Enrichir ses données avec OpenRefine (Niveau 2)

Aurélien Moisan

ENRICHIR SES DONNÉES AVEC OPENREFINE (NIVEAU 2)

MAI 2023

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

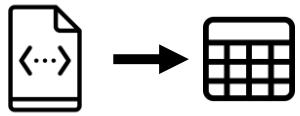
SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

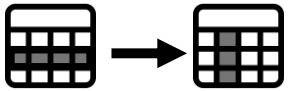
OpenRefine permet de :

Nettoyer un jeu de données

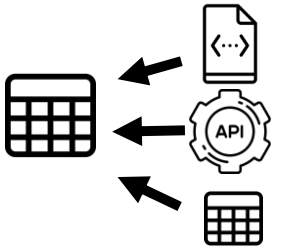
Convertir un jeu de données



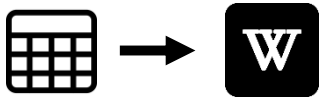
Transformer un jeu de données



Enrichir un jeu de données



Publier des données sur Wikidata



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

Sommaire

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

1

ENRICHIR VOS DONNÉES A PARTIR D'UN AUTRE PROJET

Le croisement de projet sur la base d'une colonne pivot

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 ». Avant de vous lancer dans cet démarche il faut :

- Vous assurez qu'il existe bien une colonne pivot entre vos deux projets. Sinon vous devez la créer en créant une nouvelle colonne sur la base de fusion, ou d'enrichissement (via une API par exemple) etc…
- Dans vos deux projets la colonne pivot doit être propre : supprimez les espaces superflus, vérifiez que les données ont bien été harmonisées dans la même manière etc..
- Dans votre projet de destination vérifiez qu'aucune facette/filtre inutile n'est activé. Sinon l'import ne s'effectuera que sur un jeu de données réduit

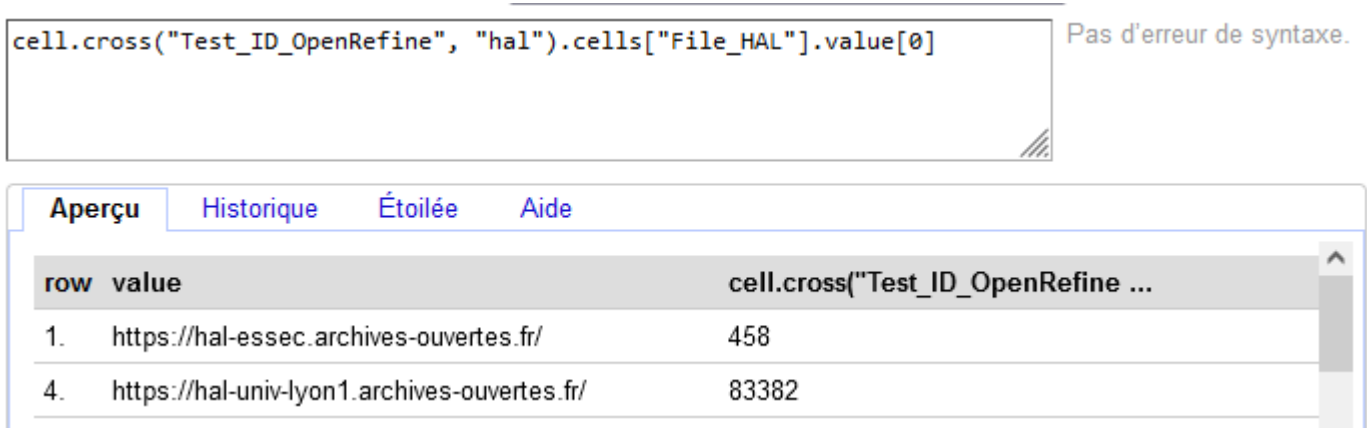
Le croisement de projet sur la base d'une colonne pivot

SORBONNE UNIVERSITÉ
BIBLIOTHÈQUE UNIVERSITAIRE

Ouvrez les deux projets dans deux onglets. Dans le projet de destination, rendez vous sur la colonne pivot, cliquez sur « ajouter une colonne » en fonction de cette colonne puis utilisez 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]



SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

2

ENRICHIR VOS DONNÉES GRACE À LA RECONCILIATION

La réconciliation

La réconciliation vous permet d'aligner des valeurs de votre de jeu de données avec une source externe via une API. Ces sources externes sont structurées et riches. La réconciliation vous permet d'associer des identifiants uniques à vos valeurs et d'enrichir votre jeu de données sur la bases de ces identifiants.

Liste des services de réconciliation disponibles : https://reconciliation-api.github.io/testbench/#/

Pour débuter votre réconciliation cliquez sur la colonne qui contient la valeur que vous souhaitez aligner > Réconcilier > Démarrer la réconciliation

