Nettoyer ses données avec OpenRefine (niveau1)

Aurélien Moisan

NETTOYER SES DONNÉES AVEC OPENREFINE (NIVEAU 1)

AURÉLIEN MOISAN
21 MAI 2024

LabRADoR
LAB DE RESSOURCES ET D'ACCOMPAGNEMENT AUX DONNÉES DE LA RECHERCHE
SORBONNE UNIVERSITÉ

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

Sommaire

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

1

INTRODUCTION

OpenRefine
Qu’est-ce que c’est ?

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

OpenRefine

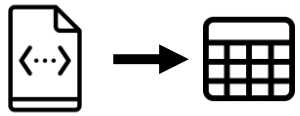
OpenRefine est un outil gratuit et open source qui permet de nettoyer, transformer, convertir, enrichir des données.

Pour plus d’informations rendez-vous sur : https://openrefine.org /

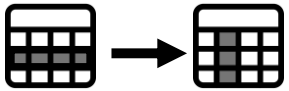
OpenRefine permet de :

Nettoyer un jeu de données

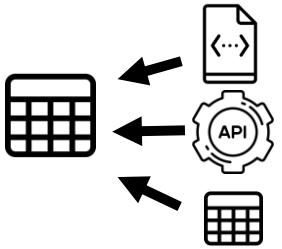
Convertir un jeu de données



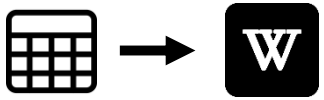
Transformer un jeu de données



Enrichir un jeu de données



Exposer des données sur Wikidata



Crédits icones : Pixel perfect; alfanz; Soni Sokell; Freepik

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 maitriser 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 stéoroïde » mais …

… ce n’est pas un tableur

SORBONNE UNIVERSITÉ
BIBLIOTHÈQUE UNIVERSITAIRE

OpenRefine ne permet pas :

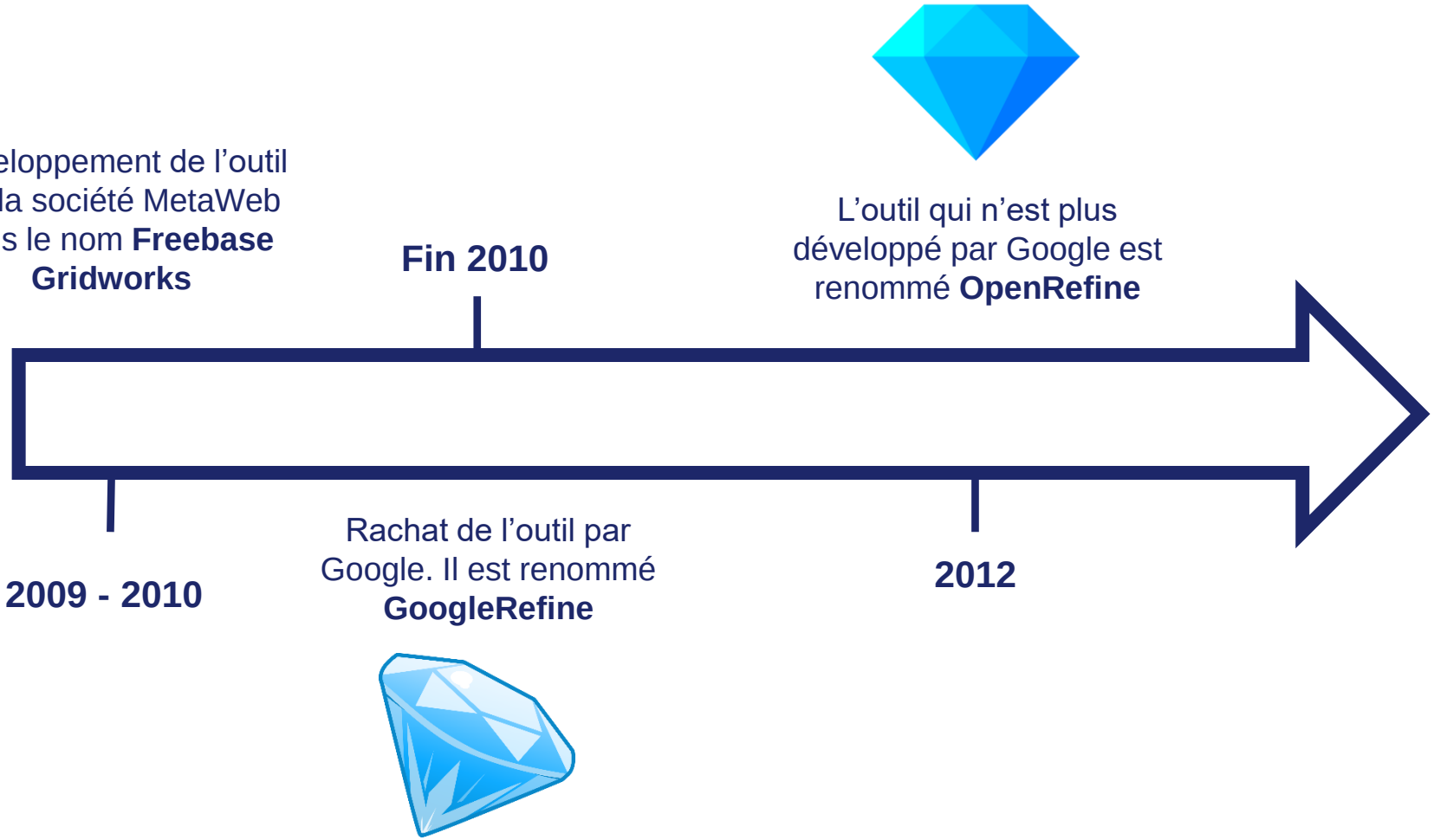
- de visualiser les données sous forme graphique
- le travail collaboratif

Il n’est pas optimisé pour :

- saisir des données
- réaliser des calculs

≠

Historique de l’outil



SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

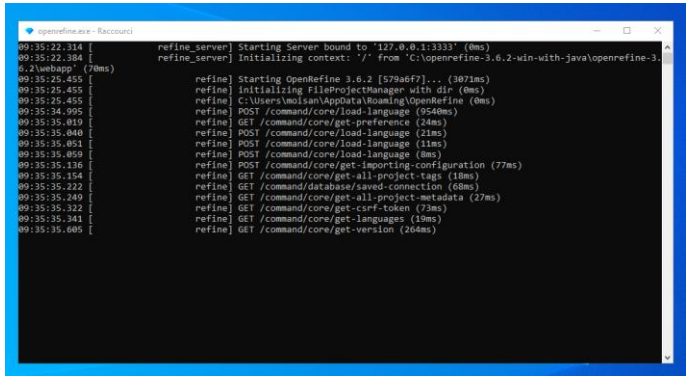
1

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.



