

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 latest 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

Google refine

A power tool for working with messy data.

Create Project

Open Project

Import Project



Version 2.5 [r2407]

[Help](#)
[About](#)

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with Google Refine extensions.

Get data from

This Computer

[Web Addresses \(URLs\)](#)

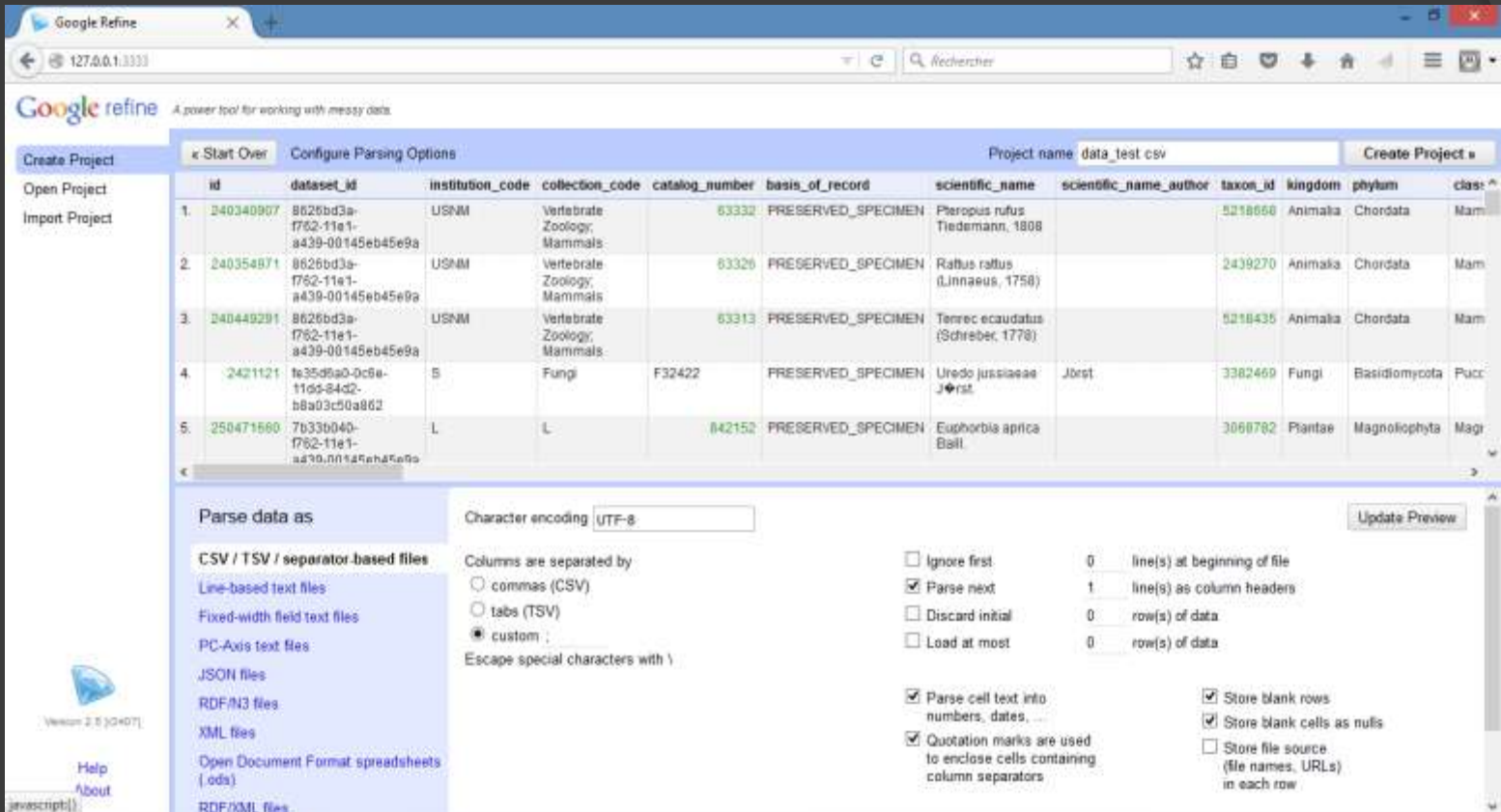
[Clipboard](#)

