Université de Lille

29/08/2025

OpenRefine
Journées CasuHAL 2025

Camille Vacher
camille.vacher@univ-lille.fr

CC BY

Université de Lille

Introduction

OpenRefine, qu’est-ce que c’est ?

- Outil de nettoyage et de transformation de données
- Outil interactif et "visuel" => travail sur des données de taille importante mais pas démesurées:
- Travail avec le web (non-couvert par la formation) :
  - téléchargement de données ;
  - fonctions sur pour manipuler du HTML ou JSON ;
  - *Réconciliation* avec des données externes (Wikidata).

Cadre et limites d’utilisation

- Ce n'est pas un tableur : pas de création de données ;
- Convient pour des données de taille « moyenne » ;
- Pas un vrai outil de Big Data, pas un ETL (Extract Transform Load) ;
- Intermédiaire entre un tableur (Excel, LibreOffice Calc …) et un ETL (Talend OpenStudio, Knime …) ;
- Convient pour un traitement externe des données de HAL, pas pour faire un traitement *dans* HAL ;
- C’est pour vous si :
  - les tableurs ne sont pas suffisant pour vos traitements ;
  - vos traitements sont trop laborieux à faire dans un tableur ;
  - Vous devez répéter régulièrement le même traitement ;
  - Vous n’utilisez pas d’ETL.

Université de Lille

Plan :

1. Création d’un projet
2. Adapter l’interface
3. Explorer les données
4. Transformations de base
5. Transformations complexes
6. Bonus

Université de Lille

Création d’un projet

Université de Lille

Fonctionnement d'OpenRefine

- Logiciel sur le modèle client/serveur :
  - Un serveur s'exécute en arrière-plan (une console sous Windows) ;
  - On consulte le serveur *via* un navigateur web ;
  - On peut traiter différents « projets » simultanément dans différents onglets.

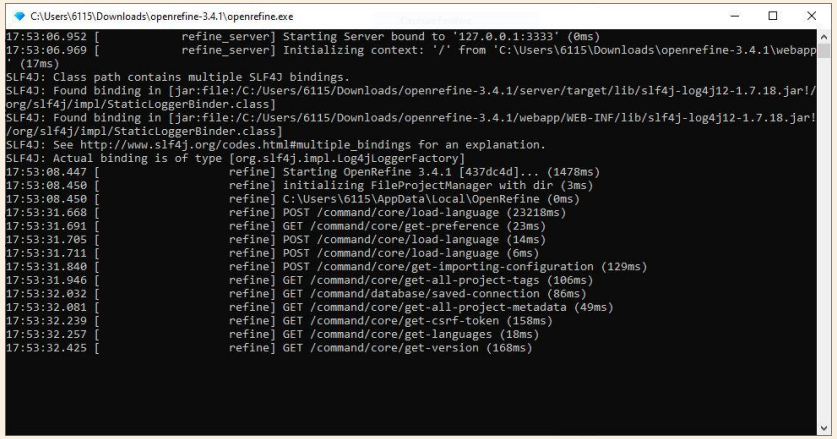
- Fonctionnement par « projet » avec des données copiées :
  - On charge un fichier (texte/tableur/JSON/XML) local ;
  - OpenRefine copie les données et travaille sur cette copie ;
  - Openrefine ne modifie jamais le fichier local initial ;
  - On exporte les données

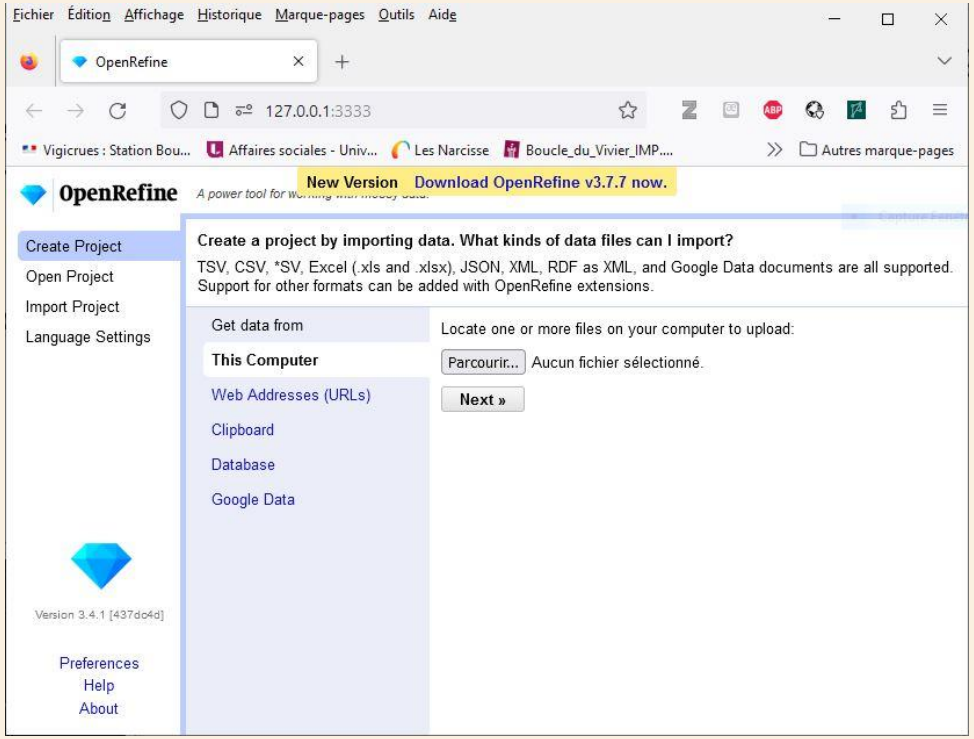
Université de Lille

Fonctionnement Client-Serveur

OpenRefine fonctionne selon un principe client/serveur

Le serveur est lancé en arrière-plan
Un navigateur permet d'accéder à l'application





Lors du lancement de l'application, le serveur est lancé dans une console et une nouvelle fenêtre de votre navigateur par défaut est lancée et se connecte directement à ce serveur.

