

NETTOYER SES DONNÉES AVEC OPENREFINE (NIVEAU 1)

AURÉLIEN MOISAN

12 JANVIER 2022



Sommaire

- 1. INTRODUCTION**
- 2. INSTALLATION ET DÉCOUVERTE DE L'INTERFACE**
- 3. FILTRES, FACETTES ET TRI**
- 4. INTRODUCTION À L'ÉDITEUR DE FORMULES GREL**
- 5. QUELQUES EXEMPLES**

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.

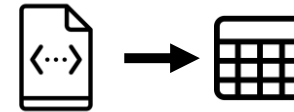
Pour plus d'informations rendez-vous sur : <https://openrefine.org/>

OpenRefine permet de :

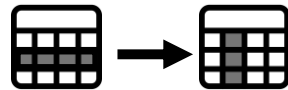
Nettoyer un jeu de données



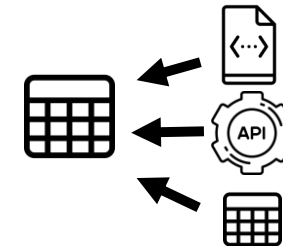
Convertir un jeu de données



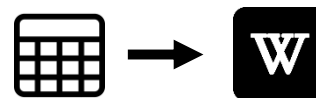
Transformer un jeu de données



Enrichir un jeu de données



Publier 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 ...

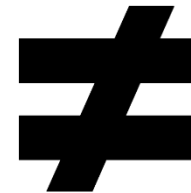
... ce n'est pas un tableur

OpenRefine ne permet pas :

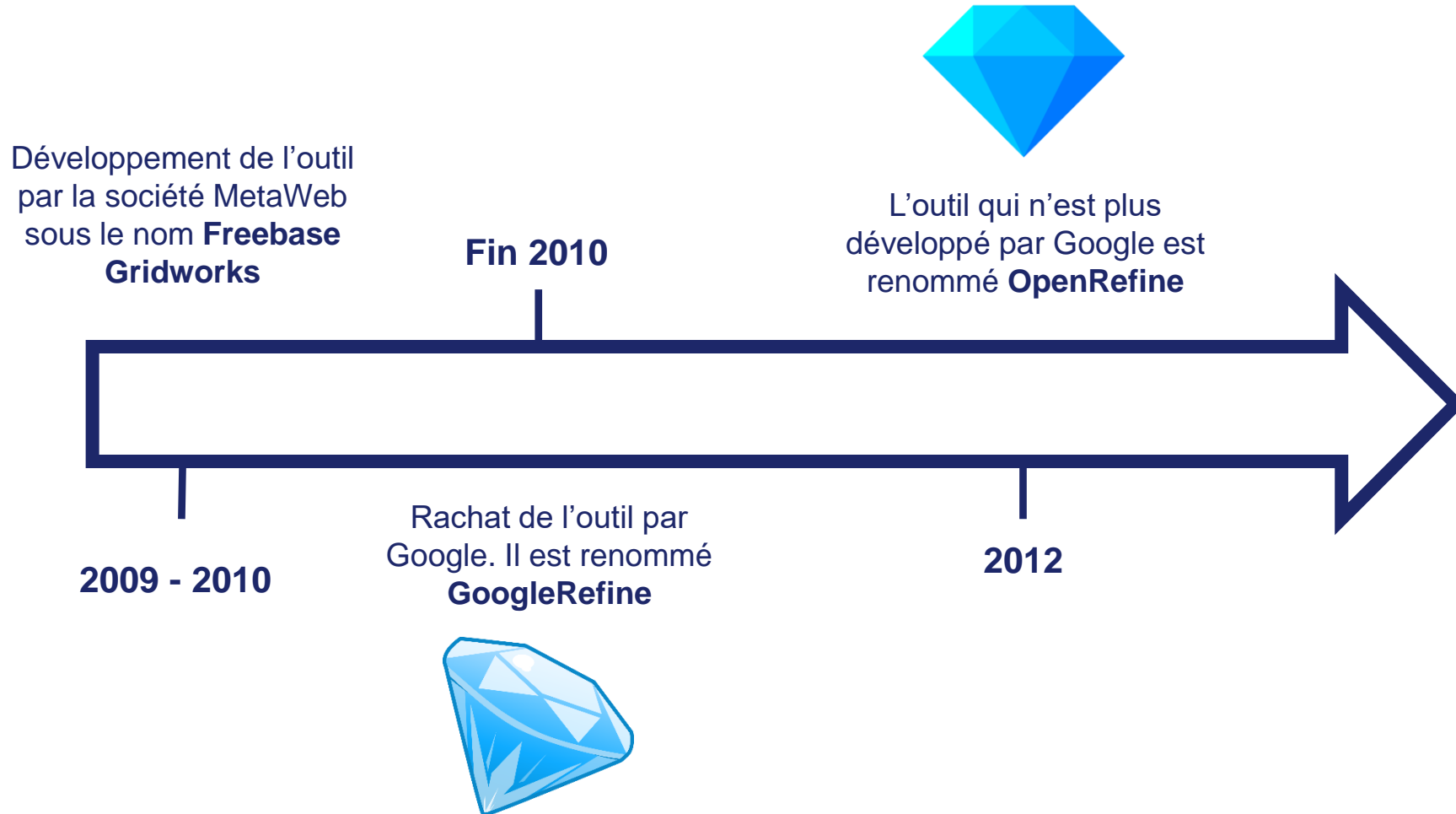
- de visualiser les données sous forme graphique
- le travail collaboratif

