

# Nettoyer et préparer des données avec OpenRefine

Formation Open Data Locale

Marseille, 9/6/2017



Mathieu Saby

[mathsabypro@gmail.com](mailto:mathsabypro@gmail.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 blue header with a search bar and social media links for GitHub and Twitter. Below the header, the main banner features the 'OpenRefine' logo in large blue letters, with 'OPEN' in a smaller, outlined font above 'Refine'. To the right of the logo is a blue diamond icon and the tagline 'A free, open source, powerful tool for working with messy data'. On the left side, there's a vertical navigation menu with links: Home, Download, Documentation, Community, and Post archive. The 'Post archive' section lists four OpenRefine News items from Spring 2016, December 2015, November 2015, and October 2015. The main content area on the right starts with a 'Welcome!' section, followed by a paragraph about OpenRefine's history and a note about Google's support. Below this is a section titled 'Using OpenRefine - The Book', which includes a small image of the book cover and a list of six topics covered in the book.

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

Google Custom Search

# OPEN Refine

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

- [Home](#)
- [Download](#)
- [Documentation](#)
- [Community](#)
- [Post archive](#)

OpenRefine News: Spring 2016

OpenRefine News: December 2015

OpenRefine News: November 2015

OpenRefine News: October 2015

## 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

[Home](#)

[Download](#)

[Documentation](#)

[Community](#)

[Post archive](#)

[OpenRefine News:](#)  
[Spring 2016](#)

[OpenRefine News:](#)

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](#).

### 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](#).

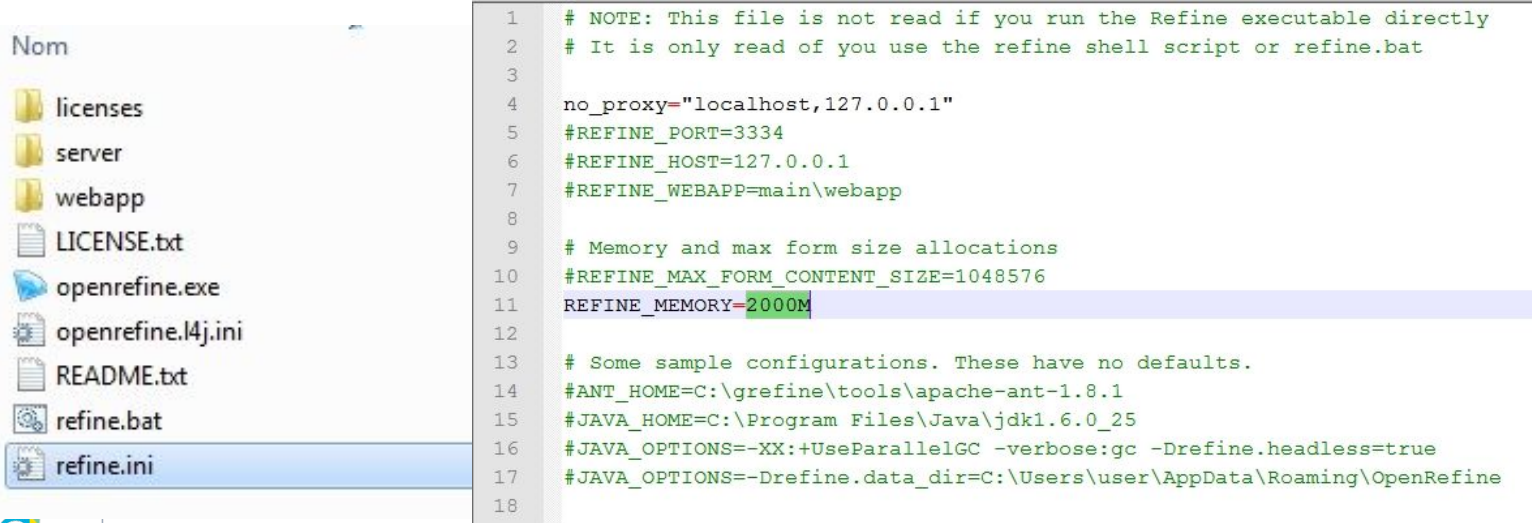
- **Windows kit**, Download, unzip, and double-click on *openrefine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it.
- **Linux kit**, Download, extract, then type `./refine`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



# 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=0B1NKejaqcJG5am5TNEphbHozalU> ou  
[https://raw.githubusercontent.com/msaby/formations/master/2017/openrefine\\_marseille/exo1.csv](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](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

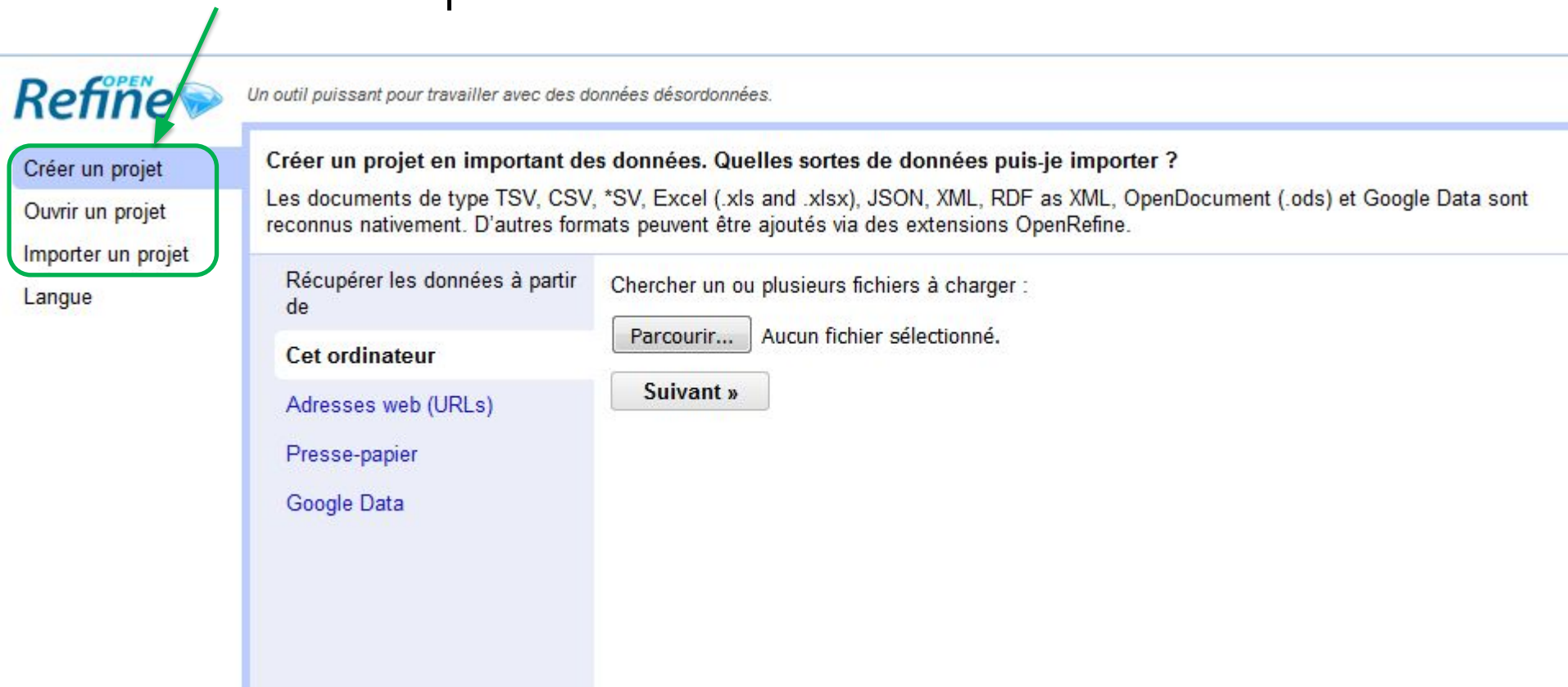
<http://opendata.regionpaca.fr/donnees/detail/annuaire-des-associations-de-la-ccabv-et-de-digne-les-bains-en-2016.html>




# 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



**Refine** OPEN  *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 :

Aucun fichier sélectionné.