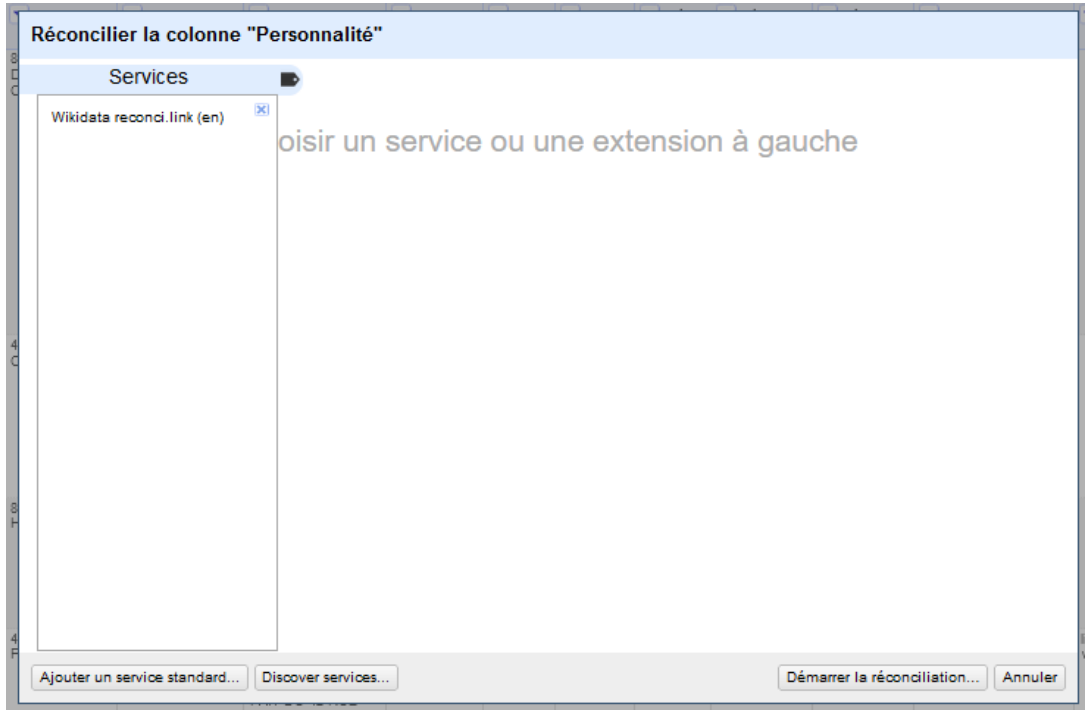
Choisir et ajouter un service de réconciliation

SORBONNE UNIVERSITÉ
BIBLIOTHÈQUE UNIVERSITAIRE

Par défaut seul Wikidata est proposé. Notez que le service de réconciliation Wikidata proposé par défaut fonctionne mal, il est donc préférable de le réimporter.

Pour cela cliquez sur «Ajouter un service standard » et ajoutez l'URL de l'API que vous pourrez récupérer dans la liste de services

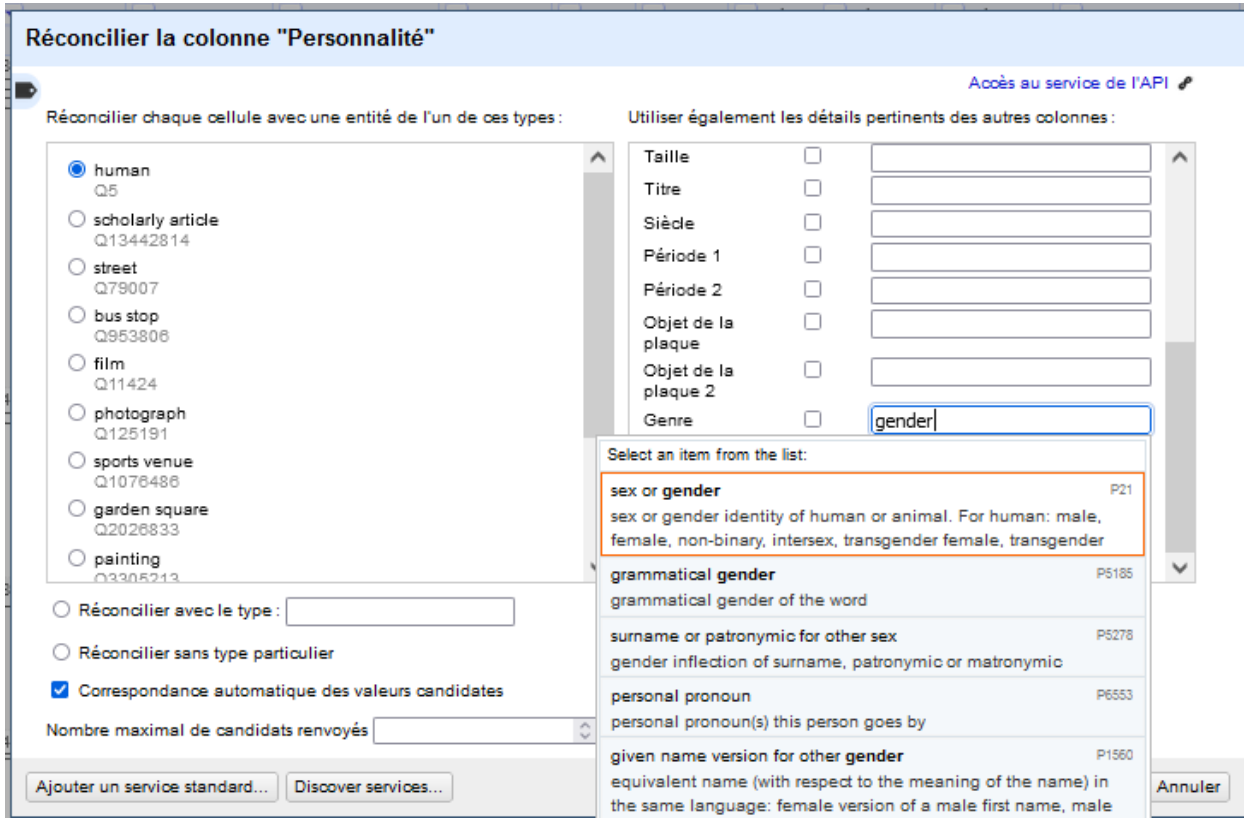
Notez que tous les services ne proposent pas les mêmes fonctionnalités. L'import de données externes via la réconciliation n'est disponible que pour les services qui ont la case « extend data » de cochée



Paramétrer sa réconciliation

La paramétrage permet d'optimiser les résultats de la réconciliation. Selon les services vous pourrez choisir :

- Un type : un sous ensemble de la base externe sur lequel travailler
- Dans la zone de droite, vous retrouvez les colonnes de votre tableau que vous pouvez aligner avec des propriétés équivalentes dans la base externe pour optimiser la désambiguïsation



Parcourir les réconciliations

A chaque correspondance trouvée, OpenRefine attribue un score.