[Google Data](#)

Locate one or more files on your computer to upload:

Next »

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



The screenshot shows the Google Refine interface with the 'Configure Parsing Options' dialog box open. The dialog is titled 'Parse data as' and is set to parse data as CSV with commas (CSV). The character encoding is set to UTF-8. The dialog includes several options for handling headers and data rows, such as 'Ignore first', 'Parse next', 'Discard initial', and 'Load at most' lines at the beginning of the file. The 'Parse next' option is checked, indicating that the first line(s) of the file are used as column headers. The 'Parse cell text into numbers, dates, ...' option is also checked, indicating that the text in the cells is parsed into numbers and dates. The 'Store blank rows' and 'Store blank cells as nulls' options are checked, indicating that blank rows and cells are stored as nulls. The 'Store file source (file names, URLs) in each row' option is unchecked, indicating that the file source is not stored in each row.

id	dataset_id	institution_code	collection_code	catalog_number	basis_of_record	scientific_name	scientific_name_author	taxon_id	kingdom	phylum	class
1.	240340907	8626bd3a-f762-11e1-a439-00145eb45e9a	USNM	Vertebrate Zoology, Mammals	63332	PRESERVED_SPECIMEN	Pteropus rufus Tiedemann, 1808	5218668	Animalia	Chordata	Mam
2.	240354871	8626bd3a-f762-11e1-a439-00145eb45e9a	USNM	Vertebrate Zoology, Mammals	63326	PRESERVED_SPECIMEN	Rattus rattus (Linnaeus, 1758)	2439270	Animalia	Chordata	Mam
3.	240449291	8626bd3a-f762-11e1-a439-00145eb45e9a	USNM	Vertebrate Zoology, Mammals	63313	PRESERVED_SPECIMEN	Tenrec caudatus (Schreber, 1778)	5218435	Animalia	Chordata	Mam
4.	2421121	fe35d9a0-2c6e-11e5-84d2-b8a03cf0a862	S	Fungi	F32422	PRESERVED_SPECIMEN	Uredo jussiaeae Jörst	3382469	Fungi	Basidiomycota	Pucc
5.	250471680	7b33b040-f762-11e1-a439-00145eb45e9a	L	L	842152	PRESERVED_SPECIMEN	Euphorbia aprica Bail.	3060782	Plantae	Magnoliophyta	Magr

- 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

Google refine Data_test2.csv Permalink Open... Export Help

Facet / Filter Undo / Redo 0 Extensions: Freebase

376 rows

Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 50 next › last »

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](#)

