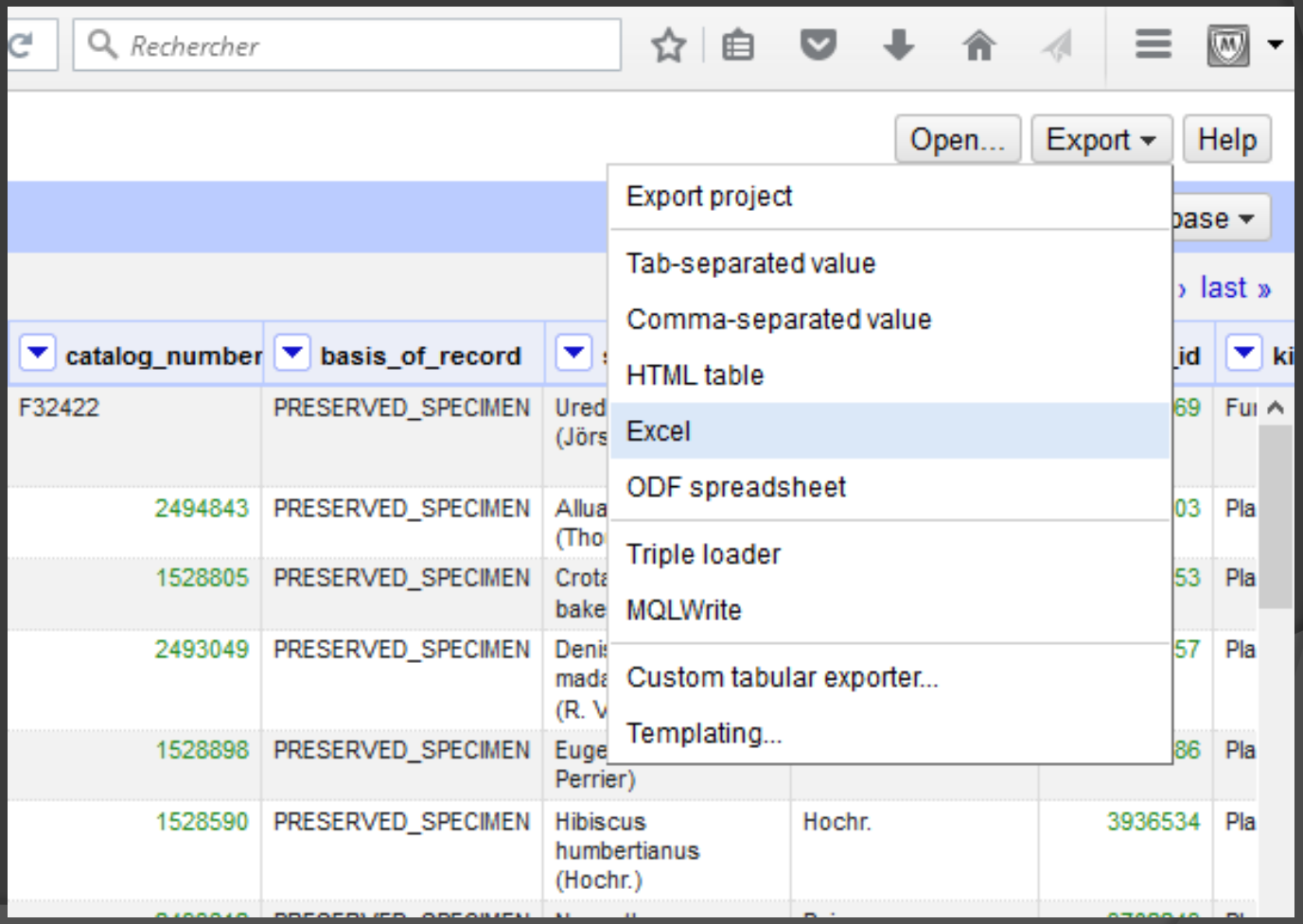
Menu déroulant sur “created” > Edit cells > Transform

Expression: value.slice(5, 7) + '/' + value.slice(8, 10) + '/' + value.slice(0, 4)



Exportation de données

Vous pouvez exporter vos données dans plusieurs formats différents, en fonction de ce que vous voulez ou ce que vous avez besoin.



Lectures complémentaires

- Les opérations décrites dans ce tutoriel sont les actions les plus fréquemment nécessaire lors du nettoyage des données pour OpenRefine.
- Le logiciel offre un ensemble beaucoup plus vaste de fonctionnalités pour nettoyer des données, il est donc utile de consulter la documentation complète à l'adresse: https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users

Misaotra!