# Importer des données

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

Depuis plusieurs **emplacements**

The screenshot shows the OpenRefine web application interface. On the left is a sidebar with navigation links: 'Créer un projet', 'Ouvrir un projet', 'Importer un projet', and 'Langue'. The main content area has a header with the OpenRefine logo and the tagline 'Un outil puissant pour travailler avec des données désordonnées.' Below this is a section titled 'Créer un projet en important des données. Quelles sortes de données puis-je importer ?'. A green box highlights the text: '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.' To the left of this text is a dropdown menu 'Récupérer les données à partir de' with a green box highlighting its options: 'Cet ordinateur', 'Adresses web (URLs)', 'Presse-papier', and 'Google Data'. To the right of the dropdown is a text input field 'Chercher un ou plusieurs fichiers à charger :', a 'Parcourir...' button, and the text 'Aucun fichier sélectionné.' Below these is a 'Suivant »' button. Green arrows point from the text above to the highlighted elements in the interface.

OPEN  
**Refine**

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 web interface. On the left is a sidebar with navigation links: 'Créer un projet', 'Ouvrir un projet', 'Importer un projet', and 'Langue'. The main content area has a header with the OpenRefine logo and the tagline 'Un outil puissant pour travailler avec des données désordonnées.' Below this is a section titled 'Créer un projet en important des données. Quelles sortes de données puis-je importer ?' which lists supported formats: TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, OpenDocument (.ods) and Google Data. The 'Importer un projet' step is highlighted with a blue bar. It contains two columns: 'Récupérer les données à partir de' and 'Chercher un ou plusieurs fichiers à charger :'. The first column has options: 'Cet ordinateur', 'Adresses web (URLs)', 'Presse-papier', and 'Google Data'. The second column shows a file 'doaj-article-sample.csv' with a 'Parcourir...' button next to it. Below the file list, there are two numbered steps: '1' and '2'. Step 1 is associated with the 'Parcourir...' button, and step 2 is associated with a 'Suivant »' button. Both buttons are highlighted with red rectangles.

**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 :**

1 **Parcourir...** doaj-article-sample.csv

2 **Suivant »**

# Importer des données

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

[« Démarrer »](#) Configurer les options pour l'analyse syntaxique

Nom du projet: or1 csv [Créer un projet »](#)

	Toutes	code_personne	date	ville	adresse	animal_preferé	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	

Considérer les données comme

Format des caractères

[Mettre à jour l'aperçu](#)

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

0

lignes du début du fichier

☒ 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

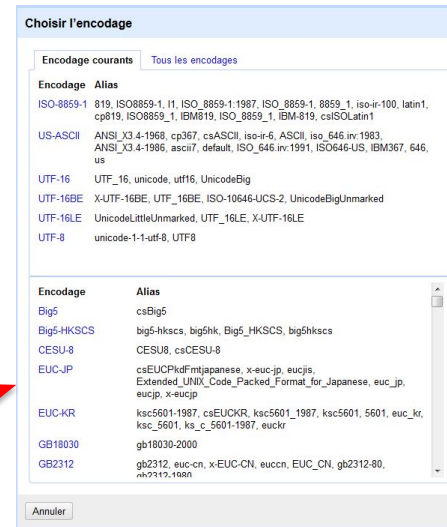
☐ Indiquer la source du fichier (noms des fichiers, URLs) dans chaque ligne

# Importer des données

## Principaux paramètres d'import

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

1



Format  
des  
caractères

Modifier si entêtes multilignes

Mettre à jour l'aperçu

3

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

- ☐ Ignorer la ou les premières 0 lignes du début du fichier
- ☒ 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...

4

- ☒ 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

5

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

# Importer des données

Pour lire et Ã©crire en franÃ§ais... choisir l'encodage correspondant Ã celui du fichier

Format  
des  
caractÃres

ISO-8859-1

1 av. St BarthÃ©lemy

Format  
des  
caractÃres

UTF-8

1 av. St BarthÃ©lemy



# 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

[Optionnel]  
Changer le nom du projet

**Créer un projet**



The image shows a user interface for creating a project. It features a light blue background. On the left, the text "Nom du projet" is displayed. To its right is a text input field containing the text "fichier\_test01". A red rectangular box highlights the input field, and a red arrow points from the text "[Optionnel] Changer le nom du projet" above to this box. To the right of the input field is a button with the text "Créer un projet »". A red rectangular box highlights the button, and a red arrow points from the text "Créer un projet" above to this box.

# L'espace de travail

Facettes  
et filtres

Historique

Lien vers le projet

Contenu du fichier

Nouveau  
projet

Export

Refine<sup>OPEN</sup> doaj article sample Permalien

Facette / Filtre Défaire / Refaire o

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](#)

10 lignes

		code_personne	date	ville	adresse	animal_preferé	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	

Extensions: « première < précédente 1 - 10 suivante > dernière »

Ouvrir... Exporter v Aide

# L'espace de travail

Numéro de ligne (automatique)

Nb lignes du fichier

Nb lignes affichées

Colonnes de données

Voir les lignes précédentes ou suivantes

Refine OPEN doaj article sample Permalien

Facette / Filtre Défaire / Refaire 0

10 lignes

Voir en: lignes entrées Afficher: 5 10 25 50 lignes

Extensions: « première < précédente 1 - 10 suivante > dernière »

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](#)

	code_personne	date	ville	adresse	animal_preferé	habillement	loisirs	logement	nt
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	

Étoiles et marques: pour isoler certaines lignes

# 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

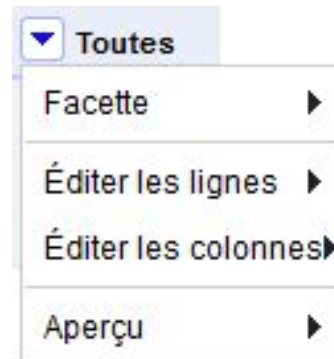
Les plus fonctions les utilisées : facettes, édition de cellules et de colonnes, tri

Fonction propre à la 1<sup>re</sup> colonne : suppression de lignes

Colonne ordinaire



Colonne « Toutes » (1<sup>re</sup> position)



# Réordonner ou supprimer des colonnes

Dans la 1<sup>re</sup> colonne *Toutes*

The image shows a software interface for managing columns. On the left, a dropdown menu is open for the 'Toutes' column. The menu options are: 'Facette', 'Éditer les lignes', 'Éditer les colonnes', and 'Aperçu'. The 'Éditer les colonnes' option is highlighted, and a red arrow points from it to a dialog box titled 'Retrier / supprimer les colonnes...'. This dialog box is titled 'Trier / Supprimer des colonnes' and contains two main sections: 'Glisser des colonnes pour les trier' and 'Déposer des colonnes ici pour les supprimer'. The 'Glisser des colonnes pour les trier' section lists various fields: Title, Authors, DOI, URL, Date, Language, Subjects, ISSN, Publisher, and Citation. A red double-headed arrow is positioned next to this list. The 'Déposer des colonnes ici pour les supprimer' section is empty. Below these sections is a 'Licence' field. At the bottom of the dialog box are 'OK' and 'Annuler' buttons.