Si vous fermez par inadvertance l'onglet ou la fenêtre, il suffit d'ouvrir un nouvel onglet et d'aller à l'URL suivante:
http://127.0.0.1:3333

Vous pouvez ouvrir différents projets simultanément dans plusieurs onglets

Chargement d’un fichier local

Création d’un « Projet »

Générer un fichier avec l'API HAL

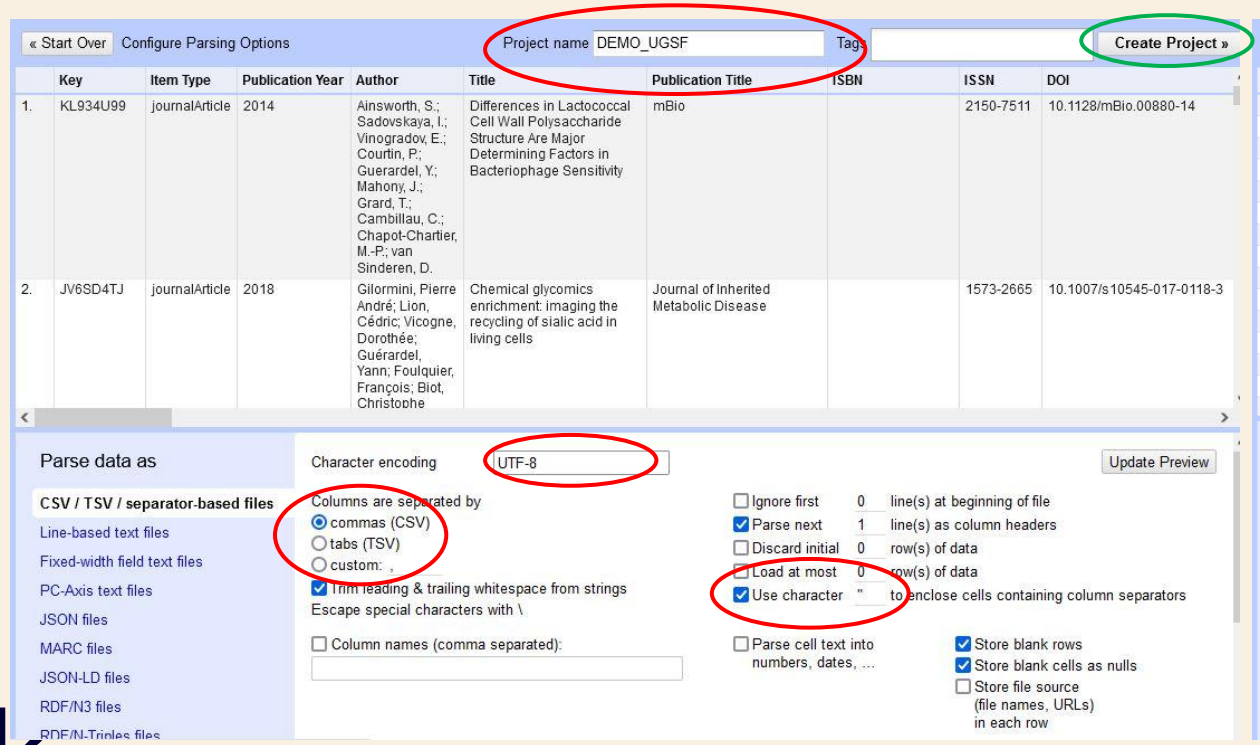
- https://filesender.renater.fr/?s=download&token=4f80ff60-61cd-453b-b694-ccb57c013278
- https://tinyurl.com/openrefinecasuhal
- Pour traiter des données HAL, les formats de sortie CSV et JSON conviennent très bien. Nous allons commencer avec du CSV pour aller plus vite.

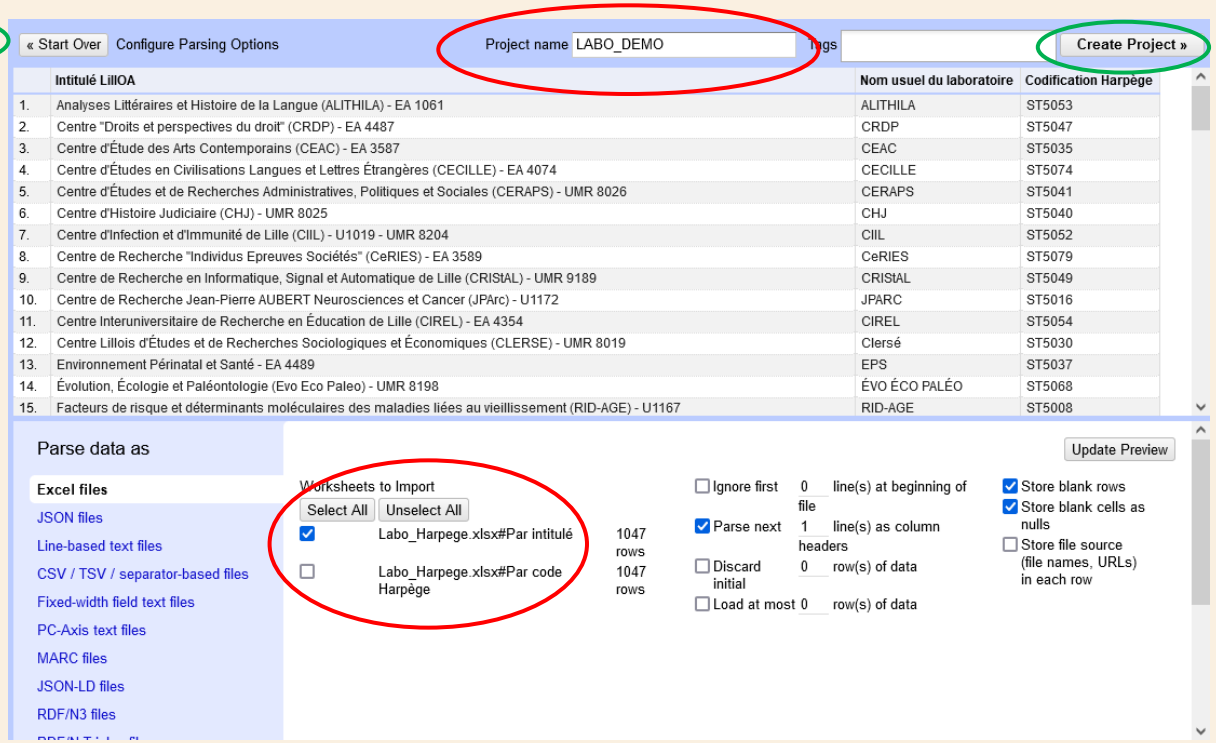
Pour la création d'un projet depuis un fichier local, il suffit de le sélectionner. La page qui s'affiche permet d'avoir un aperçu des données, de modifier les paramètres de création et de choisir le nom du projet.

Points d'attention:

- Pour les TSV/CSV: Encodage des caractères, séparateur de colonne, caractère d'échappement ;
- Pour les fichiers Excel ou ODS: choisir le feuillet voulu ;
- Pour les fichiers JSON/XML bien choisir l'unité de base ;
- Pour les XML: beaucoup de colonnes vides. À éviter si possible.