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

Paramétrage

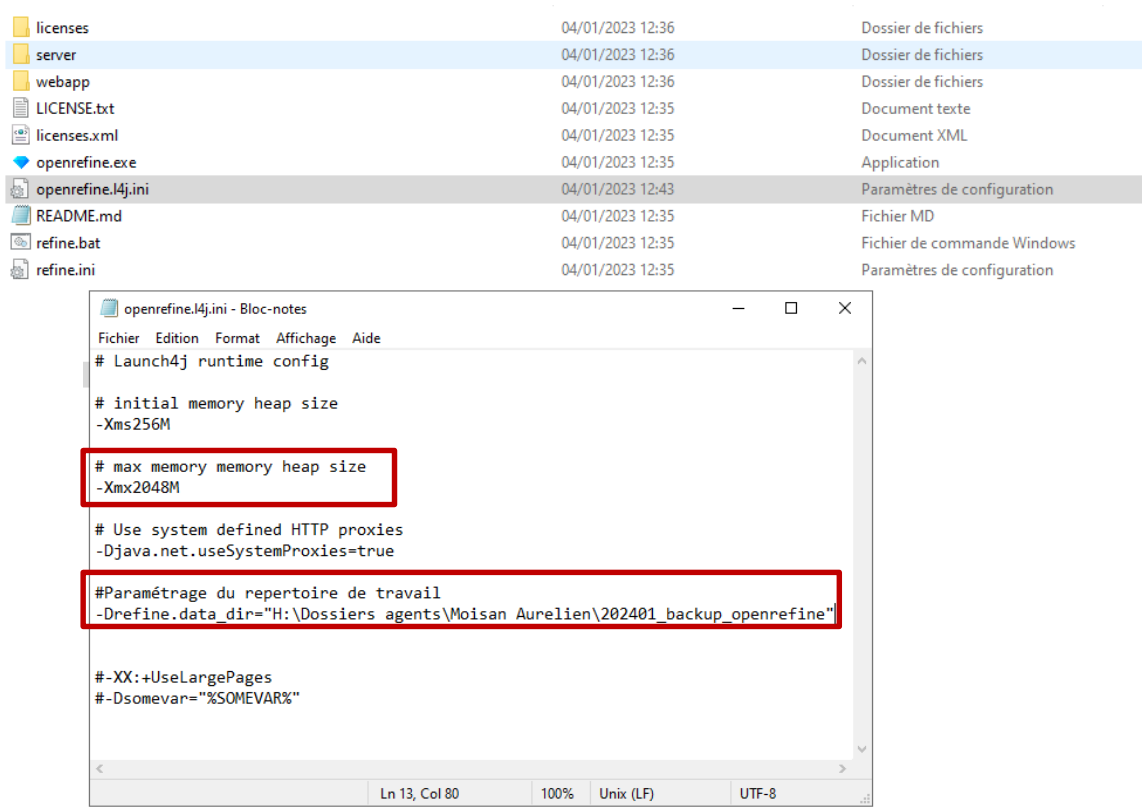
Si vous souhaitez modifier les paramètres d'OpenRefine, ouvrez le fichier openrefine.l4j.ini avec un éditeur de texte (notepad, gedit etc..); Pour :

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

max memory memory heap size
-Xmx**1024**M

- Changer le répertoire de travail, ajoutez une ligne :

-Drefine.data_dir=**Chemin absolu du repertoire**



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.