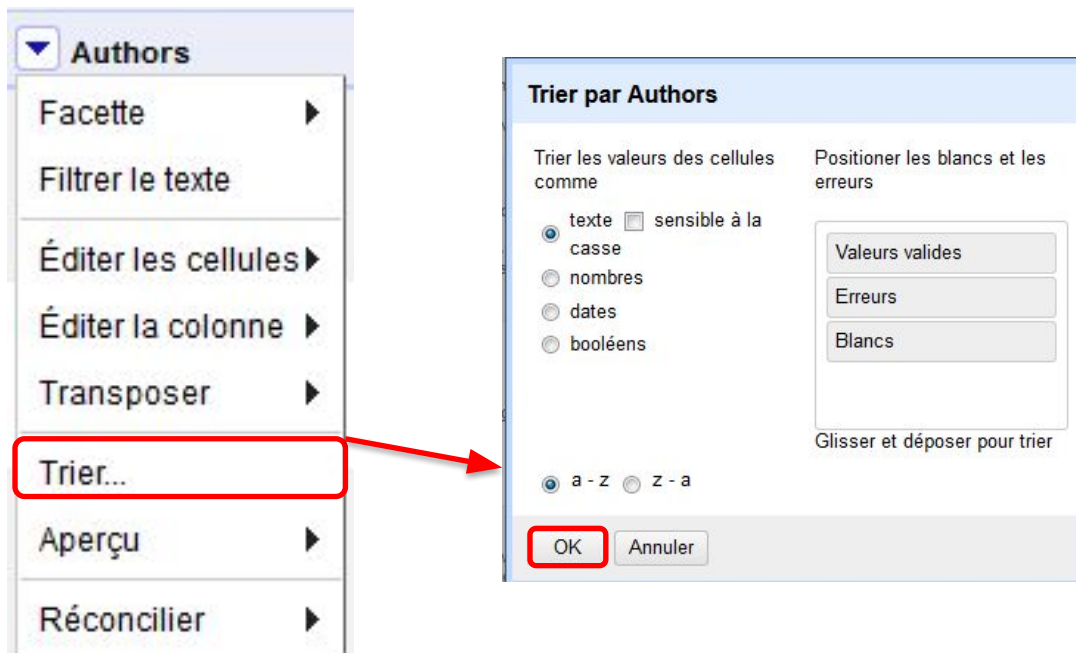
# Tris, filtres et facettes

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



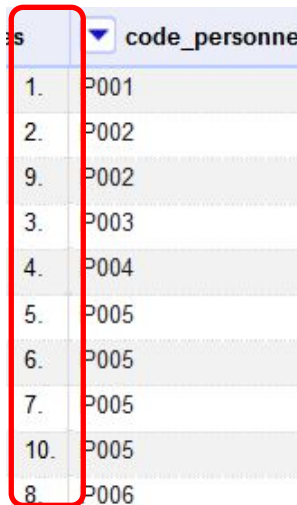
# Trier les données

**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)



The image shows a screenshot of a data table. A red rectangular box highlights the first column, which contains numerical indices, and the second column, which contains person codes. The table has a header row with a dropdown arrow and the text 'code\_personne'. The data rows are as follows:

	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 ▼

▼ Toutes

▼ code\_personne

▼ date

▼ ville

▼

Supprimer le tri

Retrier les lignes de façon permanente

Par code\_personne

						habillement	▼ loisirs	▼ logement	
☆	1.	P001	01/02/2017	NICE	1 av.		25	0,8	
☆	2.	P002	01/03/2017	CAEN					
☆	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				

# Filtrer les données

**Activité** : filtrer le fichier pour afficher les lignes dont la colonne *code\_personne* contient le mot « P005 » ET la colonne *animal\_prefer* le mot « chien »



Two filter configuration panels are shown. The first panel, titled 'code\_personne', has a text input field containing 'P005' and two checkboxes: 'sensible à la casse' and 'expression rationnelle'. The second panel, titled 'animal\_prefer', has a text input field containing 'chien' and the same two checkboxes. A red arrow points from the 'animal\_prefer' panel down to the table below.

1 matching lignes (10 total)

Voir en: lignes entrées

Afficher: 5 10 25 50 lignes

Sort ▼

▼ Toutes	▼ code_personne	▼ date	▼ ville	▼ adresse	▼ animal_prefer	▼ habillement	▼ loisirs	▼ logement	
 	10.	P005	08:01:2017	Le Havre	15 av. Jean Jaurès	Lapin, chien	200	40	800

# Filtrer les données

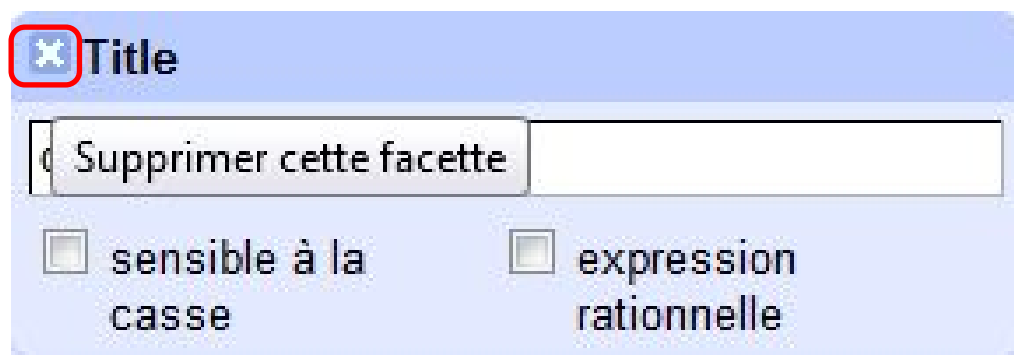
Toutes les opérations (export, nouveaux filtres, facettes, modifications groupées) s'opèreront uniquement sur les données filtrées.

Ex: modification groupée la colonne *animal\_preferre* :  
uniquement 1 lignes modifiée



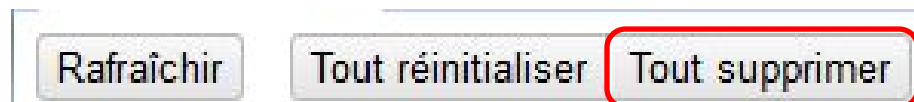
# Filtrer les données

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

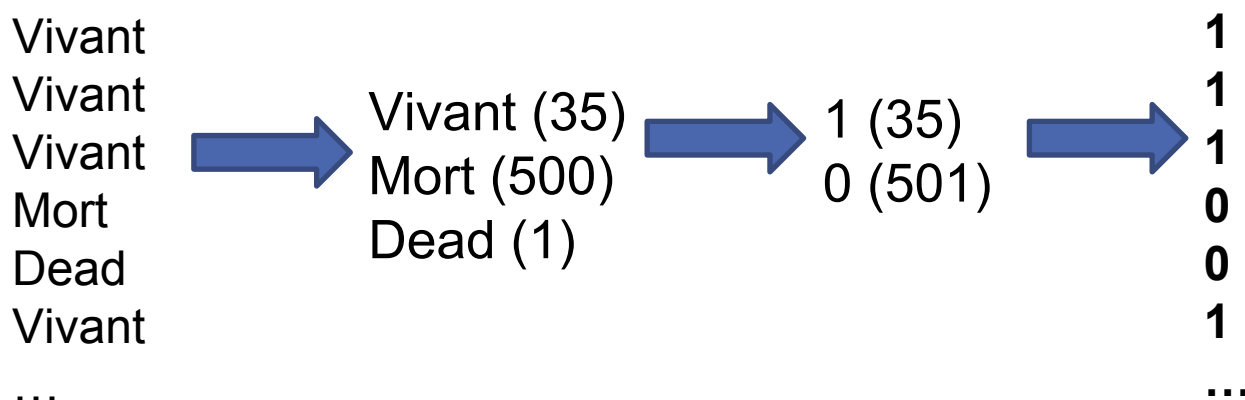


# Utiliser les facettes

**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*.



(blank) : valeur vide

Quelles anomalies repère-t-on?


Facette par nombre de choix



# 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



The screenshot shows a facet interface for the variable 'ville'. The facet title 'ville' is at the top left. Below it, the text '9 choices' is highlighted with a red box. To the right of '9 choices' is the text 'Trier par:'. Below 'Trier par:' are two options: 'nom' and 'compte', both highlighted with red boxes. To the right of these options is a button labeled 'changer', also highlighted with a red box. Below the 'changer' button is a button labeled 'Groupe', highlighted with a red box. The list of choices is displayed below the facet title, showing 'Lyon 1', 'CAEN 1', 'Caen 1', 'Havre (Le) 1', 'Le Havre 1', 'LE HAVRE 1', 'Lyon 1', 'Nice 1', 'NICE 1', and '(blank) 1'. The text 'Facette par nombre de choix' is at the bottom left of the facet.

