

Fiche pratique nettoyage de données

Utilisation de open refine

Anne-Laure Donzel

ABSTRACT Présentation de l'outil, rapide tuto et ressources pour aller plus loin

1. Description

Open Refine est un outil fait pour nettoyer, transformer et enrichir des données.

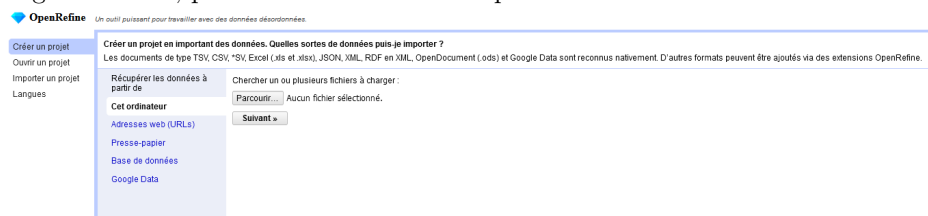
2. Découverte

Nous allons travailler à partir du fichier [depute-openrefine.csv](#)

Découverte de l'interface d'open refine et import de votre fichier

Charger un fichier

Charger le fichier, plusieurs format sont acceptés.



Si votre interface est en anglais et que la voulez en français, il faut changer la langue à ce niveau (onglet *langues*).

L'import se fait de façon très simple en parcourant vos répertoires et ensuite *suivant*.

Votre fichier va apparaître avec diverses options en bas de l'écran. Suivant les cas vous serez amené à : choisir le bon séparateur, l'encodage, la présence d'une en-tête... puis *créer le projet* en haut à gauche.

OpenRefine un outil puissant pour travailler avec des données obsolètes.

Créer un projet | Chercher un projet | Langues

« Recensement » Configurer les options d'analyse statique

Nom	date de naissance	fonction	occupation	pays de nationalité	Né le	Décédé le
1. Charles Savary	député de la Manche	personnalité politique	France	21 septembre 1841	9 septembre 1888	
2. Étienne Lamy	docteur titulaire chaire de l'université Louis	avocat et avocat	France	2 juin 1841	9 janvier 1919	
3. Françoise Gribault	député français	personnalité politique	France	26 janvier 1844	28 mai 1888	
4. Paul-Gabriel Rousselle	député de Seine-et-Marne	personnalité politique	France	27 mars 1843	1er septembre 1924	
5. Ernest Duvigneau de Neumaine	chevalier de la L.L.	personnalité politique	France	7 mars 1843	18 août 1877	
6. Georges Demerouti	grand-croix de la Légion d'honneur	politicien	France	20 septembre 1841	24 novembre 1929	
7. René Robin	député de la Seine	personnalité politique	France	23 jan 1841	13 septembre 1885	
8. Théodore Lemaître	député de la Charente	personnalité politique	France	2 janvier 1841	6 novembre 1906	
9. Gustave Trépo	député français	personnalité politique	France	1er janvier 1841	28 août 1871	
10. Céleste Bourcier	député des Deux-Sèvres-du-Pré	personnalité politique	France	26 décembre 1840	9 jan 1906	
11. Marie-Françoise-Catherine de Chabris-Thouart	député du Puy-de-Dôme	journaliste	France	18 mai 1840	26 décembre 1923	
12. Léon Clém	député des Alpes-Maritimes	personnalité politique	France	16 décembre 1839	19 janvier 1888	
13. Amédée Eugène-Louis de La-Sablon	député de la Gironde	personnalité politique	France	5 juillet 1839	2 octobre 1904	
14. René Joseph Dru	chevalier de la Légion d'honneur	personnalité politique	France	23 jan 1839	28 août 1921	
15. Arthur de Chabris-Thouart	chevalier de la L.L.	personnalité politique	France	8 jan 1839	20 février 1918	
16. Albert Thomas du Brouil de Saint-Germain	député du Cher	personnalité politique	France	3 décembre 1838	8 avril 1919	
17. Georges Buis	député français	personnalité politique	France	1er août 1838	3 août 1883	
18. Jules Rollin	député français	personnalité politique	France	20 mai 1838	21 décembre 1923	
19. Albert Deshayes	député de l'Ille	personnalité politique	France	20 mai 1838	23 janvier 1887	
20. Léon Deshayes	chevalier de la Légion d'honneur	personnalité politique	France	2 avril 1838	31 décembre 1882	
21. Louis Villiers	député français	personnalité politique	France	10 mars 1838	4 août 1911	
22. André Follin	député de la Haute-Saône	personnalité politique	France	10 mars 1838	22 mars 1905	
23. Louis de Saint	député du Puy-de-Dôme	personnalité politique	France	17 novembre 1837	27 décembre 1901	
24. Françoise Riva	député français	personnalité politique	France	17 novembre 1837	12 avril 1888	
25. Étienne Buis	député français	personnalité politique	France	10 octobre 1837	28 janvier 1886	
26. Germain Lamy	député français	personnalité politique	France	22 novembre 1837	9 décembre 1905	
27. Paul Groll	grand-croix de la Légion d'honneur	président de la République française	France	11 août 1837	25 jan 1904	
28. Céleste Clém	chevalier de la Légion d'honneur	personnalité politique	France	8 août 1837	5 mai 1911	
29. René Alexandre Besault de Bourdon	officier de la Légion d'honneur	député de la Somme	explorateur ou exploratrice	France	20 mai 1837	17 avril 1840
30. Victor de Chabris-Thouart	grand-croix de la Légion d'honneur	député de Seine-et-Marne	personnalité politique	France	23 janvier 1837	6 décembre 1915
31. Pierre Trépo	député de Lot-et-Garonne	personnalité politique	France	21 janvier 1837	20 novembre 1908	
32. Camille Rollin de Villiers	député français	personnalité politique	France	18 janvier 1837	1er août 1907	