All	id	dataset_id	institution_code	collection_code	catalog_number	basis_of_record	scientific_name	scientific_name	taxon_id	ki	
☆	1.	2421121	fe35d6a0-0c6e-11dd-84d2-b8a03c50a862	S	Fungi	F32422	PRESERVED_SPECIMEN	Uredo jussiaeae (Jörst.)	Jörst.	3382469	Fu
☆	2.	243090940	864a259a-f762-11e1-a439-00145eb45e9a	US	Botany	2494843	PRESERVED_SPECIMEN	Alluaudia humbertii (Thoux)	Thoux	5687103	Pla
☆	3.	243254560	864a259a-f762-11e1-a439-00145eb45e9a	US	Botany	1528805	PRESERVED_SPECIMEN	Crotalaria edmundi-bakeri (Viguier)	Viguier	2942253	Pla
☆	4.	243272281	864a259a-f762-11e1-a439-00145eb45e9a	US	Botany	2493049	PRESERVED_SPECIMEN	Denisophytum madagascariense (R. Vig.)	R. Vig.	2958857	Pla
☆	5.	243321088	864a259a-f762-11e1-a439-00145eb45e9a	US	Botany	1528898	PRESERVED_SPECIMEN	Eugenia scottii (H. Perrier)	H. Perrier	5416286	Pla
☆	6.	243384805	864a259a-f762-11e1-a439-00145eb45e9a	US	Botany	1528590	PRESERVED_SPECIMEN	Hibiscus humbertianus (Hochr.)	Hochr.	3936534	Pla
☆	7.	243489289	864a259a-f762-11e1-a439-00145eb45e9a	US	Botany	2493612	PRESERVED_SPECIMEN	Nepenthes madagascariensis (Poir.)	Poir.	3702243	Pla
☆	8.	243504865	864a259a-f762-11e1-a439-00145eb45e9a	US	Botany	1528628	PRESERVED_SPECIMEN	Oncostemum humbertianum (H. Perrier)	H. Perrier	3718752	Pla
☆	9.	231298451	84aca1ae-f762-11e1-a439-00145eb45e9a edit	K	Herbarium	K000300195	PRESERVED_SPECIMEN	Dypsis decipiens (Beentje & J. Dransf.)	(Becc.) Beentje & J. Dransf.	2735969	Pla
☆	10.	231322013	84aca1ae-f762-11e1-a439-00145eb45e9a	K	Herbarium	K000300237	PRESERVED_SPECIMEN	Hyphaene coriacea (Gaertn.)	Gaertn.	2735733	An
☆	11.	231830444	84aca1ae-f762-11e1-a439-00145eb45e9a	K	Herbarium	K000435668	PRESERVED_SPECIMEN	Trichomanes robinsonii (Hook. ex Baker)	Hook. ex Baker	5533416	Pla
☆	12.	230627319	84aca1ae-f762-11e1-a439-00145eb45e9a	K	Herbarium	K000394552	PRESERVED_SPECIMEN	Barleria parvispina (Benoist)	Benoist	3771519	Pla

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 apparaît 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*)

Facet / Filter

Undo / Redo

376 rows

Show as: rows records Show: 5 10 25 50 rows

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](#)

All	id	dataset_id	institution_code	collection_code	catalog_number
☆	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		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

- Facet
 - Text facet
 - Numeric facet
 - Timeline facet
 - Scatterplot facet
 - Custom text facet...
 - Custom numeric facet...
 - Customized facets
- Text filter
- Edit cells
- Edit column
- Transpose
- Sort...
- View
- Reconcile

Facet / Filter

Undo / Redo 0

Refresh

Reset All

Remove All

institution_code

change

13 choices Sort by: name count

Cluster

FishBase 1

K 21

MA 1

MNHN 275

MO 22

NY 5

RMNH 4

S 1

SANBI 1

South African Museum 6

US 7

[edit](#) [include](#)

USNM 1

ZMUC 31

Facet by choice counts

376 rows

Show as: rows records

Show: 5 10 25 50 rows

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
All	id	dataset_id	institution_code	collection_code	catalog_number	
<input type="checkbox"/>	1.	2421121	fe35d6a0-0c6e-11dd-84d2-b8a03c50a862	S	Fungi	F32422
<input type="checkbox"/>	2.	243090940	864a259a-f762-11e1-a439-00145eb45e9a	US	Botany	2494843
<input type="checkbox"/>	3.	243254560	864a259a-f762-11e1-a439-00145eb45e9a	US	Botany	1528805
<input type="checkbox"/>	4.	243272281	864a259a-f762-11e1-a439-00145eb45e9a	US	Botany	2493049
<input type="checkbox"/>	5.	243321088	864a259a-f762-11e1-a439-00145eb45e9a	US	Botany	1528898
<input type="checkbox"/>	6.	243384805	864a259a-f762-11e1-a439-00145eb45e9a	US	Botany	1528590
<input type="checkbox"/>						2493612
<input type="checkbox"/>						1528628
<input type="checkbox"/>	9.	231298451	84aca1ae-f762-11e1-a439-00145eb45e9a	K	herbarium	K000300195
<input type="checkbox"/>	10.	231322013	84aca1ae-f762-11e1-a439-00145eb45e9a	K	Herbarium	K000300237
<input type="checkbox"/>	11.	231830444	84aca1ae-f762-11e1-a439-00145eb45e9a	K	Herbarium	K000435668
<input type="checkbox"/>	12.	230627319	84aca1ae-f762-11e1-a439-00145eb45e9a	K	Herbarium	K000394552