Modifier la facette (complexe)

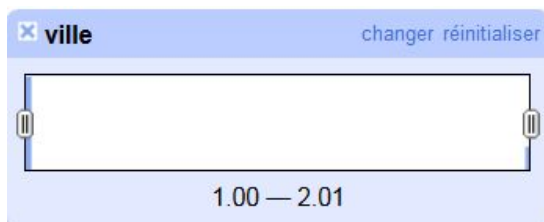
Regrouper les valeurs semblables

# Utiliser les facettes

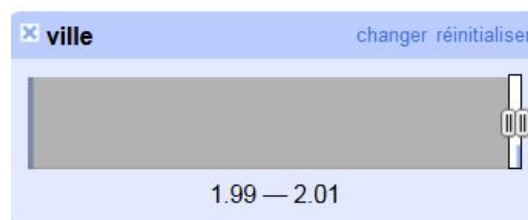
**Activité :** dans la facette *ville*, afficher les valeurs par « nombre de choix », et ne conserver que celles présentes 2 fois.



histogramme



1 seule valeur, présente 2 fois



# Utiliser les facettes

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

× ville changer

1 choices Trier par: nom compte Groupe

Havre (Le) 2 éditer include

Facette par nombre de choix

LE HAVRE

Appliquer Annuler

Valider Échap

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

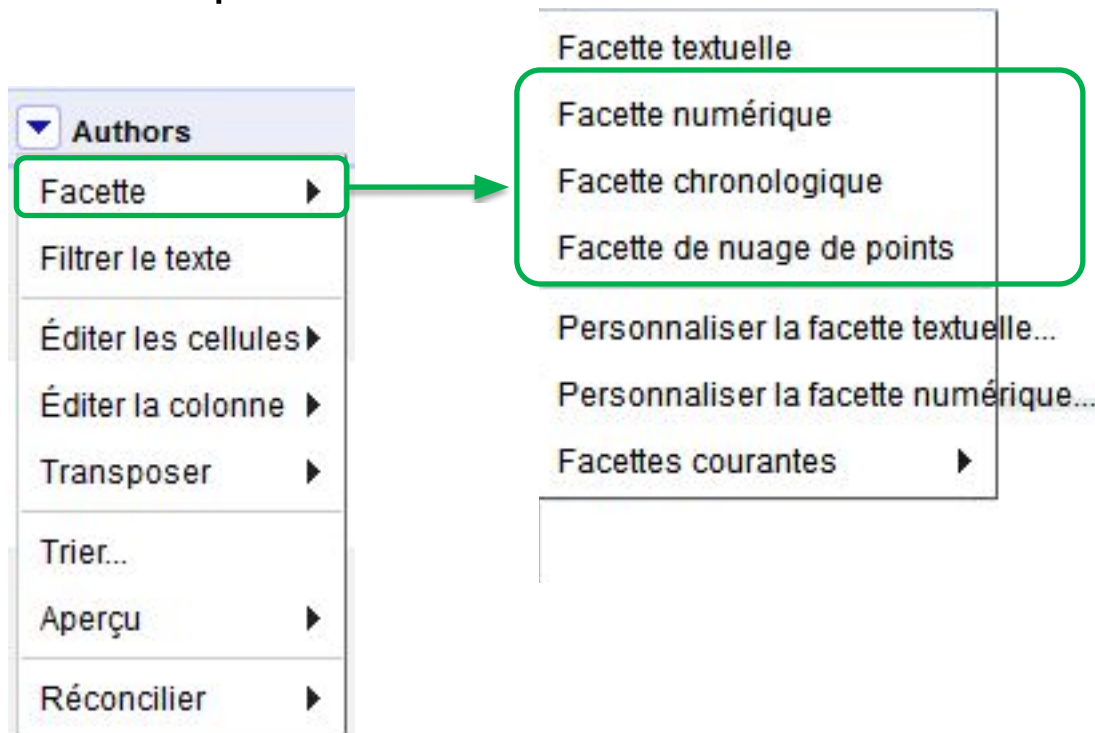


facette en **orange** : utilisé pour filtrer les données (afficher les données correspondant à la facette)

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

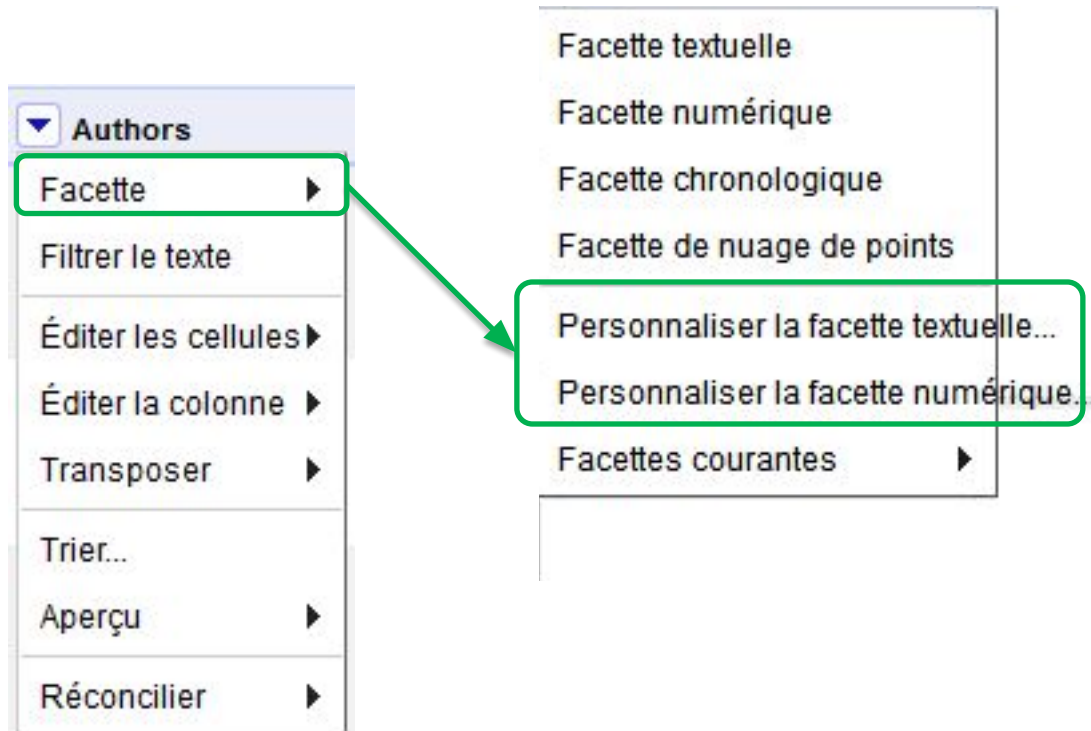
# Utiliser les facettes: pour aller plus loin

