



HAL
open science

Nettoyer ses données avec OpenRefine (niveau1)

Aurélien Moisan

► **To cite this version:**

Aurélien Moisan. Nettoyer ses données avec OpenRefine (niveau1). Printemps de la Donnée 2024, May 2024, Paris, France. 10.5281/zenodo.11263006 . hal-04684340

HAL Id: hal-04684340

<https://hal.inrae.fr/hal-04684340v1>

Submitted on 2 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

NETTOYER SES DONNÉES AVEC OPENREFINE (NIVEAU 1)

AURÉLIEN MOISAN
21 MAI 2024



Sommaire

- 1. INTRODUCTION**
- 2. INSTALLATION ET DÉCOUVERTE DE L'INTERFACE**
- 3. FILTRES, FACETTES ET TRI**
- 4. NETTOYER LES DONNÉES : LES FONCTIONS BASIQUES**
- 5. INTRODUCTION À L'ÉDITEUR DE FORMULES GREL**

1

INTRODUCTION

OpenRefine

Qu'est-ce que c'est ?



OpenRefine

OpenRefine est un outil gratuit et open source qui permet de nettoyer, transformer, convertir, enrichir des données.

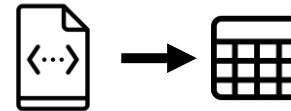
Pour plus d'informations rendez-vous sur : <https://openrefine.org/>

OpenRefine permet de :

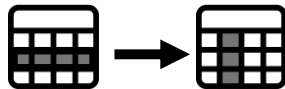
Nettoyer un jeu de données



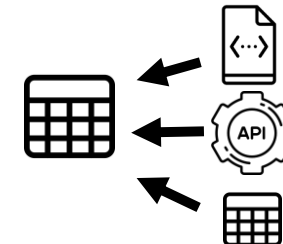
Convertir un jeu de données



Transformer un jeu de données



Enrichir un jeu de données



Exposer des données sur Wikidata



OpenRefine

Ses atouts

- Il permet de modifier des données en masse (grâce à des traitements appliqués par colonnes)
- Il permet de définir un ensemble de données sur lequel appliquer un traitement grâce à des filtres et des facettes
- Ses formules pré-enregistrées et ses extensions permettent d'effectuer des traitements simples sans maîtriser de langage de programmation
- Il enregistre l'ensemble des traitements réalisés, pour que vous puissiez les reproduire à l'identique sur un autre jeu de données
- Sa grande communauté d'utilisateurs

Il est parfois présenté dans comme « un excel sous stéroïde » mais ...

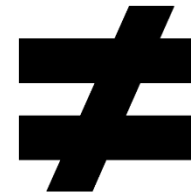
... ce n'est pas un tableur

OpenRefine ne permet pas :

- de visualiser les données sous forme graphique
- le travail collaboratif

Il n'est pas optimisé pour :

- saisir des données
- réaliser des calculs



Historique de l'outil



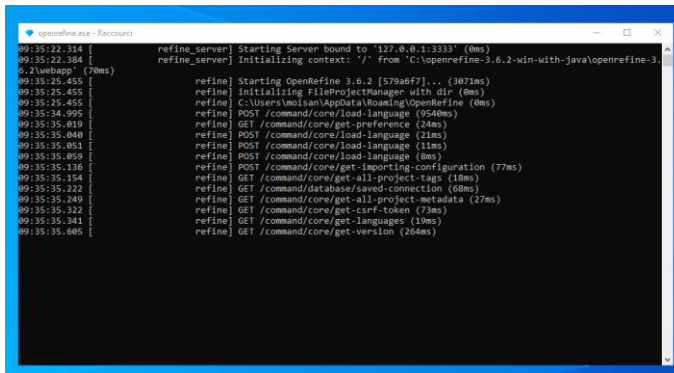
1

INSTALLATION ET DÉCOUVERTE DE L'INTERFACE

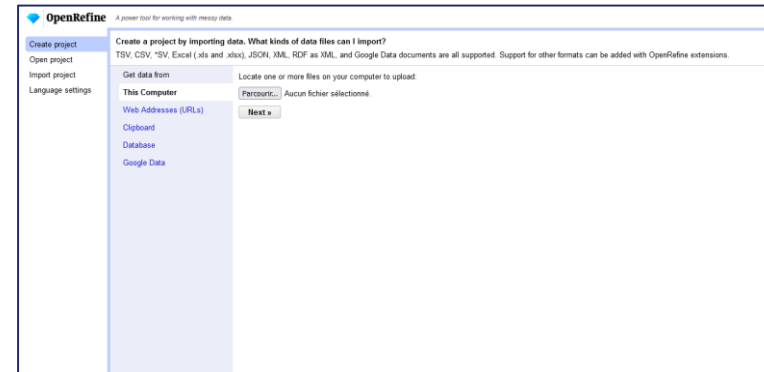
Installation et exécution

Lien de téléchargement : <https://openrefine.org/download>

- Dézippez l'archive et stockez le dossier dans le répertoire de votre choix.
- Il n'y a rien à installer, il faut simplement lancer le fichier exécutable. Ensuite :
 - Le programme s'ouvre dans une fenêtre, elle doit restée ouverte. Sinon l'interface ne répondra plus.
 - L'interface s'ouvre dans votre navigateur par défaut. OpenRefine se sert juste de votre navigateur pour afficher son interface. **C'est une application locale** rien n'est en ligne, vos données sont uniquement stockées sur votre ordinateur.



```
openrefine.exe - Backgroun
00:35:22.914 [refine_server] Starting Server bound to '127.0.0.1:3333' (0ms)
00:35:22.384 [refine_server] Initializing context: '/' from 'C:\openrefine-3.6.2-win-with-java\openrefine-3.
0.2\webapp' (70ms)
00:35:25.455 [refine] Starting OpenRefine 3.6.2 [579a6f7]... (307ms)
00:35:25.455 [refine] Initializing FileProjectManager with dir (0ms)
00:35:25.455 [refine] C:\Users\Yvoisan\AppData\Roaming\OpenRefine (0ms)
00:35:34.995 [refine] POST /command/core/load-language (9540ms)
00:35:35.019 [refine] GET /command/core/get-preference (24ms)
00:35:35.048 [refine] POST /command/core/load-language (21ms)
00:35:35.051 [refine] POST /command/core/load-language (11ms)
00:35:35.059 [refine] POST /command/core/load-language (8ms)
00:35:35.136 [refine] POST /command/core/get-importing-configuration (77ms)
00:35:35.154 [refine] GET /command/core/get-all-project-tags (18ms)
00:35:35.222 [refine] GET /command/database/saved-connection (68ms)
00:35:35.249 [refine] GET /command/core/get-all-project-metadata (27ms)
00:35:35.322 [refine] GET /command/core/get-csrf-token (72ms)
00:35:35.341 [refine] GET /command/core/get-languages (10ms)
00:35:35.605 [refine] GET /command/core/get-version (264ms)
```



- Pour fermer l'outil il faut fermer l'onglet dans le navigateur, ainsi que le programme