US

Apply Cancel

Enter Esc

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 semblent 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

Keying Function

5 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	7	<ul style="list-style-type: none">• (Drake) Drake (4 rows)• Drake (3 rows)	<input type="checkbox"/>	<input type="text" value="(Drake) Drake"/>
2	3	<ul style="list-style-type: none">• (Parr, 1928) (2 rows)• Parr, 1928 (1 rows)	<input type="checkbox"/>	<input type="text" value="(Parr, 1928)"/>
2	6	<ul style="list-style-type: none">• DC (4 rows)• DC. (2 rows)	<input type="checkbox"/>	<input type="text" value="DC"/>
2	2	<ul style="list-style-type: none">• Boeck (1 rows)• Boeck. (1 rows)	<input type="checkbox"/>	<input type="text" value="Boeck"/>
2	4	<ul style="list-style-type: none">• Renauld & Paris (3 rows)• Paris & Renauld (1 rows)	<input type="checkbox"/>	<input type="text" value="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

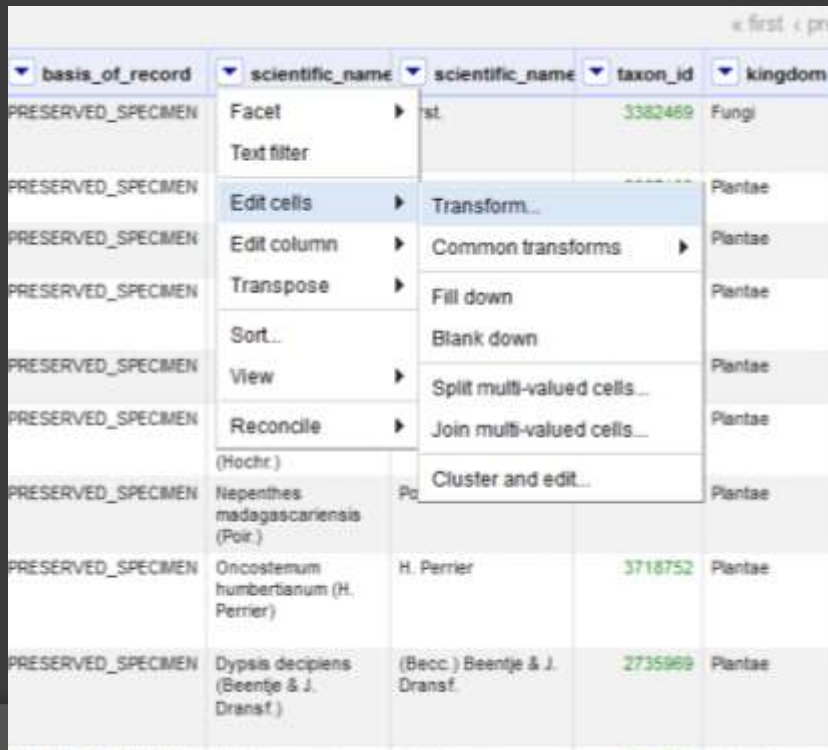
Merge Selected & Close

Close

Suppression des textes en parenthèses

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

Menu déroulant > Edit cells > Transform



The image shows a screenshot of a data table with a context menu open over the 'scientific_name' column. The table has five columns: 'basis_of_record', 'scientific_name', 'scientific_name', 'taxon_id', and 'kingdom'. The 'scientific_name' column contains several entries, some with author names in parentheses. The context menu is open over the 'scientific_name' column, showing options like 'Edit cells', 'Edit column', 'Transpose', 'Sort...', 'View', 'Reconcile', and 'Cluster and edit...'. The 'Edit cells' option is selected, and a sub-menu is open showing 'Transform...' as the first option.