**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)



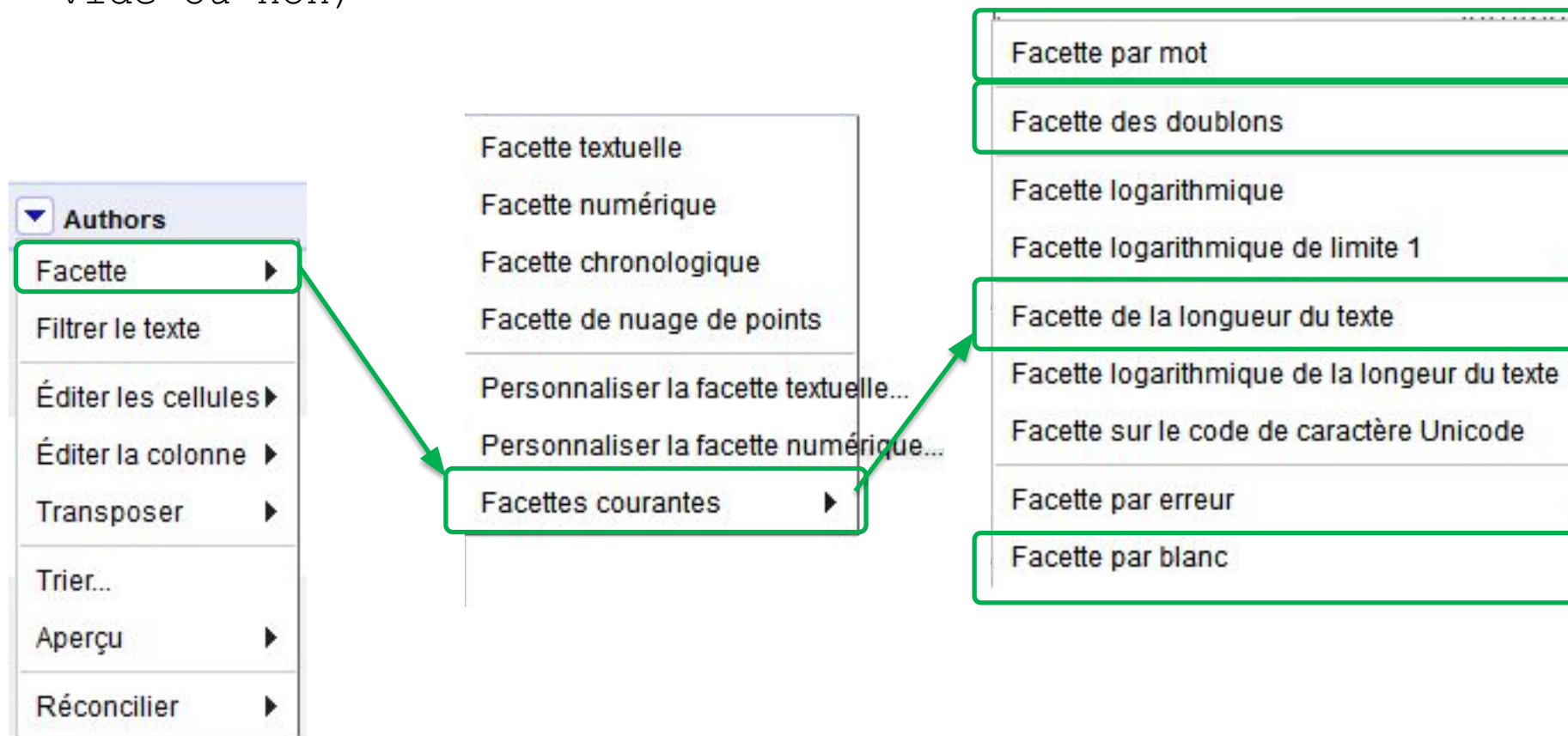
# Utiliser les facettes: pour aller plus loin

**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)



# Utiliser les facettes: pour aller plus loin

**Activité** : A partir de la colonne *animal\_prefer*, 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\_prefer** [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\_prefer** [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 choixes 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

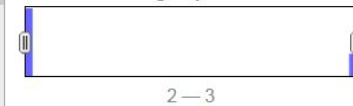
Méthode collision de clés

Fonction de codage empreinte

4 clusters trouvés

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

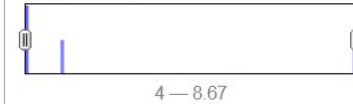
# Choix dans le groupe



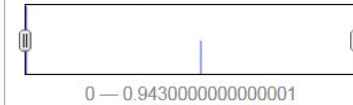
# Lignes dans le groupe



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



# 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	ville
Nice	NICE
Caen	CAEN
Lyon	LYON
Nice	NICE
Le Havre	LE HAVRE
Le Havre	LE HAVRE
Lyon	LYON
Caen	CAEN
Le Havre	LE HAVRE

# Appliquer des transformations courantes

## Modifier la casse

Editer les cellules > Transformations courantes > En majuscules

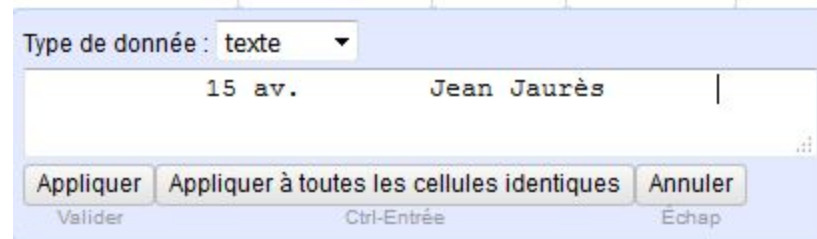


# Appliquer des transformations courantes

## 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



Type de donnée : texte

15 av. Jean Jaurès

Appliquer Appliquer à toutes les cellules identiques Annuler

Valider Ctrl-Entrée Échap

*(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



Rassembler les espaces consécutifs

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

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 apparences, ces nombres et ces dates sont considérés comme de simple suite de caractères.

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

	<input type="button" value="▼"/> Date	<input type="button" value="▼"/> habillement	<input type="button" value="▼"/> loisirs	<input type="button" value="▼"/> logement	
JJ/MM/2017	01/02/2017	100	25	0,8	Unité en k€
	01/03/2017	50.50	35,6	0,7	
	15/02/2017	10.90	70,6	700	
JJ-MM-2017	15-02-2017	400	90	600	Unité en €
	15-04-2017				
	12-02-2017				
	11/01/2017				
JJ/ MM (2017)	19/02 (2017)	Séparateur décimal .			
	16/03 (2017)	50.50	35,6	0,7	Séparateur décimal ,
JJ:MM:2017	08:01:2017	200	40	800	

# 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

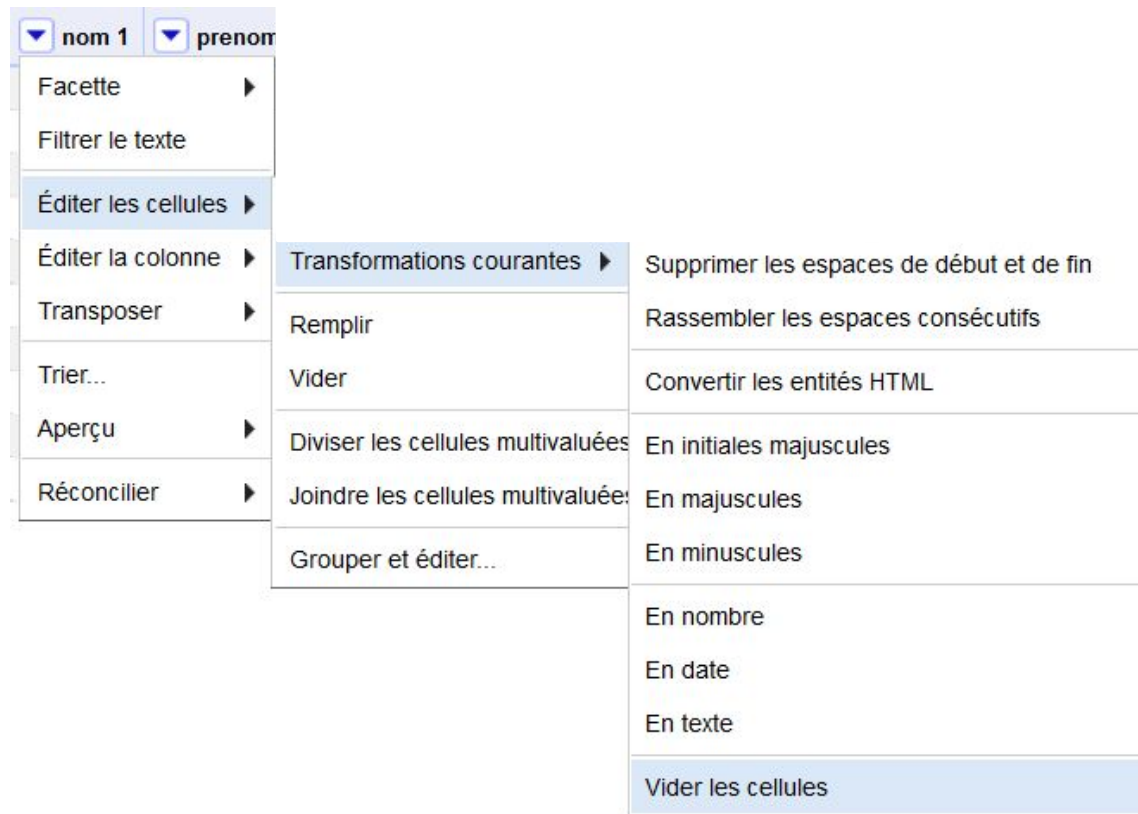
▼ Date
2017-01-02T00:00:00Z
2017-01-03T00:00:00Z
2017-02-15T00:00:00Z
2017-02-15T00:00:00Z
2017-04-15T00:00:00Z
2017-12-02T00:00:00Z
2017-11-01T00:00:00Z
19/02 (2017)
16/03 (2017)
08.01:2017