Quand tout convient, vous pouvez cliquer sur le bouton « Create Project »





Université de Lille

Adapter l’interface

Interface d’OpenRefine

- Une panneau latéral gauche avec deux onglets
  - Un onglet avec les facettes et filtres ;
  - Un onglet avec l’historique des modifications.

- Un tableur un peu particulier :
  - données non directement éditables ;
  - une colonne intitulée "Toutes" ;
  - un menu pour chaque colonne.

- Un bandeau de navigation :
  - Un mode de navigation par « ligne » ou par « entrée » ;
  - Un nombre de « lignes » ou « d’entrées » affichées ;
  - Des flèches et raccourcis pour naviguer

Université de Lille

Cliquer sur l'icône pour retourner à l'accueil

Nom du projet, modifiable en cliquant

Affichage par ligne ou entrée

Nombre d'éléments affichés simultanément

Un menu global

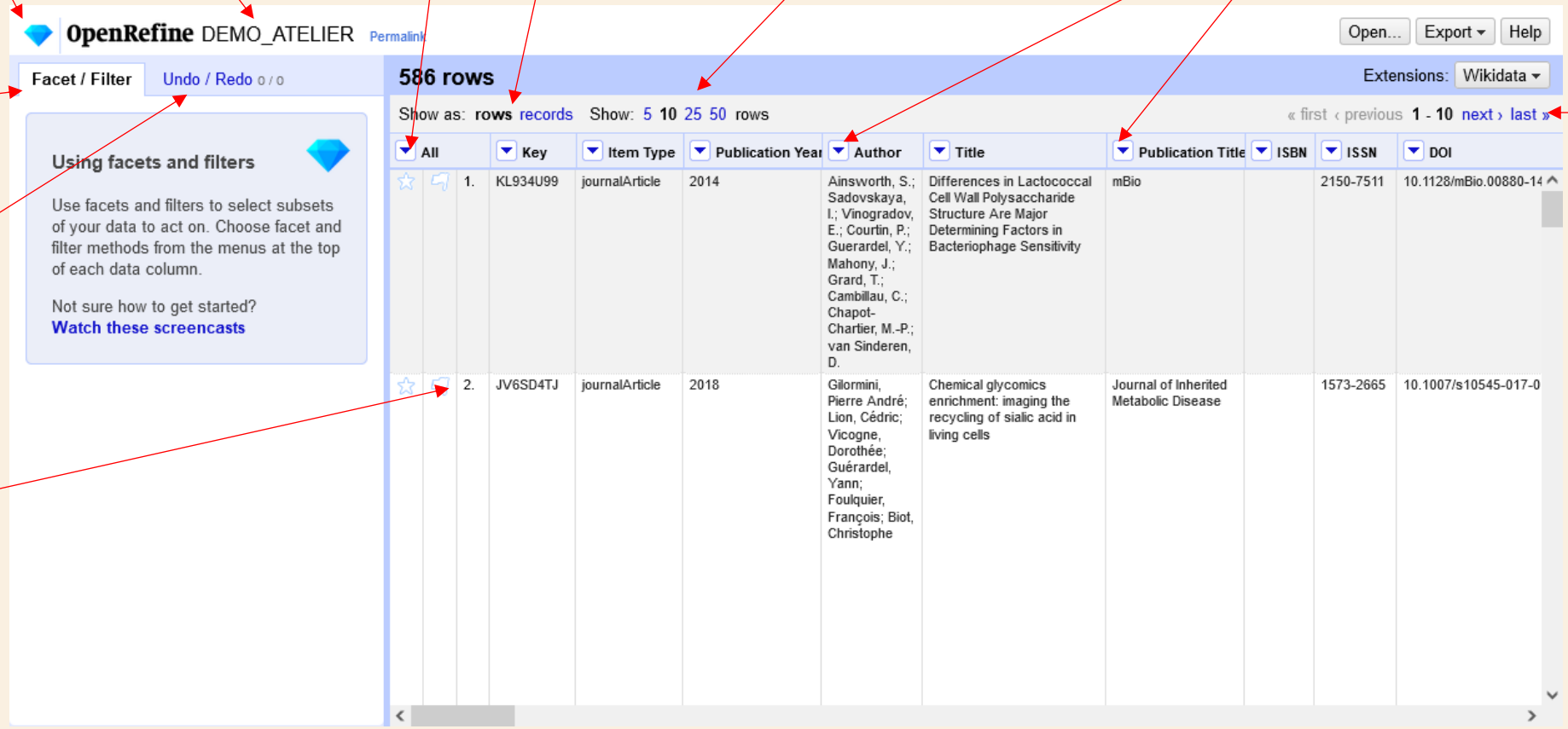
Un menu par colonne

Onglet des facettes et des filtres

Onglet de l'historique des modifications

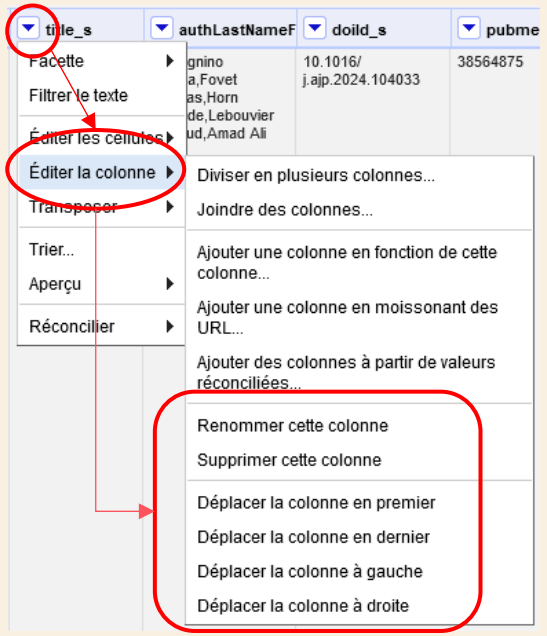
Numéro de ligne ou d'entrée

Navigation

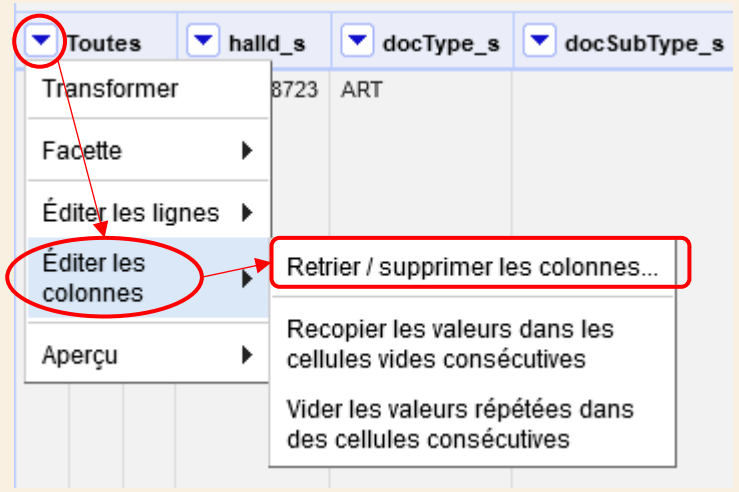


Déplacement des colonnes - 1

- On peut déplacer, supprimer ou renommer des colonnes à partir de leur menu
  - Menu de la colonne ;
  - « Éditer la colonne » ;
  - « Renommer/Supprimer cette colonne » ou « Déplacer la colonne en premier/en dernier/à gauche/à droite »