Il n'est pas optimisé pour :

- saisir des données
- réaliser des calculs



Historique de l'outil



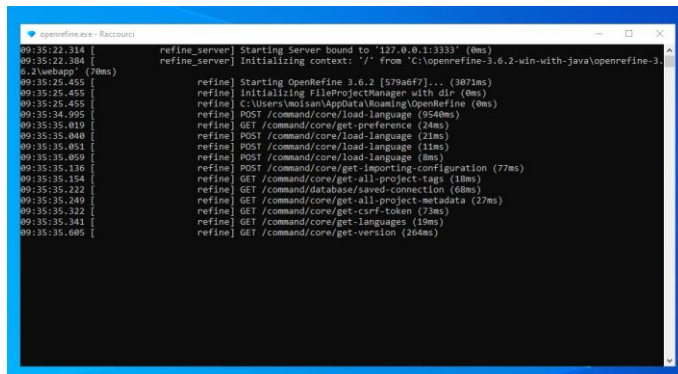
1

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.



```
09:35:22-314 [ refine_server] Starting Server bound to '127.0.0.1:3333' (8ms)
09:35:22-384 [ refine_server] Initializing context: '?' from 'C:\openrefine-3.6.2-win-with-java\openrefine-3.6.2\weapp' (78ms)
09:35:25-455 [ refine] Starting OpenRefine 3.6.2 [379a8f7]... (307ms)
09:35:25-455 [ refine] Initializing FileManager with dir: (8ms)
09:35:25-455 [ refine] C:\Users\molsan\AppData\Roaming\OpenRefine (8ms)
09:35:34-995 [ refine] POST /command/core/load-language (9548ms)
09:35:35-010 [ refine] GET /command/core/get-preference (24ms)
09:35:35-040 [ refine] POST /command/core/load-language (21ms)
09:35:35-051 [ refine] POST /command/core/load-language (11ms)
09:35:35-050 [ refine] POST /command/core/load-language (8ms)
09:35:35-136 [ refine] POST /command/core/get-importing-configuration (77ms)
09:35:35-154 [ refine] GET /command/core/get-all-project-tags (18ms)
09:35:35-222 [ refine] GET /command/database/save-connection (68ms)
09:35:35-249 [ refine] GET /command/core/get-all-project-metadata (27ms)
09:35:35-322 [ refine] GET /command/core/get-csrf-token (73ms)
09:35:35-341 [ refine] GET /command/core/get-languages (3ms)
09:35:35-605 [ refine] GET /command/core/get-version (264ms)
```