▼ habillement	▼ loisirs	▼ logement
100	25	0,8
50.5	35,6	0,7
10.9	70,6	700
400	90	600
50.5	35,6	0,7
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 image shows a data table with columns: Toutes, Ville, espece, and nombre. The 'Ville' column contains 'Nice', 'Marseille', and 'Paris'. A context menu is open over the 'Ville' column, showing options like 'Facette', 'Filtrer le texte', 'Éditer les cellules', 'Éditer la colonne', 'Transposer', 'Trier...', 'Aperçu', and 'Réconcilier'. The 'Éditer les cellules' option is selected, and a sub-menu is open showing 'Transformer...', 'Transformations courantes', 'Remplir', 'Vider', 'Diviser les cellules multivaluées...', 'Joindre les cellules multivaluées...', and 'Grouper et éditer...'. The 'Remplir' option is highlighted. Red arrows point from the 'Ville' column header and the 'Remplir' option to the 'Ville' column header of a second table on the right, which shows the result of the fill operation: all empty cells in the 'Ville' column are now filled with 'Nice'.

Toutes	Ville	espece	nombre
1.	Nice		400
2.			200
3.			10
4.			
5.			
6.			
7.			
8.			
9.			

Ville
Nice
Nice
Nice
Marseille
Marseille
Marseille
Paris
Paris
Paris

# Recopier ou supprimer des valeurs

## Opération inverse : supprimer les valeurs répétées

Editer les cellules > Vider

The image shows a data table with three columns: 'Ville', 'espece', and 'nombre'. The 'Ville' column contains repeated values: Nice, Nice, Nice, Marseille, Marseille, Marseille, Paris, Paris, Paris. A context menu is open over the 'Ville' column, with the 'Vider' option highlighted. A red arrow points from the 'Ville' column header to the menu, and another red arrow points from the 'Vider' option to the 'Ville' column header on the right side of the table.

Ville	espece	nombre
Nice	Facette	400
Nice	Filtrer le texte	200
Nice	Éditer les cellules	10
Marseille	Éditer la colonne	
Marseille	Transposer	
Marseille	Trier...	
Paris	Aperçu	
Paris	Réconcilier	

Context menu options:

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

Transformations courantes:

- Transformer...
- Transformations courantes
- Remplir
- Vider
- Diviser les cellules multivaluées
- Joindre les cellules multivaluées
- Grouper et éditer...



# 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.

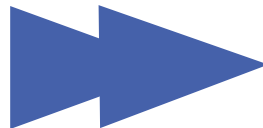
Pour créer des entrées :

- 1/ trier les données en fonction de la colonne servant de clé de regroupement
- 2/ déplacer cette colonne en 1<sup>re</sup> 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

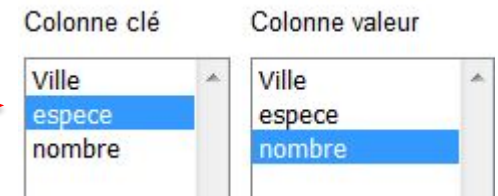
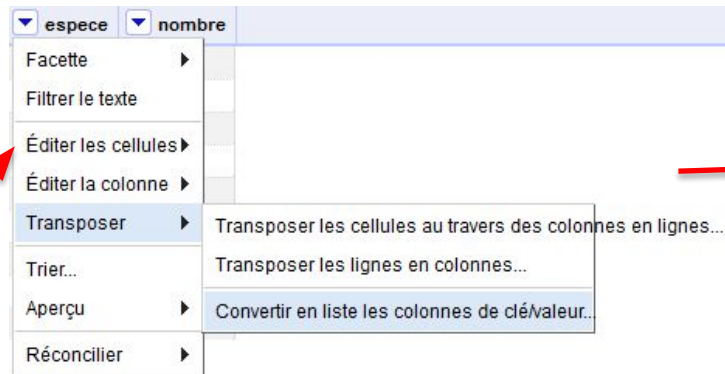
# Transformations courantes

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



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

# Restructurer les données

## Passer du format « wide » au format « long »

Transposer > Transposer les cellules au travers des colonnes en lignes

The screenshot shows the 'Transposer' menu option selected, with a red arrow pointing to it. The menu options are: Facette, Filtrer le texte, Éditer les cellules, Éditer la colonne, Transposer, Trier..., Aperçu, and Réconcilier. The 'Transposer' option is highlighted, and its sub-menu is visible, showing 'Transposer les cellules au travers des colonnes en lignes...'. A red arrow points from this option to the 'De la colonne' and 'Vers la colonne' selection interface. In this interface, 'Ville' is selected under 'De la colonne' and 'last column' is selected under 'Vers la colonne'. A red arrow points from this interface to the 'Deux nouvelles colonnes' configuration panel, which shows 'Colonne clé' set to 'espece d'origine' and 'Colonne valeur' set to 'nombre d'origine'.

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

▼ palmiers	▼ orangers	▼ bouleau
Facette	10	
Filtrer le texte	10	
Éditer les cellules	100	
Éditer la colonne		
Transposer	Transposer les cellules au travers des colonnes en lignes...	
Trier...	Transposer les lignes en colonnes...	
Aperçu	Convertir en liste les colonnes de clé/valeur...	
Réconcilier		

- ☐ Une colonne `espece_nombre`
- ☒ préfixer le nom de la colonne d'origine à chaque cellule suivie par : avant la valeur de la cellule

▼ 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

Menu contextuel pour la colonne **espece\_nombre**:

- Facette
- Filtrer le texte
- Éditer les cellules
- Éditer la colonne
  - Diviser en plusieurs colonnes...
- Transposer
- Trier...
- Aperçu
- Réconcilier

Options pour "Diviser en plusieurs colonnes..." :

- Ajouter une colonne en fonction de cette colonne..
- Ajouter une colonne en moissonnant des URL...
- Renommer cette colonne
- Supprimer cette colonne
- Déplacer la colonne en premier
- Déplacer la colonne en dernier
- Déplacer la colonne à gauche
- Déplacer la colonne à droite

**Diviser la colonne espece\_nombre en plusieurs colonnes**

**Méthode de division de la colonne**

☒ par séparateur

Séparateur :  ☐ expression rationnelle

Diviser en  colonnes au plus (laisser vide pour ne pas limiter)

☐ par les longueurs de champs

Liste les longueurs en les séparant par des virgules, par exemple 5, 7, 15

**Après la division**

☒ Deviner le type de cellule

☒ Supprimer cette colonne

Séparateur :

es	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

## 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

Répéter pour les 2 colonnes

	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

espece_nombre	espece_nombre
Facette	400
Filter le texte	200
	10
Éditer les cellules	Transformer...
Éditer la colonne	Transformations courantes
Transposer	Remplir
Trier...	Vider
Aperçu	Diviser les cellules multivaluées
Réconcilier	Joindre les cellules multivaluées..
	Grouper et éditer...