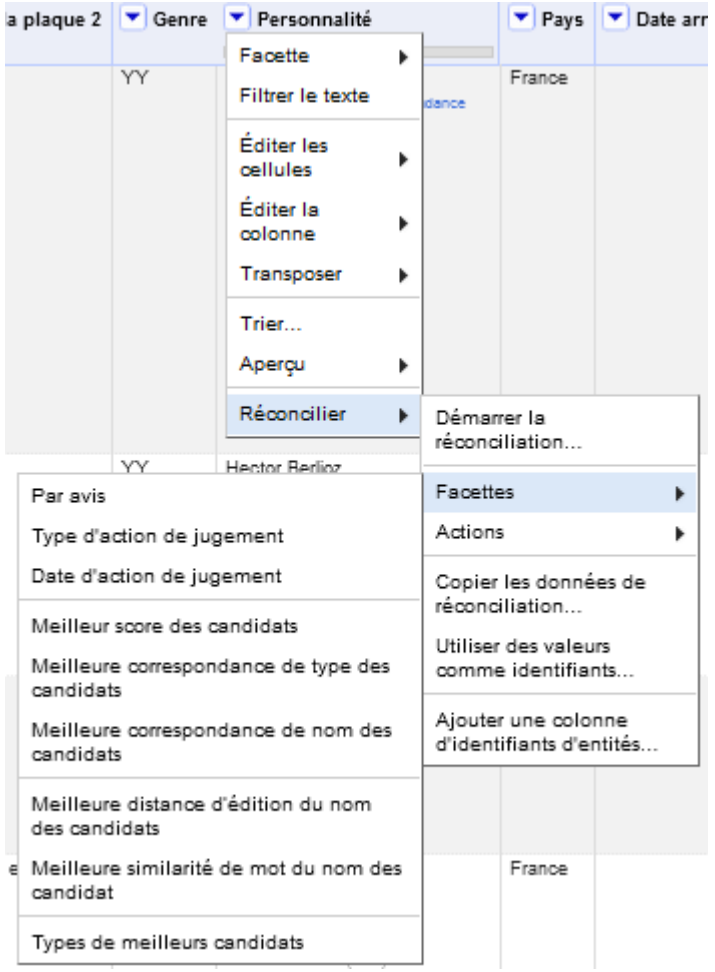
Une fois la réconciliation terminée vous pouvez parcourir le résultat avec des facettes dédiées :

La **facette « avis »** vous permet notamment d'afficher :

- Les matched : correspondance validée car elle avait un score suffisamment élevé et suffisamment supérieur aux scores des autres candidats
- Les none : le score n'a pas permis de dégagé un candidat, une validation manuelle est donc nécessaire

La **facette « jugement »** vous permet d'isoler : les résultats en attente de validation, les résultats validés manuellement, les résultats validés manuellement en masse, les résultats validés automatiquement.

D'autre facettes vous permette de parcourir les scores et les écarts de score, la date de validation etc…



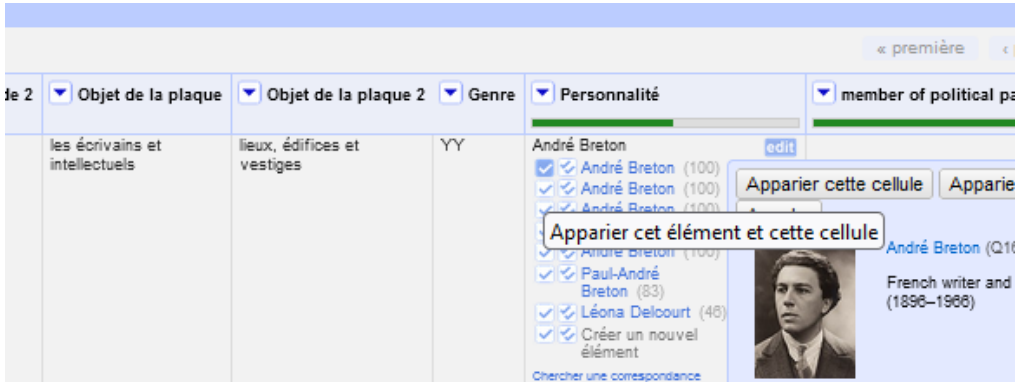
Valider les données

Vous pouvez valider une réconciliation en cliquant sur « Apparier cet éléments et cette cellule ». Vous pouvez également valider cette réconciliation pour toutes les cellules identiques

Notez que cette action est réversible.

Vous pouvez également le faire en masse via le menu dans la colonne.

Dans le cas de Wikidata vous pouvez également « créer un nouvel élément » ce qui permettrait de créer un nouvel élément si vous exportez vos données vers Wikidata

