

Nettoyer et préparer des données avec OpenRefine

Formation Open Data Locale
Marseille, 9/6/2017



Mathieu Saby
mathsabypro@gmail.com
[@27point7](https://twitter.com/27point7)

BU UNS

Plan

1. Introduction et présentation d'OpenRefine
2. Import des données et présentation de l'espace de travail
3. Tris, filtres et facettes
4. Regrouper des valeurs proches
5. Transformations courantes des valeurs
6. Restructurer des données
7. Exporter les données et les traitements
8. Appliquer des transformations personnalisées

Introduction

- 1. Introduction et présentation d'OpenRefine**
2. Import des données et présentation de l'espace de travail
3. Tris, filtres et facettes
4. Regrouper des valeurs proches
5. Transformations courantes des valeurs
6. Restructurer des données
7. Exporter les données et les traitements
8. Appliquer des transformations personnalisées

Nettoyer et préparer des données

Étapes fréquemment nécessaires avant de diffuser des données (en tant que producteur), ou de les analyser (en tant qu'utilisateur):

Nettoyage : données hétérogènes, incomplètes, erronées, bruitées, mal normalisées...

Préparation : modification du format, de l'organisation, du codage ; croisement de différents jeux de données ; enrichissement..

Quels outils ?

Tableurs



Scripts



Outils dédiés à la préparation de données



Ou intégrant la préparation et d'autres fonctions



Etc.

Positionnement d'OpenRefine

Avantages

- Fonctions absentes des tableurs traditionnels
- Interface graphique (vs. scripts)
- Version complète totalement libre et gratuite
- Installation PC, Linux, Mac
- Enregistrement des traitements réalisés
- Maîtrise des données (vs outils en *cloud*)
- Large communauté d'utilisateurs

Inconvénients

- Performance limitée (100 000 lignes maxi environ)
- Pas de fonction collaborative
- Formats d'imports limités
- Pas de connexion avec des outils « big data »
- Langage spécifique
- N'est pas intégré dans une suite d'outils
- Peu d'évolutions ces dernières années

Présentation du logiciel

- Historique
 - 2010 : création par Metaweb, pour faciliter l'alimentation de leur base de connaissance Freebase
 - 2010-2012 : rachat par Google, renommé **Google Refine**
 - 2012 : libération du code par Google, renommé **OpenRefine**
- Modèle économique
 - Logiciel opensource, donc gratuit
 - Petite communauté de développeurs
- Versions
 - Dernière version bêta : 2.7rc2 (2017). **À installer de préférence à la 2.5**
 - Dernière version « officielle » : 2.5 (2011, développée par Google). Obsolète

Présentation du logiciel

- Aspects techniques
 - Écrit en langage Java (peut compliquer l'installation sous Mac)
 - Installation mono-poste, sous PC, Mac et Linux
 - Interface accessible via un navigateur internet (adresse <http://127.0.0.1:3333/>)
 - Les données et le logiciel restent sur le PC (pas besoin de connexion Internet)

Présentation du logiciel

- Apparence d'un tableur mais ça n'en est pas un
- Fonctionnalités principales
 - **Explorer** un jeu de données : tris, facettes, regroupement de valeurs proches
 - **Modifier** des données en mode graphique ou avec des formules
 - **Enrichir** des données
 - **Garder un historique** de tous les traitements
- Adapté à des données tabulées, faiblement dynamique (pas en temps réel) et de taille faible ou moyenne.
- Extensions possible avec des plug-ins, mais ne sont pas tous compatibles avec la dernière version

Usages possibles dans le contexte de l'open data

- Par des réutilisateurs de données
- Potentiellement par des producteurs
 - Préparation et nettoyage avant mise en ligne manuelle
 - Prototypage léger avant mise en place de chaînes de traitement lourdes (outils de type ETL) pour mise en ligne automatisée de données dynamiques

Facile à installer et utiliser y compris par les services producteurs de la données (pas forcément le service informatique)

Installation

<http://openrefine.org>

The screenshot shows the OpenRefine website homepage. At the top, there's a dark header bar with a "Google Custom Search" input field, a "Search" button, and social media links for "Follow us on: Github" and "Twitter". Below the header is a large, stylized "OPEN Refine" logo where "OPEN" is in blue and white, and "Refine" is in large blue letters. To the right of the logo is a blue diamond icon and the text "A free, open source, powerful tool for working with messy data". The main content area has a "Welcome!" heading and a paragraph about OpenRefine being a powerful tool for messy data. It also mentions that Google is no longer supporting the project, which has been rebranded to OpenRefine. A "Using OpenRefine - The Book" section features an image of the book cover and a list of topics covered in the book.

Follow us on: [Github](#) [Twitter](#)

Google Custom Search [Search](#) ×

OPEN Refine

A free, open source, powerful tool for working with messy data

Welcome!

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

Please note that since October 2nd, 2012, Google is not actively supporting this project, which has now been rebranded to OpenRefine. Project development, documentation and promotion is now fully supported by volunteers. Find out more about the [history of OpenRefine](#) and how you can [help the community](#).

Using OpenRefine - The Book

Using OpenRefine, by Ruben Verborgh and Max De Wilde, offers a great introduction to OpenRefine. Organized by recipes with hands on examples, the book covers the following topics:

1. Import data in various formats
2. Explore datasets in a matter of seconds
3. Apply basic and advanced cell transformations
4. Deal with cells that contain multiple values
5. Create instantaneous links between datasets
6. Filter and partition your data easily with regular expressions

Installation

Installer la dernière version 2.7-rc2

Nécessite une version récente de Java JRE

Download OpenRefine

You will find on this page a list of OpenRefine distributions and extensions available for download. Are we missing something? Want to fix a typo? You can submit changes (pull request) [from here](#).

[Home](#)

[Download](#)

[Documentation](#)

[Community](#)

[Post archive](#)

OpenRefine News:
Spring 2016

OpenRefine News:

Official Distribution

Read the [installation instructions](#)

You can also Download All Official Releases and source from our [GITHUB RELEASES PAGE HERE](#)

OpenRefine 2.7-rc2 Release Candidate 2

An updated release on Mar 3, 2017. A change log is provided on the [release page](#).

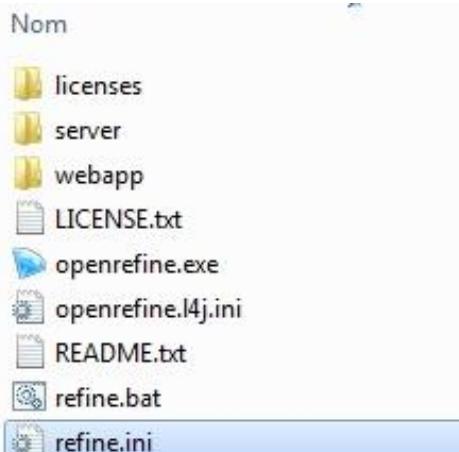
- **Windows kit**, Download, unzip, and double-click on *openrefine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it.
- **Linux kit**, Download, extract, then type *./refine*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

Installation

Par défaut OpenRefine utilise 1 Go de mémoire vive au maximum. Au besoin modifier la configuration pour allouer plus de mémoire :

Sous Windows : dans le fichier *refine.ini*, modifier la ligne **REFINE_MEMORY**

Pour que *refine.ini* soit pris en compte il faudra lancer OpenRefine avec *refine.bat* et non *openrefine.exe*



```
1  # NOTE: This file is not read if you run the Refine executable directly
2  # It is only read if you use the refine shell script or refine.bat
3
4  no_proxy="localhost,127.0.0.1"
5  #REFINE_PORT=3334
6  #REFINE_HOST=127.0.0.1
7  #REFINE_WEBAPP=main\webapp
8
9  # Memory and max form size allocations
10 #REFINE_MAX_FORM_CONTENT_SIZE=1048576
11 REFINE_MEMORY=2000M
12
13 # Some sample configurations. These have no defaults.
14 #ANT_HOME=C:\grefine\tools\apache-ant-1.8.1
15 #JAVA_HOME=C:\Program Files\Java\jdk1.6.0_25
16 #JAVA_OPTIONS=-XX:+UseParallelGC -verbose:gc -Drefine.headless=true
17 #JAVA_OPTIONS=-Drefine.data_dir=C:\Users\user\AppData\Roaming\OpenRefine
18
```

Import des données et espace de travail

1. Introduction et présentation d'OpenRefine
2. **Import des données et présentation de l'espace de travail**
3. Tris, filtres et facettes
4. Regrouper des valeurs proches
5. Transformations courantes des valeurs
6. Restructurer des données
7. Exporter les données et les traitements
8. Appliquer des transformations personnalisées

Lancer OpenRefine

Ouvrir un navigateur (Chrome ou Firefox)

Windows : lancer de préférence *refine.bat*.

Mac : chercher OpenRefine dans les applications

Si OpenRefine ne s'ouvre pas, saisir <http://localhost:3333> dans le navigateur

Les fichiers d'exercices

Télécharger sur votre bureau:

- Deux mini jeux de données fictives

Exo 1: <http://bit.ly/2sjJfGO> (URL complète

<https://drive.google.com/open?id=0B1NKejaqcJG5am5TNEphbHoza1U> ou

https://raw.githubusercontent.com/msaby/formations/master/2017/openrefine_marseille/exo1.csv)

Exo 2: <http://bit.ly/2r9wUkc> (URL complète

<https://drive.google.com/open?id=0B1NKejaqcJG5dVcxTXhCTTzoZkE> ou

https://raw.githubusercontent.com/msaby/formations/master/2017/openrefine_marseille/exo2.csv) (**exo2.csv**)

- Annuaire des associations de la CCABV et de Digne-les-Bains en 2016

<http://opendata.regionpaca.fr/donnees/detail/annuaire-des-associations-de-la-ccabv-et-de-digne-les-bains-en-2016.html>

Importer des données

Projet = un fichier de données + un ensemble de traitements

Un projet peut être réouvert, ou importé depuis une autre installation d'OpenRefine

The screenshot shows the OpenRefine web application. At the top left is the logo 'OPEN Refine'. Below it is a navigation bar with four buttons: 'Créer un projet' (highlighted with a green arrow), 'Ouvrir un projet', 'Importer un projet', and 'Langue'. The main content area has a sub-header 'Un outil puissant pour travailler avec des données désordonnées.' followed by a section titled 'Créer un projet en important des données. Quelles sortes de données puis-je importer ?'. It explains that TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, OpenDocument (.ods) and Google Data formats are supported. On the left, there's a sidebar with options: 'Récupérer les données à partir de' (with 'Cet ordinateur' selected), 'Chercher un ou plusieurs fichiers à charger :' (with a 'Parcourir...' button and a message 'Aucun fichier sélectionné.'), and buttons for 'Suivant »' and 'Précédent <'. Below the sidebar are links for 'Adresses web (URLs)', 'Presse-papier', and 'Google Data'.

Importer des données

Plusieurs **formats de fichiers** possibles
(y compris fichier zippé)

Depuis plusieurs **emplacements**



Un outil puissant pour travailler avec des données désordonnées.

Créer un projet

Ouvrir un projet

Importer un projet

Langue

Créer un projet en important des données. Quelles sortes de données puis-je importer ?

Les documents de type TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, OpenDocument (.ods) et Google Data sont reconnus nativement. D'autres formats peuvent être ajoutés via des extensions OpenRefine.

Récupérer les données à partir de

Cet ordinateur

Adresses web (URLs)

Presse-papier

Google Data

Chercher un ou plusieurs fichiers à charger :

Parcourir...

Aucun fichier sélectionné.

Suivant »

Importer des données

Créer un nouveau projet à partir du fichier exo1.csv

Importer des données

Charger le fichier dans OpenRefine le fichier précédemment enregistré sur le Bureau

The screenshot shows the OpenRefine interface with the following elements:

- Header:** OPEN Refine. Subtext: Un outil puissant pour travailler avec des données désordonnées.
- Left sidebar:** Navigation menu with options: Crée un projet (highlighted in blue), Ouvrir un projet, Importer un projet, Langue.
- Main content area:**
 - Title:** Crée un projet en important des données. Quelles sortes de données puis-je importer ?
 - Description:** Les documents de type TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, OpenDocument (.ods) et Google Data sont reconnus nativement. D'autres formats peuvent être ajoutés via des extensions OpenRefine.
 - Import methods:** Récupérer les données à partir de Cet ordinateur (option 1 highlighted with a red box), Adresses web (URLs), Presse-papier, Google Data.
 - File selection:** Chercher un ou plusieurs fichiers à charger : Parcourir... (button highlighted with a red box), doaj-article-sample.csv
 - Next step:** Suivant » (button highlighted with a red box)

Importer des données

Ecran d'import en 2 parties : aperçu des données + paramètres

« Démarrer Configurer les options pour l'analyse syntaxique

Nom : or1 csv [Créer un projet »](#)

Toutes	code_personne	date	ville	adresse	animal_prefere	habillement	loisirs	logement
1.	P001	01/02/2017	NICE	1 av. St Barthélémy	chien	100	25	0,8
2.	P002	01/03/2017	CAEN		chiens			
3.	P003	15/02/2017	Lyon	3 rue Paul Bert	chiens et chats	10.90	70,6	700
4.	P004	15-02-2017	Nice	50 avenue Saint Barthélémy	chat, cheval, poisson	400	90	600
5.	P005	15-04-2017	LE HAVRE	15 av. Jean Jaurès	CHAT			
6.	P005	12-02-2017	Havre (Le)	15 av. Jean Jaurès	chevaux			
7.	P005	11/01/2017			lapin			
8.	P006	19/02 (2017)	Lyon	1 rue Dunoir				
9.	P002	16/03 (2017)	Caen	5 rue Basse		50.50	35,6	0,7
10.	P005	08/01/2017	Le Havre	15 av. Jean Jaurès	Lapin,chien	200	40	800

Considérer les données comme Mettre à jour l'aperçu

Format des caractères

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

MARC files

RDF/N3 files

XML files

Open Document Format spreadsheets (.ods)

RDF/XML files

Excel files

Les colonnes sont séparées par : une virgule (CSV) une tabulation (TSV) autre , Protéger les caractères spéciaux avec \

Ignorer la ou les premières lignes du début du fichier 0 lignes

Analyser la ou les 1 lignes suivantes comme entêtes de colonnes

Ignorer la ou les 0 premières lignes de données

Charger au plus 0 premières lignes de données

Analyser le texte des cellules comme nombres, dates...