Indiquer le séparateur à utiliser entre les valeurs

	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

# 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

espece_nombre 1	espece_nombre
Facette	400, 200, 10
Filtrer le texte	au 200, 50, 10
Éditer les cellules	Transformer...
Éditer la colonne	Transformations courantes
Transposer	Remplir
Trier...	Vider
Aperçu	Diviser les cellules multivaluées...
Réconcilier	Joindre les cellules multivaluées...
Grouper et éditer...	

Quel séparateur sépare actuellement les valeurs ?

	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



Créer un projet

Ouvrir un projet

Importer un projet

Langue

## Rouvrir le projet exo1

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

# Restructurer les données

**Eclater toutes les valeurs de la colonne animal\_preferes dans des lignes distinctes**

Quel séparateur sépare actuellement les valeurs ?

,



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

Quel séparateur sépare actuellement les valeurs ?

et



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

# Restructurer les données

Créer une facette sur la colonne *animal\_prefer* 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

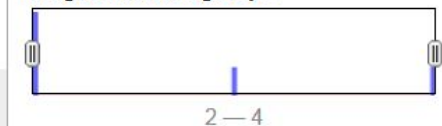
2

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

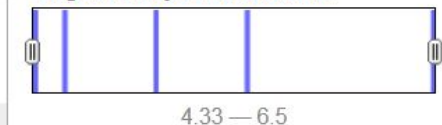
# Choix dans le groupe



# Lignes dans le groupe



Longueur moyenne des choix



Variabilité moyenne des choix



# 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

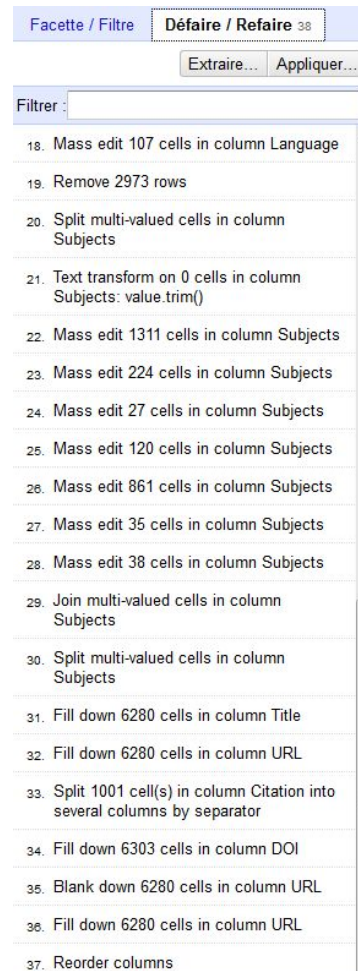
# Exporter les données transformées

## Plusieurs formats d'export



# Annuler ou rejouer un traitement

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



# 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 :

**Extraire des opérations de l'historique**

Extraire 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.

- ☒ Split multi-valued cells in column Authors
- ☒ Mass edit cells in column Authors
- ☒ Mass edit cells in column Authors
- ☒ Mass edit cells in column Authors
- ☒ Mass edit cells in column Authors
- ☒ Create column nom\_famille at index 2 based on column Authors using expression `grell.value.split(" ")[length(value.split(" ")) - 1]`
- ☒ Create column prenom at index 2 based on column Authors using expression `grell.value.split(" ").slice(0, length(value.split(" ")) - 1).join(" ")`
- ☒ Text transform on cells in column Authors using expression `grell.value.split(",").reverse().join(" ")`
- ☒ Blank down cells in column prenom
- ☒ Remove column prenom
- ☒ Remove column nom\_famille
- ☒ Create column nom at index 2 based on column Authors using expression `grell.value.split(" ")[length(value.split(" ")) - 1]`

Tout sélectionner Tout désélectionner

**2** Fermer

**3**

**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



# Réappliquer les traitements

À partir d'un jeu de données fraîchement téléchargé

Facette / Filtre   Défaire / Refaire 0

**Historique d'annulation infinie**

N'ayez pas peur de faire des erreurs. Tous les traitements que vous effectuez sont enregistrés ici et vous pouvez revenir en arrière à tout moment.

[En savoir plus »](#)

1

Extraire...   **Appliquer...**

2

3 Coller dans OpenRefine (Ctrl+V/ Cmd+V)

4

Lancer les opérations   Annuler   Valider

Ouvrir le fichier texte dans lequel les traitements ont été enregistrés

Sélectionner tout le contenu et copier dans le presse-papier (Ctrl+C / Cmd+C)

# 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 `Editor les cellules > Transformer`

Formule  
(sans = initial)

## Personnaliser la transformation du texte sur la colonne loisirs

Expression

Langue General Refine Expression Language (GREL) ▾

`value`

Pas d'erreur de syntaxe.

Aperçu

Historique

Étoilée

Aide

row value

1. 25  
2.  
3. 35,6  
4. 70,6  
5. 90  
6. null

value

25  
  
35,6  
70,6  
90  
null

Résultat de  
la formule

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

OK

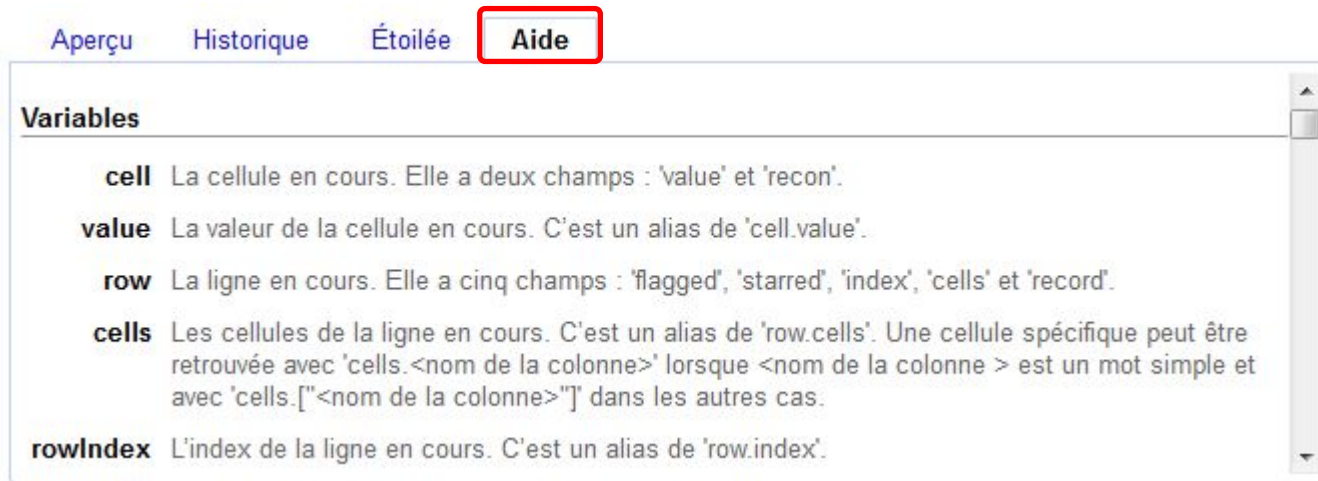
Annuler

# Appliquer des transformations personnalisées

## Deux menus utiles: aide et historique

Aide précieuse!