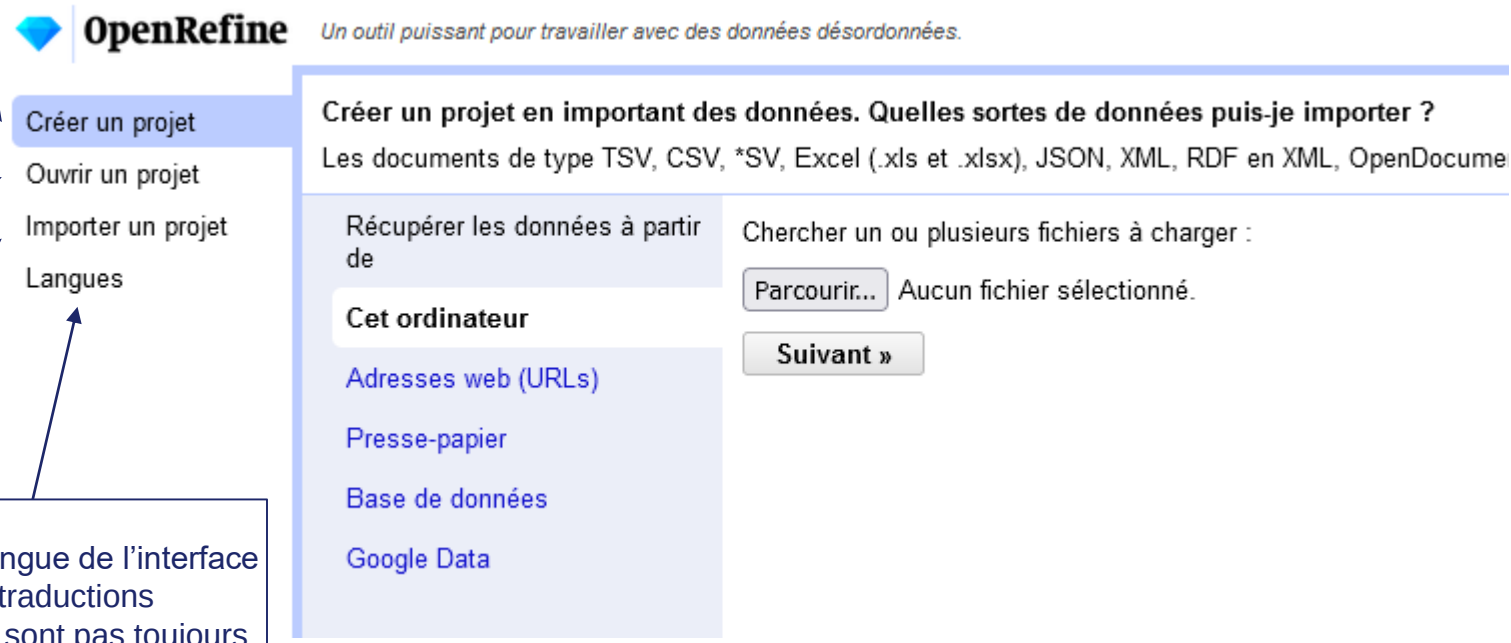
Interface
Gérer ses projets

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)

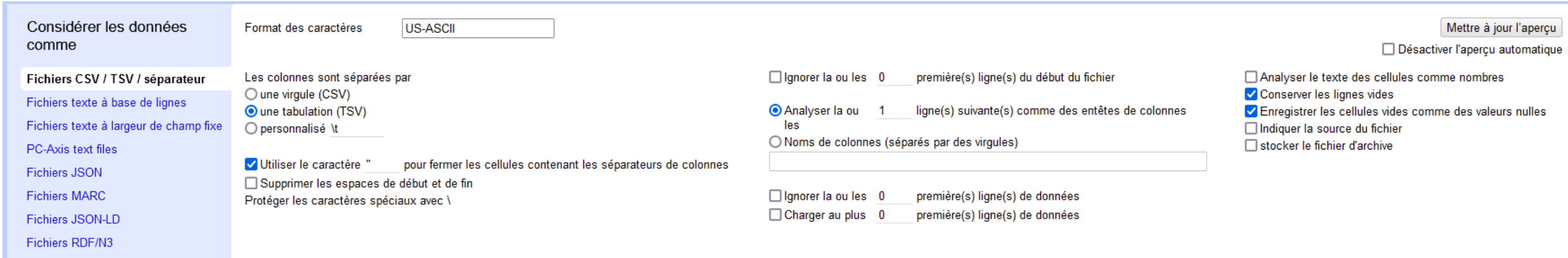


Interface
Créer un projet à partir de cet ordinateur

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

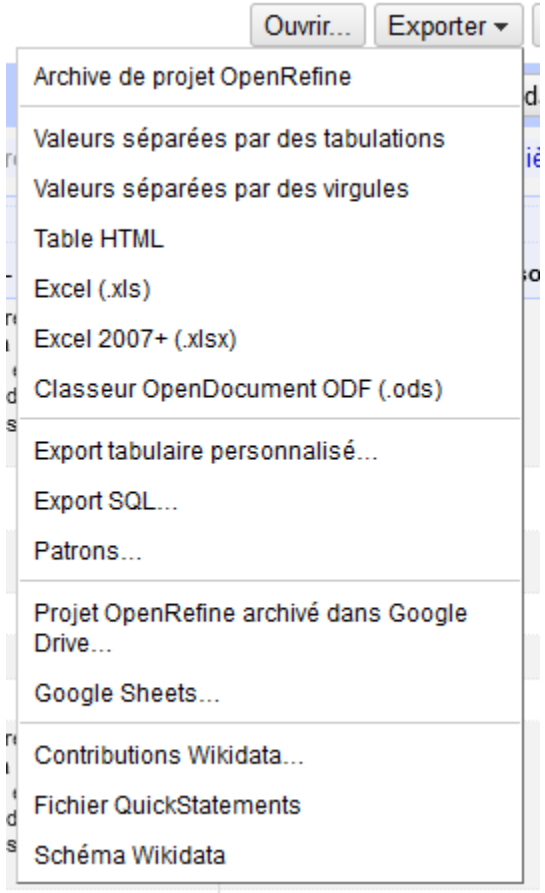
- Pour l'import d'un tableau type .xlsx ou .ods > peu de paramétrage
- Pour l'import d'un fichier avec séparateur (.tsv, .csv..) > sélectionner le séparateur, le format des caractères (l'encodage) (ex ci-dessous)
- Pour les fichiers à balises (xml, json) > sélectionner la balise racine que l'on souhaite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

Par défaut la pré-visualisation se met à jour automatiquement



Interface
Le menu

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



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
Renvoi 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

SORBONNE UNIVERSITÉ
BIBLIOTHÈQUE UNIVERSITAIRE

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.

lignes / entrées

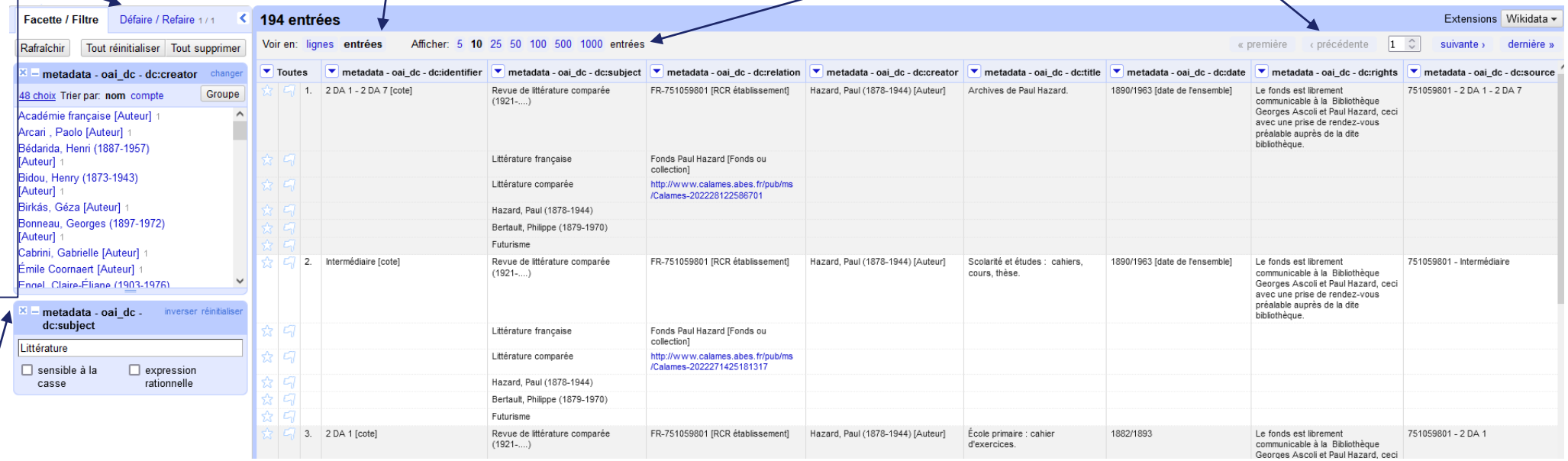
Un affichage « lignes » numérote chaque ligne et considère les lignes indépendamment les unes des autres.