Des guillemets sont utilisés pour délimiter les cellules qui contiennent des séparateurs de colonne

Conserver les lignes vides

Analyser les cellules vides comme nulles

Indiquer la source du fichier (noms des fichiers, URLs) dans chaque ligne

Importer des données

Principaux paramètres d'import

Encodage des caractères
(en général **UTF-8** ou **ISO 8859-1**)

1

Format des caractères

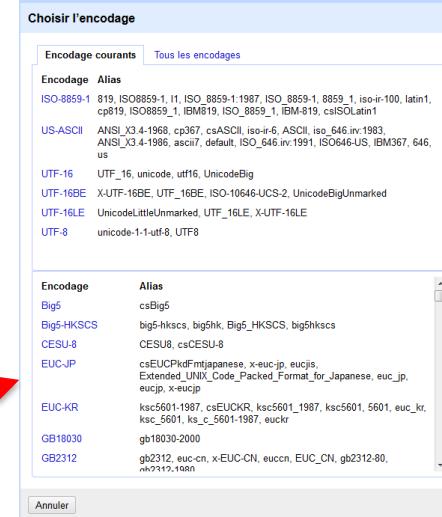
Les colonnes sont séparées par :

- une virgule (CSV)
- une tabulation (TSV)
- autre ,

Protéger les caractères spéciaux avec \

2

Séparateur de colonnes
(en général , mais parfois ; si le fichier a été créé avec une version française d'Excel)



Modifier si entêtes mult lignes

3

Mettre à jour l'aperçu

- | | | |
|--|---|--|
| <input type="checkbox"/> Ignorer la ou les premières | 0 | lignes du début du fichier |
| <input checked="" type="checkbox"/> Analyser la ou les | 1 | lignes suivantes comme entêtes de colonnes |
| <input type="checkbox"/> Ignorer la ou les | 0 | premières lignes de données |
| <input type="checkbox"/> Charger au plus | 0 | premières lignes de données |

Analyser le texte des cellules comme nombres, dates...

Des guillemets sont utilisés pour délimiter les cellules qui contiennent des séparateurs de colonne

Détection des nombres et des dates (**dans le doute, à éviter**)

En général à décocher

- Conserver les lignes vides
- Analyser les cellules vides comme nulles
- Indiquer la source du fichier (noms des fichiers, URLs) dans chaque ligne

5

Importer des données

Pour lire et écrire en français... choisir l'encodage correspondant à celui du fichier

Format
des
caractères

ISO-8859-1

1 av. St Barthélémy

Format
des
caractères

UTF-8

1 av. St Barthélemy

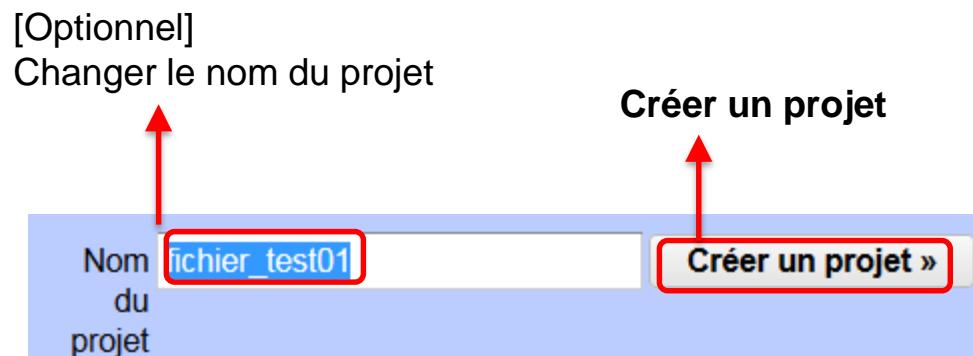
Importer des données

Prudence avec la détection automatique des nombres et des dates ! Dans le doute, désactiver l'option.
(remarque également valable pour Excel ou LibreOffice)

- Une série de chiffres n'est pas forcément un nombre.
 - Ex: Numéros de téléphone : le 0 initial doit être préservé!
- Formats de nombres et formats monétaires différents
 - Ex : 1,14 en France = 1.14 aux USA
 - Ex : 10 € mais \$ 10
- Formats de dates différents selon les pays.
 - Ex : 02-03-1979 = 2 mars 1979 en Europe
3 février 1979 aux USA

Importer des données

Une fois les paramètres d'imports choisis, lancer l'import



L'espace de travail

The screenshot shows the OpenRefine interface with several features highlighted by green arrows:

- Facettes et filtres**: Points to the facet navigation bar at the top left.
- Historique**: Points to the history navigation bar at the top left.
- Lien vers le projet**: Points to the "Permalien" button at the top center.
- Contenu du fichier**: Points to the main data grid area.
- Nouveau projet**: Points to the "Ouvrir..." button at the top right.
- Export**: Points to the "Exporter" dropdown menu at the top right.

Facet navigation bar: Facette / Filtre | Défaire / Refaire | doaj article sample

Data grid: 10 lignes | Extensions: « première < précédente 1 - 10 suivante > dernière »

	Toutes	code_personne	date	ville	adresse	animal_prefere	habillement	loisirs	logement
1.	P001	01/02/2017	NICE	1 av. St Barthélemy	chien	100	25	0,8	
2.	P002	01/03/2017	CAEN		chiens				
3.	P003	15/02/2017	Lyon	3 rue Paul Bert	chiens et chats	10.90	70,6	700	
4.	P004	15-02-2017	Nice	50 avenue Saint Barthélemy	chat, cheval, poisson	400	90	600	
5.	P005	15-04-2017	LE HAVRE	15 av. Jean Jaurès	CHAT				
6.	P005	12-02-2017	Havre (Le)	15 av. Jean Jaurès	chevaux				
7.	P005	11/01/2017			lapin				
8.	P006	19/02 (2017)	Lyon	1 rue Dunoir					
9.	P002	16/03 (2017)	Caen	5 rue Basse		50.50	35,6	0,7	
10.	P005	08.01.2017	Le Havre	15 av. Jean Jaurès	Lapin, chien	200	40	800	

Left sidebar: Utiliser les facettes et les filtres
Utiliser les facettes et les filtres pour sélectionner les sous-ensembles de données à traiter. Choisir les méthodes de facette et de filtre dans les menus situés dans les entêtes de colonne.
Vous ne savez pas par où commencer ? Regarder ces tutoriels vidéos

L'espace de travail

Numéro de ligne (automatique)

Nb lignes du fichier

Nb lignes affichées

Colonnes de données

Voir les lignes précédentes ou suivantes

Étoiles et marques: pour isoler certaines lignes

The screenshot shows the Refine data editor interface. At the top, there's a header with the Refine logo, a title 'doaj article sample', and buttons for 'Facette / Filtre' and 'Défaire / Refaire'. Below the header is a sidebar with a section titled 'Utiliser les facettes et les filtres' containing text and a link to video tutorials. The main area contains a table with 10 rows of data. The columns are labeled: 'Toutes', 'code_personne', 'date', 'ville', 'adresse', 'animal_prefere', 'habillement', 'loisirs', and 'logement'. Each row has a small icon next to the number. Above the table, there are several controls: 'Nb lignes du fichier' (set to 10), 'Nb lignes affichées' (set to 10), and 'Colonnes de données'. To the right of the table are buttons for 'Ouvrir...', 'Exporter', and 'Aide'. Below the table, there are links for navigating through the data: '< première', '< précédente', '1 - 10', 'suivante', and '> dernière'.

Toutes	code_personne	date	ville	adresse	animal_prefere	habillement	loisirs	logement
1.	P001	01/02/2017	NICE	1 av. St Barthélemy	chien	100	25	0,8
2.	P002	01/03/2017	CAEN		chiens			
3.	P003	15/02/2017	Lyon	3 rue Paul Bert	chiens et chats	10.90	70,6	700
4.	P004	15-02-2017	Nice	50 avenue Saint Barthélemy	chat, cheval, poisson	400	90	600
5.	P005	15-04-2017	LE HAVRE	15 av. Jean Jaurès	CHAT			
6.	P005	12-02-2017	Havre (Le)	15 av. Jean Jaurès	chevaux			
7.	P005	11/01/2017			lapin			
8.	P006	19/02 (2017)	Lyon	1 rue Dunoir				
9.	P002	16/03 (2017)	Caen	5 rue Basse		50.50	35,6	0,7
10.	P005	08/01/2017	Le Havre	15 av. Jean Jaurès	Lapin, chien	200	40	800

Différences avec un tableur

On ne voit pas toutes les lignes. Ce n'est pas le but de l'outil

On applique les formules à des colonnes entières, pas à des cellules

Les données sont séparées des traitements : les formules ne sont pas contenues dans les cellules

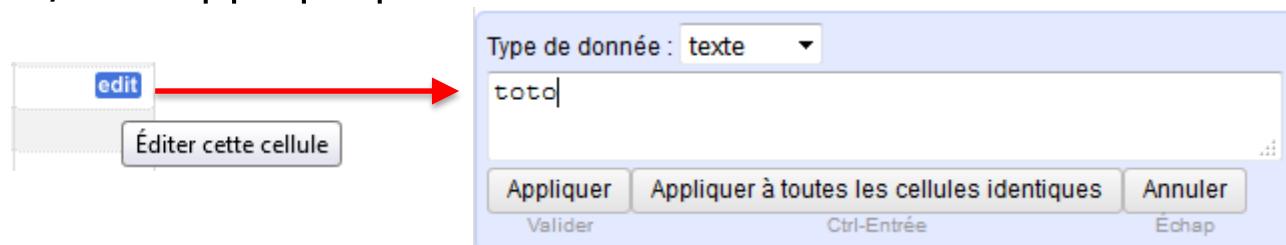
Une seule feuille

Pas de graphiques

Comment explorer et manipuler les données

Modification d'une cellule : bouton *edit* visible au survol

- Ponctuelle
- Pour toutes cellules ayant la même valeur (dans la même colonne ; ne s'applique pas aux cellules vides)



Actions globales : menu visible en cliquant sur le bouton en haut de chaque colonne

- Affichage sélectif (tris, filtres, facettes...)
- Modifications (remplacements, nouvelles colonnes...)

Les menus contextuels

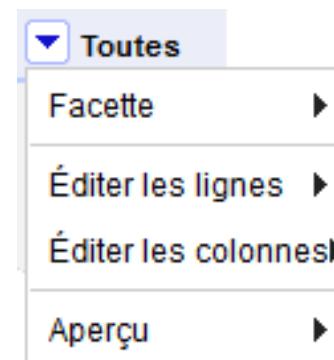
Les plus fonctions les utilisées : facettes, édition de cellules et de colonnes, tri

Fonction propre à la 1^{re} colonne : suppression de lignes

Colonne ordinaire



Colonne « Toutes » (1^{re} position)



Réordonner ou supprimer des colonnes

Dans la 1^{re} colonne *Toutes*

Toutes

Facette

Éditer les lignes

Éditer les colonnes

Aperçu

Retrier / supprimer les colonnes...

Trier / Supprimer des colonnes

Glisser des colonnes pour les trier

Déposer des colonnes ici pour les supprimer

Title

Authors

DOI

URL

Date

Language

Subjects

ISSNs

Publisher

Citation

Licence

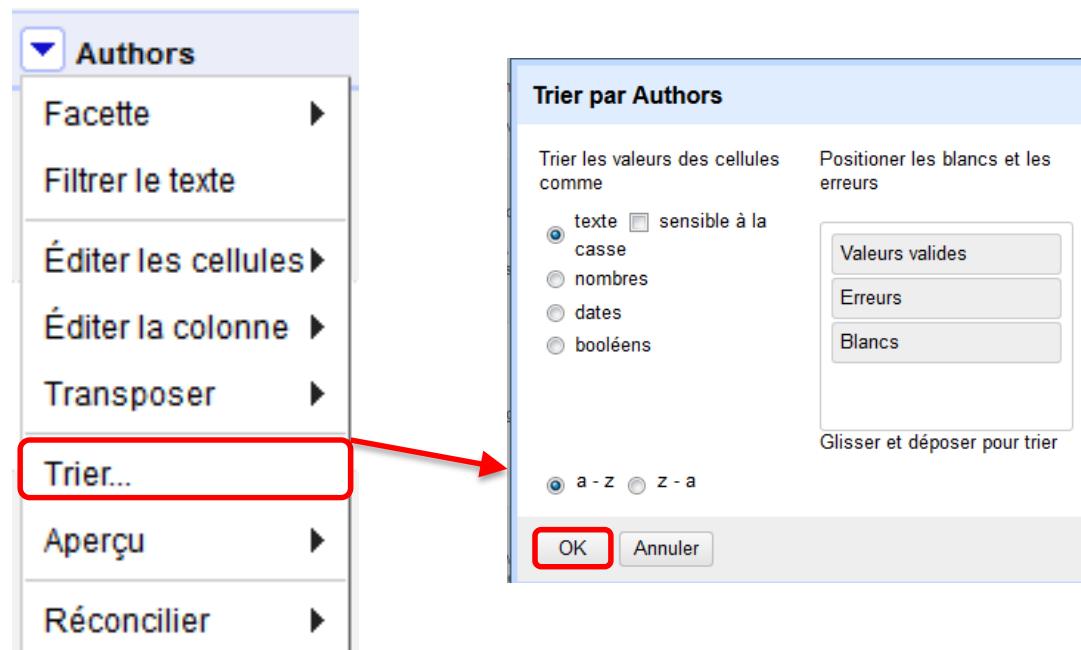
OK Annuler

Tris, filtres et facettes

1. Introduction et présentation d'OpenRefine
2. Import des données et présentation de l'espace de travail
3. **Tris, filtres et facettes**
4. Regrouper des valeurs proches
5. Transformations courantes des valeurs
6. Restructurer des données
7. Exporter les données et les traitements
8. Appliquer des transformations personnalisées