- Pour déplacer ou supprimer plusieurs colonnes, il est plus rapide d'aller dans le menu « Toutes »
  - Menu « Toutes » ;
  - « Éditer les colonnes » ;
  - « Retrier / supprimer les colonnes… »



Déplacement des colonnes - 2

Réorganiser les colonnes dans l'ordre voulu

Déposer ici les colonnes à supprimer

Masquer des colonnes

- Dans le menu => « Aperçu » ;
- Utile pour avoir un affichage plus compact ;
- Deux utilisations fréquentes :
  - Masquer une colonne avec trop de contenu ;
  - Masquer toutes les colonnes précédant ou suivant la colonne sur laquelle on travaille,
- On peut connaître le nom d'une colonne masquée en passant le curseur dessus. En cliquant dessus, on déplie la colonne
- Possible de plier ou de déplier toutes les colonnes dans le menu « Toutes »

Trier

- Possibilité de trier en fonction d'une ou de plusieurs colonnes
- Le tri est seulement un tri d'affichage : le numéro en début de lignes réfère toujours à la place réelle de la ligne dans le projet
- Pour modifier l'ordre des lignes en fonction du tri actuel, il faut aller dans le menu « Trier » qui est apparu dans le bandeau de navigation et sélectionner « Retrier les lignes de façon permanente »

Université de Lille

Explorer les données

Manipuler les facettes et les filtres

Naviguer avec les facettes

- Les facettes permettent de connaître l'ensemble des valeurs présentes dans une colonne, et le nombre d'occurrences de chacune de ces valeurs ;
- L'ordre d'apparition des facettes (alphabétique ou par nombre d'occurrence) n'est pas l'ordre d'apparition des valeurs dans la colonne ;
- On peut sélectionner de multiples valeur simultanément (avec *inclure*) ou choisir l'*inverse* d'une sélection ;
- On peut exporter l'ensemble des valeurs choisies ;
- On peut combiner des facettes sur plusieurs colonnes ;
- **Attention** : les facettes sont interdépendantes, et l'interface affiche les lignes qui correspondent à l'ensemble des facettes activées

Naviguer avec les filtres

- On peut vouloir travailler simultanément sur des valeurs distinctes mais qui ont une partie de texte en commun : on peut utiliser les filtres ;
- Par défaut, le filtre n'est pas sensible à la casse (majuscule/minuscule) mais il l'est toujours aux accents ;
- Les filtres peuvent aussi être inversés ;
- Possibilité d'utiliser des expressions régulières ;
- Astuce utile : il est possible d'ouvrir la facette textuelle d'une colonne et d'utiliser un filtre sur cette même colonne. La facette textuelle affichera l'ensemble des valeurs de la colonne qui contiennent le filtre saisi.
- **Attention** : les valeur des facettes dépendent **aussi** des filtres actifs, et l'interface affiche les lignes qui correspondent à l'ensemble des facettes **et des filtres** activés.

(Presque) Tout sur les filtres et facettes

En une seule image

Nombre de valeurs distinctes dans la facette. Cliquer dessus pour pouvoir copier les valeurs et leur nombre d'occurrence au format TSV

Élément sélectionné sur une facette

Inverser la sélection d'une facette ou le filtre

Réinitialiser la sélection d'une facette ou le filtre

Filtre en cours

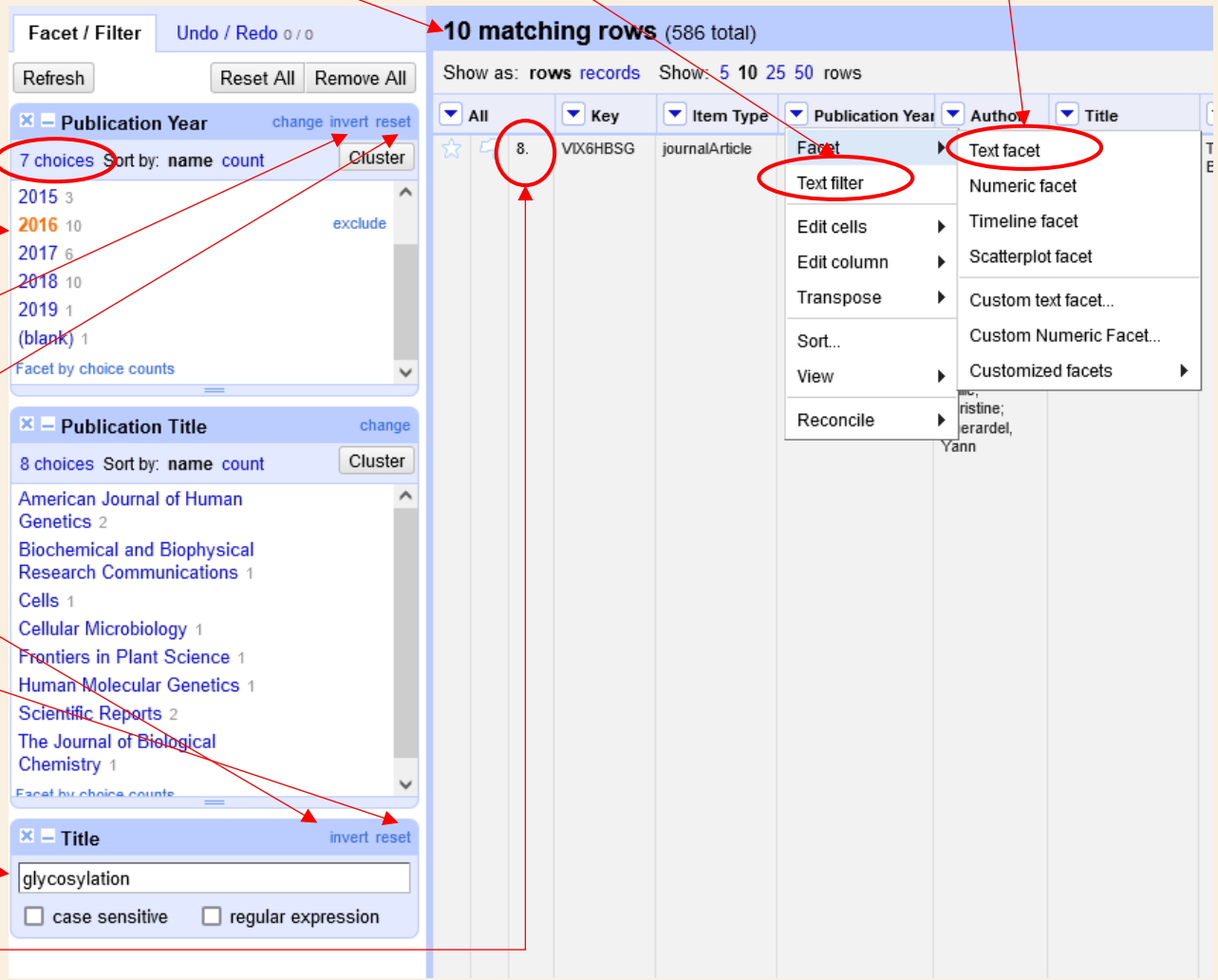
Case à cocher pour que le filtre soit sensible à la casse