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



SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

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 apparait dans l'onglet « Facette / Filtre », saisissez la valeur souhaitée, le filtre s'applique en temps réel

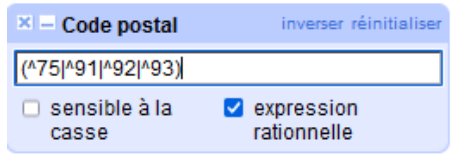
- Pour supprimez votre filtre cliquez sur la croix

Le filtre cherche la chaine 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)



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 »*

ncienne région	Adresse	Code postal
-France	12 AVENUE LÉONARD DE VINCI	92916
-France	21 rue d'Assas	75270
-France	11 AVENUE DU TREMBLAY	75012
-France	254 boulevard Raspail	75014
-France	14 RUE BONAPARTE	75006
-France	31 RUE D ULM	75240
-France	34 quai d'Austerlitz	75013

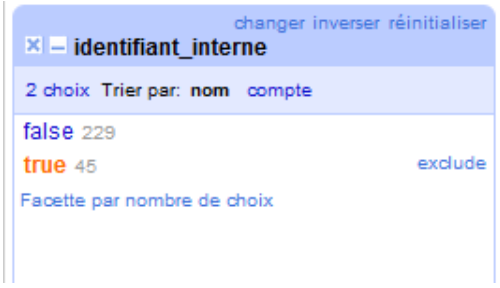
Les facette
Comment les utiliser ?

- Cliquez sur la colonne sur laquelle vous souhaitez appliquer une facette > Facette > Sélectionnez le type de facette

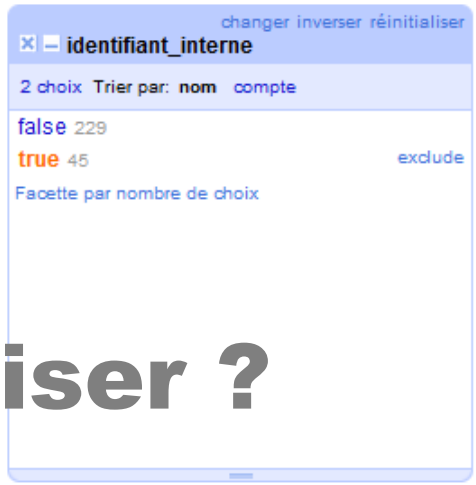
- Les facettes fonctionnent avec les types de valeur. Par exemple pour appliquer une facette chronologique il faut un champ de type « date » (pour modifier le type d'un champ voir diapo)

- 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)

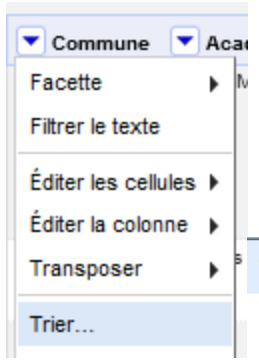


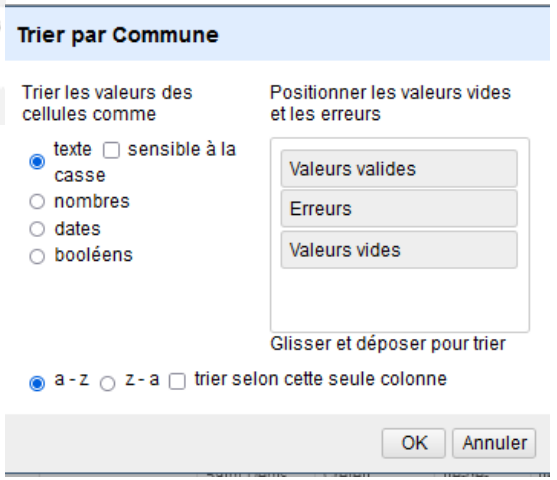
Les facettes
Dans quel cas les utiliser ?



- **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 chaines de caractères. Cela pourra être utilisé pour normaliser les valeurs d'un champ. (voir diapo)

Le tri
Comment l'utiliser ?

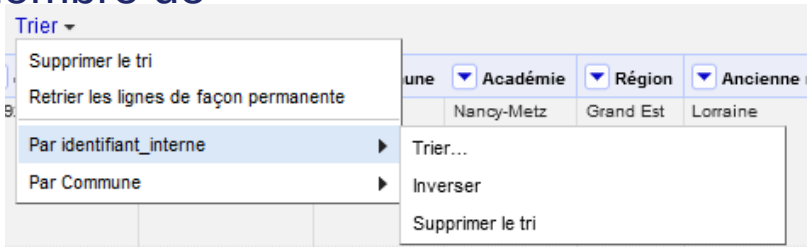




- Cliquez sur la colonne sur laquelle vous souhaitez appliquer un tri > Trier > Sélectionnez le type de tri

- 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

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.