Trier les données

Activité : trier les données en fonction des valeurs de la colonne *Nom* (de A à Z, sans tenir compte des majuscules)



Trier les données

Activité : trier les données en fonction des valeurs de la colonne *code_personne* (de A à Z, sans tenir compte des majuscules)

	code_personne
1.	P001
2.	P002
9.	P002
3.	P003
4.	P004
5.	P005
6.	P005
7.	P005
10.	P005
8.	P006

Pour l'instant l'ordre original est préservé (le tri concerne juste l'affichage)

Trier les données

Retrier de façon permanente

Voir en: lignes entrées Afficher: 5 10 25 50 lignes					Sort ▾	habillement	loisirs	logement
	Toutes	code_personne	date	ville				
1.	P001	01/02/2017	NICE	1 av.	Supprimer le tri		25	0,8
2.	P002	01/03/2017	CAEN		Retrier les lignes de façon permanente			
9.	P002	16/03 (2017)	Caen	5 rue basse	Par code_personne	50.50	35,6	0,7
3.	P003	15/02/2017	Lyon	3 rue Paul Bert	chiens et chats	10.90	70,6	700
4.	P004	15-02-2017	Nice	50 avenue Saint Barthélemy	chat, cheval, poisson	400	90	600
5.	P005	15-04-2017	LE HAVRE	15 av. Jean Jaurès	CHAT			
6.	P005	12-02-2017	Havre (Le)	15 av. Jean Jaurès	chevaux			
7.	P005	11/01/2017			lapin			
10.	P005	08/01/2017	Le Havre	15 av. Jean Jaurès	Lapin,chien	200	40	800
8.	P006	19/02 (2017)	Lyon	1 rue Dunoir				

Filtrer les données

Activité : filtrer le fichier pour afficher les lignes dont la colonne *code_personne* contient le mot « P005 » ET la colonne *animal_prefere* le mot « chien »

The screenshot shows a data filtering interface with two main sections:

- code_personne**:
 - Facette
 - Filtrer le texte** (highlighted with a red box)
 - Éditer les cellules
 - Éditer la colonne
 - Transposer
 - Trier
 - Aperçu
 - Réconcilier
- animal_prefere**:
 - Facette
 - Filtrer le texte** (highlighted with a red box)
 - Éditer les cellules
 - Éditer la colonne
 - Transposer
 - Trier...
 - Aperçu
 - Réconcilier

Two filter panels are open on the right:

- code_personne**:
 - P005
 - sensible à la casse expression rationnelle
- animal_prefere**:
 - chien
 - sensible à la casse expression rationnelle

A red arrow points from the bottom of the 'animal_prefere' panel towards the results table.

Results:

1 matching ligne(s) (10 total)									
Voir en:	lignes	entrées	Afficher:	5	10	25	50	lignes Sort ▾	
	Toutes	code_personne	date	ville	adresse	animal_prefere	habillement	loisirs	logement
	10.	P005	08:01:2017	Le Havre	15 av. Jean Jaurès	Lapin, chien	200	40	800

Filtrer les données

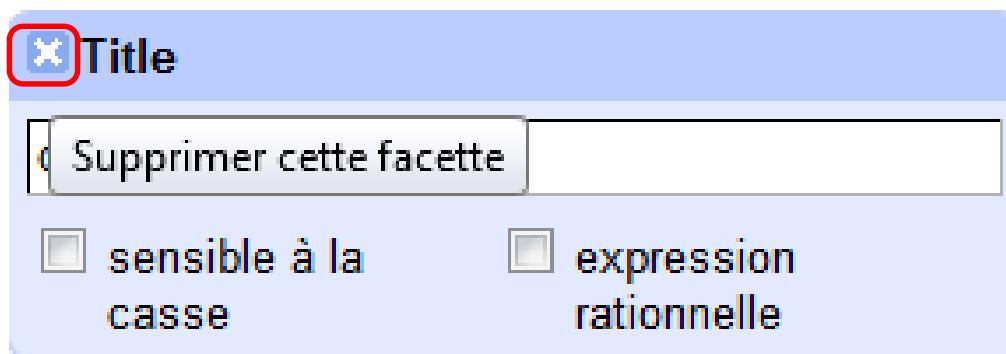
Toutes les opérations (export, nouveaux filtres, facettes, modifications groupées) s'opèreront uniquement sur les données filtrées.

Ex: modification groupée la colonne *animal_prefere* :
uniquement 1 lignes modifiée



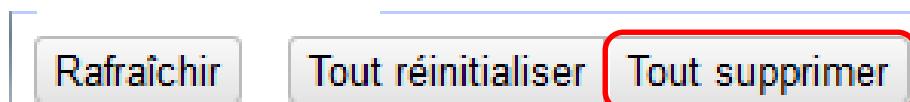
Filtrer les données

Pour annuler un filtre, cliquez sur la croix dans le coin supérieur gauche du filtre



Nous allons annuler tous les filtres

Pour cela, cliquez sur *Tout supprimer* au dessus des filtres

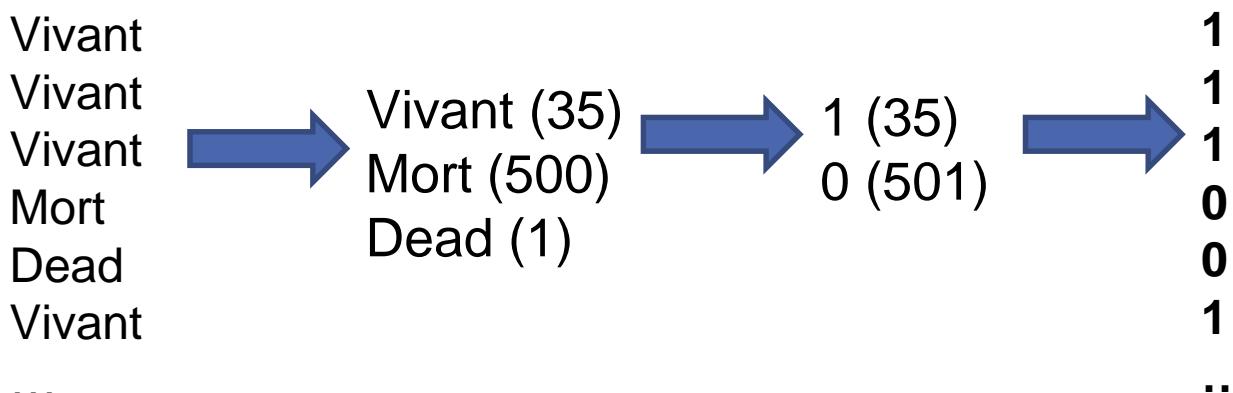


Utiliser les facettes

Les facettes permettent d'avoir un aperçu synthétique sur le contenu d'une colonne.

Utile pour repérer des anomalies, isoler des valeurs à modifier, modifier globalement un codage...

Ex : modifier et corriger un codage



Utiliser les facettes

Activité : afficher les facettes textuelles correspondant au contenu de la colonne *ville*.

The screenshot shows a software interface for managing data facets. On the left, a sidebar for the 'Ville' column lists various options: Facette (selected), Filtrer le texte, Éditer les cellules, Éditer la colonne, Transposer, Trier..., Aperçu, and Réconcilier. A red arrow points from the 'Facette' option to a dropdown menu. This menu contains: Facette textuelle (highlighted with a red box), Facette numérique, Facette chronologique, Facette de nuage de points, Personnaliser la facette textuelle..., Personnaliser la facette numérique, and Facettes courantes. An arrow points from this menu to the right-hand facet list.

The right-hand panel displays the selected facet: **ville**. It shows 8 choices, sorted by count: Lyon 1, CAEN 1, Caen 1, Havre (Le) 2, Le Havre 1, Lyon 1, Nice 1, NICE 1, and (blank) 1. There is a 'Groupe' button at the top right. Below the list, the text '(blank) : valeur vide' is displayed.

Quelles anomalies repère-t-on?

Utiliser les facettes

Les options d'une facette

Récupérer la liste Tri alphabétique (défaut) ou par nombre d'occurrences
valeurs vides (blank) toujours à la fin

ville

9 choices Trier par: nom compte changer

Groupe

Lyon 1
CAEN 1
Caen 1
Havre (Le) 1
Le Havre 1
LE HAVRE 1
Lyon 1
Nice 1
NICE 1
(blank) 1

Facette par nombre de choix

Utiliser les facettes

Activité : dans la facette *ville*, afficher les valeurs par « nombre de choix », et ne conserver que celles présentes 2 fois.

Facette ville

9 choices Trier par: nom compte Groupe

- Lyon 1
- CAEN 1
- Caen 1
- Havre (Le) 1
- Le Havre 1
- LE HAVRE 1
- Lyon 1
- Nice 1
- NICE 1
- (blank) 1

Facette par nombre de choix

histogramme

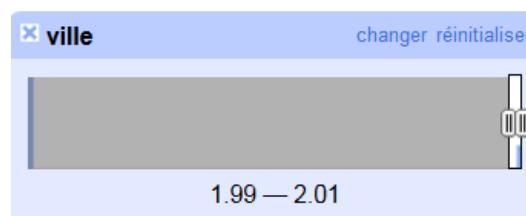
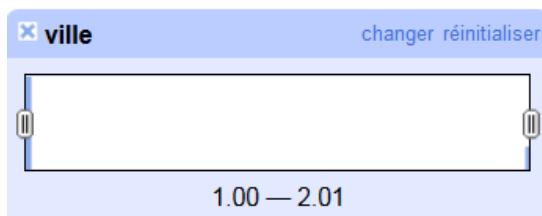
Facette ville

1 choices Trier par: nom compte Groupe

- Havre (Le) 2

Facette par nombre de choix

1 seule valeur, présente 2 fois



Utiliser les facettes

Activité : dans la facette *ville*, remplacer la valeur « *Havre (Le)* » par *LE HAVRE*

The screenshot shows a software interface for managing a 'ville' facet. At the top, there's a header with a close button (X), the facet name 'ville', a 'changer' button, and a 'Groupe' button. Below this, it says '1 choices Trier par: nom compte'. A choice named 'Havre (Le) 2' is listed with 'éditer include' and 'Facette par nombre de choix' buttons. A red arrow points from the text 'remplacer la valeur « Havre (Le) » par LE HAVRE' to the 'Havre (Le)' text in the list. A modal dialog box is open at the bottom, containing the text 'LE HAVRE' in a input field, with 'Appliquer' and 'Annuler' buttons below it. The background shows a yellow bar at the bottom with the text 'Mass edit 2 cells in column ville Défaire'.

Utiliser les facettes

Activité : utiliser la facette *ville* pour afficher les lignes dont la ville n'est PAS Lyon

The screenshot shows a search interface with a facet panel for the term 'ville'. The facet panel has a blue header bar with the text 'ville' and three buttons: 'changer', 'inverser', and 'réinitialiser'. Below this, there is a button 'Groupe' and a link 'Facette par nombre de choix'. The main list contains the following items:

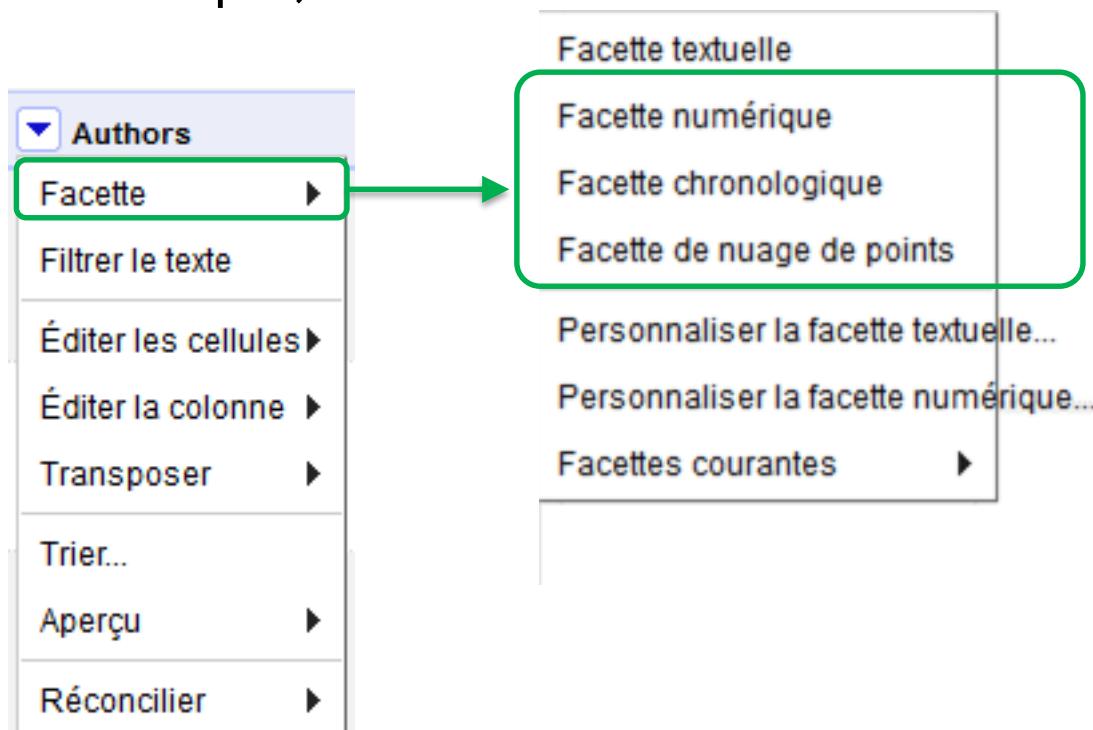
Ville	Nombre d'occurrences	Action
Lyon	-1	include
CAEN	-1	
Caen	-1	
Le Havre	-1	
LE HAVRE	-2	
Lyon	-1	include
Nice	-1	
NICE	-1	
(blank)	-1	

facette en orange : utilisé pour filtrer les données (afficher les données correspondant à la facette)

facette en noir barré : filtre inversé (afficher les données ne correspondant pas à la facette)