basis_of_record	scientific_name	scientific_name	taxon_id	kingdom
PRESERVED_SPECIMEN	Facet	st.	3382469	Fungi
PRESERVED_SPECIMEN				
PRESERVED_SPECIMEN				Plantae
PRESERVED_SPECIMEN				Plantae
PRESERVED_SPECIMEN				Plantae
PRESERVED_SPECIMEN				Plantae
PRESERVED_SPECIMEN				Plantae
PRESERVED_SPECIMEN	(Hochr.)			Plantae
PRESERVED_SPECIMEN	Nepenthes madagascariensis (Poir.)	Po		Plantae
PRESERVED_SPECIMEN	Oncostemum humbertianum (H. Perrier)	H. Perrier	3718752	Plantae
PRESERVED_SPECIMEN	Dypsis decipiens (Beentje & J. Dransf.)	(Becc.) Beentje & J. Dransf.	2735969	Plantae

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

Expression:

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

Custom text transform on column scientific_name

Expression Language

No syntax error.

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

row	value	value.split("(")[0].strip()
1.	Uredo jussiaeae (Jörst.)	Uredo jussiaeae
2.	Alluudia humbertii (Thoux)	Alluudia humbertii
3.	Crotalaria edmundi-bakeri (Viguier)	Crotalaria edmundi-bakeri
4.	Denisophytum madagascariense (R. Vig.)	Denisophytum madagascariense
5.	Eugenia scottii (H. Perrier)	Eugenia scottii
6.	Hibiscus humbertianus (Hochr.)	Hibiscus humbertianus

On error keep original set to blank store error Re-transform up to times until no change

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)`

" data-bbox="172 353 790 989"/>

Custom text transform on column created

Expression: `value.slice(5, 7) + '/' + value.slice(8, 10) + '/' + value.slice(0, 4)` Language: Google Refine Expression Language (GREL) No syntax error.