SORBONNE UNIVERSITÉ

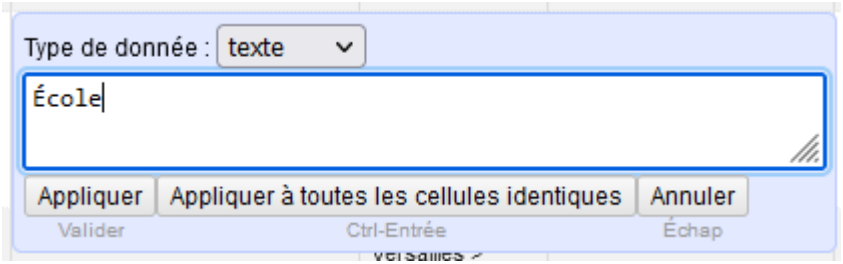
BIBLIOTHÈQUE UNIVERSITAIRE

3

NETTOYER LES DONNÉES : LES FONCTIONS BASIQUES

Editer les cellules
La modification simple

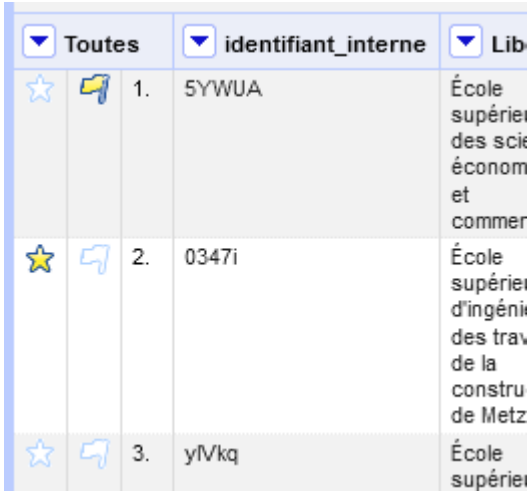
Lorsque vous survolez une cellule avec votre souris, un bouton « edit » apparait. Il vous permet de modifier la valeur, ou d' « Appliquer la modification à toutes les cellules identiques ». Cela modifiera toutes les cellules qui ont la même valeur.

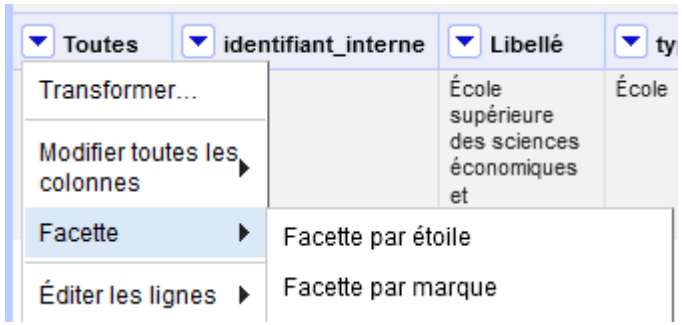


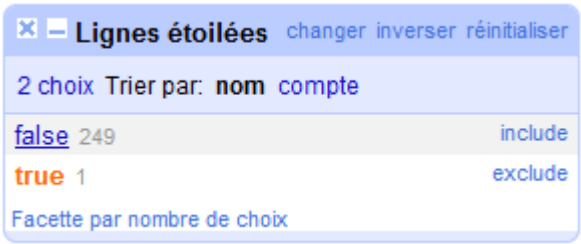
« Appliquer la modification à toutes les cellules identiques » ne s'appliquera que sur la plage sélectionnée. Aussi, pensez bien à supprimer vos facettes et vos filtres si vous souhaitez appliquer la modification à l'ensemble de votre projet.

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 colonnes « Toutes » : Facette > Facette par étoile / Facette par marque

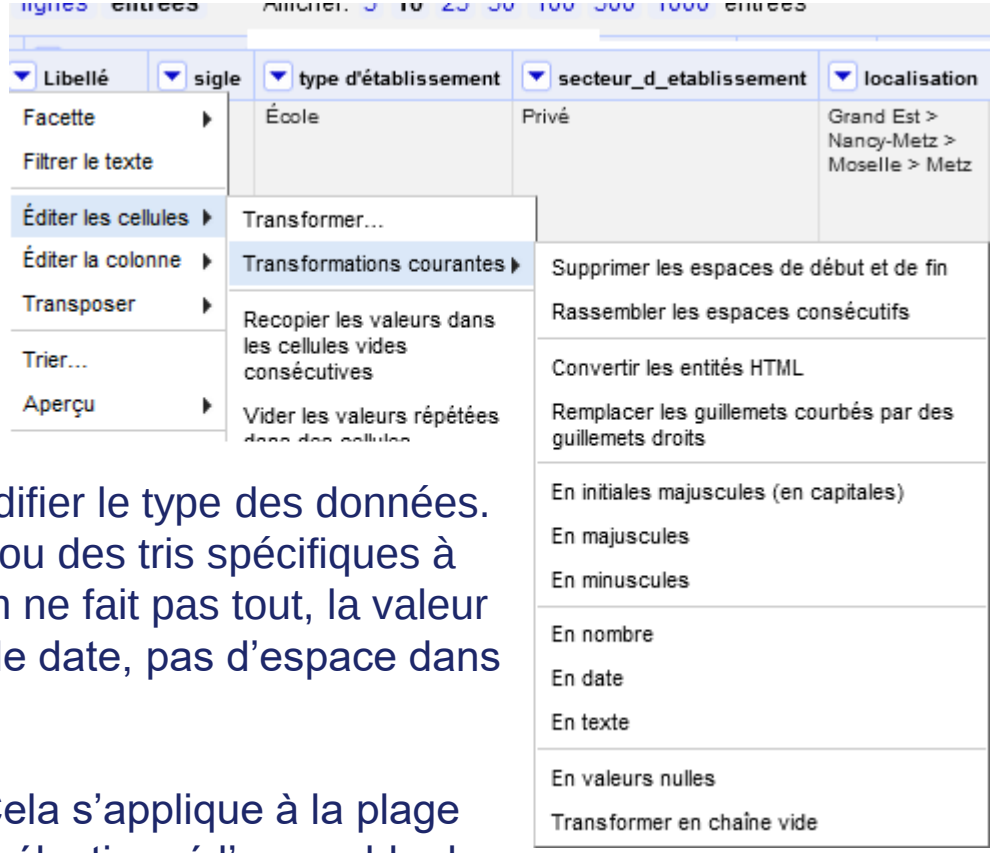






Editer les cellules
Les transformations courantes

- **Supprimer les espaces de début et de fin / rassembler les espaces consécutifs:** Ces fonctions permettent de supprimer les espaces indésirables. Quand vous manipulez des données, que 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.