Le numéro de ligne correspond toujours à la ligne dans l'ensemble du document

Nombre d'éléments répondants aux divers filtres et sélections dans les facettes

Créer un filtre pour une colonne: Menu de la colonne -> Text filter

Créer une facette pour une colonne: Menu de la colonne -> facet -> text facet



Astuce des facettes et des filtres

- L’export (en CSV ou XLS) n’exporte que les lignes ou entrées sélectionnées par les facettes et filtres en cours.

Université de Lille

Transformations de base

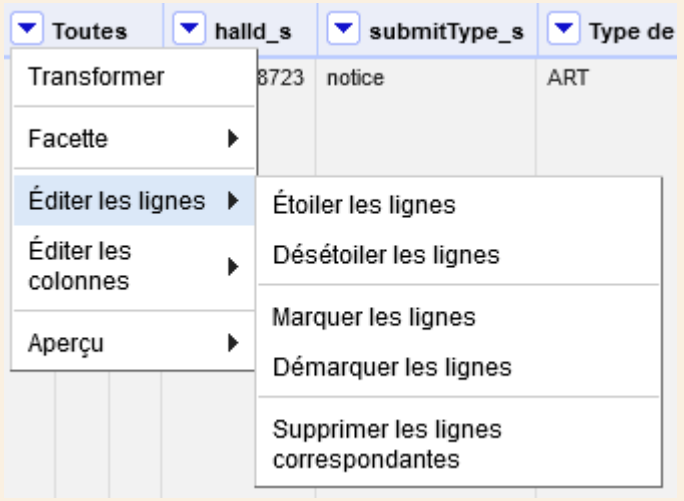
Transformations standards

- On peut éditer directement une cellule ;
- Le menu de colonne propose des transformations de nettoyage de base :
  - Menu => « Éditer les cellules » => « Transformations courantes »
  - Suppression des espaces consécutifs, ou traînant en début ou fin de texte ;
  - Transformation en tout minuscule, tout majuscule ou « title case ».

Transformations par colonne

- Dans Menu => « Éditer la colonne »
- Fusionner (joindre) des colonnes, en précisant un éventuel séparateur
- Séparer des colonnes en précisant le séparateur, et éventuellement le nombre limite

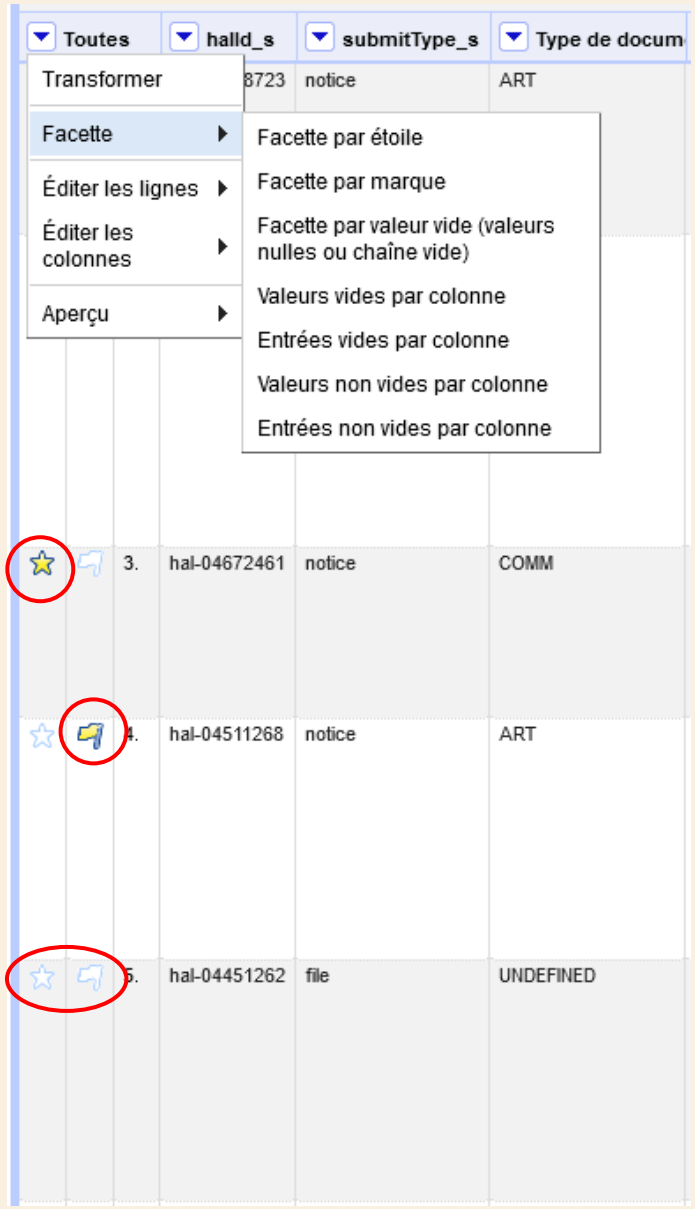
Suppression de lignes



- Menu « Toutes » => « Éditer des lignes » => « Supprimer les lignes correspondantes » ;
- Supprime toutes les lignes qui répondent aux critères des facettes et filtres actifs ;
- **Attention** : sans facette ni filtre, la commande supprime donc toutes les lignes du projet.