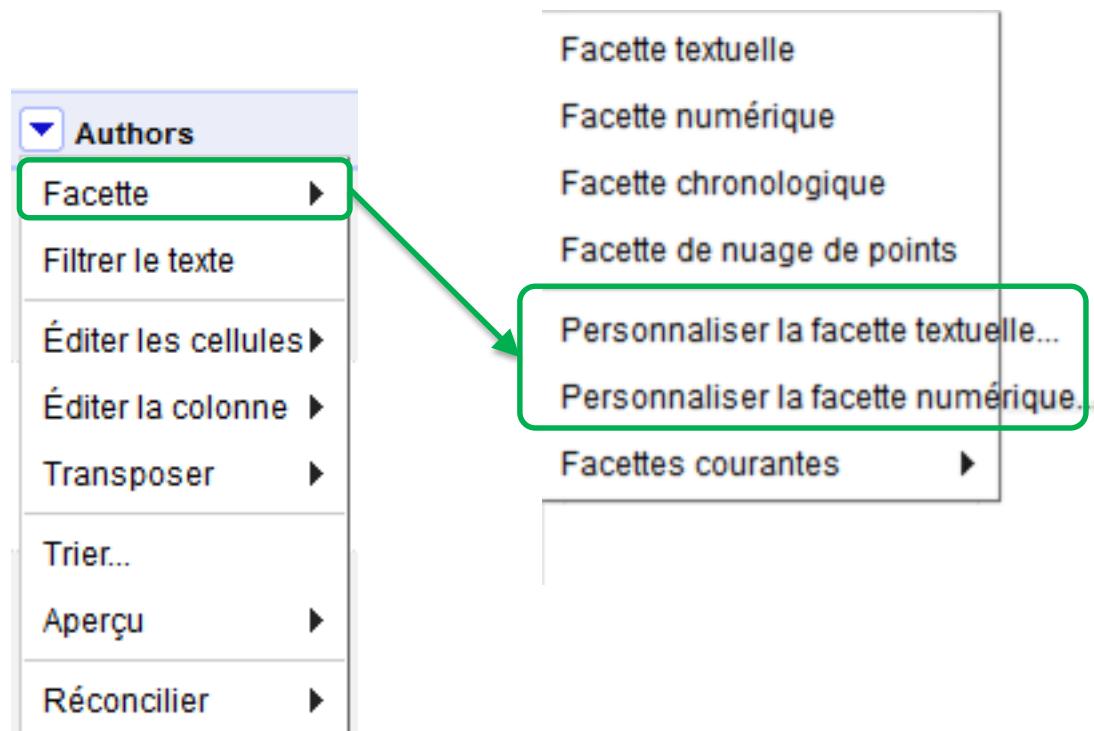
Utiliser les facettes: pour aller plus loin

Facettes numériques, chronologiques, en nuage de point : suppose d'avoir des données reconnues par OpenRefine comme des dates ou des nombres (pas le cas dans notre exemple)



Utiliser les facettes: pour aller plus loin

Facettes personnalisées : suppose une utilisation du langage GREL (voir plus loin)



Utiliser les facettes: pour aller plus loin

Facettes courantes: plusieurs options souvent utiles: par mot, par doublons, par longueur de texte, par blanc (valeur vide ou non)

The screenshot shows a software interface with a sidebar on the left containing a tree view of data structures. The 'Authors' node is expanded, revealing a context menu with the following items:

- Facette
- Filtrer le texte
- Éditer les cellules
- Éditer la colonne
- Transposer
- Trier...
- Aperçu
- Réconcilier

The 'Facette' item is highlighted with a green box and has a green arrow pointing to it from the main title. The 'Facette' menu itself contains the following options:

- Facette textuelle
- Facette numérique
- Facette chronologique
- Facette de nuage de points
- Personnaliser la facette textuelle...
- Personnaliser la facette numérique...
- Facettes courantes

The 'Facettes courantes' item is highlighted with a green box and has a green arrow pointing to it from the main title. A large green box surrounds the entire list of facet options on the right.

Facette par mot
Facette des doublons
Facette logarithmique
Facette logarithmique de limite 1
Facette de la longueur du texte
Facette logarithmique de la longueur du texte
Facette sur le code de caractère Unicode
Facette par erreur
Facette par blanc

Utiliser les facettes: pour aller plus loin

Activité : A partir de la colonne *animal_prefere*, appliquer une facette textuelle ordinaire, puis une facette « par mots ».

Quelle différence? Sur quels critères les mots ont-ils été isolés dans la facette par mots ?

Facette textuelle

animal_prefere changer

8 choices Trier par: nom compte Groupe

- CHAT 1
- chat, cheval, poisson 1
- chevaux 1
- chien 1
- chiens 1
- chiens et chats 1
- lapin 1
- Lapin,chien 1
- (blank) 2

Facette par nombre de choix

Facette par mot

animal_prefere changer

11 choices Trier par: nom compte

- CHAT 1
- chat, 1
- chats 1
- cheval, 1
- chevaux 1
- chien 1
- chiens 2
- et 1
- lapin 1
- Lapin,chien 1
- poisson 1

Tris, filtres et facettes

1. Introduction et présentation d'OpenRefine
2. Import des données et présentation de l'espace de travail
3. Tris, filtres et facettes
- 4. Regrouper des valeurs proches**
5. Transformations courantes des valeurs
6. Restructurer des données
7. Exporter les données et les traitements
8. Appliquer des transformations personnalisées

Regrouper des valeurs proches

Activité : Créer des facettes textuelles pour la colonne *ville*, puis grouper les résultats pour repérer des variantes d'orthographe ou de présentation.

Regrouper & éditer une colonne "ville"

Cet outil vous aide à identifier des groupes de cellules ayant des valeurs différentes mais qui peuvent correspondre à des représentations alternatives de la même valeur. Par exemple, les deux chaînes "New York" et "new york" n'ont qu'une différence de casse et font très certainement référence à la même ville. "Gödel" et "Godel" se réfèrent probablement à la même personne. [En trouver davantage...](#)

ville	changer
9 choices Trier par: nom compte	Groupe
Lyon 1	
CAEN 1	
Caen 1	
Havre (Le) 1	
Le Havre 1	
LE HAVRE 1	
Lyon 1	
Nice 1	
NICE 1	
(blank) 1	

Méthode collision de clés Fonction de codage empreinte 4 clusters trouvé

Taille du groupe	Nombre de lignes	Valeurs dans le groupe	Fusionner ?	Nouvelle valeur pour la cellule
3	3	<ul style="list-style-type: none">Havre (Le) (1 rows)LE HAVRE (1 rows)Le Havre (1 rows)	<input type="checkbox"/>	Havre (Le)
2	2	<ul style="list-style-type: none">CAEN (1 rows)Caen (1 rows)	<input type="checkbox"/>	CAEN
2	2	<ul style="list-style-type: none">NICE (1 rows)Nice (1 rows)	<input type="checkbox"/>	NICE
2	2	<ul style="list-style-type: none">Lyon (1 rows)Lyon (1 rows)	<input type="checkbox"/>	Lyon

Choix dans le groupe
Lignes dans le groupe
Longueur moyenne des choix
Variabilité moyenne des choix

Tout sélectionner Tout désélectionner Exporter les groupes Fusionner la sélection & regrouper Fusionner la sélection & fermer Fermer

Regrouper des valeurs proches

Activité : Créer des facettes textuelles pour la colonne *ville*, puis grouper les résultats pour repérer des variantes d'orthographe ou de présentation.

Plusieurs options correspondant à différents algorithmes. À tester en fonction de ses données. **Attention!** ces algorithmes peuvent faire des rapprochements non pertinents.

Résultat : on passe de 9 villes à 4 villes

ville

4 choices Trier par: nom compte Groupe

Caen 2
Le Havre 3
Lyon 2
Nice 2
(blank) 1

Facette par nombre de choix

Transformations courantes

1. Introduction et présentation d'OpenRefine
2. Import des données et présentation de l'espace de travail
3. Tris, filtres et facettes
4. Regrouper des valeurs proches
5. **Transformations courantes des valeurs**
6. Restructurer des données
7. Exporter les données et les traitements
8. Appliquer des transformations personnalisées

Appliquer des transformations courantes

Modifier la casse

Activité: Transformer les valeurs de la colonne *ville* pour obtenir passer tous les noms en majuscules



ville
Nice
Caen
Lyon
Nice
Le Havre
Le Havre
Lyon
Caen
Le Havre

ville
NICE
CAEN
LYON
NICE
LE HAVRE
LE HAVRE
LYON
CAEN
LE HAVRE

Appliquer des transformations courantes

Modifier la casse

Editer les cellules > Transformations courantes > En majuscules

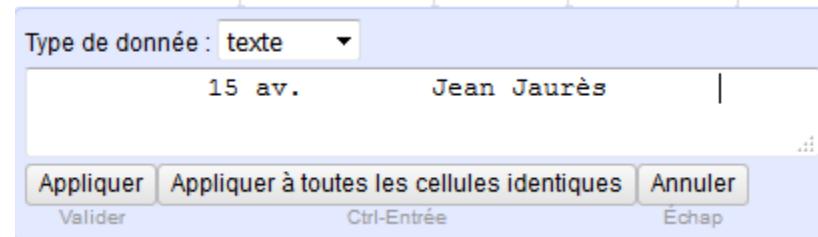


Appliquer des transformations courantes

Supprimer les espaces superflus

Activité: Ajouter manuellement plusieurs espaces au début et à l'intérieur d'une cellule, puis supprimer ces espaces en utilisant le menu.

1/ ajout des espaces



(les espaces ne seront pas visibles dans l'affichage général des données, mais ils ont bien été ajoutés)

Appliquer des transformations courantes

Supprimer les espaces superflus

2/ suppression

Deux opérations

Supprimer les espaces de début et de fin



Text transform on 5 cells in column Title: value.trim()

Rassembler les espaces consécutifs

Text transform on 3 cells in column Title:
value.replace(/\s+/,' ') Défaire

Appliquer des transformations courantes

Transformer du texte en nombres ou en dates

Malgré les apparence, ces nombres et ces dates sont considérés comme de simple suite de caractères.

Mais l'hétérogénéité des données peut rendre leur reconnaissance délicate.

JJ/MM/2017	Date	habillement	loisirs	logement	Unité en k€
	01/02/2017	100	25	0,8	
	01/03/2017	50.50	35,6	0,7	
	15/02/2017	10.90	70,6	700	
	15-02-2017	400	90	600	Unité en €
	15-04-2017				
	12-02-2017				
	11/01/2017				
JJ-MM-2017					
JJ / MM (2017)					
JJ:MM:2017					
	Séparateur décimal .				
		50.50	35,6	0,7	
		200	40	800	Séparateur décimal ,



Appliquer des transformations courantes

Transformer du texte en nombres ou en dates

Editer les cellules > Transformations courantes > En nombre / En date



Appliquer des transformations courantes

Transformer du texte en nombres ou en dates

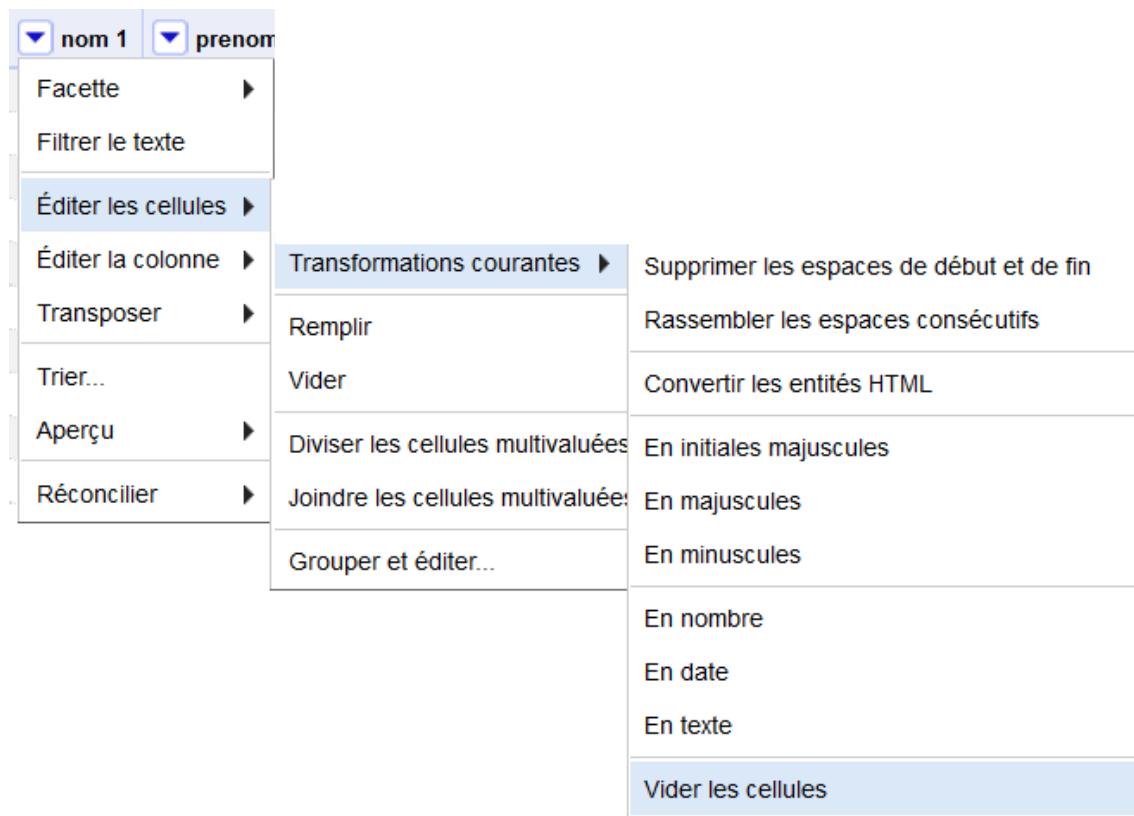
Cette approche ne suffit pas à résoudre les cas complexes
(structure atypique, variation d'unités, séparateur décimal ,)
→ besoin d'utiliser le langage GREL