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

Editer les cellules
Remplacer

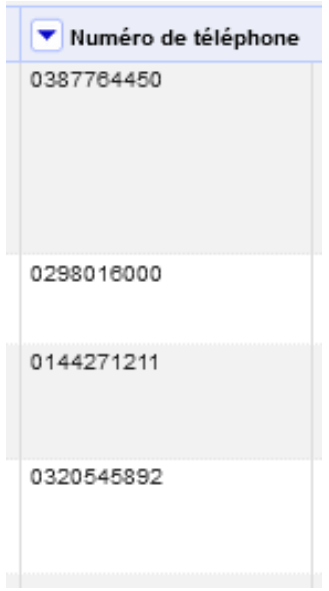
Cette fonctionnalité permet de remplacer une chaine de caractères par une autre.

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

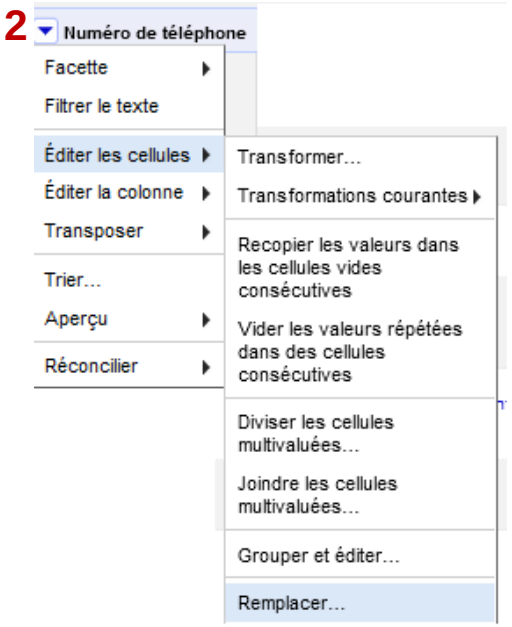
Vous avez la possibilité de remplacer des caractères par une valeur nulle, ce qui aura pour effet de supprimer le caractère ou la chaine de caractères sélectionnée.