Paramétrage

Si vous souhaitez modifier les paramètres d'OpenRefine, ouvrez le fichier openrefine.l4j.ini avec un éditeur de texte (notepad, gedit etc.); Pour :

- Augmenter la mémoire allouée à OpenRefine, modifiez la ligne :

```
# max memory memory heap size  
-Xmx1024M
```

- Changer le répertoire de travail, ajoutez une ligne :

```
-Drefine.data_dir=Chemin absolu du répertoire
```

nom	date	type
licenses	04/01/2023 12:36	Dossier de fichiers
server	04/01/2023 12:36	Dossier de fichiers
webapp	04/01/2023 12:36	Dossier de fichiers
LICENSE.txt	04/01/2023 12:35	Document texte
licenses.xml	04/01/2023 12:35	Document XML
openrefine.exe	04/01/2023 12:35	Application
openrefine.l4j.ini	04/01/2023 12:43	Paramètres de configuration
README.md	04/01/2023 12:35	Fichier MD
refine.bat	04/01/2023 12:35	Fichier de commande Windows
refine.ini	04/01/2023 12:35	Paramètres de configuration

```
openrefine.l4j.ini - Bloc-notes  
Fichier Edition Format Affichage Aide  
# Launch4j runtime config  
  
# initial memory heap size  
-Xms256M  
  
# max memory memory heap size  
-Xmx2048M  
  
# Use system defined HTTP proxies  
-Djava.net.useSystemProxies=true  
  
#Paramétrage du répertoire de travail  
-Drefine.data_dir="H:\Dossiers agents\Moisan Aurelien\202401_backup_openrefine"  
  
#-XX:+UseLargePages  
#-Dsomevar="%SOMEVAR%"  
  
Ln 13, Col 80 100% Unix (LF) UTF-8
```

Attention : ces modifications peuvent entrainer un dysfonctionnement de l'outil. Si vous n'êtes pas sûr de vous, effectuez une copie du fichier de paramétrage initial avant toute modification pour pouvoir le restaurer en cas d'erreur.

Interface

Gérer ses projets

Créer un projet

Créer un nouveau projet à partir :

- d'un document stocké sur son ordinateur (.csv, .odt, .xls, .xml, .txt ...) ([voir diapo](#))
- d'une url de téléchargement
- du presse-papier
- d'une base SQL
- d'un google sheet

Ouvrir un projet

- Accéder à ses projets
- Enrichir / modifier les métadonnées d'un projet
- Supprimer un projet

Importer un projet

Créer un nouveau projet à partir d'un projet précédemment exporté d'OpenRefine (pour exporter un projet voir [diapo](#))

Langues

Choix de la langue de l'interface (attention les traductions françaises ne sont pas toujours optimales)



OpenRefine

Un outil puissant pour travailler avec des données désordonnées.

Créer un projet

Ouvrir un projet

Importer un projet

Langues

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

Les documents de type TSV, CSV, *SV, Excel (.xls et .xlsx), JSON, XML, RDF en XML, OpenDocume

Récupérer les données à partir de

Cet ordinateur

Adresses web (URLs)

Presse-papier

Base de données

Google Data

Chercher un ou plusieurs fichiers à charger :

Parcourir...

Aucun fichier sélectionné.

Suivant »

Interface

Créer un projet à partir de cet ordinateur

Le plus souvent le format de fichier est détecté automatiquement par OpenRefine.

- Pour l'import d'un tableau type .xlsx ou .ods > peu de paramétrage
- Pour l'import d'un fichier avec séparateur (.tsv, .csv..) > sélectionner le séparateur, le format des caractères (l'encodage) (ex ci-dessous)
- Pour les fichiers à balises (xml, json) > sélectionner la balise racine que l'on souhaite importer (ci-contre)

Par défaut la pré-visualisation se met à jour automatiquement

Cliquer sur le premier élément XML correspondant à la première entrée à charger.

```

<ListRecords xmlns:dcterms="http://purl.org/dc/terms/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <record>
    <header>
      <identifier>oai:calames.abes.fr:Calames-202228122586701</identifier>
      <timestamp />
      <setSpec>751059801</setSpec>
    </header>
    <metadata>
      <oa1_dc>
        <dc:title xmlns:dc="http://purl.org/dc/elements/1.1/title">Archives de Paul Hazard.</dc:title>
        /ListRecords/record/metadata xmlns:dc="http://purl.org/dc/elements/1.1/identifier"> DA 1 - 2 DA 7 [cote]</dc:identif
        <dc:relation xmlns:dc="http://purl.org/dc/elements/1.1/relation">FR-751059801 [RCR établissement]</dc:r
        <dc:relation xmlns:dc="http://purl.org/dc/elements/1.1/relation">http://www.calames.abes.fr/pub/ms/Cal
        <dc:creator xmlns:dc="http://purl.org/dc/elements/1.1/creator">Hazard, Paul (1878-1944) [Auteur]</dc:cr
        <dc:subject xmlns:dc="http://purl.org/dc/elements/1.1/subject">Revue de littérature comparée (1921-....
        <dc:subject xmlns:dc="http://purl.org/dc/elements/1.1/subject">Littérature française</dc:subject>
        <dc:subject xmlns:dc="http://purl.org/dc/elements/1.1/subject">Littérature comparée</dc:subject>
        <dc:subject xmlns:dc="http://purl.org/dc/elements/1.1/subject">Hazard, Paul (1878-1944)</dc:subject>
        <dc:subject xmlns:dc="http://purl.org/dc/elements/1.1/subject">Bertault, Philippe (1879-1970)</dc:subje
        <dc:subject xmlns:dc="http://purl.org/dc/elements/1.1/subject">Futurisme</dc:subject>
        <dc:date xmlns:dc="http://purl.org/dc/elements/1.1/date">1890/1963 [date de l'ensemble]</dc:date>
  
```

Considérer les données

Considérer les données
comme

Fichiers CSV / TSV / séparateur

Fichiers texte à base de lignes

Fichiers texte à largeur de champ fixe

PC-Axis text files

Fichiers JSON

Fichiers MARC

Fichiers JSON-LD

Fichiers RDF/N3

Format des caractères

US-ASCII

Les colonnes sont séparées par

- une virgule (CSV)
 une tabulation (TSV)
 personnalisé

Utiliser le caractère " pour fermer les cellules contenant les séparateurs de colonnes

Supprimer les espaces de début et de fin

Protéger les caractères spéciaux avec \

Ignorer la ou les 0 première(s) ligne(s) du début du fichier

Analyser la ou 1 ligne(s) suivante(s) comme des entêtes de colonnes les

Noms de colonnes (séparés par des virgules)

Ignorer la ou les 0 première(s) ligne(s) de données

Charger au plus 0 première(s) ligne(s) de données

Mettre à jour l'aperçu

Désactiver l'aperçu automatique

Analyser le texte des cellules comme nombres

Conserver les lignes vides

Enregistrer les cellules vides comme des valeurs nulles

Indiquer la source du fichier

stocker le fichier d'archive

Interface

Le menu

Lorsqu'on a créé ou ouvert un projet, un menu en haut à droite apparait

Ouvrir... Exporter ▾ Aide



Exporter

Permet d'afficher les différents formats d'export.