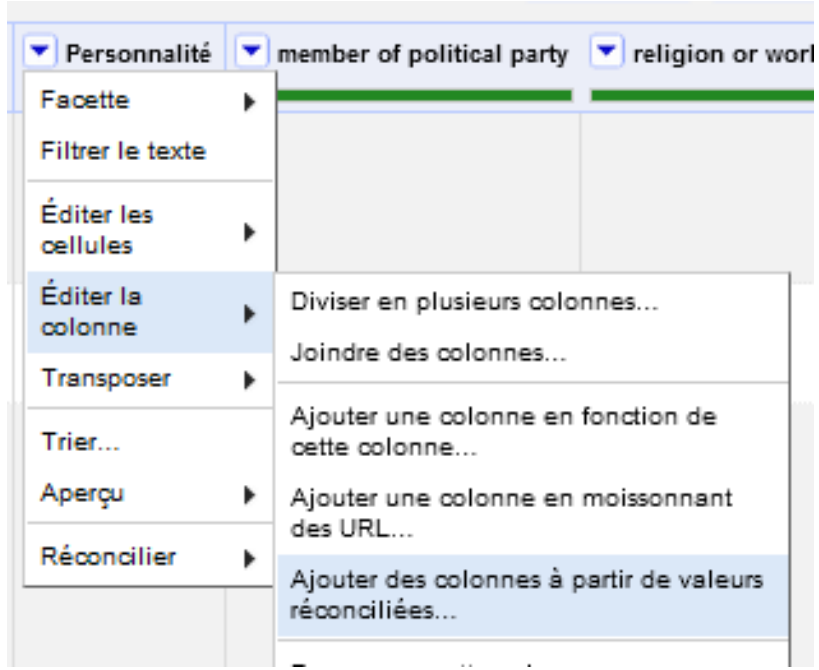


Enrichir votre jeu de données à partir des réconciliations

SORBONNE UNIVERSITÉ
BIBLIOTHÈQUE UNIVERSITAIRE

Lorsque votre service de réconciliation propose la fonctionnalité « Extend data », vous pouvez l'utiliser via le bouton « Ajouter une colonne à partir des valeurs réconcilier ». Ainsi vous pouvez :

- enrichir votre jeu de données une fois les réconciliations vérifiées et validées
- vérifier des réconciliations validées à l'aide de données complémentaires



SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

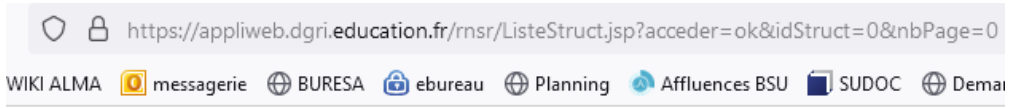
2

ENRICHIR VOS DONNÉES À PARTIR D'UNE PAGE WEB

Identifier un lien entre vos données et l'URL des pages web à moissonner

Si l'URL n'est pas déjà présente dans votre jeu de données, il faut que les URL des pages web que vous souhaitez moissonner aient une structure fixe, et que l'élément qui varie d'une page à une autre soit présent dans votre jeu de données. Il faut utiliser un lien profond :

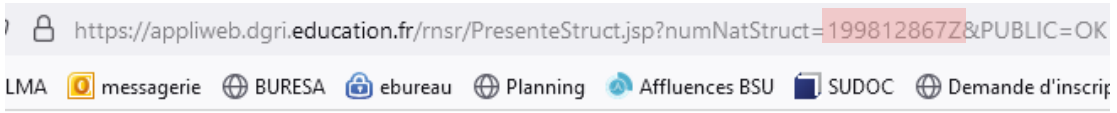
Exemple avec deux structures de liens qui renvoient tous les deux vers la même page du RNSR :



DE L'ENSEIGNEMENT SUPÉRIEUR,
HERCHE ET DE L'INNOVATION

199812867Z : LMD Laboratoire de météorologie dynamique
Unité de recherche (situation 2023)

Le lien qui ne pointe pas spécifiquement sur la page (inexploitable)



L'ENSEIGNEMENT SUPÉRIEUR,
RCHE ET DE L'INNOVATION

199812867Z : LMD Laboratoire de météorologie dynamique
Unité de recherche (situation 2023)

Lien profond qui contient un identifiant (exploitable)

Générer une URL à partir de vos données

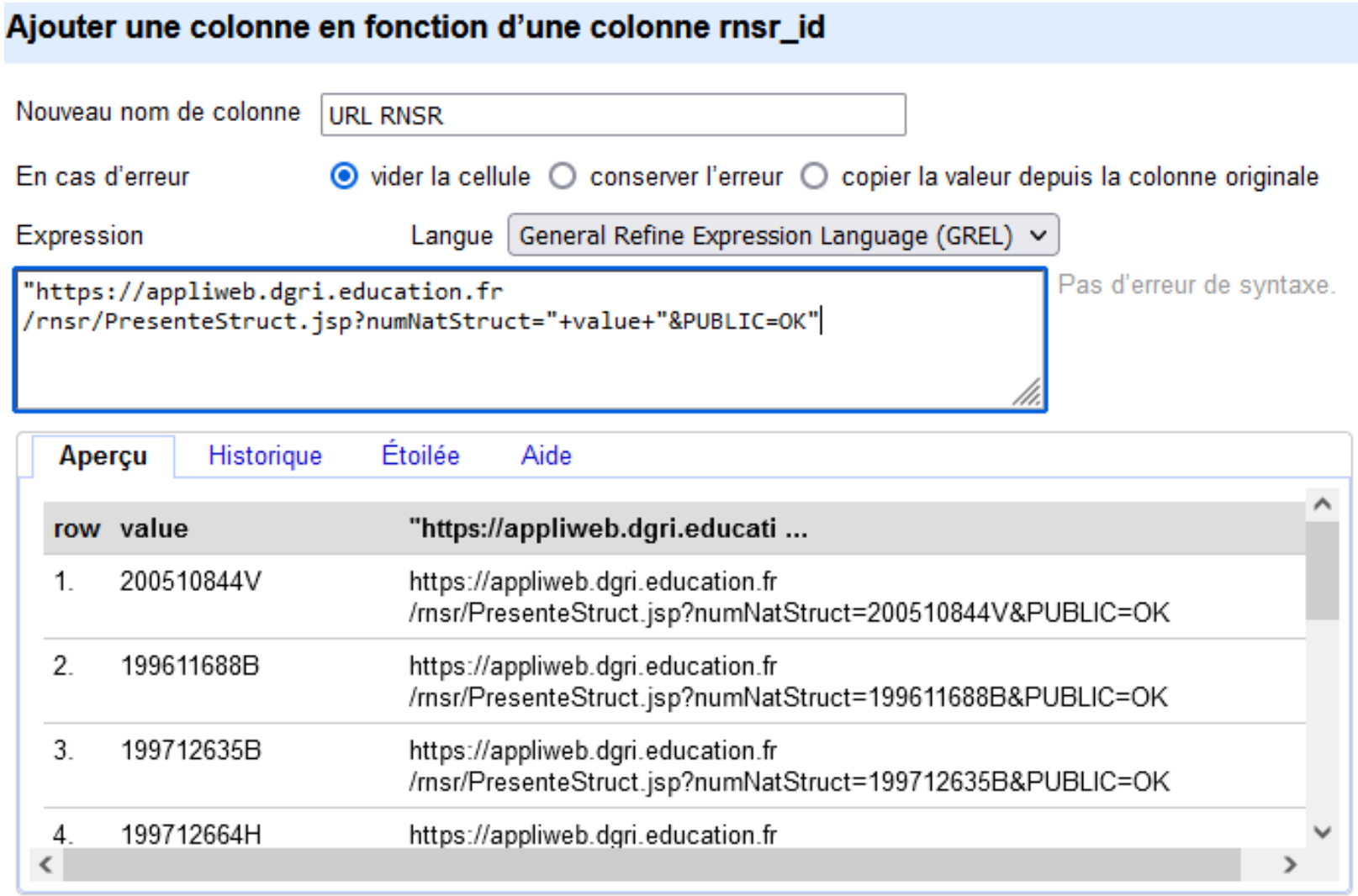
L'import de l'HTML d'une page web se fera à partir de l'URL. Pour cela vous allez devoir recréer l'URL de la page à partir de vos données. Pour cela le plus simple est d'utiliser la concaténation