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

1

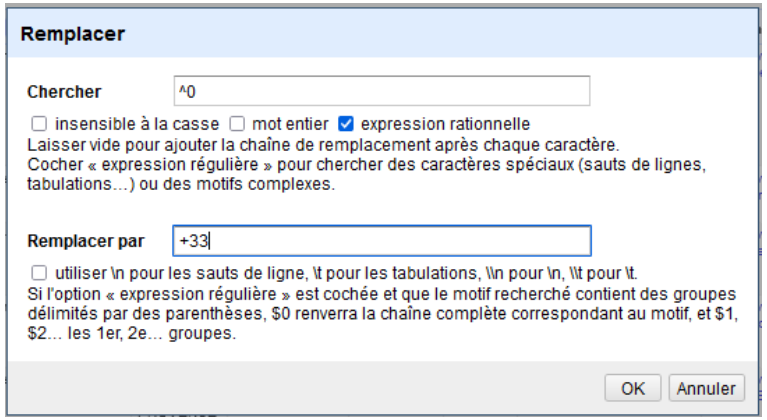


2



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

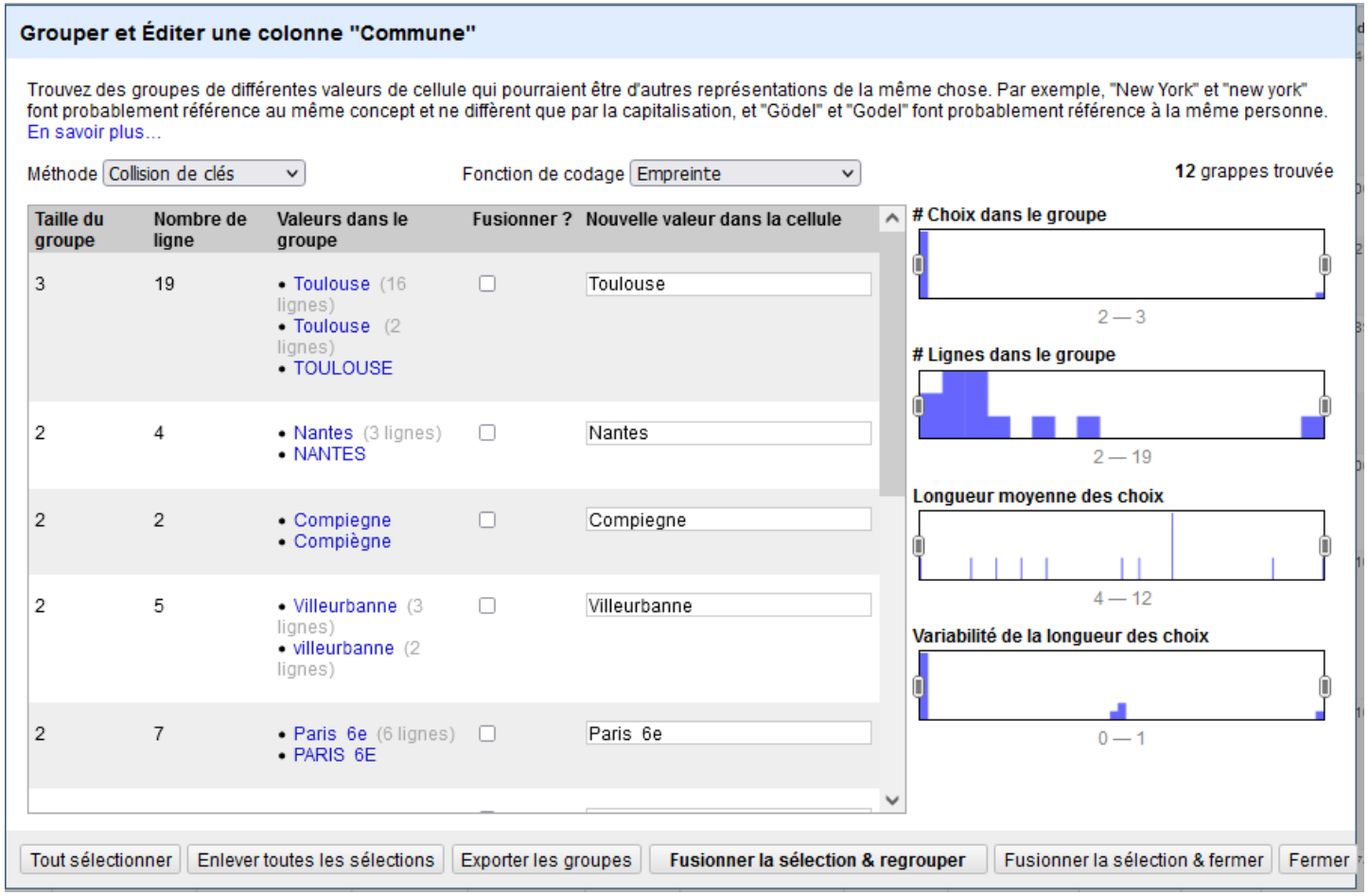


4

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)



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

SORBONNE UNIVERSITÉ
BIBLIOTHÈQUE UNIVERSITAIRE

- **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

SORBONNE UNIVERSITÉ
BIBLIOTHÈQUE UNIVERSITAIRE

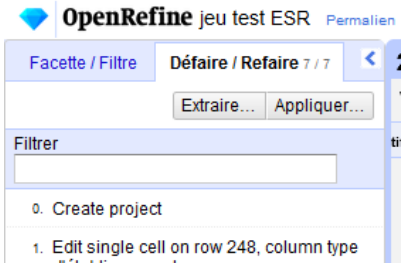
- **Diviser en plusieurs colonnes :** Vous pouvez diviser votre colonne sur la base d'un séparateur ou d'une longueur. Notez que contrairement à un tableur classique :
  - **Vous pouvez utiliser un séparateur de plusieurs caractères, ou basé sur une expression régulière.**
  - **La division créer de 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)

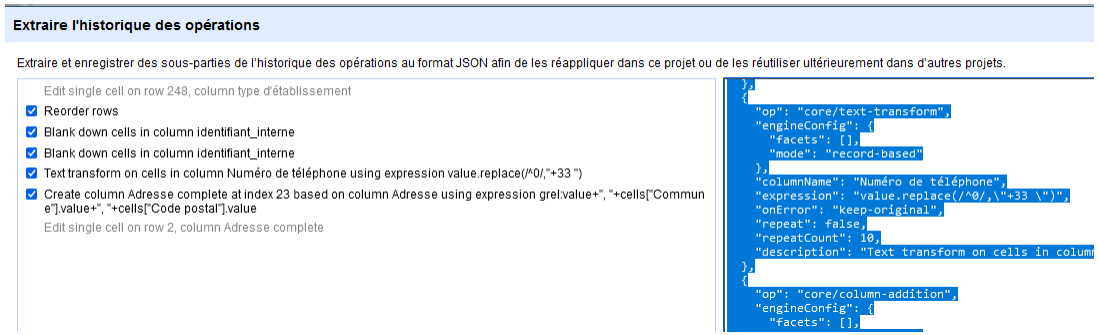
Extraire et appliquer des traitements

- 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'apparaitra 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



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

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

4

INTRODUCTION A L'ÉDITEUR DE FORMULE GREL

Les formules dans OpenRefine

SORBONNE UNIVERSITÉ
BIBLIOTHÈQUE UNIVERSITAIRE

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 :

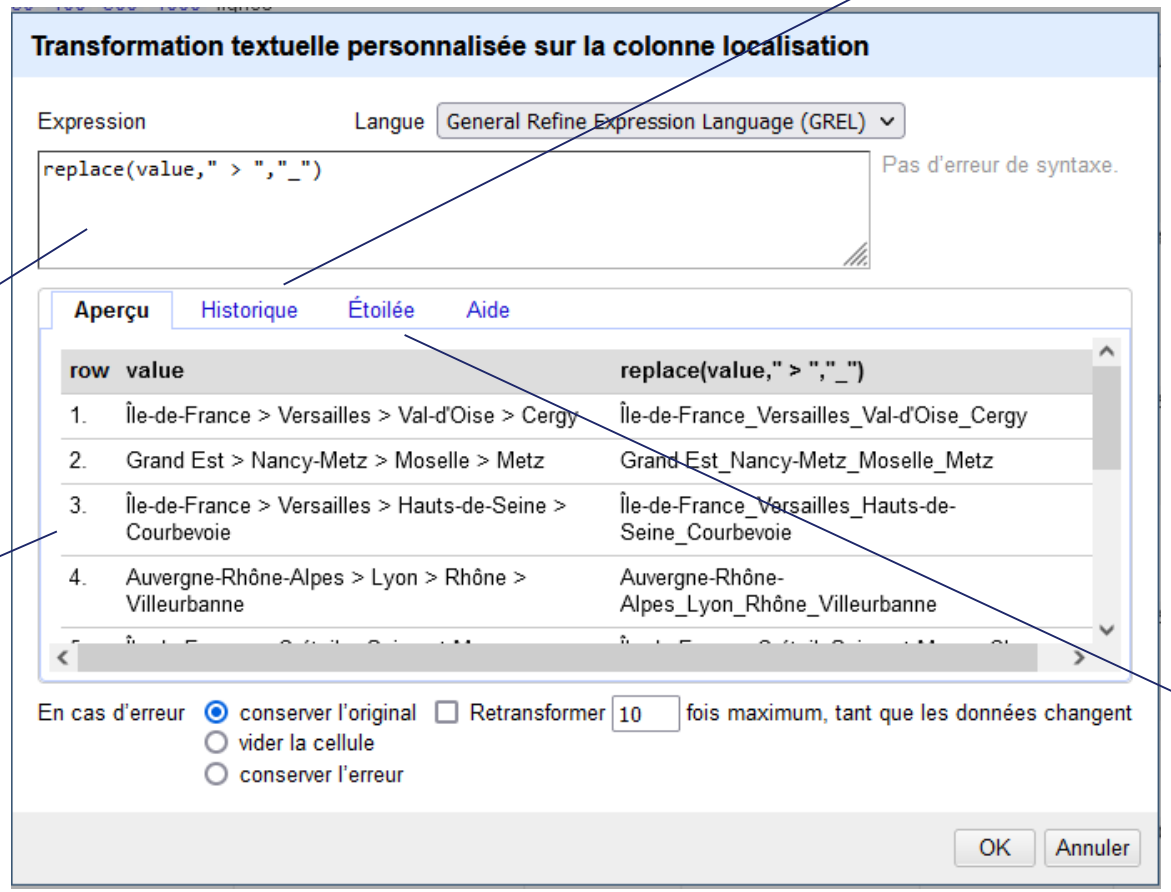
- **Editer les cellules > Transformer** : permet d'appliquer des modifications en masse dans la colonne sélectionnée
- **Editer la colonne > Ajouter une colonne en fonction de cette colonne** : permet d'appliquer les modifications en masse dans une nouvelle colonne, ainsi la colonne sélectionnée n'est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