Date	habillement	loisirs	logement
2017-01-02T00:00:00Z	100	25	0,8
2017-01-03T00:00:00Z	50,5	35,6	0,7
2017-02-15T00:00:00Z	10,9	70,6	700
2017-02-15T00:00:00Z	400	90	600
2017-04-15T00:00:00Z			
2017-12-02T00:00:00Z			
2017-11-01T00:00:00Z			
19/02 (2017)			
16/03 (2017)	50,5	35,6	0,7
08:01:2017	200	40	800

Appliquer des transformations courantes

Vider des cellules d'une colonne

Si des lignes ont été sélectionnées avec une facette, ne s'applique qu'à la sélection



Recopier ou supprimer des valeurs

Créer un nouveau projet à partir du fichier **exo2.csv**

		Toutes	Ville	espece	nombre
		1.	Nice	palmiers	400
		2.		orangers	200
		3.		bouleau	10
		4.	Marseille	palmiers	200
		5.		orangers	50
		6.		bouleau	10
		7.	Paris	palmiers	10
		8.		orangers	10
		9.		bouleau	100

Recopier ou supprimer des valeurs

Dans la colonne *ville*, recopier automatiquement le nom de chaque ville dans les cellules vides.

Editer les cellules > Remplir

The screenshot shows a spreadsheet interface with a vertical column labeled "Ville" on the left. The first three cells contain "Nice", "Marseille", and "Paris" respectively. The fourth cell is empty. A red arrow points from the top-left towards this empty cell. A context menu is open over the empty cell, listing options numbered 1 through 9. Option 4, "Éditer les cellules" (Edit cells), is highlighted with a blue background. Other options include "Facette", "Filtrer le texte", "Transformer...", "Transformations courantes", "Remplir" (which is also highlighted with a blue background), "Vider", "Diviser les cellules multivaluées...", "Joindre les cellules multivaluées...", and "Réconcilier". To the right of the menu, the rest of the "Ville" column is visible, showing the values "Nice", "Nice", "Nice", "Marseille", "Marseille", "Marseille", "Paris", "Paris", and "Paris".

Recopier ou supprimer des valeurs

Opération inverse : supprimer les valeurs répétées

Editer les cellules > Vider

The screenshot shows a spreadsheet interface with three columns: Ville, espèce, and nombre. The Ville column contains the following data: Nice, Nice, Nice, Marseille, Marseille, Marseille, Paris, Paris, Paris. A context menu is open over the first 'Nice' cell in the Ville column. The menu options are: Facette, Filtrer le texte, Éditer les cellules (which is selected), Éditer la colonne, Transposer, Trier..., Aperçu, Réconcilier, Transformer..., Transformations courantes, Remplir, Vider (which is highlighted in blue), Diviser les cellules multivaluées, Joindre les cellules multivaluées, and Grouper et éditer... On the far left, there is a vertical list labeled 'Ville' with the same data points. On the far right, there is a vertical list also labeled 'Ville' with the same data points.

Recopier ou supprimer des valeurs

La suppression des valeurs répétées peut être nécessaire pour d'autres opérations :

- Supprimer des doublons (suppression de la valeur répétée > facette sur la colonne> suppression des lignes)
- Regrouper dans une seule cellule des valeurs présentes dans des lignes successives (suppose de créer des « entrées »)

Lignes et entrées

Des lignes peuvent être regroupées en « entrées » (*records*) si elles se rapportent à un même objet.

Travailler avec des entrées permet des traitements avancés.

Pour créer des entrées :

- 1/ trier les données en fonction de la colonne servant de clé de regroupement
- 2/ déplacer cette colonne en 1^{re} position du tableau
- 3/ supprimer les valeurs répétées dans cette colonne

3 entrées

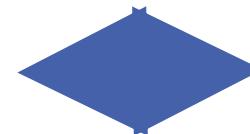
Voir en: [lignes](#) [entrées](#)

s	Ville	espece	nombre
1.	Nice	palmiers	400
		orangers	200
		bouleau	10
2.	Marseille	palmiers	200
		orangers	50
		bouleau	10
3.	Paris	palmiers	10
		orangers	10
		bouleau	100

9 lignes

Voir en: [lignes](#) [entrées](#)

s	Ville	espece	nombre
1.	Nice	palmiers	400
2.		orangers	200
3.		bouleau	10
4.	Marseille	palmiers	200
5.		orangers	50
6.		bouleau	10
7.	Paris	palmiers	10
8.		orangers	10
9.		bouleau	100



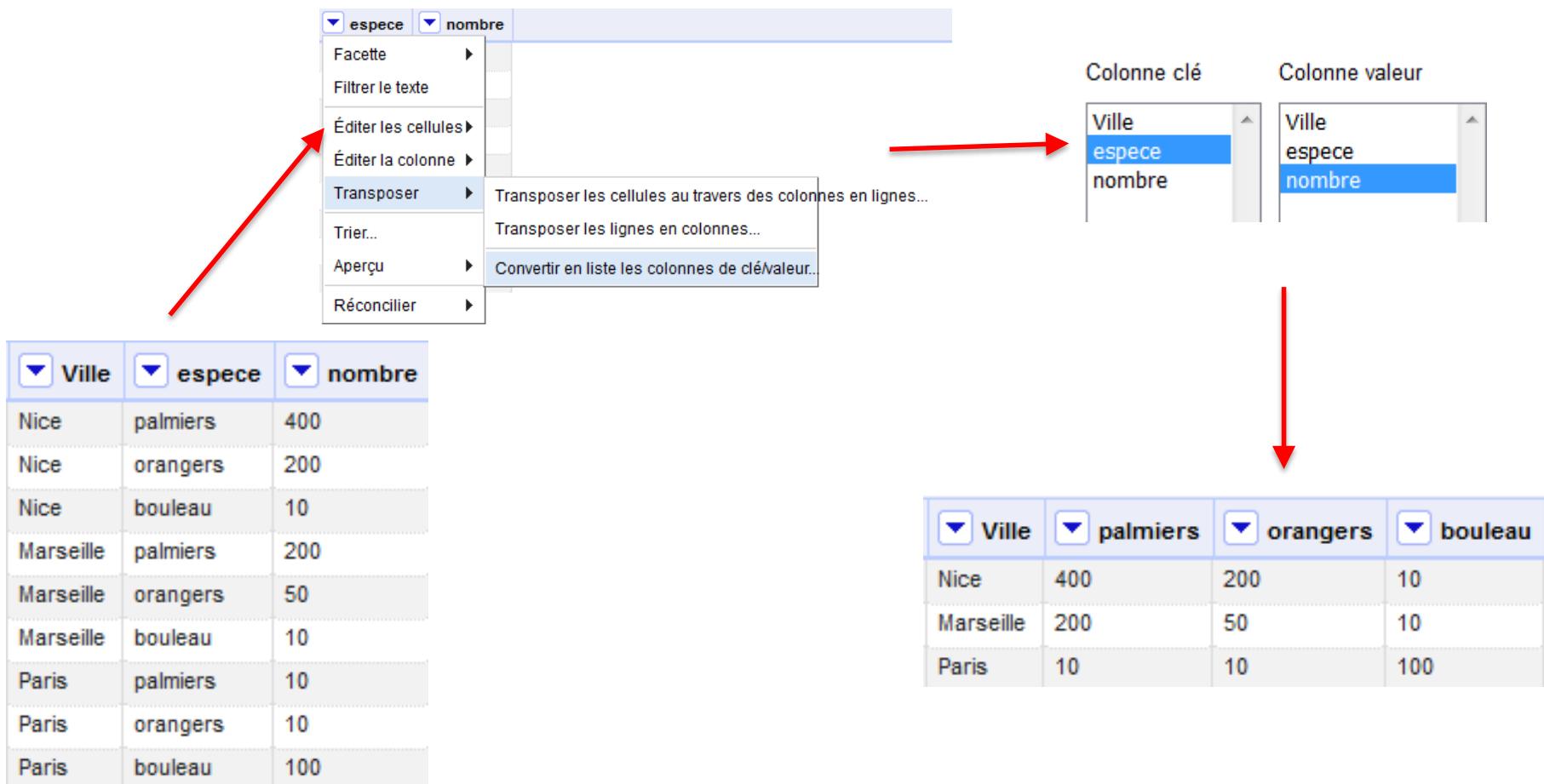
Transformations courantes

1. Introduction et présentation d'OpenRefine
2. Import des données et présentation de l'espace de travail
3. Tris, filtres et facettes
4. Regrouper des valeurs proches
5. Transformations courantes des valeurs
6. **Restructurer des données**
7. Exporter les données et les traitements
8. Appliquer des transformations personnalisées

Restructurer les données

Passer du format « long » au format « wide »

Transposer > Convertir en liste les colonnes de clé/valeur



Restructurer les données

Passer du format « wide » au format « long »

Transposer > Transposer les cellules au travers des colonnes en lignes

The screenshot shows a spreadsheet interface with three main sections:

- Left Panel:** A sidebar with dropdown menus for "palmiers", "orangers", and "bouleau".
- Middle Panel:** A context menu for the selected cell containing:
 - "Transposer" (highlighted in blue) with the sub-option "Transposer les cellules au travers des colonnes en lignes...".
 - "Trier..."
 - "Aperçu"
 - "Réconcilier"
- Right Panel:** A configuration panel for "Vers la colonne" (To Column) with two options:
 - Deux nouvelles colonnes
 - Colonne clé (key column)
 - Colonne valeur (value column)

Two red arrows point from the "Transposer" menu option to the resulting tables below. One arrow points to the "De la colonne" table, and another points to the "Vers la colonne" table.

Ville	palmiers	orangers	bouleau
Nice	400	200	10
Marseille	200	50	10
Paris	10	10	100

Ville	espece	nombre
Nice	palmiers	400
Nice	orangers	200
Nice	bouleau	10
Marseille	palmiers	200
Marseille	orangers	50
Marseille	bouleau	10
Paris	palmiers	10
Paris	orangers	10
Paris	bouleau	100

Ville	palmiers	orangers	bouleau
Nice	400	200	10
Marseille	200	50	10
Paris	10	10	100

Ville	espece	nombre
Nice	palmiers	400
Nice	orangers	200
Nice	bouleau	10
Marseille	palmiers	200
Marseille	orangers	50
Marseille	bouleau	10
Paris	palmiers	10
Paris	orangers	10
Paris	bouleau	100



Restructurer les données

Regrouper les paires clés/valeurs dans les mêmes cellules

Transposer > Transposer les cellules au travers des colonnes en lignes



Ville	palmiers	orangers	bouleau
Nice	400	200	10
Marseille	200	50	10
Paris	10	10	100

Toutes	Ville	espece_nombre
1.	Nice	palmiers:400
2.		orangers:200
3.		bouleau:10
4.	Marseille	palmiers:200
5.		orangers:50
6.		bouleau:10
7.	Paris	palmiers:10
8.		orangers:10
9.		bouleau:100

Restructurer les données

Éclater une colonne en plusieurs colonnes

Éditer la colonne > Diviser en plusieurs colonnes

The screenshot illustrates the process of splitting a single column into multiple columns in a spreadsheet application.

Initial State: A table with three columns: 'es', 'Ville', and 'espece_nombre'. The 'espece_nombre' column contains entries like 'palmiers:400', 'orangers:200', etc.

Context Menu: A context menu is open over the 'espece_nombre' column, with the 'Diviser en plusieurs colonnes...' option highlighted by a red arrow.

Division Dialog: The 'Diviser la colonne espece_nombre en plusieurs colonnes' dialog is shown. It has two methods:

- par séparateur**: The 'Séparateur' field is set to ':', indicated by a red box and arrow. The 'expression rationnelle' checkbox is unchecked.
- par les longueurs de champs**: This method is also available but not selected.

Checkboxes for 'Deviner le type de cellule' and 'Supprimer cette colonne' are present on the right.