Exemple :
Si j'ai une colonne qui contient des identifiants RNSR pour les transformer en URL je dois :

- Ajouter « *https://appliweb.dgri.education.fr/rnsr/PresenteStruct.jsp?numNatStruct=* » comme préfixe
- Ajouter « *&PUBLIC=OK* » comme suffixe

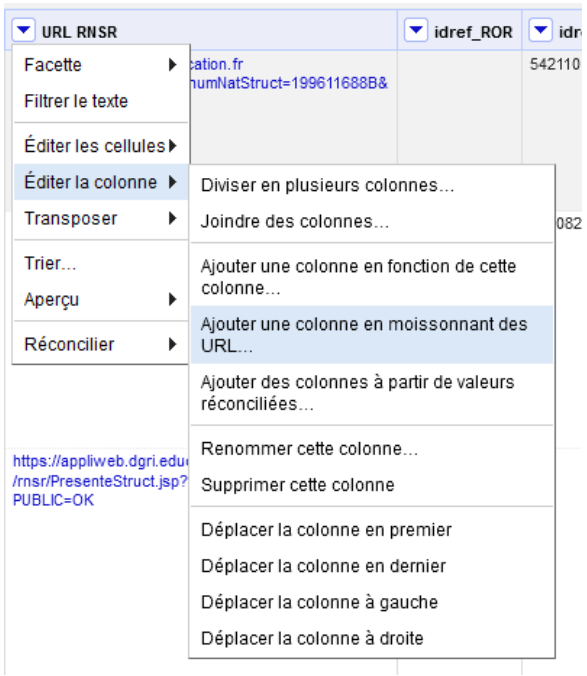
Ainsi pour chacune des lignes vous allez générer un lien de la forme :
« https://appliweb.dgri.education.fr/rnsr/PresenteStruct.jsp?numNatStruct=IDENTIFIANT_RNSR&PUBLIC=OK »