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

Paramétrage

Ouvrez le fichier openrefine.l4j.ini avec un éditeur de texte (notepad, gedit etc.); Pour :

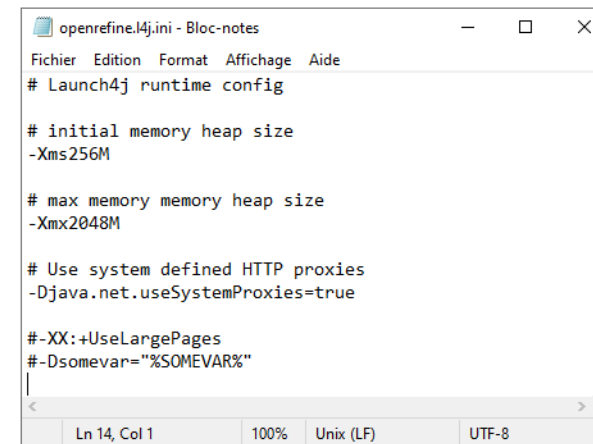
- Augmenter la mémoire allouée à OpenRefine, modifiez la ligne :

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

- Changer le répertoire de travail, ajoutez une ligne :

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

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

#-XX:+UseLargePages
#-Dsomevar="%SOMEVAR%"

Ln 14, Col 1    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



BIBLIOTHÈQUE
UNIVERSITAIRE

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

Chercher un ou plusieurs fichiers à charger :

Parcourir...

Aucun fichier sélectionné.

Suivant »

Cet ordinateur

Adresses web (URLs)

Presse-papier

Base de données

Google Data

Interface

Créer un projet à partir de cet ordinateur

Le plus souvent le format de fichier est détecté automatiquement par OpenRefine. Selon le format la création d'un projet demandera plus ou moins de paramétrage :