Utiliser Ctrl+F pour rechercher une commande

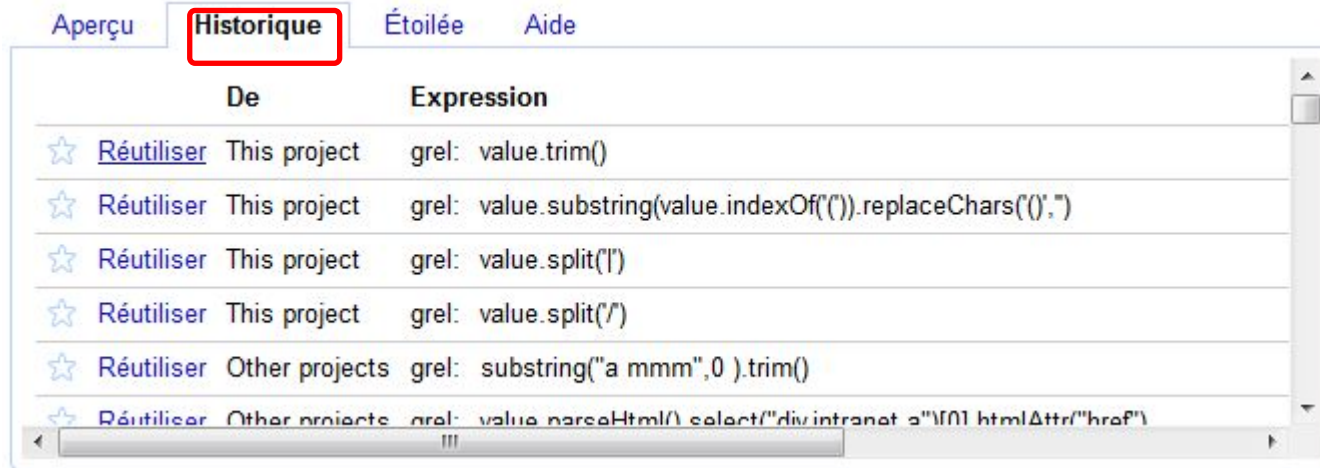


# Appliquer des transformations personnalisées

Deux menus utiles: aide et historique

Historique

Permet de réutiliser une formule



# 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<sup>er</sup> élément du tableau)

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

Ex : `"01/12/2015".split('/') [2]` -> « 2015 » (3<sup>er</sup> 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)

`value.replace(",",".")` Pas d'erreur de syntaxe.

**Aperçu** Historique Étoilée Aide

row	value	<code>value.replace(",",".")</code>
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  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



loisirs	loisirs	loisirs
25	25	25
35,6	35.6	35.6
70,6	70.6	70.6
90	90	90
40	40	40

# 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) ▼ 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  fois maximum, tant que les données changent  
☐ vider la cellule  
☐ conserver l'erreur

← 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 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

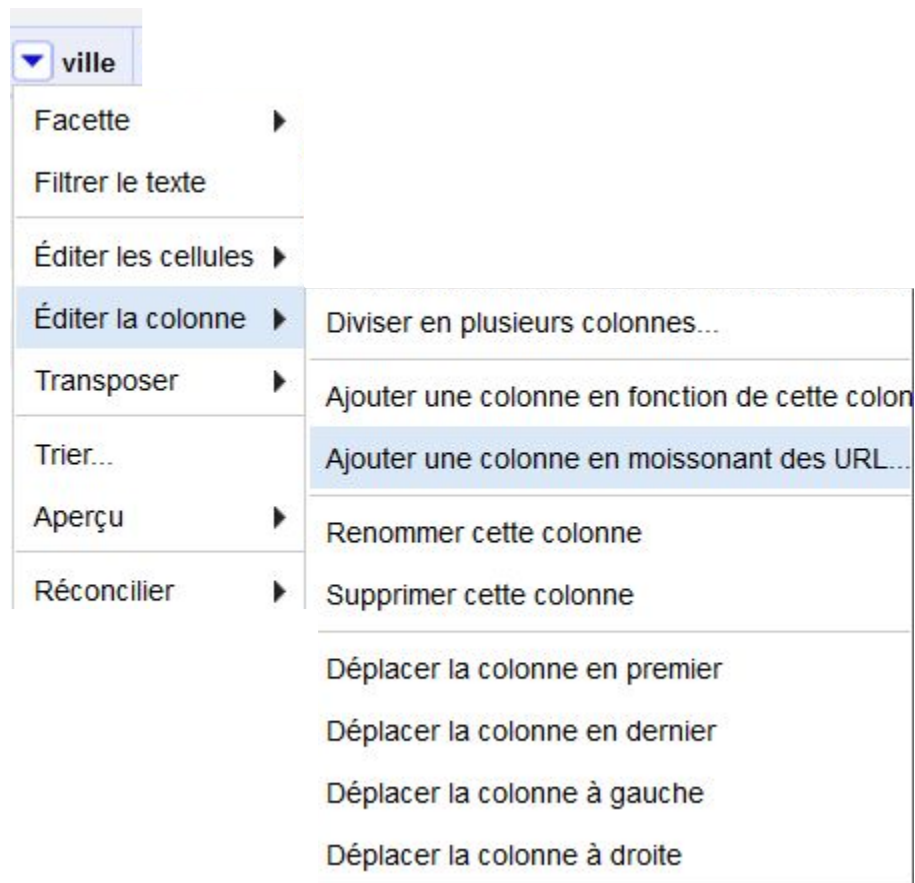
OK Annuler

Prévisualisation  
du résultat

Cette fois « 2017 » est copié dans la colonne  
[0] permet d'accéder au 1<sup>er</sup> 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  Délai de récupération  millisecondes

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

**Indiquer les URL à moissonner :**

Expression  Langue  Pas d'erreur de syntaxe.

**Aperçu** Historique Étoilée Aide

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

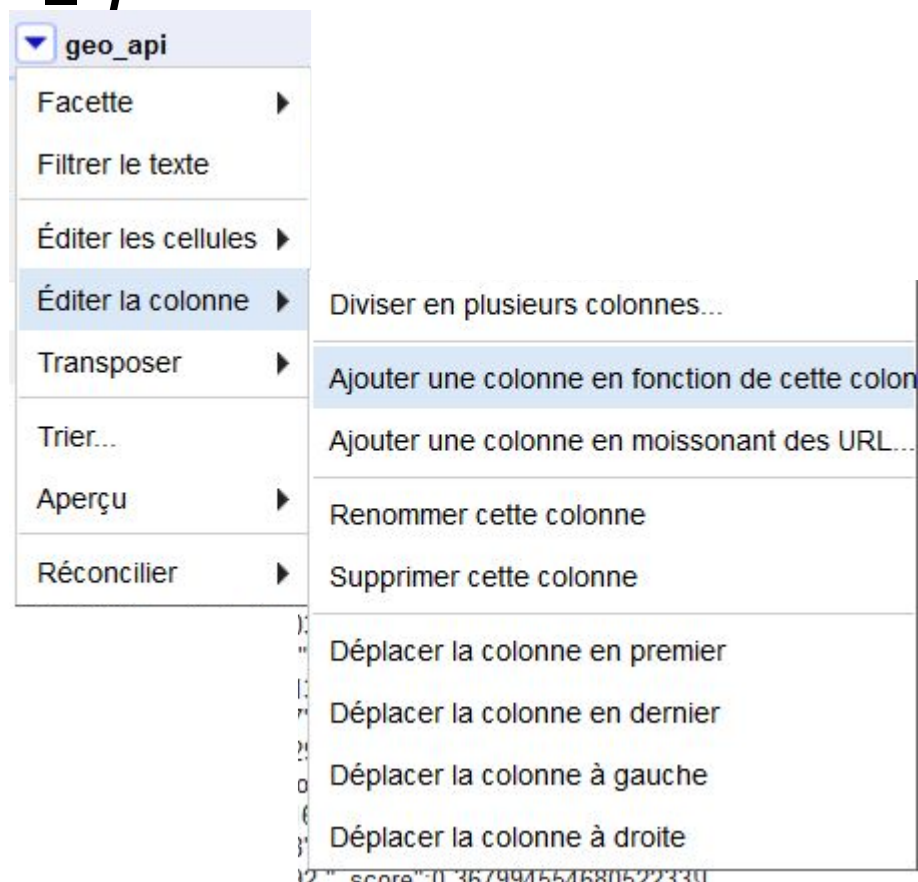
# Récupérer des données sur le web

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

ville	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": 0.4103555309756283}, {"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": 0.39461243833144605}, {"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	<a href="#">edit</a>
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": 0.4103555309756283}, {"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": 0.39461243833144605}, {"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}]

# 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

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

Pas d'erreur de syntaxe.

ville	geo_api	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": 0.41035555309756283}, {"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}]	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

edit



# 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