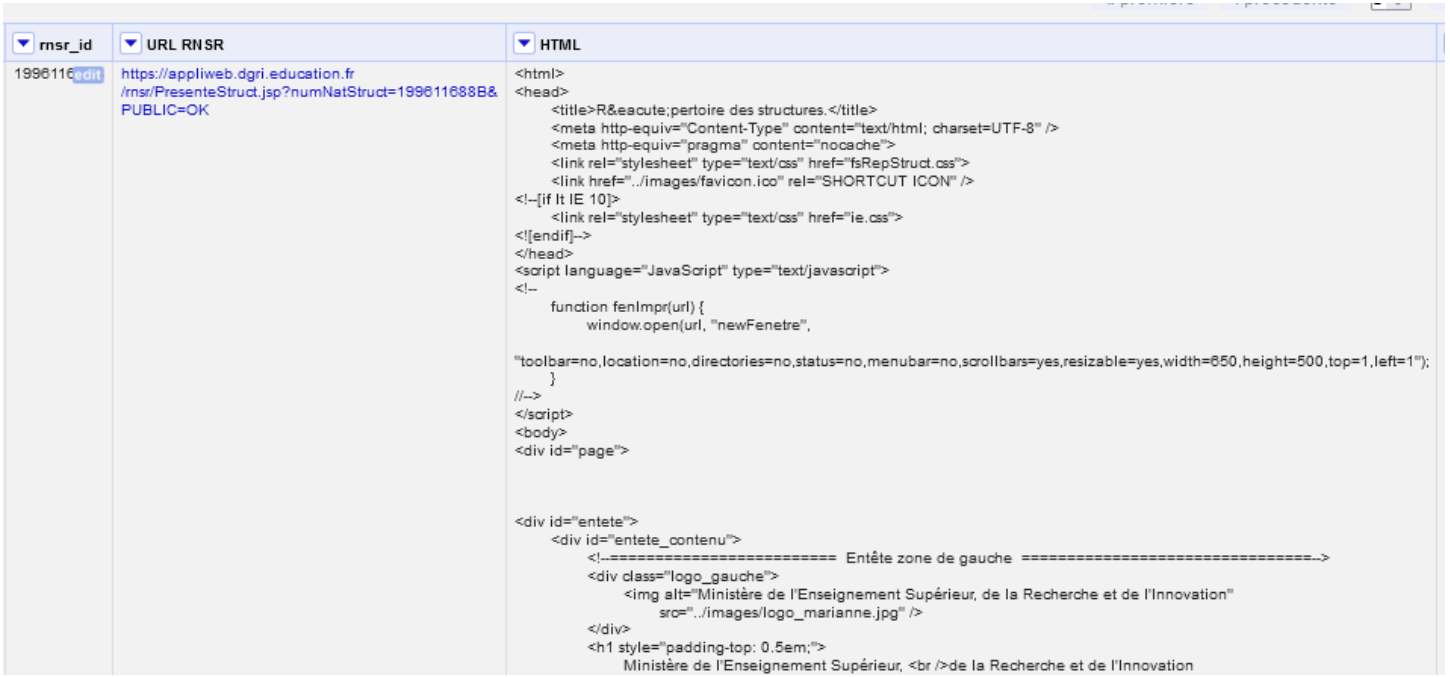
Générer une URL à partir de vos données



Importer l'HTML

Une fois l'URL générée vous avez simplement à utiliser la fonction « Ajouter une colonne en moissonnant des URL » sur la colonne qui contient les URL. L'HTML de chaque page sera alors importé dans une nouvelle colonne. Cette étape peut prendre du temps surtout sur des gros jeux de données :





Parser les données HTML

Une fois l'HTML récupéré il faut parser les données pour aller pointer sur la balise qui contient les données qui vous intéresse. Pour cela vous allez utiliser la fonction value.parseHtml( ) et utiliser l'***arborescence*** de l'HTML qui va aller pointer la donnée qui vous intéresse :

value.parseHtml().select('***html body div[id=page] div[id=contenu] table tr td table tr td***')[0].ownText()

```
2. <html>
<head>…
</head>
<body>
<div id="page">
<div id="entete">…
</div>
    <div class="menu_container">…
    </div>
    <div id="contenu">
        <p class="erreur">…
        </p>

        <div align="center" style="display:inline-block; text-align:center;min-width:700px;">…
        </div>

        <table width="95%" align="center" style="display:clear;">
            <tr>
                <td>

                    <table border="0" cellpadding="5" width="100%">
                        <tr>
                            <td valign="middle"><h2>Responsable(s)</h2>

                Directrice-adjointe - Nathalie DOSTATNI  &agrave; partir du 01/09/2016
                Directrice - Angela TADDEI  &agrave; partir du 01/09/2015
                            </td>
```

Transformation textuelle personnalisée sur la colonne HTML

Expression — Langue: General Refine Expression Language (GREL)

```
value.parseHtml().select('html body div[id=page] div[id=contenu]
table tr td table tr td' )[0].ownText()
```

Pas d'erreur de syntax

Aperçu | Historique | Étoilée | Aide

row	value	value.parseHtml().select('html ...
2.	<html> <head> <title>Répertoire des structures. </title> <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" /> <meta http-equiv="pragma" content="nocache"> <link rel="stylesheet" type="text/css"	Directrice-adjointe - Nathalie DOSTATNI à partir du 01/09/2016 Directrice - Angela TADDEI à partir du 01/09/2015

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

3

ENRICHIR VOS DONNÉES AVEC UNE API

Interroger l'API à partir de vos données

Sur le même principe que les URL vous allez devoir construire une requête API qui pointe vers le résultat que vous souhaitez. Si vous souhaitez obtenir d'un seul résultat et un résultat fiable vous devrez utiliser un identifiant pérenne ou à défaut une clé d'identification forte.

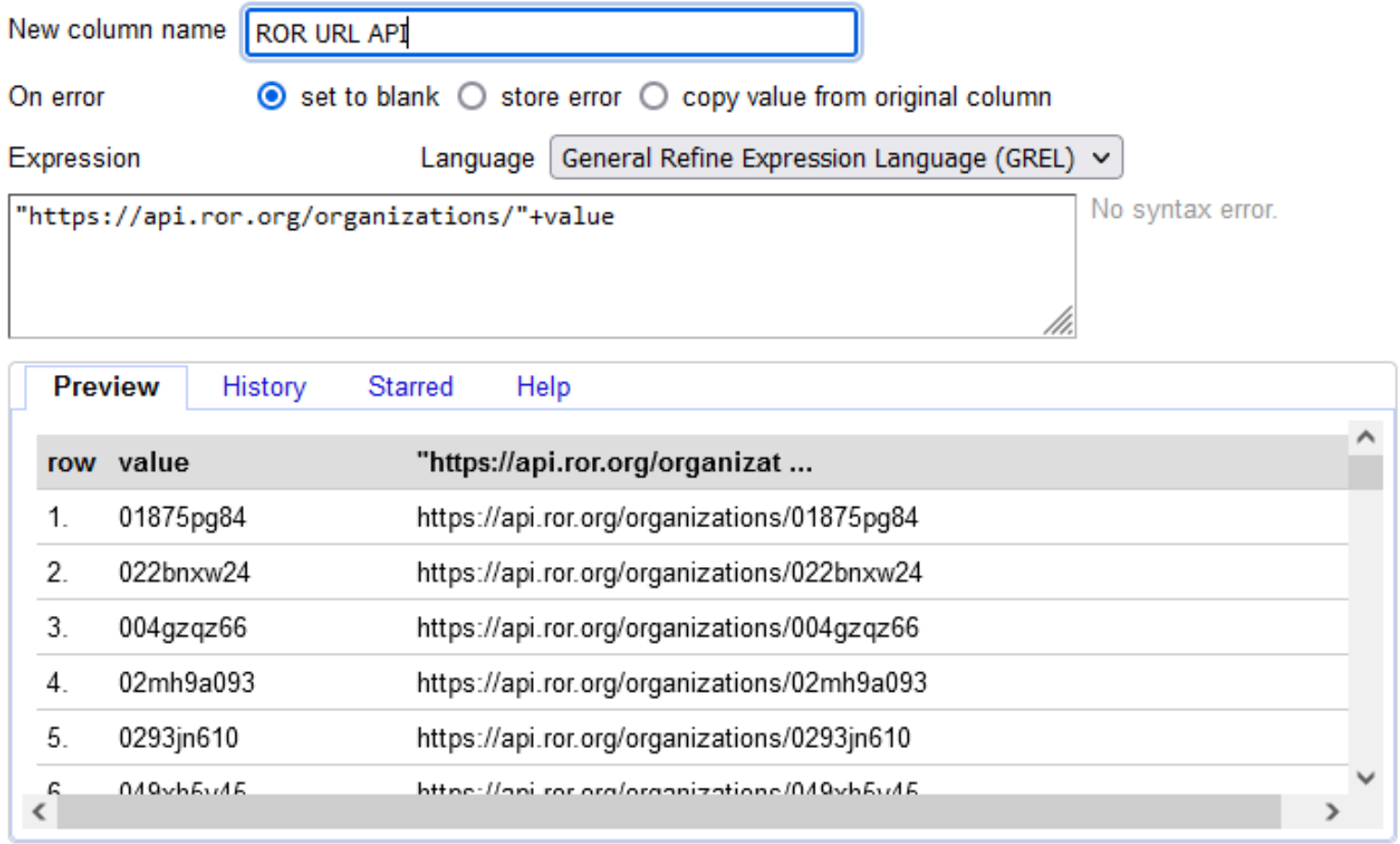
Exemple :