Archive de projet OpenRefine permet d'exporter la totalité de votre projet (données + historique des traitements) pour le réimporter dans une autre instance OpenRefine

Ouvrir

Permet d'ouvrir un nouvel onglet pour ouvrir un projet en parallèle. C'est notamment utile lorsque l'on souhaite importer une colonne à partir d'un autre projet (voir [diapo](#))

Aide

Renvoie vers la documentation complète de l'outil.

Notez que la communauté OpenRefine est très active, vous pourrez donc également trouver de l'aide sur des forums

Interface

L'espace de travail

lignes / entrées

Un affichage « lignes » numérote chaque ligne et considère les lignes indépendamment les unes des autres.

Un affichage « entrée » se base sur la première colonne (l'identifiant unique) pour définir des « entrées » qui peuvent contenir plusieurs lignes (exemple ci-dessous).

Navigation dans les entrées

Vous avez la possibilité de paramétrer le nombre de résultats affichés et de naviguer dans les différentes pages.

Notez qu'OpenRefine est un outil pour effectuer des traitements en masse sur la totalité de vos entrées, vous n'avez donc pas besoin de toutes les voir. Pour vérifier si vos traitements ont bien fonctionné, vous pouvez utiliser les facettes pour afficher la liste des valeurs d'un champ

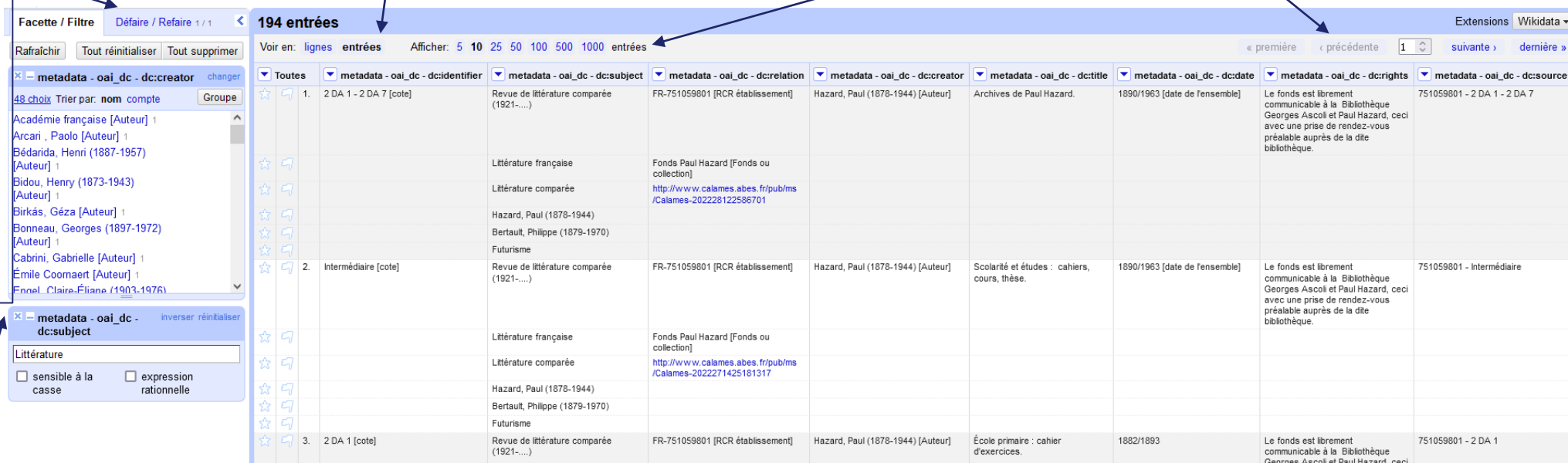
Defaire / Refaire

C'est l'historique des traitements effectués. Vous pouvez naviguer dans cet historique pour annuler ou réappliquer des traitements.

Vous pouvez également extraire les traitements effectués pour les appliquer à un autre jeu de données. A l'inverse vous pouvez appliquer des traitements réalisés sur un autre jeu (voir [diapo](#))

Facette / Filtre

Dans cet onglet vous pourrez paramétrer et supprimer vos facettes/filtres.



The screenshot shows the OpenRefine interface with 194 entries displayed in a table. The table has columns for various metadata fields: identifier, subject, relation, creator, title, date, rights, and source. The entries are grouped into three main categories: '2 DA 1 - 2 DA 7 [cote]', 'Intermédiaire [cote]', and '2 DA 1 [cote]'. The interface includes a top navigation bar with options for 'lignes' and 'entrées', a 'Facette / Filtre' sidebar on the left, and a 'Defaire / Refaire' panel on the right. The 'Facette / Filtre' panel shows a facet for 'metadata - oai_dc - dc:subject' with a list of subjects like 'Littérature' and 'Littérature comparée'. The 'Defaire / Refaire' panel shows a list of 48 actions, including 'Académie française [Auteur]', 'Arcari, Paolo [Auteur]', etc.

2

FILTRES, FACETTES ET TRI

Les filtres

Comment les utiliser ?

- Cliquez sur la colonne sur laquelle vous souhaitez appliquer un filtre > Filtrer le texte



- Votre filtre apparaît dans l'onglet « Facette / Filtre », saisissez la valeur souhaitée, le filtre s'applique en temps réel
- Pour supprimer votre filtre cliquez sur la croix



Facette / Filtre Défaire / Refaire 0 / 0

Rafraîchir Tout réinitialiser Tout supprimer

× Commune inverser réinitialiser

Aix

sensible à la casse expression rationnelle

3 matching entrées (271 total)

Voir en: [lignes](#) [entrées](#) Afficher:

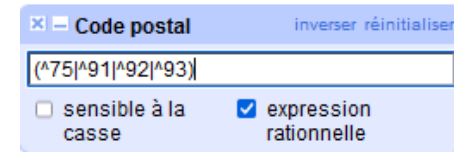
eture	Commune	Académie	Ré
	Aix-en-Provence	Aix-Marseille	Prover Alpes-d'Azur
	Roubaix	Lille	Hauts-France
	Aix-en-Provence	Aix-Marseille	Prover Alpes-d'Azur

Le filtre cherche la chaîne de caractère exacte ! Faites attention notamment quand vous filtrez avec des chiffres

Les filtres

Dans quel cas les utiliser ?