Resulting Table: The table after division shows the results. The 'espece_nombre' column has been split into two columns: 'espece' (containing 'palmiers', 'orangers', 'bouleau') and 'nombre' (containing '400', '200', '10').

es	Ville	espece_nombre
1.	Nice	palmiers:400
2.		orangers:200
3.		bouleau:10
4.	Marseille	palmiers:200
5.		orangers:50
6.		bouleau:10
7.	Paris	palmiers:10
8.		orangers:10
9.		bouleau:100

es	Ville	espece	nombre
1.	Nice	palmiers	400
2.		orangers	200
3.		bouleau	10
4.	Marseille	palmiers	200
5.		orangers	50
6.		bouleau	10
7.	Paris	palmiers	10
8.		orangers	10
9.		bouleau	100

Restructurer les données

Regrouper les valeurs d'une colonne sur une seule ligne par entrée

1. Organiser le tableau en « entrées » (cf. plus haut)
2. Editer les cellules > Joindre les cellules multivaluées
3. Choisir un séparateur

	Ville	espece_nombr	espece_nombr
1.	Nice	palmiers	400
2.		orangers	200
3.		bouleau	10
4.	Marseille	palmiers	200
5.		orangers	50
6.		bouleau	10
7.	Paris	palmiers	10
8.		orangers	10
9.		bouleau	100

Répéter pour les 2 colonnes

A screenshot of a data editing interface. A context menu is open over a cell containing three values: 400, 200, and 10. The menu options include Facette, Filtrer le texte, Éditer les cellules, Éditer la colonne, Transposer, Trier..., Aperçu, Réconcilier, Transformer..., Transformations courantes, Remplir, Vider, Diviser les cellules multivaluées, Joindre les cellules multivaluées, and Grouper et éditer... A red arrow points from the text "Répéter pour les 2 colonnes" to the "Éditer les cellules" option. Another red arrow points from a text box labeled "Indiquer le séparateur à utiliser entre les valeurs" to the "Joindre les cellules multivaluées" option. The text box contains a vertical bar separator (|).

	Ville	espece_nombr 1	espece_nombr
1.	Nice	palmiers orangers bouleau	400 200 10
2.	Marseille	palmiers orangers bouleau	200 50 10
3.	Paris	palmiers orangers bouleau	10 10 100

Restructurer les données

Opération inverse : éclater une colonne sur plusieurs lignes

1. Editer les cellules > Diviser les cellules multivaluées
2. Choisir le séparateur

Répéter pour les 2 colonnes

	Ville	espece_nombre 1	espece_nombre
1.	Nice	palmiers orangers bouleau	400 200 10
2.	Marseille	palmiers orangers bouleau	200 50 10
3.	Paris	palmiers orangers bouleau	10 10 100

A screenshot of a data editing interface. On the left, a table has three rows with data in the 'espece_nombre' column: '400, 200, 10', '200, 50, 10', and '10, 10, 100'. A context menu is open over the first row, with 'Diviser les cellules multivaluées...' highlighted. A red arrow points from the text 'Répéter pour les 2 colonnes' to this menu item. To the right, a tooltip window asks 'Quel séparateur sépare actuellement les valeurs?' with a placeholder '|'. Two red arrows point from the bottom right towards this tooltip and the menu item.

	Ville	espece_nombre	espece_nombre
1.	Nice	palmiers	400
2.		orangers	200
3.		bouleau	10
4.	Marseille	palmiers	200
5.		orangers	50
6.		bouleau	10
7.	Paris	palmiers	10
8.		orangers	10
9.		bouleau	100

Restructurer les données



Rouvrir le projet exo1

[Créer un projet](#)[Ouvrir un projet](#)[Importer un projet](#)[Langue](#)

	Toutes	code_personne	date	ville	adresse	animal_prefere	habillement	loisirs	logement
1.	P001	01/02/2017	NICE	1 av. St Barthélemy	chien	100	25	0,8	
2.	P002	01/03/2017	CAEN		chiens				
9.	P002	16/03 (2017)	Caen	5 rue Basse		50.50	35,6	0,7	
3.	P003	15/02/2017	Lyon	3 rue Paul Bert	chiens et chats	10.90	70,6	700	
4.	P004	15-02-2017	Nice	50 avenue Saint Barthélemy	chat, cheval, poisson	400	90	600	
5.	P005	15-04-2017	LE HAVRE	15 av. Jean Jaurès	CHAT				
6.	P005	12-02-2017	Havre (Le)	15 av. Jean Jaurès	chevaux				
7.	P005	11/01/2017			lapin				
10.	P005	08:01:2017	Le Havre	15 av. Jean Jaurès	Lapin, chien	200	40	800	
8.	P006	19/02 (2017)	Lyon	1 rue Dunoir					

Restructurer les données

Eclater toutes les valeurs de la colonne animal_prefere dans des lignes distinctes

Quel séparateur sépare actuellement les valeurs ?

,



Quel séparateur sépare actuellement les valeurs ?

et



chien
chiens
chiens et chats
chat, cheval, poisson
CHAT
chevaux
lapin
Lapin, chien

chien
chiens
chiens et chats
chat
cheval
poisson
CHAT
chevaux
lapin
Lapin
chien

Restructurer les données

Créer une facette sur la colonne *animal_prefere* et regrouper les valeurs proches

Trouver un paramétrage détectant le plus de doublons

Méthode plus proche voisin ▾ Fonction distance : Levenshtein ▾ Rayon 2 Bloc de caractères 2

Taille du groupe	Nombre de lignes	Valeurs dans le groupe	Fusionner ?	Nouvelle valeur pour la cellule
3	3	<ul style="list-style-type: none">chat (1 rows)CHAT (1 rows)chats (1 rows)	<input type="checkbox"/>	chat
2	2	<ul style="list-style-type: none">chat (1 rows)chats (1 rows)	<input type="checkbox"/>	chat
2	2	<ul style="list-style-type: none">Lapin (1 rows)lapin (1 rows)	<input type="checkbox"/>	Lapin
2	4	<ul style="list-style-type: none">chiens (2 rows)chien (2 rows)	<input type="checkbox"/>	chiens
2	2	<ul style="list-style-type: none">cheval (1 rows)chevaux (1 rows)	<input type="checkbox"/>	cheval

Choix dans le groupe

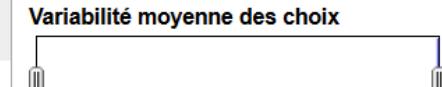
2 — 3

Lignes dans le groupe

2 — 4

Longueur moyenne des choix

4.33 — 6.5

Variabilité moyenne des choix

0 — 0.5

Restructurer les données

Rejoindre les valeurs de la colonne dans une seule ligne par entrée

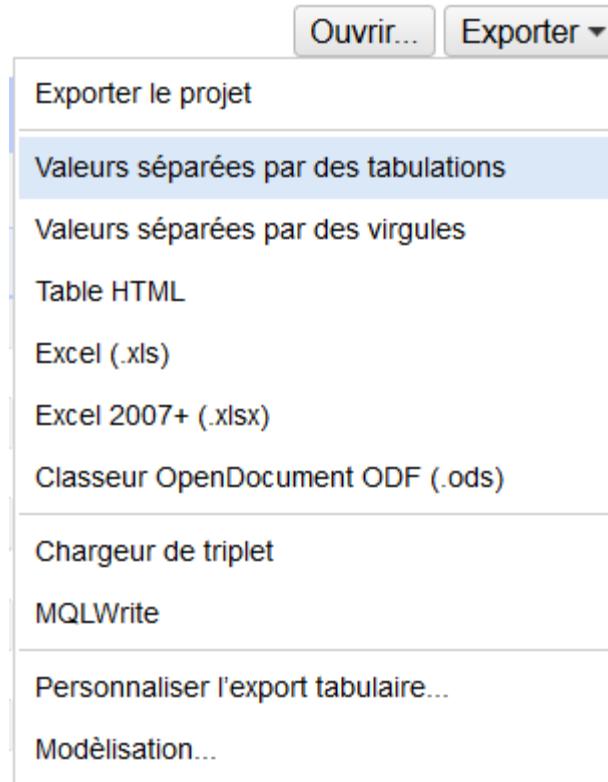
chien
chien
chien chat
chat cheval poisson
chat
cheval
lapin
lapin chien

Exporter données et traitements

1. Introduction et présentation d'OpenRefine
2. Import des données et présentation de l'espace de travail
3. Tris, filtres et facettes
4. Regrouper des valeurs proches
5. Transformations courantes des valeurs
6. **Restructurer des données**
7. Exporter les données et les traitements
8. Appliquer des transformations personnalisées

Exporter les données transformées

Plusieurs formats d'export



Annuler ou rejouer un traitement

Historique permettant d'annuler (« défaire ») ou rejouer (« refaire) les traitements sans limites

Facette / Filtre Défaire / Refaire 38

Extraire... Appliquer...

Filtrer :

18. Mass edit 107 cells in column Language
19. Remove 2973 rows
20. Split multi-valued cells in column Subjects
21. Text transform on 0 cells in column Subjects: value.trim()
22. Mass edit 1311 cells in column Subjects
23. Mass edit 224 cells in column Subjects
24. Mass edit 27 cells in column Subjects
25. Mass edit 120 cells in column Subjects
26. Mass edit 861 cells in column Subjects
27. Mass edit 35 cells in column Subjects
28. Mass edit 38 cells in column Subjects
29. Join multi-valued cells in column Subjects
30. Split multi-valued cells in column Subjects
31. Fill down 6280 cells in column Title
32. Fill down 6280 cells in column URL
33. Split 1001 cell(s) in column Citation into several columns by separator
34. Fill down 6303 cells in column DOI
35. Blank down 6280 cells in column URL
36. Fill down 6280 cells in column URL
37. Reorder columns

Exporter les traitements

Les traitements peuvent être exportés et réappliqués au jeu de données ou à un autre jeu présentant la même structure.

Extraire l'historique pour enregistrer les traitements :

1

Extrire... Appliquer...

2

Fermer

3

Copier dans le presse-papier (Ctrl+C / Cmd+C)

4

Créer un fichier texte sur l'ordinateur
Ouvrir avec un éditeur de texte
Coller le contenu du presse papier (Ctrl+V / Cmd+V)
Enregistrer le fichier

Extrire des opérations de l'historique

Extrire et enregistrer des sous-parties de l'historique des opérations au format JSON afin de les réappliquer dans ce projet ou de les réutiliser ultérieurement dans d'autres projets.

18. Mass edit 107 cells in column Language
19. Remove 2973 rows
20. Split multi-valued cells in column Subjects
21. Text transform on 0 cells in column Subjects: value.trim()
22. Mass edit 1311 cells in column Subjects
23. Mass edit 224 cells in column Subjects
24. Mass edit 27 cells in column Subjects
25. Mass edit 120 cells in column Subjects
26. Mass edit 861 cells in column Subjects
27. Mass edit 35 cells in column Subjects
28. Mass edit 38 cells in column Subjects
29. Join multi-valued cells in column Subjects
30. Split multi-valued cells in column Subjects
31. Fill down 6280 cells in column Title
32. Fill down 6280 cells in column URL
33. Split 1001 cell(s) in column Citation into several columns by separator
34. Fill down 6303 cells in column DOI
35. Blank down 6280 cells in column URL
36. Fill down 6280 cells in column URL
37. Reorder columns

[] Split multi-valued cells in column Authors
[] Mass edit cells in column Authors
[] Create column nom_famille at index 2 based on column Authors using expression grel:value.split(" ")[length(value.split(" "))-1]
[] Create column prenoms at index 2 based on column Authors using expression grel:value.split(" ").slice(0,length(value.split(" "))-1).join(" ")
[] Text transform on cells in column Authors using expression grel:value.split(".").reverse().join("")
[] Blank down cells in column prenoms
[] Remove column prenoms
[] Remove column nom_famille
[] Create column nom at index 2 based on column Author s using expression grel:value.split(" ")[length(value.split(" "))-1]

[] {
[] "op": "core/fill-down",
[] "description": "Fill down cells in column",
[] "engineConfig": {
[] },
[] "mode": "row-based",
[] "facets": []
[] },
[] {
[] "columnName": "URL"
[] },
[] {
[] "op": "core/column-reorder",
[] "description": "Reorder columns",
[] "columnNames": [
[] "DOI",
[] "URL",
[] "Subjects"
[]],
[] },
[] {
[] "op": "core/column-reorder",
[] "description": "Reorder columns",
[] "columnNames": [
[] "URL",
[] "DOI",
[] "Subjects"
[]]
[] }

Tout sélectionner Tout désélectionner

Réappliquer les traitements

À partir d'un jeu de données fraîchement téléchargé

The diagram illustrates a four-step process for reapplying treatments from a JSON file in OpenRefine:

- 1**: On the left, the OpenRefine interface shows the "Historique d'annulation infini" (Infinite cancellation history) panel. A red arrow points from this panel to the "Appliquer..." button (step 1).
- 2**: Below the history panel, the text "Ouvrir le fichier texte dans lequel les traitements ont été enregistrés" (Open the text file where the treatments were recorded) is displayed.
- 3**: On the right, the "Appliquer la liste des opérations" (Apply the list of operations) dialog box is shown. It contains the instruction "Coller dans OpenRefine (Ctrl+V/ Cmd+V)".
- 4**: Below the dialog box, the text "Sélectionner tout le contenu et copier dans le presse-papier (Ctrl+C / Cmd+C)" (Select all content and copy to the clipboard (Ctrl+C / Cmd+C)) is displayed.

Appliquer des transformations personnalisées

1. Introduction et présentation d'OpenRefine
2. Import des données et présentation de l'espace de travail
3. Tris, filtres et facettes
4. Regrouper des valeurs proches
5. Transformations courantes des valeurs
6. Restructurer des données
7. Exporter les données et les traitements
8. **Appliquer des transformations personnalisées**

Appliquer des transformations personnalisées

Pendant quelques secondes, une information s'affiche après une modification des données réalisée via le menu:

Ex:

Text transform on 5 cells in column Title: value.trim()

Text transform on 3 cells in column Title:
value.replace(/\s+/, ' ') Défaire

Elle indique la **formule** utilisée par OpenRefine.

Ici :

`value.trim()` supprime les espaces initiaux et finaux

`value.replace(/\s+/, '')` simplifie les espaces répétés

Ces formules se retrouvent aussi dans l'historique des traitements.

Appliquer des transformations personnalisées

Les transformations personnalisées reposent sur des formules de ce type, saisies manuellement.

Elles utilisent le langage **GREL** (*Google Refine Expression Language*, ou *General Refine Expression Language*).

Documentation :

<https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language>

Appliquer des transformations personnalisées

Dans la colonne *loisirs*, ouvrir le menu Editer les cellules > Transformer

Personnaliser la transformation du texte sur la colonne loisirs

Formule (sans = initial) → Expression Langue General Refine Expression Language (GREL) Pas d'erreur de syntaxe.

value

Aperçu	Historique	Étoilée	Aide
row	value	value	
1.	25	25	↑
2.			↓
3.	35,6	35,6	☰
4.	70,6	70,6	
5.	90	90	
6.	null	null	

Résultat de la formule

En cas d'erreur conserver l'original Retransformer 10 fois maximum, tant que les données changent
 vider la cellule
 conserver l'erreur

OK Annuler

Appliquer des transformations personnalisées

Deux menus utiles: aide et historique

Aide précieuse!

Utiliser Crtrl+F pour rechercher une commande

The screenshot shows a software window with a menu bar at the top. The 'Aide' (Help) tab is highlighted with a red box. Below the menu, there is a section titled 'Variables' containing the following definitions:

- cell**: La cellule en cours. Elle a deux champs : 'value' et 'recon'.
- value**: La valeur de la cellule en cours. C'est un alias de 'cell.value'.
- row**: La ligne en cours. Elle a cinq champs : 'flagged', 'starred', 'index', 'cells' et 'record'.
- cells**: Les cellules de la ligne en cours. C'est un alias de 'row.cells'. Une cellule spécifique peut être retrouvée avec 'cells.<nom de la colonne>' lorsque <nom de la colonne > est un mot simple et avec 'cells.[<nom de la colonne>]' dans les autres cas.
- rowIndex**: L'index de la ligne en cours. C'est un alias de 'row.index'.

Appliquer des transformations personnalisées

Deux menus utiles: aide et historique

Historique

Permet de réutiliser une formule

De	Expression
Réutiliser This project	grel: value.trim()
Réutiliser This project	grel: value.substring(value.indexOf('(')).replaceChars('(', ')')
Réutiliser This project	grel: value.split('(')
Réutiliser This project	grel: value.split('/')
Réutiliser Other projects	grel: substring("a mmm",0).trim()
Réutiliser Other projects	grel: value.parseHtml().select("div.intranet a").first().attr("href")

Appliquer des transformations personnalisées

Syntaxe générale

- Pas de = avant les fonctions
- + permet de concaténer deux valeurs. Ex: "a"+ "b" -> "ab"
- Nom des fonctions sensible aux majuscules
- *value* désigne le contenu d'une cellule
- Les fonctions sont suivies de (parametre1, parametre2...), ou de () en absence de paramètre
- Une fonction peut s'écrire de deux manières :
 - *value.nom_de_la_fonction(parametres)*. Ex: *value.trim ()*
 - ou
 - *nom_de_la_fonction(value, parametres)*. Ex: *trim (value)*

Appliquer des transformations personnalisées

Exemple de fonctions utiles

length () longueur de la chaîne de caractère

trim () supprime les espaces initiaux et finaux

toUpperCase() passe en majuscules (y compris lettres accentuées)

Ex: "école".*toUpperCase* () -> « ÉCOLE »

toLowerCase () passe en minuscules

indexOf (x) renvoie la position de x

(Attention, la numérotation commence à 0)

Ex: "bleu".*indexOf* ('b') -> 0

substring (pos1, pos2) extrait les caractères entre pos1 et pos2

(Attention, la numérotation commence à 0
et pos2 est exclu)

Ex: "bleu".*substring* (1, 3) -> « le » (lettres de position 1 et 2)

Appliquer des transformations personnalisées

Exemple de fonctions utiles

replace (x, y) remplace la chaîne de caractère x par y.

Ex : "zorro".replace ('zo', 'x') -> "xrro"

replaceChars (x,y) replace les caractères contenus dans x par z

Ex : "zorro".replaceChars ('zo', 'x') -> "xrx"

split (x) décompose la valeur en tableau, en utilisant x comme séparateur. Les éléments du tableau sont accessibles par [n] (numérotation à partir de 0)

Ex : "01/12/2015".split ('/') [0] -> « 01 » (1^{er} élément du tableau)

Ex : "01/12/2015".split ('/') [1] -> « 12 » (2^{er} élément du tableau)

Ex : "01/12/2015".split ('/') [2] -> « 2015 » (3^{er} élément du tableau)

join (t, s) inverse de split : agrège les éléments d'un tableau t en utilisant comme séparateur la chaîne de caractères s.

Appliquer des transformations personnalisées

Exemple de fonctions utiles

cross (cell c, Nom_projet2, Nom_colonne) permet de croiser deux projets : retourne un tableau de 0, 1 ou + lignes du projet Nom_projet2 pour lesquelles les cellules de la colonne Nom_colonne ont le même contenu que la cellule c.

parseJson (s) : analyse la chaîne s et renvoie un objet manipulable

Ex : value.parseJson () ["element-niv1"] ["element-niv2"]

parseHtml (s) : analyse la chaîne s et renvoie un objet manipulable avec d'autres fonctions

Ex:

value.parseHtml () .select ("div#content") [0] .select ("tr") .toString ()

Appliquer des transformations personnalisées

Activité : dans la colonne *loisirs*, appliquer la formule
value.replace(",",".")

Personnaliser la transformation du texte sur la colonne loisirs

Expression Langue General Refine Expression Language (GREL) Pas d'erreur de syntaxe.

Aperçu Historique Étoilée Aide

row	value	value.replace(',', '.')
1.	25	25
2.	null	Erreur: replace expects 3 strings, or 1 string, 1 regex, and 1 string
3.	35,6	35.6
4.	70,6	70.6
5.	90	90
6.	null	Erreur: replace expects 3 strings, or 1 string, 1 regex, and 1 string

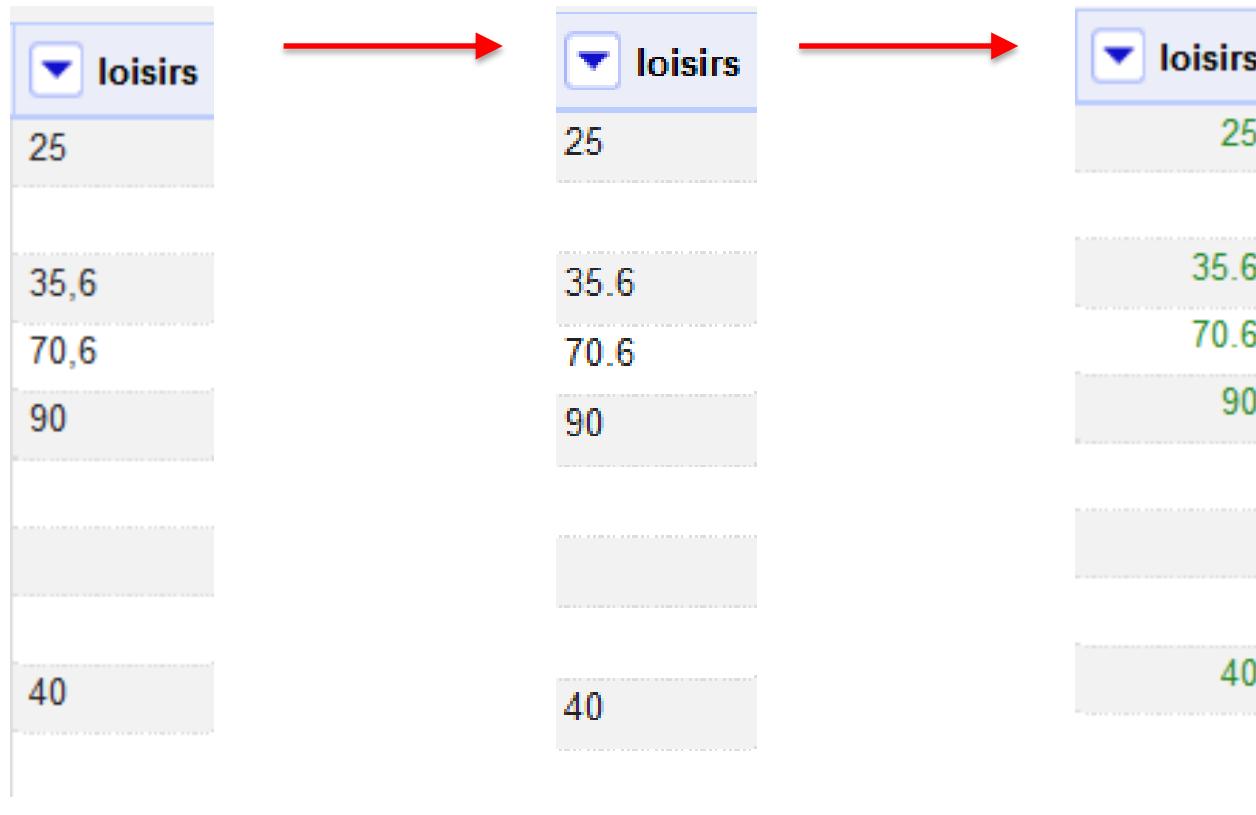
En cas d'erreur conserver l'original Retransformer 10 fois maximum, tant que les données changent
 vider la cellule
 conserver l'erreur

OK Annuler

Prévisualisation du résultat

Appliquer des transformations personnalisées

Les valeurs peuvent maintenant être interprétées comme des nombres par OpenRefine



Appliquer des transformations personnalisées

Activité : dans la colonne *date*, appliquer la formule

```
value.match(/.*(\d{4}).*/)
```

(Expression régulière :
suite de caractères + 4
chiffres + suite de
caractères ; capturer les
4 chiffres)

Personnaliser la transformation du texte sur la colonne date

Expression Langue General Refine Expression Language (GREL)
`value.match(/.*(\d{4}).*/)` Pas d'erreur de syntaxe.

Aperçu Historique Étoilée Aide

row	value	value.match(/.*(\d{4}).*/)
1.	01/02/2017	["2017"]
2.	01/03/2017	["2017"]
3.	16/03 (2017)	["2017"]
4.	15/02/2017	["2017"]
5.	15-02-2017	["2017"]
6.	15-04-2017	["2017"]

En cas d'erreur conserver l'original Retransformer 10 fois maximum, tant que les données changent
 vider la cellule
 conserver l'erreur

OK Annuler

Prévisualisation du résultat

Rien ne se passe !

["2017"] est un tableau à une colonne. Openrefine ne peut pas l'afficher tel quel.

Appliquer des transformations personnalisées

Activité : dans la colonne *date*, appliquer la formule
value.match(/.*(\d{4}).*/)[0]

Personnaliser la transformation du texte sur la colonne date

Expression : value.match(/.*(\d{4}).*/)

Langue : General Refine Expression Language (GREL)

Prévisualisation du résultat :

row	value	value.match(/.*(\d{4}).*/)
1.	01/02/2017	["2017"]
2.	01/03/2017	["2017"]
3.	16/03 (2017)	["2017"]
4.	15/02/2017	["2017"]
5.	15-02-2017	["2017"]
6.	15-04-2017	["2017"]

En cas d'erreur :

- conserver l'original
- Retransformer 10 fois maximum, tant que les données changent
- vider la cellule
- conserver l'erreur

OK Annuler

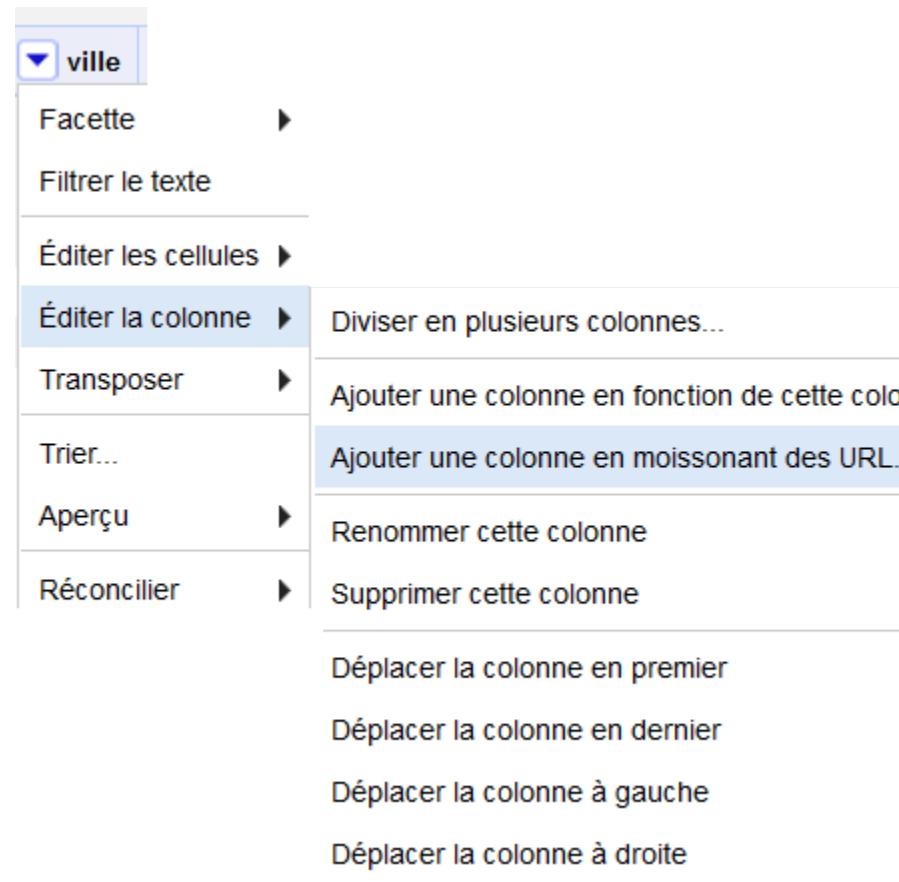
row	value	value.match(/.*(\d{4}).*/)
1.	01/02/2017	["2017"]
2.	01/03/2017	["2017"]
3.	16/03 (2017)	["2017"]
4.	15/02/2017	["2017"]
5.	15-02-2017	["2017"]
6.	15-04-2017	["2017"]

Cette fois « 2017 » est copié dans la colonne [0] permet d'accéder au 1^{er} élément du tableau



Récupérer des données sur le web

Activité : créer une nouvelle colonne récupérant des données depuis l'API <https://geo.api.gouv.fr/> et le nom de la chaque ville



Récupérer des données sur le web

Formule :

"<https://geo.api.gouv.fr/communes?nom=value>
50 millisecondes de délai

Ajouter une colonne en moissonnant les données depuis les URL d'une colonne ville

Nouveau nom de colonne geo_api Délai de récupération 50 millisecondes

En cas d'erreur vider la cellule conserver l'erreur

Indiquer les URL à moissonner :

Expression Langue General Refine Expression Language (GREL)

```
"https://geo.api.gouv.fr/communes?nom="+value
```

Pas d'erreur de syntaxe.