Une requête api basée sur un identifiant pérenne:
https://api.openalex.org/institutions/I39804081

Une requête basé sur une clé d'identification forte : titre + année de publication + orcid
https://api.openalex.org/works?filter=title.search:''QUANTUM%20ESPRESSO: a modular and open-source software project for quantum simulations of materials'',publication_year:2009,authorships.author.orcid:0000-0002-9635-3227

Générer une URL à partir de vos données

Comme pour le moissonnage HTLM, l'objectif va donc être d'utiliser la concaténation pour générer une requête API à partir de vos données



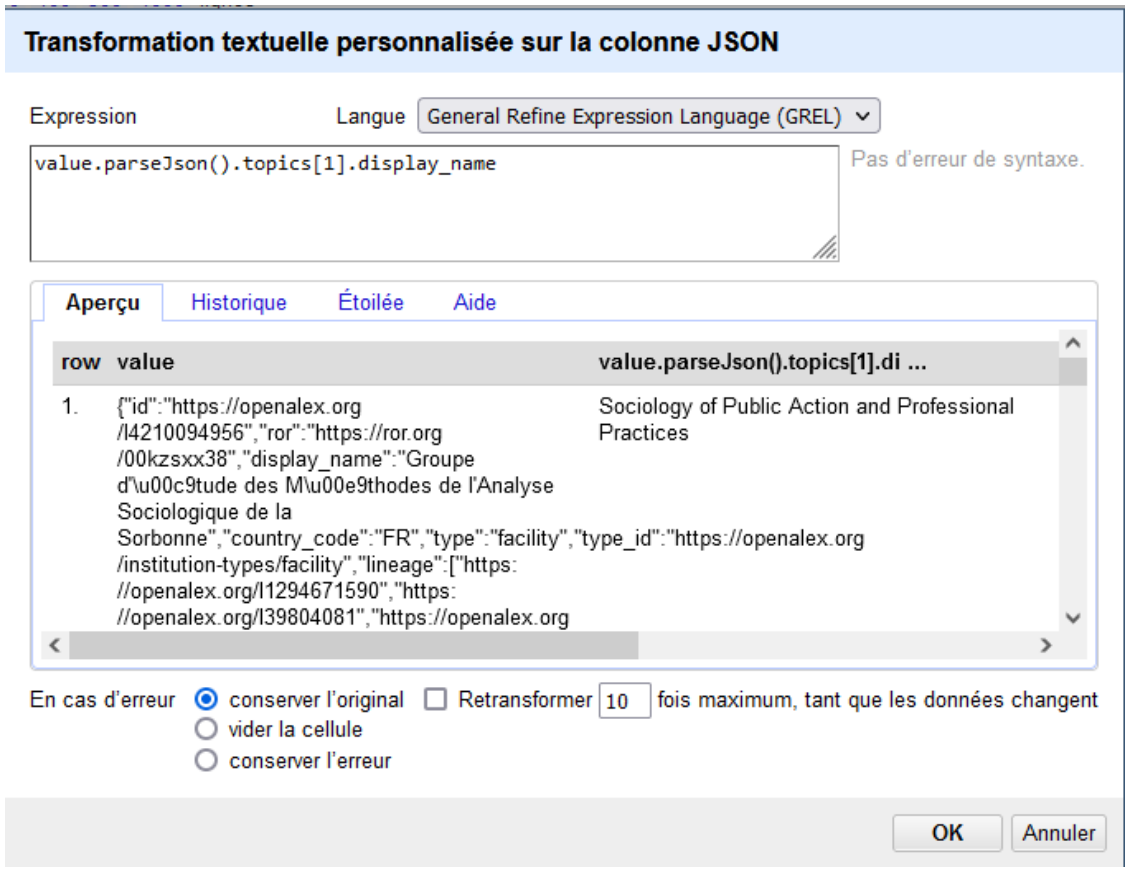
Parser les données Json

Les API proposent le plus souvent des résultats en JSON (pour parser les données XML/HTML voir diapo).

Pour parser le json utilisez la fonction value.parseJson( )

Puis comme pour l'html vous déclarer le chemin où se trouve la donnée que vous souhaitez afficher en séparant chaque niveau de la hiérarchie JSON par un «.»