- **Définir une plage:** dans OpenRefine on applique des traitements en masse. Parfois on souhaite simplement traiter un sous-ensemble
- **Explorer les données :** bien que ce ne soit pas la fonction première d'OpenRefine, un filtre peu permette de savoir si une valeur est présente dans un champ ou non
- **Filtrer des valeurs inconnues :** Il est possible de construire un filtre à partir de REGEX (expression régulière). Par exemple le filtre «[^]01» sur un champ numéro de téléphone permettra de filtrer tous les numéros commençant pas l'indicateur « 01 ». Ce qui serait trop fastidieux à partir d'une facette (exemple ci contre)



Code postal inverser réinitialiser

sensible à la casse expression rationnelle

Ici un filtre sur le champ « Code postal » qui permet d'afficher les codes postaux commençant par 75, 91, 92 ou 93

- Cochez « **expression rationnelle** » si vous utilisez des opérateurs booléens ou des REGEX
- Pour utiliser l'opérateur booléens OR, il faut utiliser sa version informatique | (AltGr + 6)
- [^] est la REGEX utilisée pour « commence par »

ancienne région	Adresse	Code postal
-France	12 AVENUE LÉONARD DE VINCI	92918
-France	21 rue d'Assas	75270
-France	11 AVENUE DU TREMBLAY	75012
-France	254 boulevard Raspail	75014
-France	14 RUE BONAPARTE	75008
-France	31 RUE D ULM	75240
-France	34 quai d'Austerlitz	75013

Les facette

Comment les utiliser ?



- Cliquez sur la colonne sur laquelle vous souhaitez appliquer une facette > Facette > Sélectionnez le type de facette
- Les facettes fonctionnent avec les types de valeur. Par exemple pour appliquer une facette chronologique il faut un champ de type « date » (pour modifier le type d'un champ voir [diapo](#))
- Une fois la facette appliquée, elle s'affiche dans l'onglet « Facette /Filtre ». Vous pouvez alors sélectionner la ou les valeur(s) à filtrer
- Les facettes peuvent lister les valeurs de votre colonne (ex: facette textuelle à gauche) ou être booléennes (facette doublons à droite)



Les facettes

Dans quel cas les utiliser ?

- **Définir une plage**: dans OpenRefine on applique des traitements en masse. Parfois on souhaite simplement traiter un sous-ensemble ;
- **Explorer les données** : une facette permet d'afficher toutes les valeurs d'un champ. Ainsi, on peut facilement identifier si des valeurs ne sont pas normées (facette textuelle; exemple bas), s'il y a des doublons (facette doublons; exemple haut), des valeurs erronées..
- **Modifier des valeurs** : Vous pouvez éditer les facettes textuelles, pour corriger des coquilles ou fusionner des valeurs. Cela permet de modifier l'ensemble des enregistrements correspondants à la valeur d'un seul coup ;
- **Clusteriser** : La fonction « cluster » (« grouper » en français), permet d'identifier des valeurs similaires grâce à des algorithmes basés sur les chaînes de caractères. Cela pourra être utilisé pour normaliser les valeurs d'un champ. (voir [diapo](#))



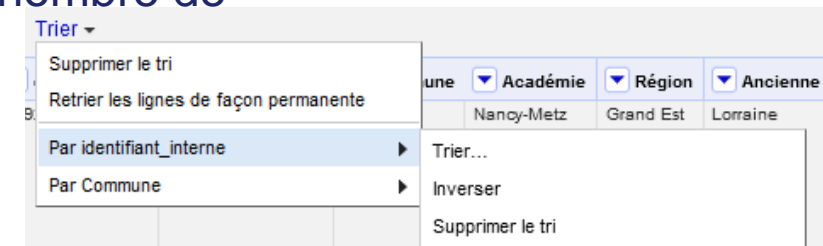
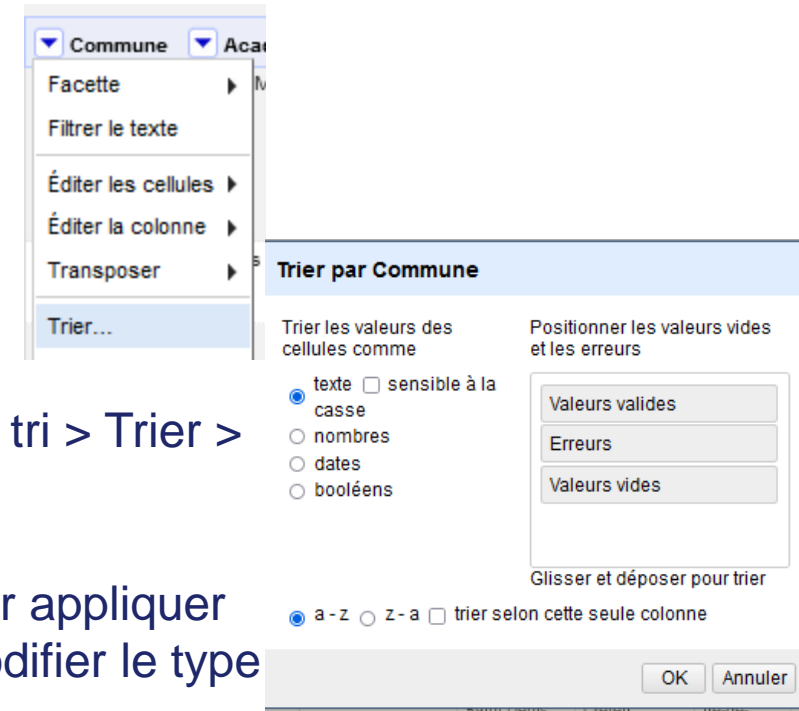
Toutes	identifiant_interne	Lib
☆	21.	15kv5
☆	287.	15kv5
☆	15.	59da6
☆	281.	59da6
☆	20.	5eUBS



Le tri

Comment l'utiliser ?

- Cliquez sur la colonne sur laquelle vous souhaitez appliquer un tri > Trier > Sélectionnez le type de tri
- Les tris fonctionnent avec les types de valeur. Par exemple pour appliquer un tri « nombre » il faut un champ de type « nombre » (pour modifier le type d'un champ voir [diapo](#))
- Une fois le tri effectué, un bouton « Trier », apparait à côté du nombre de résultats par page. Il permet de modifier ou supprimer un tri
- Si vous appliquez plusieurs tris, ils s'appliqueront dans l'ordre
- Par défaut un tri est temporaire, vous pouvez le supprimer. Mais vous pouvez l'appliquer de manière permanente, cela entrainera une renumérotation de vos lignes et vous ne pourrez plus le supprimer



Le tri

Dans quel cas l'utiliser ?

Au delà de la réorganisation des valeurs, le tri est surtout utilisé pour appliquer les deux traitements suivants :

- **Vider des valeurs répétées dans des cellules consécutives** : sur un champ trié, cette fonction permet de conserver uniquement la première occurrence d'une valeur et de supprimer toutes les autres. Ce qui est utile pour supprimer des doublons (exemple ci-contre) ou réorganiser son tableau
- **Recopier les valeurs dans les cellules vides consécutives** : si une valeur d'une entrée est attribuée à plusieurs lignes, vous pouvez la dupliquer sur chaque ligne de l'entrée

1

Facette / Filtre Défaire / Refaire 3 / 3 45 matching lignes (274 tc)

Rafrâichir Tout réinitialiser Tout supprimer

changer inverser réinitialiser

identifiant_interne

2 choix Trier par: nom compte

false 229

true 45

Facette par nombre de choix

2

identifiant_interne Libellé sigle ty

Facette

Filtrer le texte

Éditer les cellules

Éditer la colonne

Transposer

Trier...

Aperçu

Réconcilier

Transformer...