- Pour l'import d'un tableau type .xlsx ou .ods il y a peu de paramétrage
- Pour l'import d'un fichier avec séparateur (.tsv, .csv..) il faudra notamment sélectionner le séparateur, le format des caractères (l'encodage) (ci-dessous)
- Pour les fichiers à balises (xml) il faudra 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>
      <identifiant>oai:oaicalames.abes.fr:Calames-202228122586701</identifiant>
      <timestamp />
      <setSpec>751059801</setSpec>
    </header>
    <metadata>
      <oaai_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/identifiant">2 DA 1 - 2 DA 7 [cite]</dc:identif
        <dc:relation xmlns:dc="http://purl.org/dc/elements/1.1/relation">FR-751059801 [RCR établissement]</dc:cr
        <dc:relation xmlns:dc="http://purl.org/dc/elements/1.1/relation">Fonds Paul Hazard [Fonds ou collection
        <dc:relation xmlns:dc="http://purl.org/dc/elements/1.1/relation">http://www.calames.abes.fr/pub/ms/Cala
        <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>
      </oaai_dc>
    </metadata>
  </record>
</ListRecords>
```

Considérer les données

Considérer les données
comme

Format des caractères

Désactiver l'aperçu automatique

Fichiers CSV / TSV / séparateur

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 première(s) ligne(s) du début du fichier

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

Noms de colonnes (séparés par des virgules)

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

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

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

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

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.

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

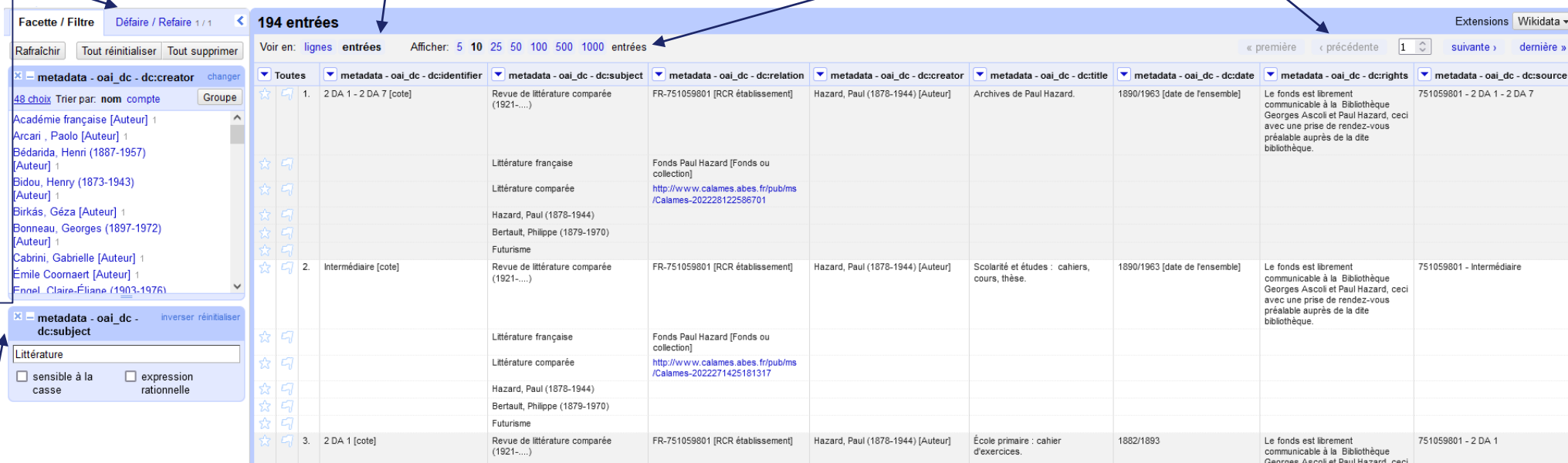
Défaire / Refaire

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 'lignes', 'entrées', and 'Afficher' (5, 10, 25, 50, 100, 500, 1000). The entries are sorted by 'dc:subject'. A facet for 'dc:subject' is visible on the left, showing a list of subjects like 'Littérature', 'Littérature comparée', 'Futurisme', and 'Intermédiaire'. The facet is currently set to 'sensible à la casse' and 'expression rationnelle'.

lignes	entrées	Afficher	dc:subject	dc:relation	dc:creator	dc:title	dc:date	dc:rights	dc:source
1.	2 DA 1 - 2 DA 7 [cote]		Revue de littérature comparée (1921-....)	FR-751059801 [RCR établissement]	Hazard, Paul (1878-1944) [Auteur]	Archives de Paul Hazard.	1890/1963 [date de l'ensemble]	Le fonds est librement communicable à la Bibliothèque Georges Ascoli et Paul Hazard, ceci avec une prise de rendez-vous préalable auprès de la dite bibliothèque.	751059801 - 2 DA 1 - 2 DA 7
2.	Intermédiaire [cote]		Revue de littérature comparée (1921-....)	FR-751059801 [RCR établissement]	Hazard, Paul (1878-1944) [Auteur]	Scolarité et études : cahiers, cours, thèse.	1890/1963 [date de l'ensemble]	Le fonds est librement communicable à la Bibliothèque Georges Ascoli et Paul Hazard, ceci avec une prise de rendez-vous préalable auprès de la dite bibliothèque.	751059801 - Intermédiaire
3.	2 DA 1 [cote]		Revue de littérature comparée (1921-....)	FR-751059801 [RCR établissement]	Hazard, Paul (1878-1944) [Auteur]	École primaire : cahier d'exercices.	1882/1893	Le fonds est librement communicable à la Bibliothèque Georges Ascoli et Paul Hazard, ceci	751059801 - 2 DA 1

2

FILTRES, FACETTES ET TRI

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 ;
- **Lorsqu'un champ comporte un nombre important de valeurs :** si on sait quelle valeur on souhaite filtrer, ce sera plus rapide d'utiliser un filtre que de parcourir la liste des valeurs dans une facette ;
- **Filtrer à partir de valeurs inconnues :** Il est possible de construire un filtre à partir de REGEX (expressions régulières). 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 ;
- **Lorsqu'on cherche une valeur dans un champ libre**

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 ;
- **Identifier des coquilles :** en affichant toutes les valeurs d'un champ vous pouvez facilement identifier si des valeurs ne sont pas normées, s'il y a des doublons, des valeurs erronées..
- **Modifier des valeurs :** Vous pouvez éditer les facettes, 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 basé sur les chaînes de caractères. Cela pourra être utilisé pour normaliser les valeurs d'un champ. (voir [diapo](#))

Le tri

Dans quel cas l'utiliser ?

Le tri permet de réordonner les valeurs d'un champ (et donc les entrées de votre tableau).

Par défaut le tri est temporaire, vos entrées garderont leur numérotation initiale.

Vous pouvez cependant faire le choix d'appliquer un tri de manière permanente ce qui entraîne une renumérotation de vos entrées.

Le tri permanent permet notamment d'utiliser les fonctions suivantes :

- **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 et 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

3

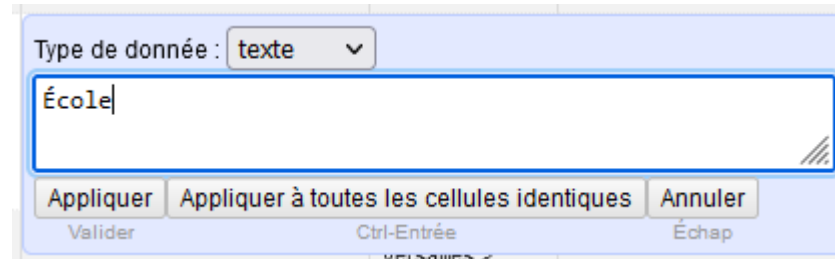
NETTOYER LES DONNÉES : LES FONCTIONS
BASQUES

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 « 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



Type de donnée :

École|

Appliquer Appliquer à toutes les cellules identiques Annuler







Valider Ctrl-Entrée Échap

« 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 jeu de données.

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 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 »**

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 les caractères par une valeur nulle, ce qui aura pour effet de supprimer le caractère ou la chaîne de caractères sélectionnée.

Nous pouvez élégamment utiliser des REGEX

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 à 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 de 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 à 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 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 selection
 - 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 changer l'ordre des colonnes et de supprimer des colonnes
 - Recopier / vider les valeur 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
- **Supprimer cette colonne**
- **Déplacer la colonne...** : permet modifier la position d'une colonne. Notez qu'il est plus simple d'utiliser la colonne la colonne « Toutes » pour travailler sur la réorganisation 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 un 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. Vous pouvez faire le choix de supprimer ou non la colonne initiale
- **Joindre de colonnes** : permet de fusionner deux ou 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 résultats dans la colonne active ou dans une nouvelle colonne

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 qu'une grande partie des fonctionnalités de base sont en réalité 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 basé sur 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) ▼ Pas d'erreur de syntaxe.

Aperçu **Historique** **Étoilée** **Aide**

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(/.*\((.*)\).*/)

Étoilée: 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

Par exemple, si je veux générer une URL à partir d'un identifiant IDREF. Il faut que j'ajoute « https://idref.fr/ » devant l'identifiant. J'utilise donc la formule : "**https://idref.fr/**" + valeur

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

Pas d'erreur de syntaxe.

row	value	"https://idref.fr/" + valeur
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

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>
- Lancez-vous, testez, recherchez des solutions sur des forums !

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)
- 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 »

Services autour de la science ouverte à la BSU

Cellule Données de la recherche & Humanités numériques :
data-bsu@sorbonne-universite.fr

Département Publications et Open Access :
hal@sorbonne-universite.fr



Service des Archives et du Recueil des Actes :
sara-archives@sorbonne-universite.fr

Vos questions

Par chat ou de vive voix !



BIBLIOTHÈQUE
UNIVERSITAIRE



MERCI !

MERCI

Cellule données et humanités numériques
data-bsu@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