Aperçu	Historique	Étoilée	Aide
row value 1. Nice https://geo.api.gouv.fr/communes?nom=Nice 2. Caen https://geo.api.gouv.fr/communes?nom=Caen 3. Caen https://geo.api.gouv.fr/communes?nom=Caen 4. Lyon https://geo.api.gouv.fr/communes?nom=Lyon 5. Nice https://geo.api.gouv.fr/communes?nom=Nice 6. Le Havre https://geo.api.gouv.fr/communes?nom=Le Havre			

OK Annuler

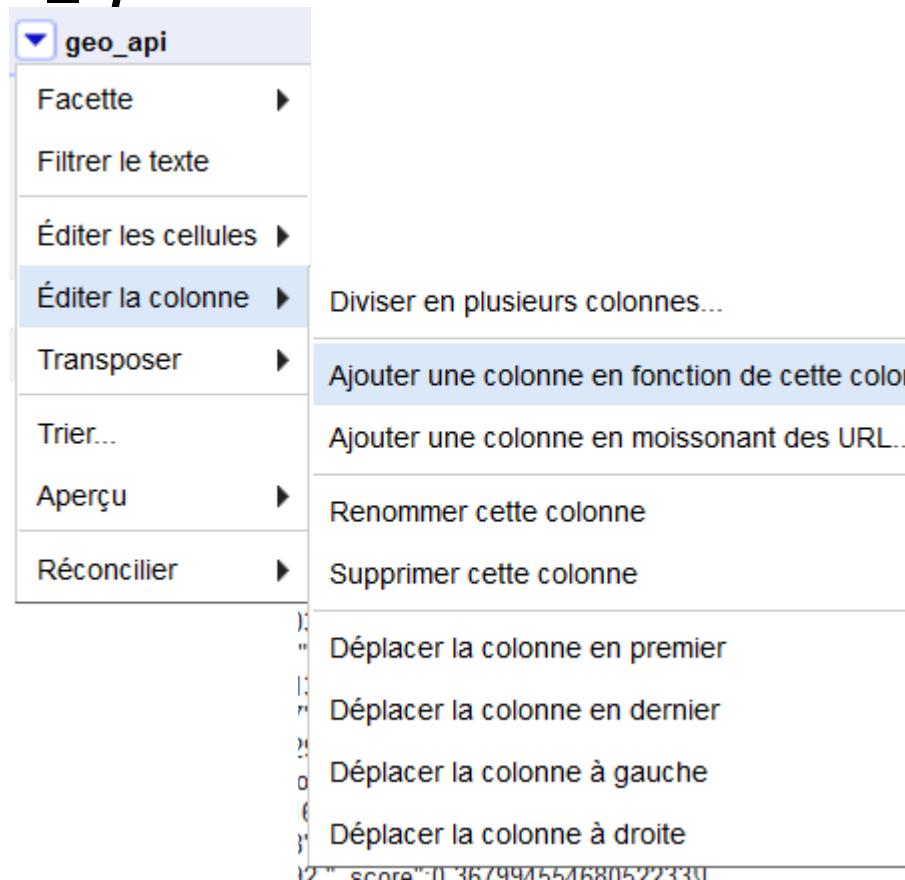
Récupérer des données sur le web

Résultat : données en JSON. Problème avec « Le Havre » (l'API attendait « Havre »)

<input checked="" type="checkbox"/> ville	<input checked="" type="checkbox"/> geo_api
Nice	[{"nom": "Nice", "code": "06088", "codeDepartement": "06", "codeRegion": "93", "codesPostaux": ["06000", "06100", "06200", "06300"], "population": 342295, "score": 0.7599648464478079}, {"nom": "Nicey", "code": "21454", "codeDepartement": "21", "codeRegion": "27", "codesPostaux": ["21330"], "population": 117, "score": 0.6499641776002426}, {"nom": "Nicey-sur-Aire", "code": "55384", "codeDepartement": "55", "codeRegion": "44", "codesPostaux": ["55260"], "population": 112, "score": 0.4844392152437498}]
Caen	[{"nom": "Caen", "code": "14118", "codeDepartement": "14", "codeRegion": "28", "codesPostaux": ["14000"], "population": 107229, "score": 1}]
Caen	[{"nom": "Caen", "code": "14118", "codeDepartement": "14", "codeRegion": "28", "codesPostaux": ["14000"], "population": 107229, "score": 1}]
Lyon	[{"nom": "Lyon", "code": "69123", "codeDepartement": "69", "codeRegion": "84", "codesPostaux": ["69001", "69002", "69003", "69004", "69005", "69006", "69007", "69008", "69009"], "population": 500715, "score": 0.5823860248317049}, {"nom": "Cognat-Lyonne", "code": "03080", "codeDepartement": "03", "codeRegion": "84", "codesPostaux": ["03110"], "population": 703, "score": 0.42635312754935634}, {"nom": "Lyons-la-Forêt", "code": "27377", "codeDepartement": "27", "codeRegion": "28", "codesPostaux": ["27480"], "population": 742, "score": 0.41035555309756283}, {"nom": "Rives de l'Yon", "code": "85213", "codeDepartement": "85", "codeRegion": "52", "codesPostaux": ["85310"], "population": 4030, "score": 0.4082054198789028}, {"nom": "Chazelles-sur-Lyon", "code": "42059", "codeDepartement": "42", "codeRegion": "84", "codesPostaux": ["42140"], "population": 5136, "score": 0.40786102429884924}, {"nom": "Beauvoir-en-Lyons", "code": "76067", "codeDepartement": "76", "codeRegion": "28", "codesPostaux": ["76220"], "population": 629, "score": 0.39461243833144605}, {"nom": "Sainte-Foy-lès-Lyon", "code": "69202", "codeDepartement": "69", "codeRegion": "84", "codesPostaux": ["69110"], "population": 21646, "score": 0.3872196772079782}, {"nom": "Beauficel-en-Lyons", "code": "27048", "codeDepartement": "27", "codeRegion": "28", "codesPostaux": ["27480"], "population": 192, "score": 0.36799455468052233}]
Nice	[{"nom": "Nice", "code": "06088", "codeDepartement": "06", "codeRegion": "93", "codesPostaux": ["06000", "06100", "06200", "06300"], "population": 342295, "score": 0.7599648464478079}, {"nom": "Nicey", "code": "21454", "codeDepartement": "21", "codeRegion": "27", "codesPostaux": ["21330"], "population": 117, "score": 0.6499641776002426}, {"nom": "Nicey-sur-Aire", "code": "55384", "codeDepartement": "55", "codeRegion": "44", "codesPostaux": ["55260"], "population": 112, "score": 0.4844392152437498}]
Le Havre	
Le Havre	
Le Havre	
Lyon	[{"nom": "Lyon", "code": "69123", "codeDepartement": "69", "codeRegion": "84", "codesPostaux": ["69001", "69002", "69003", "69004", "69005", "69006", "69007", "69008", "69009"], "population": 500715, "score": 0.5823860248317049}, {"nom": "Cognat-Lyonne", "code": "03080", "codeDepartement": "03", "codeRegion": "84", "codesPostaux": ["03110"], "population": 703, "score": 0.42635312754935634}, {"nom": "Lyons-la-Forêt", "code": "27377", "codeDepartement": "27", "codeRegion": "28", "codesPostaux": ["27480"], "population": 742, "score": 0.41035555309756283}, {"nom": "Rives de l'Yon", "code": "85213", "codeDepartement": "85", "codeRegion": "52", "codesPostaux": ["85310"], "population": 4030, "score": 0.4082054198789028}, {"nom": "Chazelles-sur-Lyon", "code": "42059", "codeDepartement": "42", "codeRegion": "84", "codesPostaux": ["42140"]}]

Récupérer des données sur le web

Exploitation des données : créer une nouvelle colonne à partir de *geo_api*



Récupérer des données sur le web

Formule :

```
value.parseJson () [0] ["population"]
```

Nouveau nom de colonne

- vider la cellule conserver l'erreur copier la valeur depuis la colonne originale

Expression

Langue General Refine Expression Language (GREL) ▾

```
value.parseJson () [0] ["population"]
```

Pas d'erreur de syntaxe.

	<input type="checkbox"/> ville <input type="checkbox"/> geo_api	<input type="checkbox"/> population
Nice	[{"nom": "Nice", "code": "06088", "codeDepartement": "06", "codeRegion": "93", "codesPostaux": ["06000", "06100", "06200", "06300"], "population": 342295, "score": 0.7599648464478079}, {"nom": "Nicey", "code": "21454", "codeDepartement": "21", "codeRegion": "27", "codesPostaux": ["21330"], "population": 117, "score": 0.6499641776002426}, {"nom": "Nicey-sur-Aire", "code": "55384", "codeDepartement": "55", "codeRegion": "44", "codesPostaux": ["55260"], "population": 112, "score": 0.4844392152437498}]	342295
Caen	[{"nom": "Caen", "code": "14118", "codeDepartement": "14", "codeRegion": "28", "codesPostaux": ["14000"], "population": 107229, "score": 1}]	107229
Caen	[{"nom": "Caen", "code": "14118", "codeDepartement": "14", "codeRegion": "28", "codesPostaux": ["14000"], "population": 107229, "score": 1}]	107229
Lyon	[{"nom": "Lyon", "code": "69123", "codeDepartement": "69", "codeRegion": "84", "codesPostaux": ["69001", "69002", "69003", "69004", "69005", "69006", "69007", "69008", "69009"], "population": 500715, "score": 0.5823860248317049}, {"nom": "Cognat-Lyonne", "code": "03080", "codeDepartement": "03", "codeRegion": "84", "codesPostaux": ["03110"], "population": 703, "score": 0.42635312754935634}, {"nom": "Lyons-la-Forêt", "code": "27377", "codeDepartement": "27", "codeRegion": "28", "codesPostaux": ["27480"], "population": 742, "score": 0.4103555309756283}, {"nom": "Rives de l'Yon", "code": "85213", "codeDepartement": "85", "codeRegion": "52", "codesPostaux": ["85310"], "population": 4030, "score": 0.4082054198789028}, {"nom": "Chazelles-sur-Lyon", "code": "42059", "codeDepartement": "42", "codeRegion": "84", "codesPostaux": ["42140"], "population": 5136, "score": 0.40786102429884924}, {"nom": "Beauvoir-en-Lyons", "code": "76067", "codeDepartement": "76", "codeRegion": "28", "codesPostaux": ["76220"], "population": 629, "score": 0.39461243831344605}, {"nom": "Sainte-Foy-lès-Lyon", "code": "69202", "codeDepartement": "69", "codeRegion": "84", "codesPostaux": ["69110"], "population": 21646, "score": 0.3872196772079782}, {"nom": "Beauficel-en-Lyons", "code": "27048", "codeDepartement": "27", "codeRegion": "28", "codesPostaux": ["27480"], "population": 192, "score": 0.36799455468052233}]	500715
Nice	[{"nom": "Nice", "code": "06088", "codeDepartement": "06", "codeRegion": "93", "codesPostaux": ["06000", "06100", "06200", "06300"], "population": 342295, "score": 0.7599648464478079}, {"nom": "Nicey", "code": "21454", "codeDepartement": "21", "codeRegion": "27", "codesPostaux": ["21330"], "population": 117, "score": 0.6499641776002426}, {"nom": "Nicey-sur-Aire", "code": "55384", "codeDepartement": "55", "codeRegion": "44", "codesPostaux": ["55260"], "population": 112, "score": 0.4844392152437498}]	342295

Pour aller plus loin

Documentation officielle

- [Site](#)
- [Documentation](#) (wiki)

Quelques tutoriels et retours d'expérience

- M. Bourdic, [*OpenRefine, "Excel aux hormones" pour nettoyage de données*](#), 2017
- A. Courtin, [*"Reconcilier" une liste de nom d'architectes avec Wikidata en utilisant OpenRefine*](#), 2017
- Karen H, [*Using OpenRefine to Reconcile Name Entities*](#), 2017
- Leçons du programme Library Carpentry. Open Refine for Librarians, 2016
- Leçons du programme Data Carpentry, 2015 ; variante [*Open Refine for Ecology*](#)
- T. Padilla, [*Getting Started with OpenRefine*](#), 2015
- S. van Hooland , R. Verborgh et M. De Wilde, [*Cleaning Data with OpenRefine*](#), 2013
- T. Hirst, [*Merging Datasets with Common Columns in Google Refine*](#), 2011
- A. Falcone, [*Google Refine CheatSheets*](#), 2011