Étoiles et marques

- Chaque dispose d'une étoile et d'une marque (drapeau) ;
- On peut cliquer sur l'un ou l'autre pour l'activer ;
- Le menu « Toutes » permet d'ajouter des facettes par étoile ou par marque ;
- Dans le sous-menu « Éditer les lignes » il est aussi possible d'étoiler ou de marquer toutes les lignes actuellement sélectionnées.

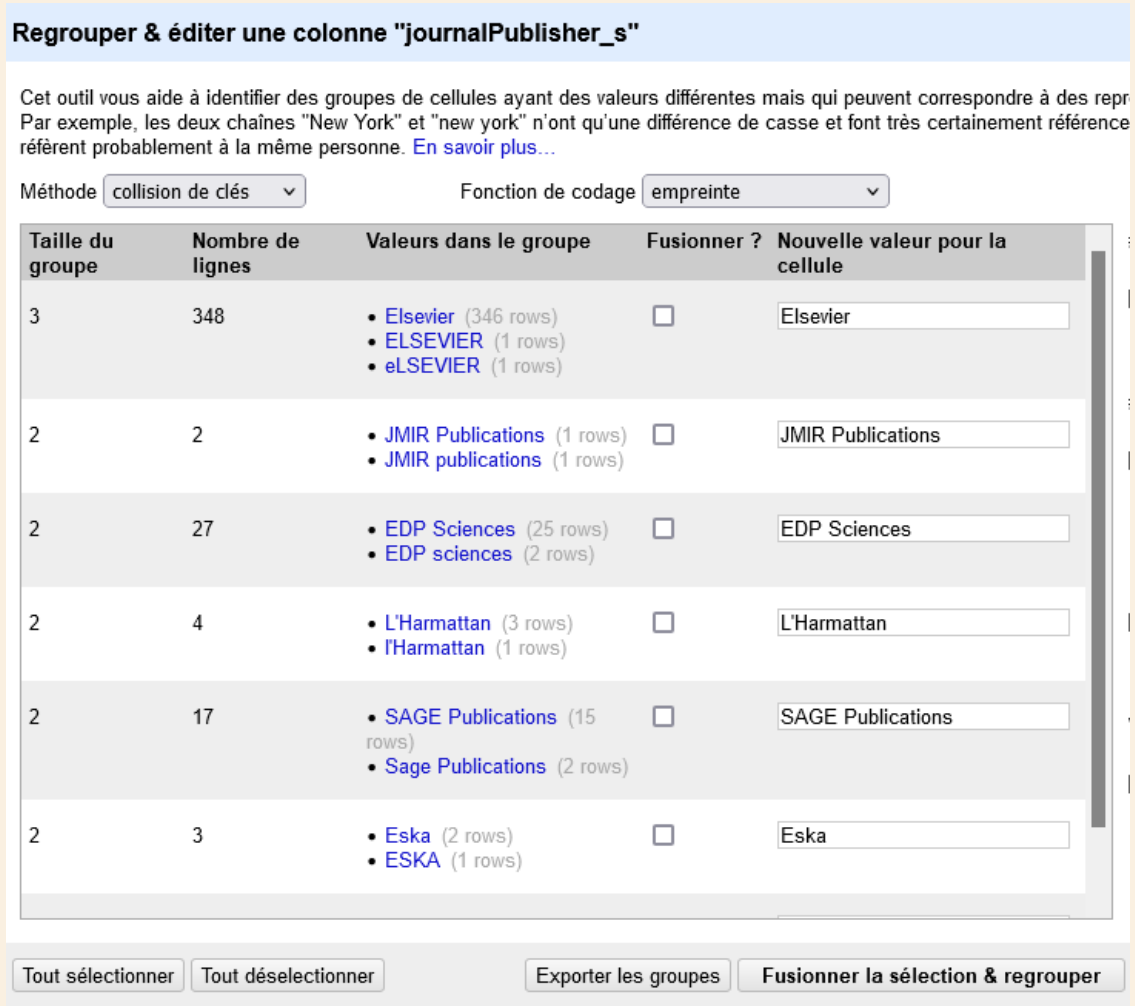


Transformations en masse simples

- Depuis une cellule, on peut modifier en masse toutes les valeurs identiques ;
- Il est possible d'éditer des valeurs directement dans une facette :
  - Toutes les cellules de la colonne concernée avec la valeur concernée vont être modifiées et prendre la nouvelle valeur ;
  - **Attention** : si d'autres facettes ou filtres sont actifs, seules les lignes avec les filtres actifs vont être modifiés.

Grouper et éditer : le « clustering »

- Dans le menu d'une colonne => « Éditer les cellules » => « Grouper et éditer… »
- OpenRefine détecte tout seul des « clusters » de termes proches ;
- Il propose de les fusionner, en proposant par défaut le terme le plus fréquent ;
- Il suffit de cliquer sur un des termes pour demander de les fusionner avec la valeur choisie ;
- Ne pas hésiter à varier la méthode et les différents paramètres ;
- Tout savoir sur le clustering: https://openrefine.org/docs/technical-reference/clustering-in-depth



Université de Lille

Transformations complexes

Transformations en masse complexes - 1

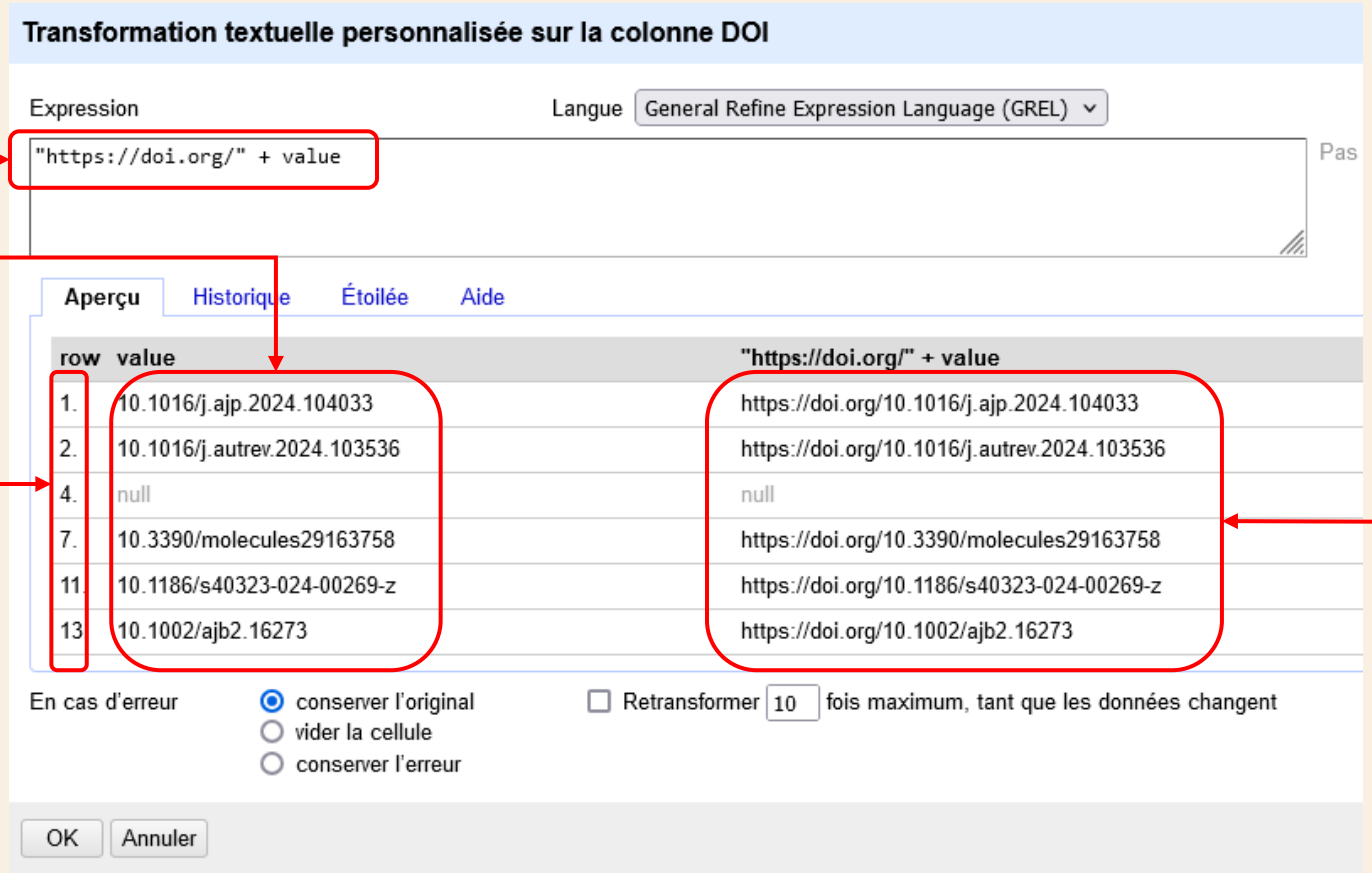
- Accessible depuis les menus :
  - « Éditer les cellules » => « Transformer… »
  - « Éditer la colonne » => « Ajouter une colonne en fonction de cette colonne »
- Utilisation d'un langage de script propre à OpenRefine : GREL ;
- Une expression GREL peut utiliser des fonctions mais ça peut aussi être simplement un nombre ou du texte entre guillemets, simples ou double : ‘Exemple’ ou "Exemple"

Transformations en masse complexes - 2

L'expression GREL à effectuer sur la colonne

Les valeurs initiales à transformer avec l'expression

Attention, la transformation ne se fait que sur les lignes sélectionnées par les facettes et filtres actifs



Un aperçu du résultat des transformations sur chaque ligne

Contenu conçu à partir d'une présentation
de Mathieu Saby sous licence CC-BY-NC

Syntaxe de GREL

- Le texte doit être écrit entre guillemets simples ou doubles: `'Du texte'` ou `"Du texte"`
- Le contenu de chaque cellule est accessible par le mot-clé *`value`*.
- Le contenu de la cellule de la même ligne mais d'une autre colonne est accessible par le mot-clé *`cells`* suivi du nom de la colonne (deux syntaxes : *`cells`*`['nom_colonne'].value` ou *`cells`*`.nom_colonne.value`)
- Une fonction peut s'écrire de deux manières:
  - *`value`*`.nom_de_la_fonction(parametres).` Ex: *value.*`trim ()`
  - `nom_de_la_fonction(`*`value`*`, parametres).` Ex: `trim` (*value*)
- Concaténation de plusieurs valeurs avec le signe +
  - Ex: `"a"+"b"` -> "ab"
  - Ex: *`value`*`+" ;"+cells.colonne1 +" ;"+ cells.colonne2`
- Nom des fonctions sensible aux majuscules.
- Les fonctions sont toujours suivies de parenthèses (), contenant 0 ou x paramètres
- Les fonctions peuvent être chaînées. Ex: *`value`*`.split('|').uniques().join('|')`
- Tout savoir sur GREL : https://openrefine.org/docs/manual/grel