Etoilée: Permet d'accéder à vos formules favorites

Quelques notions de bases
La concaténation

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

"**valeur à ajouter**"+value

Exemple :

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

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

```
"https://idref.fr/"+value
```

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"

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

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

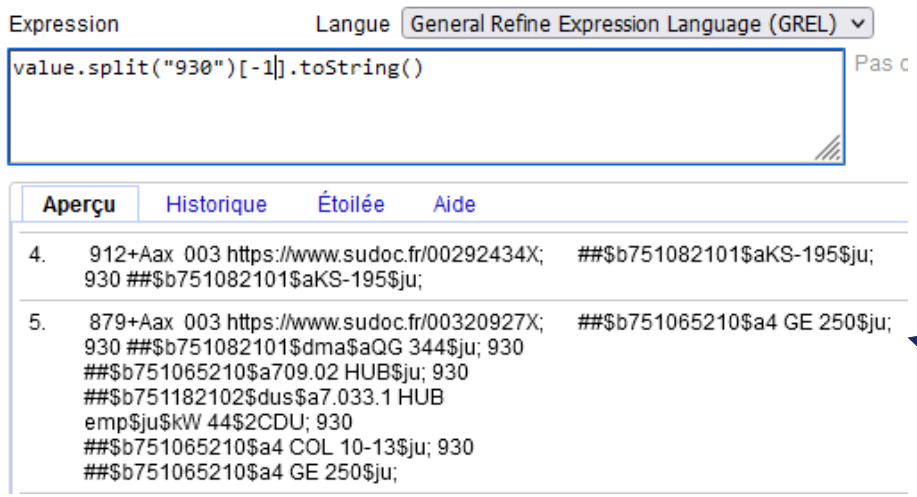
Quelques notions de bases
Extraire une partie d'une cellule (division)

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()



0 879+Aax 003 https://www.sudoc.fr/00320927X; 930 ##$b751082101$dma$aQG 344$ju; 930
1
2 ##$b751065210$a709.02 HUB$ju; 930
3 ##$b751182102$dus$a7.033.1 HUB
4 empjukW 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 empjukW 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)

SORBONNE UNIVERSITÉ
BIBLIOTHÈQUE UNIVERSITAIRE

OpenRefine vous permet d'extraire une chaine de caractère d'une cellule avec la formule value.find(). Pour cela vous devez connaitre la structure de la chaine 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 chaine de caractères en ajoutant .toString() à la fin de votre formule

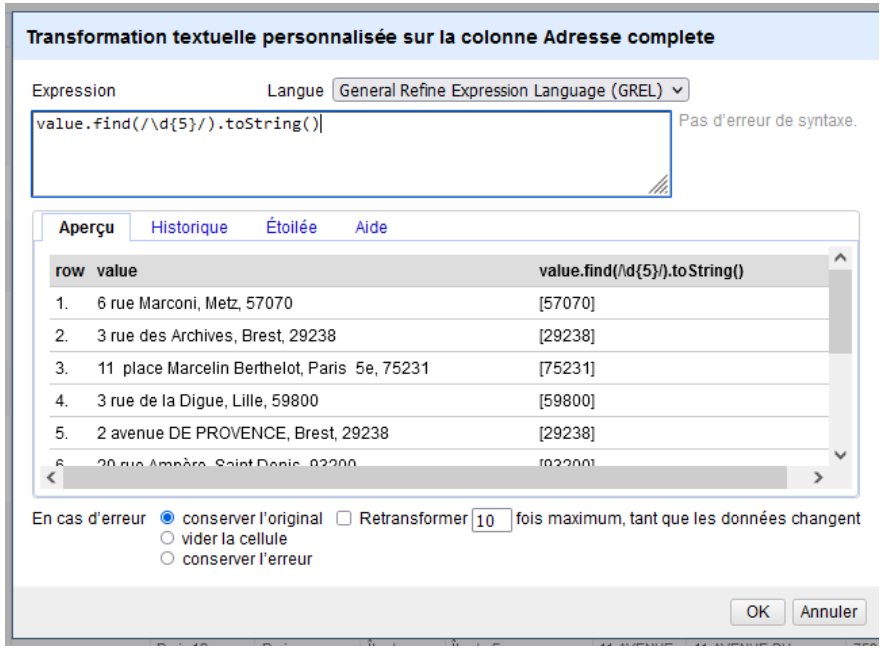
value.find(/**REGEX de la chaine de caractère à extraire**/).toString()

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*
- *\d{5} signifie 5 chiffres successifs*



Quelques notions de bases
Le croisement de colonnes

SORBONNE UNIVERSITÉ
BIBLIOTHÈQUE UNIVERSITAIRE

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.

Aperçu | Historique | Étoilée | Aide

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

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

- 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

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

LabRADoR

LAB DE RESSOURCES ET D'ACCOMPAGNEMENT AUX DONNÉES DE LA RECHERCHE SORBONNE UNIVERSITÉ

labrador@sorbonne-universite.fr

MERCI

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

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE