



HAL
open science

Enrichir ses données avec OpenRefine (Niveau 2)

Aurélien Moisan

► To cite this version:

Aurélien Moisan. Enrichir ses données avec OpenRefine (Niveau 2). Printemps de la Donnée 2024, INRAE; Université Haute-Alsace; Université de Strasbourg; INSA; PNDB; AgroParisTech; Université de Lille; Sorbonne Université; Data Terra, May 2024, Paris, France. 10.5281/zenodo.11449613 . hal-04684375

HAL Id: hal-04684375

<https://hal.inrae.fr/hal-04684375v1>

Submitted on 2 Sep 2024

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

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



Distributed under a Creative Commons Attribution 4.0 International License

ENRICHIR SES DONNÉES AVEC OPENREFINE (NIVEAU 2)

MAI 2023

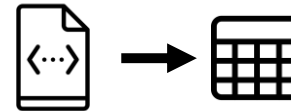


OpenRefine permet de :

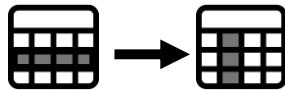
Nettoyer un jeu de données



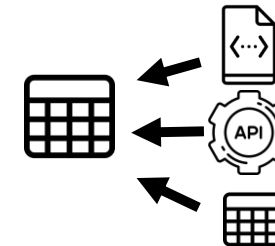
Convertir un jeu de données



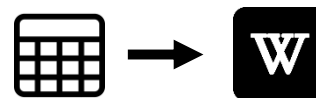
Transformer un jeu de données



Enrichir un jeu de données



Publier des données sur Wikidata



Sommaire

- 1. ENRICHIR VOS DONNÉES À PARTIR D'UN AUTRE PROJET**
- 2. ENRICHIR VOS DONNÉES GRÂCE À LA RÉCONCILIATION**
- 3. ENRICHIR VOS DONNÉES À PARTIR D'UNE PAGE WEB**
- 4. ENRICHIR VOS DONNÉES À PARTIR D'UNE API**

1

ENRICHIR VOS DONNÉES A PARTIR D'UN AUTRE
PROJET

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

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 cette 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

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]`

```
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

2

ENRICHIR VOS DONNÉES GRACE À LA
RECONCILIATION

La réconciliation

La réconciliation vous permet d'aligner des valeurs de votre 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

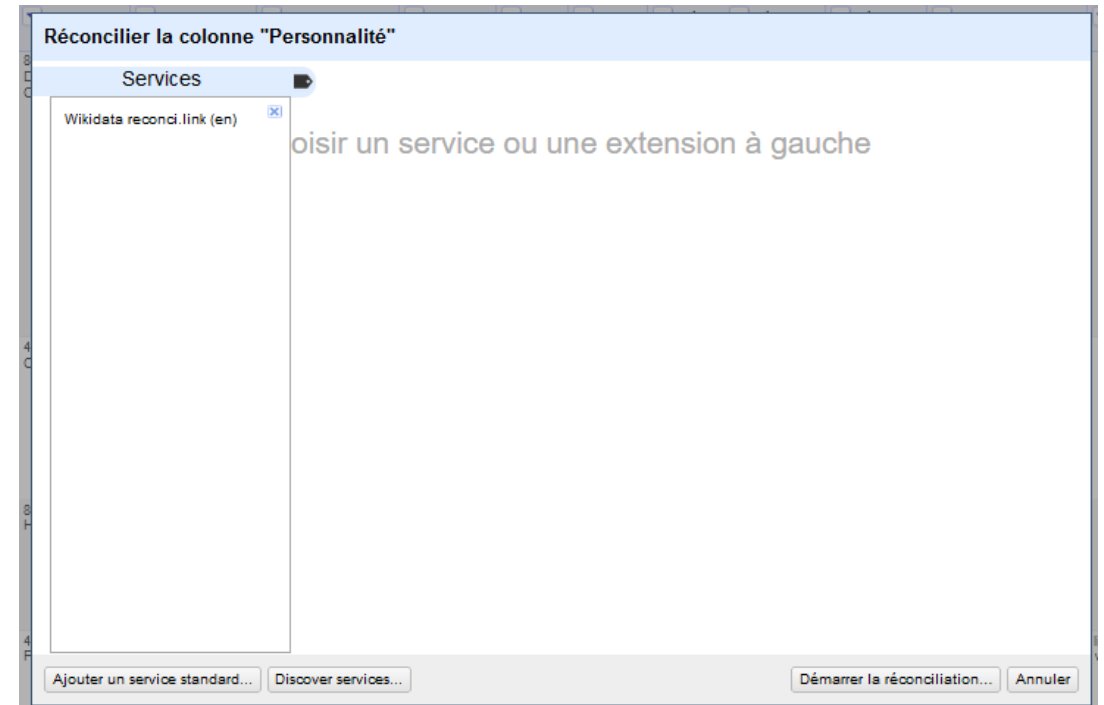


Choisir et ajouter un service de réconciliation

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](#)

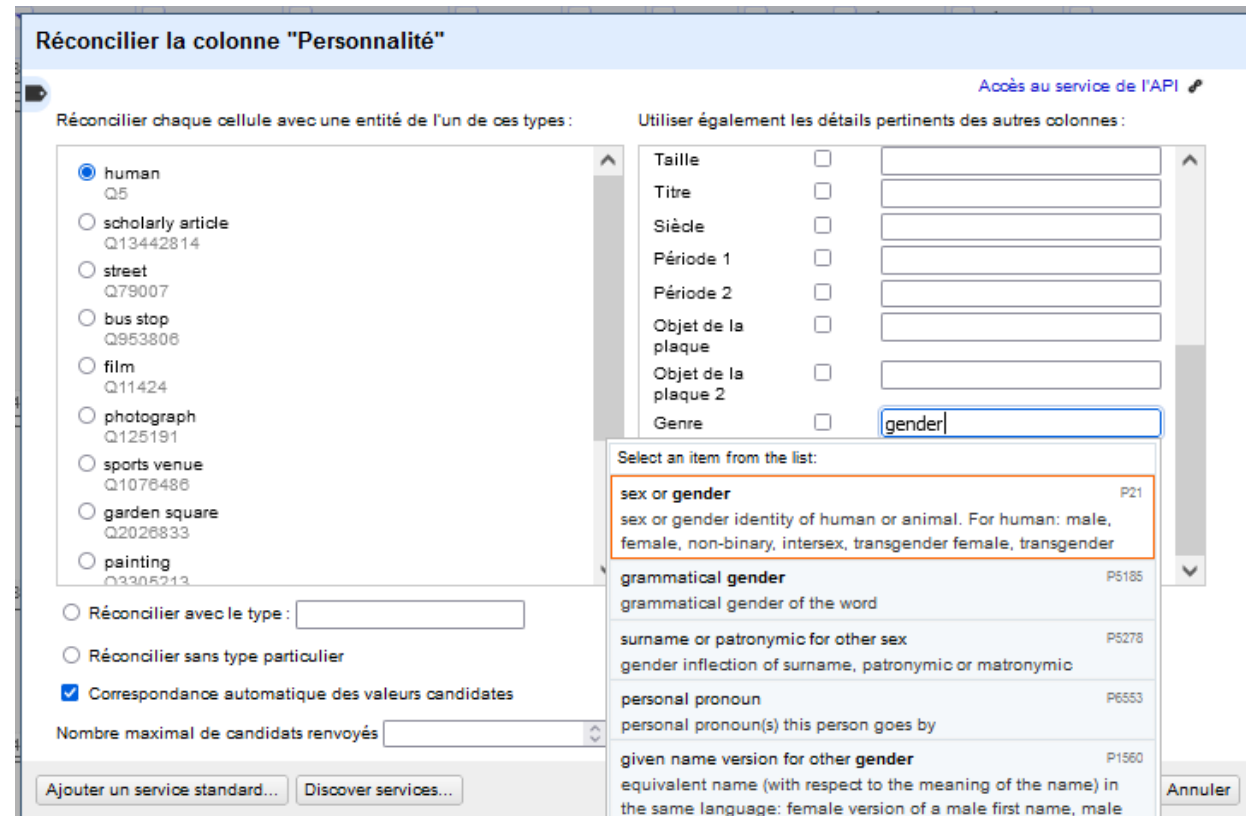
Notez que tous les services ne proposent pas les mêmes fonctionnalités. L'import de données externes via la réconciliation n'est disponible que pour les services qui ont la case « extend data » de cochée



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



Réconcilier la colonne "Personnalité" [Accès au service de l'API](#)

Réconcilier chaque cellule avec une entité de l'un de ces types :

- human (Q5)
- scholarly article (Q13442814)
- street (Q79007)
- bus stop (Q953806)
- film (Q11424)
- photograph (Q125191)
- sports venue (Q1076488)
- garden square (Q2026833)
- painting (Q3305213)

Réconcilier avec le type :

Réconcilier sans type particulier

Correspondance automatique des valeurs candidates

Nombre maximal de candidats renvoyés

Utiliser également les détails pertinents des autres colonnes :

- Taille
- Titre
- Siècle
- Période 1
- Période 2
- Objet de la plaque
- Objet de la plaque 2
- Genre

Select an item from the list:

- sex or gender** (P21)
sex or gender identity of human or animal. For human: male, female, non-binary, intersex, transgender female, transgender
- grammatical gender (P5185)
grammatical gender of the word
- surname or patronymic for other sex (P5278)
gender inflection of surname, patronymic or matronymic
- personal pronoun (P6553)
personal pronoun(s) this person goes by
- given name version for other gender (P1560)
equivalent name (with respect to the meaning of the name) in the same language: female version of a male first name, male

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 dégager 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'autres facettes vous permettent de parcourir les scores et les écarts de score, la date de validation etc...

The screenshot shows a table with columns: 'la plaque 2', 'Genre', 'Personnalité', 'Pays', and 'Date arr'. The 'Genre' column has a facet 'YY'. The 'Personnalité' column has a facet 'Hector Berlioz'. The 'Pays' column has a facet 'France'. A context menu is open over the 'Personnalité' column, listing actions: 'Facette', 'Filtrer le texte', 'Éditer les cellules', 'Éditer la colonne', 'Transposer', 'Trier...', 'Aperçu', and 'Réconcilier'. The 'Réconcilier' option is highlighted, and a sub-menu is visible with options: 'Démarrer la réconciliation...', 'Facettes', 'Actions', 'Copier les données de réconciliation...', 'Utiliser des valeurs comme identifiants...', 'Ajouter une colonne d'identifiants d'entités...', 'Meilleur score des candidats', 'Meilleure correspondance de type des candidats', 'Meilleure correspondance de nom des candidats', 'Meilleure distance d'édition du nom des candidats', 'Meilleure similarité de mot du nom des candidats', and 'Types de meilleurs candidats'.

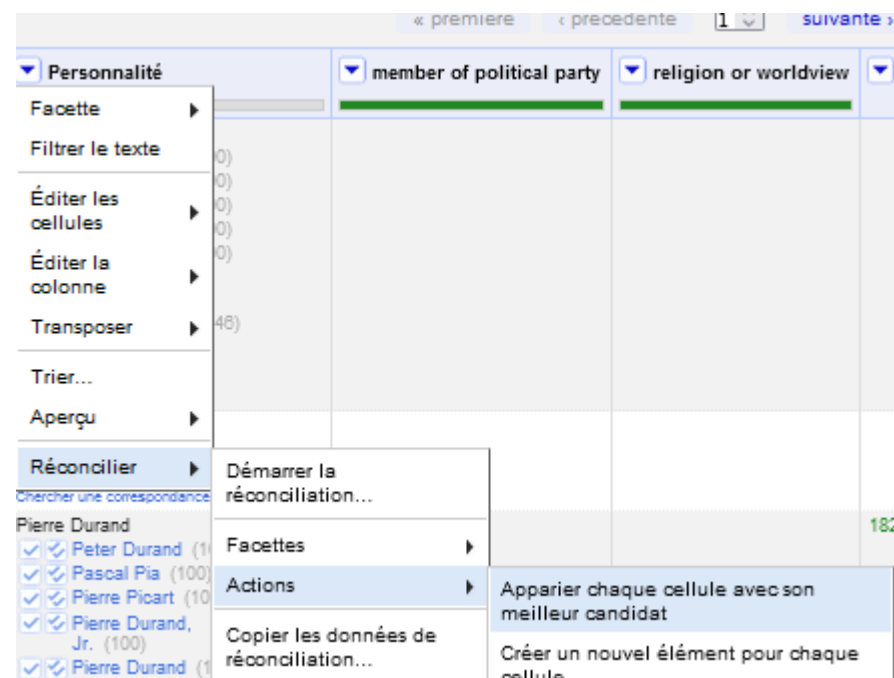
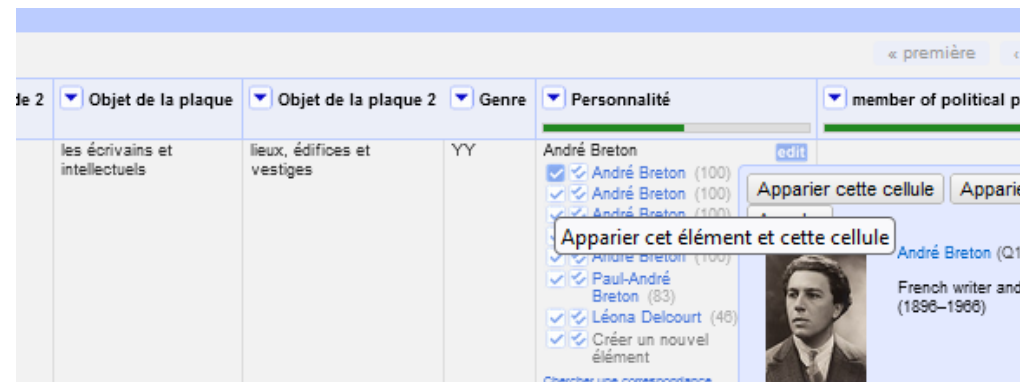
Valider les données

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.

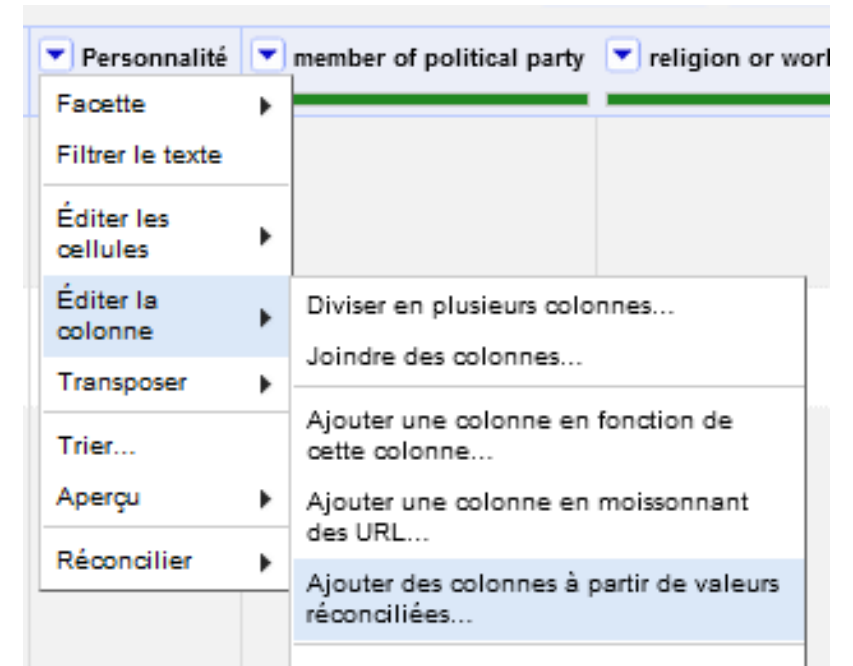
Dans le cas de Wikidata vous pouvez également « créer un nouvel élément » ce qui permettrait de créer un nouvel élément si vous exportez vos données vers Wikidata



Enrichir votre jeu de données à partir des réconciliations

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



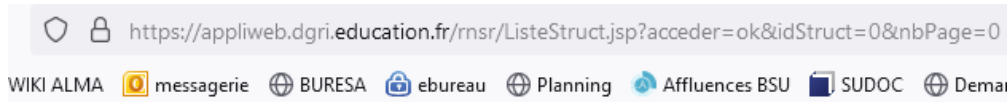
2

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 :

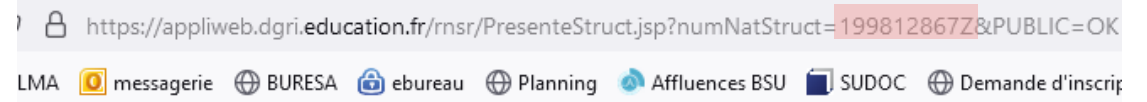
Exemple avec deux structures de liens qui renvoient tous les deux vers la même page du RNSR :



MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR,
RECHERCHE 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)



MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR,
RECHERCHE 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

Ajouter une colonne en fonction d'une colonne rnsr_id

Nouveau nom de colonne

En cas d'erreur vider la cellule conserver l'erreur copier la valeur depuis la colonne originale

Expression Langue

```
"https://appliweb.dgri.education.fr  
/rnsr/PresenteStruct.jsp?numNatStruct="+value+"&PUBLIC=OK"
```

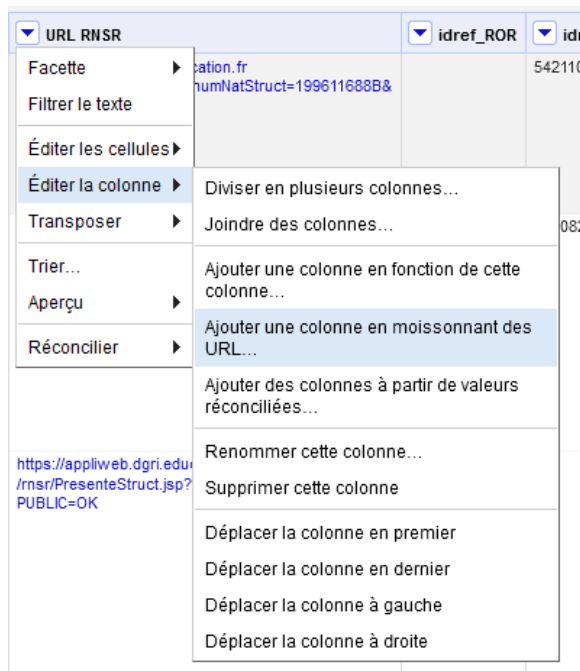
Pas d'erreur de syntaxe.

Aperçu Historique Étoilée Aide

row	value	"https://appliweb.dgri.educati ..."
1.	200510844V	https://appliweb.dgri.education.fr /rnsr/PresenteStruct.jsp?numNatStruct=200510844V&PUBLIC=OK
2.	199611688B	https://appliweb.dgri.education.fr /rnsr/PresenteStruct.jsp?numNatStruct=199611688B&PUBLIC=OK
3.	199712635B	https://appliweb.dgri.education.fr /rnsr/PresenteStruct.jsp?numNatStruct=199712635B&PUBLIC=OK
4.	199712664H	https://appliweb.dgri.education.fr

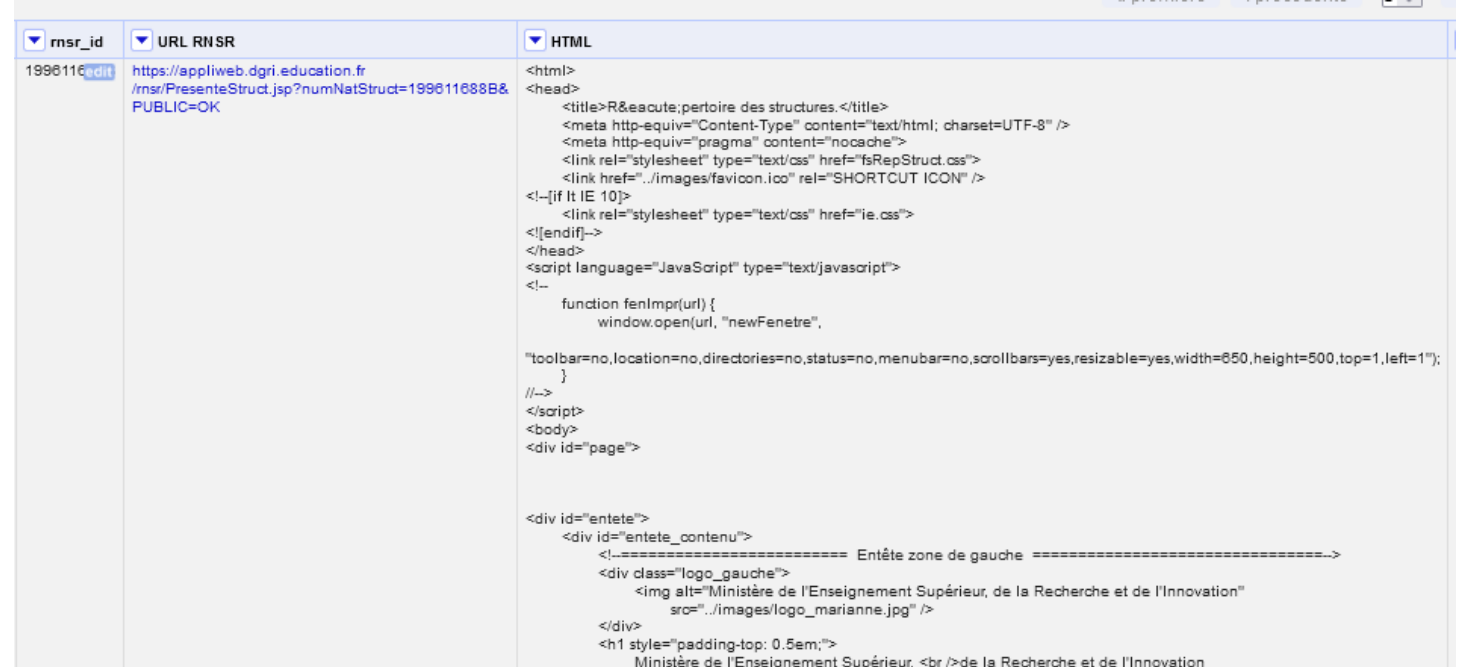
Importer l'HTML

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 :



The screenshot shows a table with three columns: 'URL RNSR', 'idref_ROR', and 'idr'. The 'URL RNSR' column contains the URL 'https://appliweb.dgri.education.fr/rnsr/PresenteStruct.jsp?numNatStruct=199611688B&PUBLIC=OK'. A context menu is open over this cell, listing various actions. The option 'Ajouter une colonne en moissonnant des URL...' is highlighted in blue.

URL RNSR	idref_ROR	idr
https://appliweb.dgri.education.fr/rnsr/PresenteStruct.jsp?numNatStruct=199611688B&PUBLIC=OK		542110



The screenshot shows a table with three columns: 'rnsr_id', 'URL RNSR', and 'HTML'. The 'URL RNSR' column contains the same URL as in the previous screenshot. The 'HTML' column contains the HTML code of the page, including the title 'Répertoire des structures.', meta tags, and a JavaScript function 'fenImpr'.

rnsr_id	URL RNSR	HTML
199611688B	https://appliweb.dgri.education.fr/rnsr/PresenteStruct.jsp?numNatStruct=199611688B&PUBLIC=OK	<html> <head> <title>Répertoire des structures.</title> <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" /> <meta http-equiv="pragma" content="no-cache"> <link rel="stylesheet" type="text/css" href="fsRepStruct.css"> <link href="..../images/favicon.ico" rel="SHORTCUT ICON" /> <!--[if IE 10]> <link rel="stylesheet" type="text/css" href="ie.css"> <![endif--> </head> <script language="JavaScript" type="text/javascript"> <!-- function fenImpr(url) { window.open(url, "newFenetre", "toolbar=no,location=no,directories=no,status=no,menubar=no,scrollbars=yes,resizable=yes,width=650,height=500,top=1,left=1"); } //--> </script> <body> <div id="page"> <div id="entete"> <div id="entete_contenu"> <!--===== Entête zone de gauche =====> <div class="logo_gauche"> </div> <h1 style="padding-top: 0.5em;"> Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation

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>
</tr>
</table>
</td>
</tr>
</table>
</div>
</body>
</html>
```

Transformation textuelle personnalisée sur la colonne HTML

Expression Langue Pas d'erreur de syntaxe

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

```
<br/> | <br/>
```

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/139804081>

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](https://api.openalex.org/works?filter=title.search:)

Générer une URL à partir de vos données

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

New column name

On error set to blank store error copy value from original column

Expression Language No syntax error.

Preview History Starred Help

row	value	"https://api.ror.org/organizat ...
1.	01875pg84	https://api.ror.org/organizations/01875pg84
2.	022bnxw24	https://api.ror.org/organizations/022bnxw24
3.	004gzqz66	https://api.ror.org/organizations/004gzqz66
4.	02mh9a093	https://api.ror.org/organizations/02mh9a093
5.	0293jn610	https://api.ror.org/organizations/0293jn610
6.	049xb5v45	https://api.ror.org/organizations/049xb5v45

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 2^e), puis sur le sous-élément *display_name*

Transformation textuelle personnalisée sur la colonne JSON

Expression Langue Pas d'erreur de syntaxe.

Aperçu Historique Étoilée Aide

row	value	value.parseJson().topics[1].di ...
1.	<pre>{ "id": "https://openalex.org/14210094956", "ror": "https://ror.org/00kzsxx38", "display_name": "Groupe d\u00c9tude des M\u00e9thodes de l'Analyse Sociologique de la Sorbonne", "country_code": "FR", "type": "facility", "type_id": "https://openalex.org/institution-types/facility", "lineage": ["https://openalex.org/11294671590", "https://openalex.org/139804081", "https://openalex.org</pre>	Sociology of Public Action and Professional Practices

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

OK Annuler

4

PARSER DES DONNÉES MULTIVALUÉES

Parser les données

Sélectionner plusieurs balises

```
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]
```

```
<?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>
```

5

COMPARER DES COLONNES A L'AIDE DE
FACETTES PERSONNALISÉES

Créer des facettes textuelles personnalisées

Comparer deux colonnes

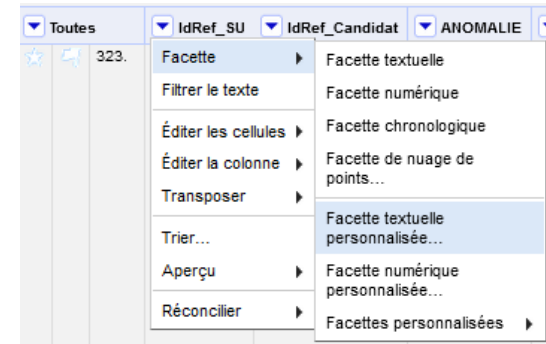
Les facettes textuelles personnalisées vous permettent notamment de comparer des chaînes de caractères :

- 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`



Facette personnalisée sur la colonne IdRef_SU

Expression Langue General Refine Expression Language (GREL)

`value == cells["IdRef_Candidat"].value` Pas d'e

Aperçu Historique Étoilée Aide


row	value	value == cells["IdRef_Candidat ...
2.	243020287	true
3.	243020287	true
4.	243020287	true
9.	130949868	true
25.	23760468X	true
28.	132206177	true

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



The screenshot shows a search interface with a facet for 'IdRef_SU' and a table of 8 matching lines. The facet is titled 'Facette / Filtre' and has a 'Défaire / Refaire 2 / 2' button. It contains a 'Rafrâichir' button, 'Tout réinitialiser', and 'Tout supprimer' buttons. The facet is currently set to '2 choix' and is sorted by 'nom' and 'compte'. The facet shows 8 'false' results and 2483 'true' results. The table shows 8 matching lines with columns for 'Toutes', 'IdRef_SU', and 'IdRef_Candidat'. The first two rows are visible, showing values 323, 082866309, 074207059 and 324, 082866310, 074207059.

Toutes	IdRef_SU	IdRef_Candidat
323.	082866309	074207059
324.	082866310	074207059

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-donnees-avec-openrefine-partie-2/99618030>

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

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

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

Site : <https://labrador.sorbonne-universite.fr/>

Mail : labrador@sorbonne-universite.fr



MERCI

labrador@sorbonne-universite.fr



BIBLIOTHÈQUE
UNIVERSITAIRE



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