Preview History Starred Help

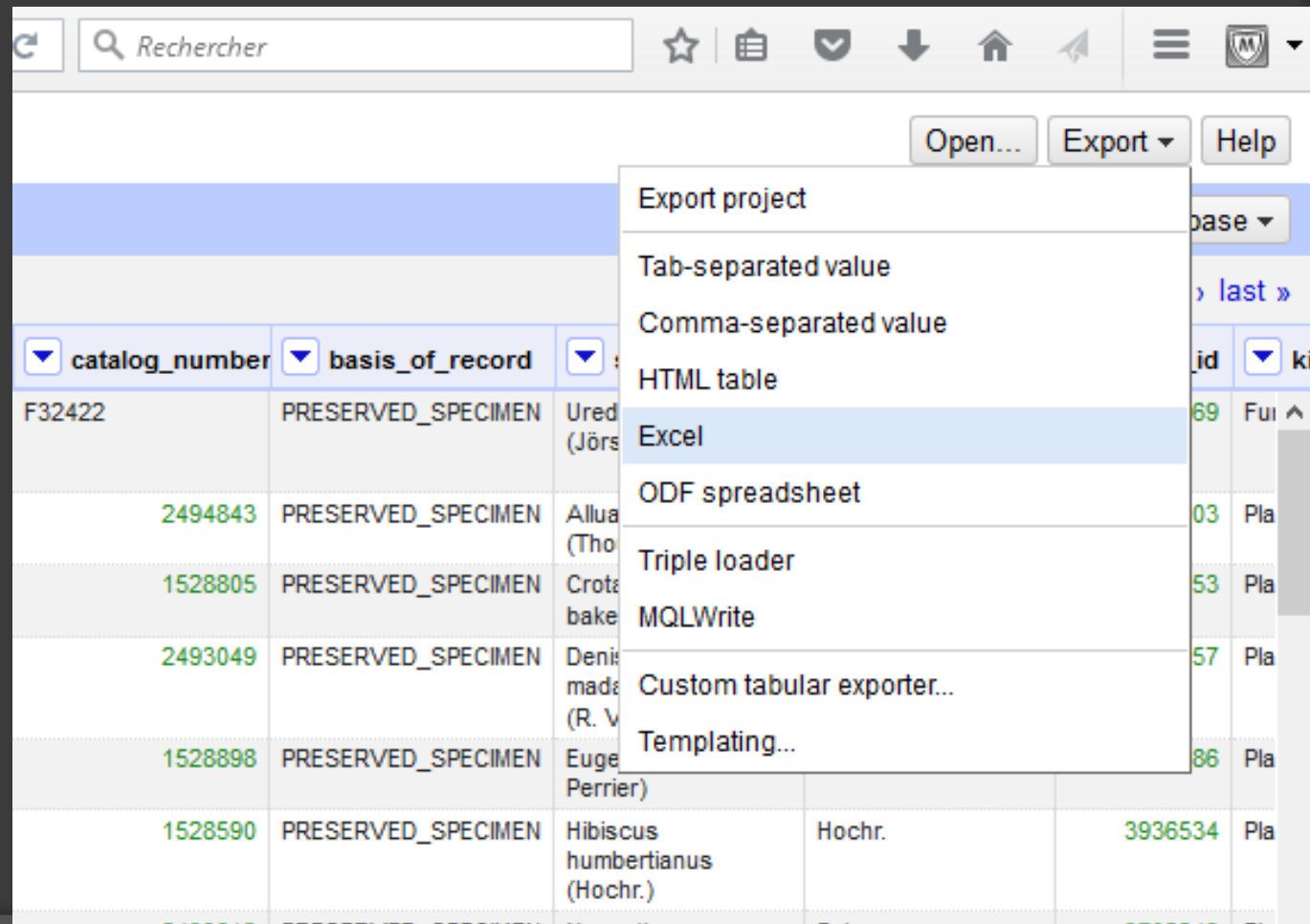
row	value	<code>value.slice(5, 7) + '/' + value.slice(8, 10) + '/' + value.slice(0, 4)</code>
1.	2007-03-10T00:55Z	03/10/2007
2.	2010-08-15T15:19Z	08/15/2010
3.	2010-08-15T16:47Z	08/15/2010
4.	2010-08-15T16:56Z	08/15/2010
5.	2010-08-15T17:21Z	08/15/2010
6.	2010-08-15T17:54Z	08/15/2010

On error: keep original set to blank store error Re-transform up to times until no change

OK Cancel

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.



The screenshot shows a web application interface with a search bar at the top containing the text "Rechercher". Below the search bar is a navigation bar with buttons for "Open...", "Export", and "Help". The main content area displays a table with columns for "catalog_number", "basis_of_record", and other fields. The "Export" dropdown menu is open, showing the following options:

- Export project
- Tab-separated value
- Comma-separated value
- HTML table
- Excel
- ODF spreadsheet
- Triple loader
- MQLWrite
- Custom tabular exporter...
- Templating...

catalog_number	basis_of_record				
F32422	PRESERVED_SPECIMEN	Ured	(Jörs		
2494843	PRESERVED_SPECIMEN	Allua	(Tho		03 Pla
1528805	PRESERVED_SPECIMEN	Crota	bake		53 Pla
2493049	PRESERVED_SPECIMEN	Deni	made	(R. V	57 Pla
1528898	PRESERVED_SPECIMEN	Euge	Perrier)		86 Pla
1528590	PRESERVED_SPECIMEN	Hibiscus	humbertianus	Hochr.	3936534 Pla

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!