Transformations courantes

Recopier les valeurs dans les cellules vides consécutives

Vider les valeurs répétées dans des cellules consécutives

3

identifiant_interne	Libellé	sigle	type d'établissement
7.	Institut national du sport, de l'expertise et de la performance	INSEP	Grand établissement
8.	Institut national du sport, de l'expertise et de la performance	INSEP	Grand établissement
37.	École nationale supérieure des sciences de l'information et des bibliothèques	ENSSIB	Grand établissement
38.	École nationale supérieure des sciences de l'information et des bibliothèques	ENSSIB	Grand établissement

4

Toutes	identifiant_interne	Lib
7.	15kV5	Institut national du sport, de l'expert de la perform
37.	59da6	École

Après avoir effectué un tri alphabétique permanent sur la colonne `identifiant_interne` (voir diapo précédente) :

- j'applique une « facette doublons » pour afficher les doublons
- je vide les valeur répétées dans les cellules consécutives. Ce qui me permet de supprimer les valeurs en doublons.

Ensuite je n'ai plus qu'à supprimer les lignes (voir [diapo](#)) dont la valeur est nulle pour supprimer mes lignes en double.

3

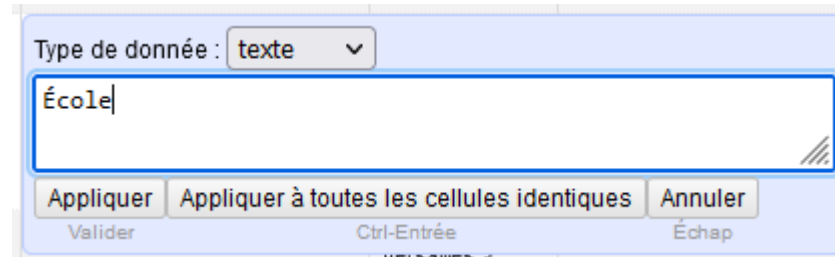
NETTOYER LES DONNÉES : LES FONCTIONS
BASQUES

Editer les cellules

La modification simple

Lorsque vous survolez une cellule avec votre souris, un bouton « edit » apparaît. Il vous permet de modifier la valeur, ou d'« Appliquer la modification à toutes les cellules identiques ». Cela modifiera toutes les cellules qui ont la même valeur.

	et commerciales		
	École supérieure	École	edit Pri









« Appliquer la modification à toutes les cellules identiques » ne s'appliquera que sur la plage sélectionnée. Aussi, pensez bien à supprimer vos facettes et vos filtres si vous souhaitez appliquer la modification à l'ensemble de votre projet.

Editer les cellules

Marquer des entrées

Dans la colonne « Toutes », au début de chaque ligne, vous pourrez marquer vos entrées avec des étoiles ou des drapeaux. Il s'agit de repères (par exemple en vue d'une suppression). Vous pourrez ensuite afficher les entrées marquées avec des facettes dédiées accessibles dans la colonne « Toutes » : Facette > Facette par étoile / Facette par marque

▼ Toutes	▼ identifiant_interne	▼ Lib
  1.	5YWUA	École supérieure des sciences économiques et commerciales
  2.	0347i	École supérieure d'ingénierie des travaux de la construction de Metz
  3.	yIVkq	École supérieure

▼ Toutes	▼ identifiant_interne	▼ Libellé	▼ ty
Transformer...		École supérieure des sciences économiques et	École
Modifier toutes les colonnes			
Facette		Facette par étoile	
Éditer les lignes		Facette par marque	

×	Lignes étoilées	changer	inverser	réinitialiser
2 choix	Trier par: nom	compte		
false	249		include	
true	1		exclude	
Facette par nombre de choix				

Editer les cellules

Les transformations courantes

- **Supprimer les espaces de début et de fin / rassembler les espaces consécutifs:** Ces fonctions permettent de supprimer les espaces indésirables. Quand vous fusionnez des cellules, il n'est pas rare que des espaces indésirables s'invitent dans vos cellules. Dans l'idéal effectuez « rassembler » puis « supprimer » les espaces avant de commencer à manipuler les données. Faites-le également après avoir terminé de manipuler vos données.
- **En nombre / En date / En texte:** Ces fonctions permettent de modifier le type des données. Elles sont notamment utiles si vous souhaitez utiliser des facettes ou des tris spécifiques à un type de données (ex : facette TimeLine). Attention la conversion ne fait pas tout, la valeur initiale devra déjà correspondre à un formalise particulier (format de date, pas d'espace dans les nombres..).
- **En valeur nulle:** Permet de supprimer les données concernées. Cela s'applique à la plage de données sélectionnées. Si aucun filtre ou aucune facette n'est sélectionné l'ensemble des valeurs de la colonne est supprimé.

Libellé	sigle	type d'établissement	secteur_d_etablissement	localisation
Facette		École	Privé	Grand Est > Nancy-Metz > Moselle > Metz

- Éditer les cellules ▶ Transformer...
- Éditer la colonne ▶ Transformations courantes ▶
 - Supprimer les espaces de début et de fin
 - Rassembler les espaces consécutifs
 - Convertir les entités HTML
 - Remplacer les guillemets courbés par des guillemets droits
 - En initiales majuscules (en capitales)
 - En majuscules
 - En minuscules
 - En nombre
 - En date
 - En texte
 - En valeurs nulles
 - Transformer en chaîne vide
- Transposer ▶
- Trier... ▶
- Aperçu ▶

Ces modification peuvent être appliquées à plusieurs colonnes si elles sont effectuées à partir de la colonne « Toutes » (voir [diapo](#))

Editer les cellules

Remplacer

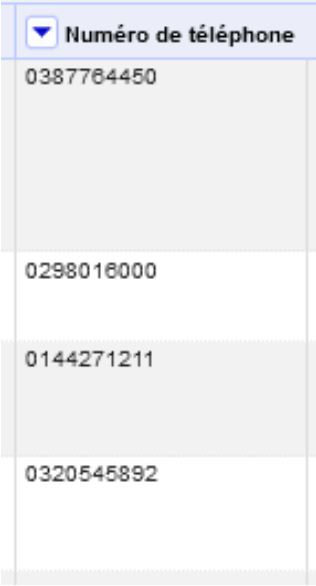
Cette fonctionnalité permet de remplacer une chaîne de caractères par une autre.

Notez qu'elle fonctionne avec n'importe quel caractère y compris les espaces.

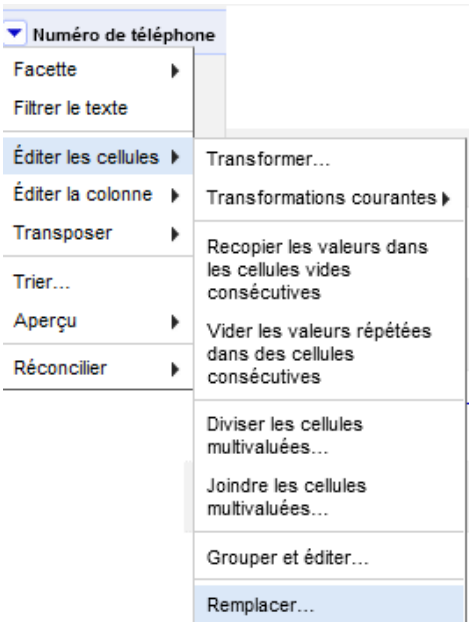
Vous avez la possibilité de remplacer des caractères par une valeur nulle, ce qui aura pour effet de supprimer le caractère ou la chaîne de caractères sélectionnée.

Vous pouvez élégamment utiliser des REGEX pour identifier des chaînes de caractères à remplacer (exemple ci-contre)

1



2



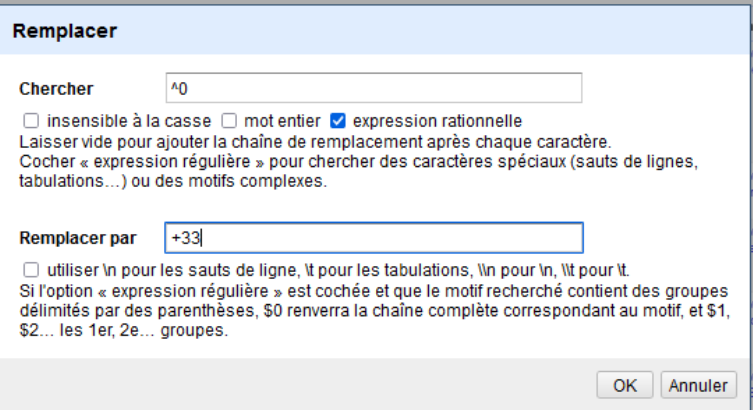
Numéro de téléphone

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

- Transformer...
- Transformations courantes
- Recopier les valeurs dans les cellules vides consécutives
- Vider les valeurs répétées dans des cellules consécutives
- Diviser les cellules multivaluées...
- Joindre les cellules multivaluées...
- Grouper et éditer...
- Remplacer...

Ici un exemple avec une REGEX pour remplacer les 0 en début de ligne par +33 pour modifier les numéros de téléphone en ajoutant l'indicatif français +33

3



Remplacer

Chercher

insensible à la casse mot entier expression rationnelle

Laisser vide pour ajouter la chaîne de remplacement après chaque caractère.
Cocher « expression régulière » pour chercher des caractères spéciaux (sauts de lignes, tabulations...) ou des motifs complexes.


Remplacer par

utiliser \n pour les sauts de ligne, \t pour les tabulations, \n pour \n, \t pour \t.

Si l'option « expression régulière » est cochée et que le motif recherché contient des groupes délimités par des parenthèses, \$0 renverra la chaîne complète correspondant au motif, et \$1, \$2... les 1er, 2e... groupes.

OK Annuler

4



Numéro de téléphone

+33 387764450

+33 298016000

+33 144271211

Editer les cellules

Grouper et éditer (clusteriser)

Grouper et éditer permet d'identifier des valeurs similaires qui ne seraient pas orthographiées de la même manière. Cette fonctionnalité est très utile pour normaliser les valeurs d'un champ. Il existe différents algorithmes qui permettent d'identifier des doublons potentiels.

Cette fonctionnalité est aussi accessible via les facettes (voir [diapo](#))

Grouper et Éditer une colonne "Commune"

Trouvez des groupes de différentes valeurs de cellule qui pourraient être d'autres représentations de la même chose. Par exemple, "New York" et "new york" font probablement référence au même concept et ne diffèrent que par la capitalisation, et "Gödel" et "Godel" font probablement référence à la même personne.
[En savoir plus...](#)

Méthode **Collision de clés** Fonction de codage **Empreinte** 12 grappes trouvée

Taille du groupe	Nombre de ligne	Valeurs dans le groupe	Fusionner ?	Nouvelle valeur dans la cellule
3	19	<ul style="list-style-type: none">Toulouse (16 lignes)Toulouse (2 lignes)TOULOUSE	<input type="checkbox"/>	Toulouse
2	4	<ul style="list-style-type: none">Nantes (3 lignes)NANTES	<input type="checkbox"/>	Nantes
2	2	<ul style="list-style-type: none">CompiègneCompiègne	<input type="checkbox"/>	Compiègne
2	5	<ul style="list-style-type: none">Villeurbanne (3 lignes)villeurbanne (2 lignes)	<input type="checkbox"/>	Villeurbanne
2	7	<ul style="list-style-type: none">Paris 6e (6 lignes)PARIS 6E	<input type="checkbox"/>	Paris 6e

Choix dans le groupe

Lignes dans le groupe

Longueur moyenne des choix

Variabilité de la longueur des choix

Tout sélectionner Enlever toutes les sélections Exporter les groupes Fusionner la sélection & regrouper Fusionner la sélection & fermer Fermer

Editer les cellules / Transposer

Les cellules multivaluées

- **Joindre les cellules multivaluées** : Pour une même entrée, lorsqu'un champ a plusieurs valeurs sur plusieurs lignes, cette fonction permet de regrouper les valeurs avec un séparateur sur une seule ligne.
- **Diviser les cellules multivaluées** : Sur la base d'un séparateur, vous créez autant de lignes que de valeurs présentes dans votre cellule initiale.
- **Transposer les cellules de plusieurs colonnes en ligne** : permet de regrouper des cellules de plusieurs colonnes dans une seule colonne avec des cellules multivaluées.
- **Transposer les cellules en colonne séparée** : Pour une même entrée, lorsqu'un champ a plusieurs valeurs sur plusieurs lignes, cette fonction permet de transposer les valeurs dans des colonnes. Attention vous devrez définir en amont le nombre de colonne, il faut que chaque entrée ait le même nombre de valeurs, ou connaitre l'entrée qui a le nombre de valeur le plus élevé.
- **Convertir en liste des colonne de clé/valeur** : Permet de créer des colonnes sur la base d'une liste de valeurs d'un champ, et d'alimenter ces colonnes avec la valeur d'un autre champ

Les modifications groupées sur l'ensemble des colonnes

La colonne « Toutes » permet de travailler sur l'ensemble des colonnes

- **Modifier toutes les colonnes** : faire des transformations courantes ([voir diapo](#)) sur plusieurs colonnes
- **Editer les lignes** :
 - Marquer / Etoiler : permet d'ajouter / de supprimer des étoiles ou des drapeaux pour la sélection
 - Supprimer les lignes correspondantes : permet de supprimer la sélection
- **Facettes** :
 - Par étoile / par marque : permet d'afficher les entrées marquées (voir ci-dessus et [diapo](#))
 - Par valeur vide : permet d'afficher les lignes entièrement vides
 - Valeurs / Entrées vides/non vides par colonne : permet d'afficher une facette avec chaque nom de colonne, qui affichera les cellules / entrées vides pour chaque colonne.
- **Editer les colonnes** :
 - Retrier / Supprimer : permet de réordonner ou supprimer des colonnes
 - Recopier / vider les valeurs dans les cellules consécutives (voir [diapo](#))

Editer les colonnes

Renommer, supprimer, déplacer des colonnes

- **Renommer cette colonne** : permet de renommer la colonne concernée. Notez que contrairement à un tableur, vous ne pouvez pas nommer deux colonnes de la même manière. Si vous avez prévu d'utiliser des formules, utilisez un nommage simple et explicite.
- **Supprimer cette colonne**
- **Déplacer la colonne...** : permet modifier la position d'une colonne. Notez qu'il est plus simple d'utiliser la colonne « Toutes » pour travailler sur la réorganisation et la suppression des colonnes (voir [diapo](#))
- **Ajouter un colonne ?** OpenRefine ne permet pas d'ajouter de nouvelles colonnes vierges. Vous pouvez le faire de manière détournée en utilisant la fonctionnalité « Ajouter une colonne en fonction de cette colonne » sur une colonne lambda et en remplaçant « value » par des doubles guillemets.

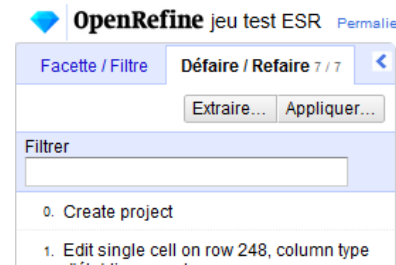
Editer les colonnes

Joindre et diviser des colonnes

- **Diviser en plusieurs colonnes** : Vous pouvez diviser votre colonne sur la base d'un séparateur ou d'une longueur. Notez que contrairement à un tableur classique :
 - Vous pouvez utiliser un séparateur de plusieurs caractères, ou basé sur une expression régulière.
 - La division créer de nouvelles colonnes, elle ne remplace pas les valeurs dans des colonnes déjà existantes
- **Joindre de colonnes** : permet de fusionner plusieurs colonnes sur la base d'un ordre et d'un séparateur. Vous pouvez ou non prendre en compte les valeurs nulles, et ajouter le résultat dans la colonne active ou dans une nouvelle colonne. Pour des jonctions plus complexes vous pouvez aussi utiliser une formule (voir [diapo](#))

Extraire et appliquer des traitements

- Vous avez la possibilité d'extraire les traitements réalisés au format JSON. Ne seront extraits que les traitements de masse (ex : si vous faites de l'édition simple sur une cellule cela n'apparaîtra pas dans le fichier JSON)



- Pour cela rendez-vous dans votre historique (Défaire/Refaire) et cliquez sur « Extraire »

- Vous pouvez sélectionner les traitements à exporter

Extraire l'historique des opérations

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.

- Edit single cell on row 248, column type d'établissement
- Reorder rows
- Blank down cells in column identifiant_interne
- Blank down cells in column identifiant_interne
- Text transform on cells in column Numéro de téléphone using expression value.replace(/^0/, "+33 ")
- Create column Adresse complete at index 23 based on column Adresse using expression grelvalue+" "+cells["Commune"].value+" "+cells["Code postal"].value
- Edit single cell on row 2, column Adresse complete

```
{
  "op": "core/text-transform",
  "engineConfig": {
    "facets": [ ]
  },
  "mode": "record-based"
},
{
  "columnName": "Numéro de téléphone",
  "expression": "value.replace(/^0/, "+33 ")",
  "onError": "keep-original",
  "repeat": false,
  "repeatCount": 10,
  "description": "Text transform on cells in column"
},
{
  "op": "core/column-addition",
  "engineConfig": {
    "facets": [ ]
  }
}
```

- Vous pourrez ensuite réappliquer ces traitements sur un tableau qui a une structure similaire en collant le JSON dans « Appliquer »

4

INTRODUCTION A L'ÉDITEUR DE FORMULE GREL

Les formules dans OpenRefine

OpenRefine fonctionne avec un langage qui lui est propre, le GREL « Google/General Refine Expression Language ».

Pour les cas les plus complexes il est possible d'utiliser Jython ou Clojure.

Notez que les fonctionnalités de base sont en grande partie disponibles via les menus « Editer les cellules » et « Editer les colonnes »

Pour accéder à l'éditeur de formules GREL :

- **Editer les cellules > Transformer** : permet d'appliquer des modifications en masse dans la colonne sélectionnée
- **Editer la colonne > Ajouter une colonne en fonction de cette colonne** : permet d'appliquer les modifications en masse dans une nouvelle colonne, ainsi la colonne sélectionnée n'est pas modifiée

L'éditeur de formule GREL

Interface

Historique: Historique des formules utilisées au cours du projet. Un bouton « Réutiliser » permet de copier la formule dans l'interface de saisie. L'Etoile permet d'identifier des formules « favorites »

Interface de saisie :
Interface de saisie de la formule. On vous indique en bout de ligne si la syntaxe est correcte ou non

L'aperçu: A gauche il s'agit du contenu actuel de la colonne, et droite du contenu une fois la formule appliquée

Transformation textuelle personnalisée sur la colonne localisation

Expression Langue General Refine Expression Language (GREL) ▼

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

row	value	replace(value," > ","_")
1.	Île-de-France > Versailles > Val-d'Oise > Cergy	Île-de-France_Versailles_Val-d'Oise_Cergy
2.	Grand Est > Nancy-Metz > Moselle > Metz	Grand_Est_Nancy-Metz_Moselle_Metz
3.	Île-de-France > Versailles > Hauts-de-Seine > Courbevoie	Île-de-France_Versailles_Hauts-de-Seine_Courbevoie
4.	Auvergne-Rhône-Alpes > Lyon > Rhône > Villeurbanne	Auvergne-Rhône-Alpes_Lyon_Rhône_Villeurbanne

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

OK Annuler

Aperçu Historique Étoilée Aide

- ☆ Réutiliser This project grel: value.parseXml().select('rec
- ☆ Réutiliser This project grel: value.parseHtml().select('rec
- ☆ Réutiliser This project grel: value.parseHtml().select('rec
- ★ Réutiliser This project grel: "https://www.idref.fr/"+value+
- ☆ Réutiliser This project grel: "026916886"
- ☆ Réutiliser This project grel: value
- ☆ Réutiliser This project grel: replace(value,"json","xml")

Aperçu Historique Étoilée Aide

Expression

Supprimer Reuse grel: "https://www.idref.fr/"+value+".xml"

Supprimer Reuse grel: value.match(/.*\((.*)\).*/)

Etoilée: Permet d'accéder à vos formules favorites

Quelques notions de bases

La concaténation

L'une des fonctions de base pour les chaînes de caractères est la concaténation. Pour cela vous articulez vos « valeur » initiales avec d'autres chaînes de caractères grâce à des « + » :

"**valeur à ajouter**" + valeur

Exemple :

"**https://idref.fr/**" + valeur

Pour générer une URL à partir d'un identifiant IDREF. Il faut ajouter « https://idref.fr/ » devant l'identifiant.

Pas d'erreur de syntaxe.

Aperçu			
	Historique	Étoilée	Aide
row	value	\"https://idref.fr/\"+value	
1.	028029429	https://idref.fr/028029429	
4.	026402823	https://idref.fr/026402823	
7.	034817670	https://idref.fr/034817670	
8.	026453932	https://idref.fr/026453932	

Quelques notions de bases

Remplacer le contenu d'une cellule par une cellule d'une autre colonne

Cette formule permet de transférer les valeurs d'une colonne dans une autre colonne

cells["**colonne à transférer**"].value

Exemple :

cells["**test_ID**"].value

Si on applique cette formule sur une colonne, le contenu de chaque cellule de cette colonne sera remplacé par le contenu de la cellule de la colonne "test_ID"

Transformation textuelle personnalisée sur la colonne Adresse

Expression Langue Pas d'erreur de syntaxe.

Aperçu Historique Étoilée Aide

row	value	value+", "+cells[\"Commune\"].va ...
1.	AVENUE BERNARD HIRSCH	AVENUE BERNARD HIRSCH, Cergy, 95021
2.	6 rue Marconi	6 rue Marconi, Metz, 57070
3.	12 AVENUE LÉONARD DE VINCI	12 AVENUE LÉONARD DE VINCI, Courbevoie, 92916
4.	43 boulevard du 11 Novembre 1918	43 boulevard du 11 Novembre 1918, Villeurbanne, 69622

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

Un second exemple où l'on reconstitue une cellule à partir de cellules existantes. Ici on reconstitue une adresse postale complète en ajoutant le contenu de la colonne "Commune" puis de la colonne "Code postal" à la colonne "Adresse" initiale. Le tout séparé par des virgules

value+", "+cells[\"Commune\"].value+", "+cells[\"Code postal\"].value

Quelques notions de bases

Extraire une partie d'une cellule (division)

Avec cette formule vous choisissez un séparateur pour diviser votre cellule et vous sélectionnez la partie que vous souhaitez conserver. La numérotation des positions commence à 0 (pour sélectionner le premier segment il faut choisir la position « 0 »). Vous pouvez aussi utiliser une numérotation inversée (pour sélectionner le premier segment en partant de la fin il faut choisir la position « -1 ») :

`value.split("Séparateur")[position de la chaine à conserver].toString()`

Exemple :

`value.split("930")[-1].toString()`

Expression Langue

`value.split("930")[-1].toString()`

Pas c

	Aperçu	Historique	Étoilée	Aide
4.	912+Aax 003 https://www.sudoc.fr/00292434X; 930 ##\$b751082101\$aKS-195\$ju;			
5.	879+Aax 003 https://www.sudoc.fr/00320927X; 930 ##\$b751082101\$dma\$aQG 344\$ju; 930 ##\$b751065210\$a709.02 HUB\$ju; 930 ##\$b751182102\$dus\$a7.033.1 HUB emp\$ju\$kW 44\$2CDU; 930 ##\$b751065210\$a4 COL 10-13\$ju; 930 ##\$b751065210\$a4 GE 250\$ju;			

```

0 879+Aax 003 https://www.sudoc.fr/00320927X;
  930 ##$b751082101$dma$aQG 344$ju; 930
2 ##$b751065210$a709.02 HUB$ju; 930
3 ##$b751182102$dus$a7.033.1 HUB
4 emp$ju$kW 44$2CDU; 930
5 ##$b751065210$a4 COL 10-13$ju; 930
6 ##$b751065210$a4 GE 250$ju;
  
```

OU

```

-7 879+Aax 003 https://www.sudoc.fr/00320927X;
   930 ##$b751082101$dma$aQG 344$ju; 930
-5 ##$b751065210$a709.02 HUB$ju; 930
-4 ##$b751182102$dus$a7.033.1 HUB
-3 emp$ju$kW 44$2CDU; 930
-2 ##$b751065210$a4 COL 10-13$ju; 930
-1 ##$b751065210$a4 GE 250$ju;
  
```


Quelques notions de bases

Extraire une partie d'une cellule (REGEX)

OpenRefine vous permet d'extraire une chaîne de caractère d'une cellule avec la formule `value.find()`. Pour cela vous devez connaître la structure de la chaîne de caractères à exporter et l'indiquer via une expression régulière. Notez que dans une formule vous devez mettre les REGEX entre slash (/). Par défaut les résultats de cette formule sont stockés dans une liste, vous pouvez convertir ce résultat en chaîne de caractères en ajoutant `.toString()` à la fin de votre formule

`value.find(/REGEX de la chaîne de caractère à extraire/).toString()`

Exemple :

`value.find(/\\d{5}/).toString()`

Ici un exemple pour récupérer les suites de 5 chiffres dans un champ « Adresse complète ». Ce qui permet de récupérer les codes postaux.

- `\\d` est la REGEX pour les chiffres
- `{5}` signifie 5 chiffres successifs

Transformation textuelle personnalisée sur la colonne Adresse complète

Expression Langue General Refine Expression Language (GREL)

`value.find(/\\d{5}/).toString()` Pas d'erreur de syntaxe.

Aperçu Historique Étoilée Aide

row	value	value.find(/\\d{5}/).toString()
1.	6 rue Marconi, Metz, 57070	[57070]
2.	3 rue des Archives, Brest, 29238	[29238]
3.	11 place Marcelin Berthelot, Paris 5e, 75231	[75231]
4.	3 rue de la Digue, Lille, 59800	[59800]
5.	2 avenue DE PROVENCE, Brest, 29238	[29238]
6.	20 rue Ampère, Saint-Denis, 93200	[93200]

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

OK Annuler

Quelques notions de bases

Le croisement de colonnes

OpenRefine vous permet d'importer des colonnes d'un autre projet OpenRefine sur la base d'une valeur pivot grâce à la fonction « cross » avec la formule suivante :

```
cell.cross("Nom du projet duquel on souhaite importer la colonne", "colonne pivot").cells["colonne à importer"].value[0]
```

Exemple :

```
cell.cross("Test_ID_OpenRefine", "hal").cells["File_HAL"].value[0]
```

```
cell.cross("Test_ID_OpenRefine", "hal").cells["File_HAL"].value[0]
```

Pas d'erreur de syntaxe.

row	value	cell.cross("Test_ID_OpenRefine ...
1.	https://hal-essec.archives-ouvertes.fr/	458
4.	https://hal-univ-lyon1.archives-ouvertes.fr/	83382

De la documentation complète sur GREL

- Sur le site d'OpenRefine : <https://openrefine.org/docs/manual/grelfunctions>
- Mathieu Saby – Mémo : Programmer dans Openrefine avec GREL, 2019 : <https://fr.slideshare.net/27point7/programmer-dans-openrefine-avec-grel>
- Les tutoriels vidéo du Réseau Bases de Données : <https://www.canal-u.tv/chaines/rbdd/tes-premiers-pas-avec-openrefine-0>
- Lancez-vous, testez, recherchez des solutions sur des forums !

Nous Contacter



labrador@sorbonne-universite.fr

MERCI

Cellule données et humanités numériques
labrador@sorbonne-universite.fr



BIBLIOTHÈQUE
UNIVERSITAIRE



Sauf mention contraire, cette présentation est mise
à disposition selon les termes de la Licence
Creative Commons Attribution 2.0 France.
Icônes : freepik