Dans l'exemple ci-contre, je pointe sur l'éléments *topics* (comme j'ai plusieurs éléments topics je choisis lequel je veux pointer, attention l'indexation commence à 0 donc ici je pointe sur le 2e), puis sur le sous-élément *display_name*



SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

4

PARSER DES DONNÉES MULTIVALUÉES

Parser les données

Sélectionner plusieurs balises

SORBONNE UNIVERSITÉ
BIBLIOTHÈQUE UNIVERSITAIRE

idref2id

```
<?xml version="1.0" encoding="UTF-8"?>
<sudoc service="idref2source">
<query><ppn>19219884X</ppn><result><source>RNSR</source>
<identifiant>199812880N</identifiant></result></query><query><ppn>19219884X</ppn>
<result><source>HAL</source><identifiant>541668</identifiant></result></query><query>
<ppn>19219884X</ppn><result><source>HAL</source><identifiant>541692</identifiant>
</result></query><query><ppn>19219884X</ppn><result><source>HAL</source>
<identifiant>542129</identifiant></result></query>
</sudoc>
```

Formule

```
forEach(filter(value.parseXml().select("query"),
e,e.select("source")[0].ownText()=="HAL"), f,
f.select("identifiant")[0].ownText())
```

idhal

```
[541668,
541692,
542129]
```

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

5

COMPARER DES COLONNES A L'AIDE DE FACETTES PERSONNALISÉES

Créer des facettes textuelles personnalisées
Comparer deux colonnes

SORBONNE UNIVERSITÉ
BIBLIOTHÈQUE UNIVERSITAIRE

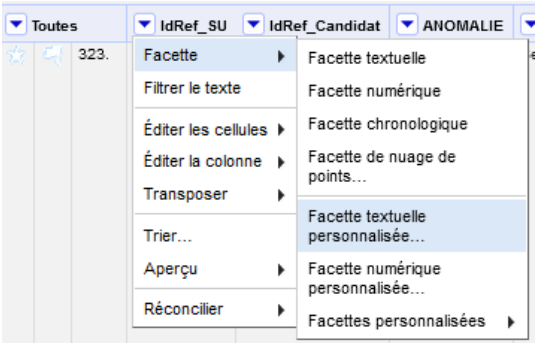
Les facettes textuelles personnalisées vous permettent notamment de comparer des chaines de caractères :

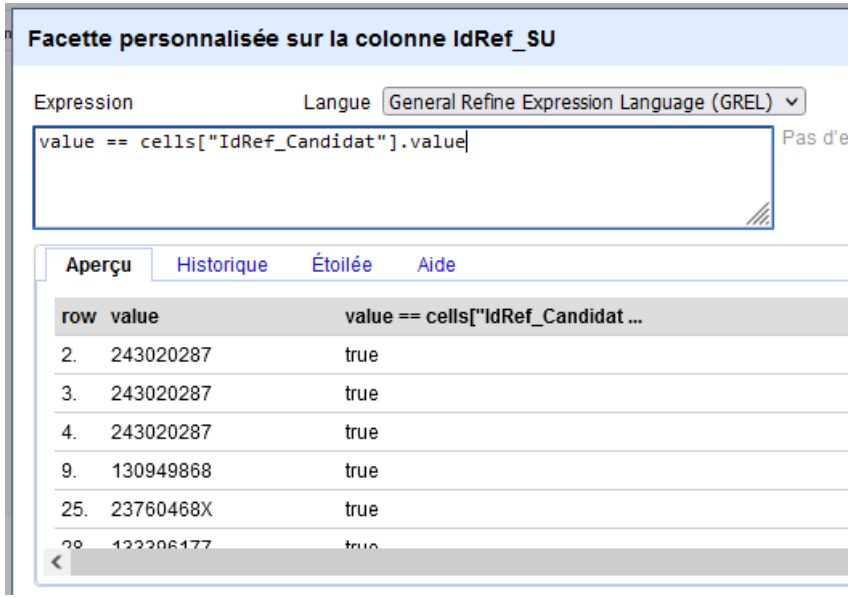
- Sur l'une des colonnes à comparer : Facettes > Facette textuelle personnalisée
- Utilisez la formule :

value == cells["**colonne à comparer**"].value

L'exemple ci-dessous compare les valeurs de la colonne IdRef_SU avec la colonne IdRef_Candidat

value == cells[« **IdRef_Candidat**"].value





Créer des facettes textuelles personnalisées
Comparer deux colonnes

Cela permet de créer une facette booléenne :

- **true** : affiche les lignes pour lesquelles les valeurs de IdRef_SU et IdRef_Candidat sont similaires
- **false** : affiche les lignes pour lesquelles les valeurs de IdRef_SU et IdRef_Candidat sont différentes

Pour aller plus loin

- VIB-Bits : un plug'in qui fourni une interface graphique pour croiser vos projets
- Un tutoriel OpenRefine sur la réconciliation Wikidata : https://www.wikidata.org/wiki/Wikidata:Tools/OpenRefine/Editing/Tutorials/Video/fr
- Le Mooc Wikidata de Wikimedia France pour mieux appréhender la structure des données dans Wikidata : https://www.wikidata.org/wiki/Wikidata:Tools/OpenRefine/Editing/Tutorials/Video/fr
- La formation de Mathieu Saby : https://fr.slideshare.net/slideshow/nettoyer-et-transformer-ses-donnes-avec-openrefine-partie-2/99618030

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

Des questions ? Des besoins de formation ? D'accompagnement ?

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE

Le **LAB de Ressources et d'Accompagnement aux DOnnées de la Recherche :**

Site : https://labrador.sorbonne-universite.fr/
Mail : labrador@sorbonne-universite.fr

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

MERCI

labrador@sorbonne-universite.fr

SORBONNE UNIVERSITÉ

BIBLIOTHÈQUE UNIVERSITAIRE