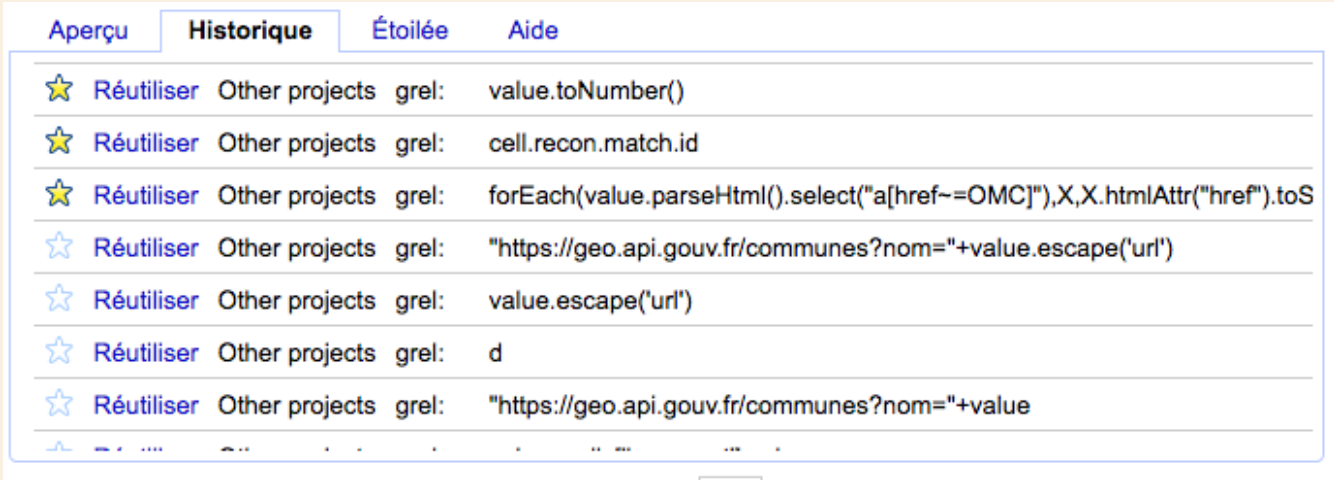
Université de Lille

L'écran de saisie des formules

Contenu conçu à partir d'une présentation
de Mathieu Saby sous licence CC-BY-NC

Trois onglets utiles: Historique, Étoilée, Aide

Permet de réutiliser une formule



Université de Lille

L'écran de saisie des formules

Contenu conçu à partir d'une présentation
de Mathieu Saby sous licence CC-BY-NC

Trois onglets utiles: Historique, Étoilée, Aide

Liste les expressions qui ont été étoilées dans l'historique

Aperçu | Historique | **Étoilée** | Aide

			Expression
Supprimer	Reuse	grel:	forEach(value.parseHtml().select("a[href~=OMC]"),X,X.htmlAttr("href").toString()).join('::')
Supprimer	Reuse	grel:	cell.recon.match.id
Supprimer	Reuse	grel:	"http://data.bnf.fr/sparql?default-graph-uri=&query=SELECT+DISTINCT+%3FidArk %0D%0AWHERE+%7B%0D%0A++%3FidArk+bnf-onto%3AFRBNF+%22"+value+" %22%5E%5Exsd%3Ainteger.%0D%0A%7D&format=application%2Fsparql-results%2Bjson&timeout=0&should-sponge=&debug=on"
Supprimer	Reuse	grel:	value.parseJson().results.bindings[0].idArk.value
Supprimer	Reuse	grel:	value.match("FRBNF(.{8}).*")

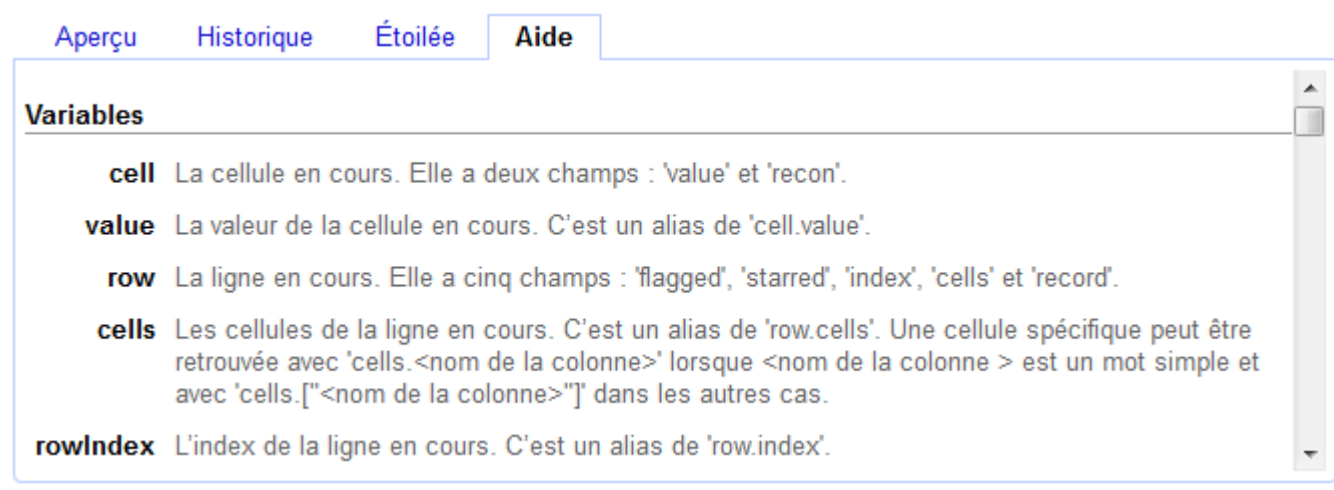
Université de Lille

L'écran de saisie des formules

Contenu conçu à partir d'une présentation de Mathieu Saby sous licence CC-BY-NC

Trois onglets utiles: Historique, Étoilée, Aide

Liste les variables et fonctions
Utiliser Crtrl+F pour rechercher une commande



Manipulations de chaînes de caractère en GREL - 1

- Concaténation avec + :
  - `"a"+"b"` => "ab"

- Sélection d'un caractère avec les crochets [] :
  - "Salut" [0] => "S"
  - "Salut" [1] => "a"
  - "Salut" [-1] => "t"

- Fonction length() :
  - "Salut".length() => 5

- Fonction substring :
  - "Salut".substring(2) => "lut"
  - "Salut".substring(2,4) => "lu"
  - "Salut".substring(1,-1) => "alu"

Manipulations de chaînes de caractère en GREL - 2

- Fonction contains() :
  - "Salut".contains("alu") => true
  - "Salut".contains("sal") => false

- Fonctions startsWith() et endsWith() :
  - "Salut".startsWith("Sa") => true
  - "Salut".startsWith("ut") => false
  - "Salut".endsWith("Sa") => false
  - "Salut".endsWith("ut") => true

- Fonction replace() :
  - "Salut".replace("alu", "por") => "Sport"
  - "Salut".replace("l", "") => "Saut"

Manipulations de tableaux en GREL

- Il y a plusieurs manières de générer des tableaux dans GREL, mais la manière la plus commune fait suite à l'utilisation de la fonction split() :
  - "Salut, comment ça va ?".split(",") => ["Salut", " comment ça va ?"]
  - "Salut, comment ça va ?".split(" ") => ["Salut,", "comment", "ça", "va", "?"]

- On peut accéder aux différents éléments d'un tableau ou connaître sa taille de la même manière qu'avec une chaîne de caractère :
  - "Salut, comment ça va ?".split(",").length() => 2 ;
  - "Salut, comment ça va ?".split(" ").length() => 5 ;
  - "Salut, comment ça va ?".split(",")[1] => " comment ça va?" ;
  - "Salut, comment ça va ?".split(" ")[-2] => "va" ;

- Fonction join() :
  - ["Salut","à","toi"].join(" ") => "Salut à toi" ;

Université de Lille

BONUS !

Aller plus loin … mais sans support !

Université de Lille

Affichage et travail par entrée

Université de Lille

Fill/Blank/Changement d’ordre d’entrée

Université de Lille

Exploiter un projet différent

Université de Lille

Références

- Contact : camille.vacher@univ-lille.fr
- Documentation officielle : https://openrefine.org/docs
- Tutoriel en français assez complet, longue bibliographie avec de nombreux exemples : https://msaby.gitlab.io/tutoriel-openrefine/