Considérer les données comme : **Format des caractères :**

Fichiers CSV / TSV / séparateur : ☒ une virgule (CSV) ☐ une tabulation (TSV) ☐ personnel ☐ personnel

Fichiers texte à largeur de champ fixe : ☒ Utiliser le caractère " " pour fermer les cellules contenant des séparateurs de colonnes ☐ Supprimer les espaces de début et de fin ☐ Polir les caractères spéciaux avec

☐ Ignorer la ou les 0 : ☐ première(s) ligne(s) du début du fichier ☐ première(s) ligne(s) de données

☐ Ignorer la ou les 0 : ☐ première(s) ligne(s) de données ☐ première(s) ligne(s) de données

☐ Décharger l'aperçu automatique

☐ Analyser le texte des cellules comme nombres ☒ Conserver les lignes vides ☒ Remplacer les cellules vides comme des valeurs nulles ☐ Indiquer la source du fichier ☐ Stocker le fichier d'archive

Bon à savoir : Open Refine ne modifie pas directement votre fichier (comme Excel par exemple), il crée un projet et toutes vos modifications sont enregistrées. Vos projets sont accessibles à droite *Ouvrir un projet* (ou en cliquant sur le logo) et une fois ouvert, l'onglet *Défaire/refaire* vous montre toutes vos actions de traitement, et vous pouvez remonter très simplement à un état antérieur du fichier.

La découverte des facettes

Le contenu de chaque colonne peut être visualiser sous la forme de facette. C'est une liste des termes présents dans une colonne. Cela permet de facilement voir des erreurs ou un manque d'harmonisation.

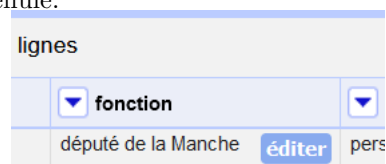
Les facettes (et la plupart des possibilités de traitement sur une colonne) sont accessibles en haut de chaque colonne avec le triangle.



Réaliser des modifications sur le contenu des cellules

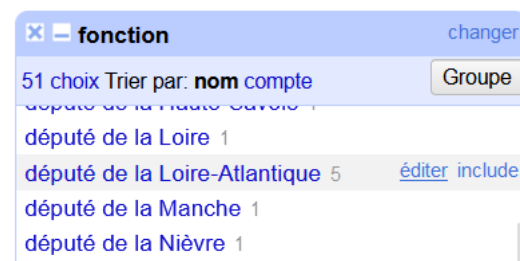
De façon unitaire

Vous pouvez changer le contenu d'une cellule en cliquant sur *éditer* sur chaque cellule.



Par facettes

Vous pouvez également modifier l'ensemble des occurrences d'une facette.

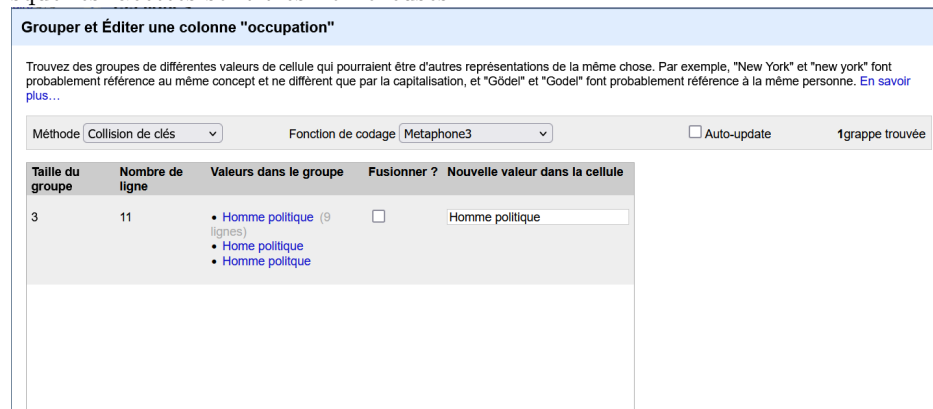


Cluster

La fonction cluster permet de faire du rapprochement sémantique, elle va permettre d'identifier des éléments proches mais pas identique.



C'est une fonctionnalité très intéressante sur des gros volumes de données, lorsque les facettes sont très nombreuses.



Expressions régulières

Enfin, et là cela se complique, vous allez pouvoir faire des transformations très poussées via des expressions régulières.

Pour accéder aux expressions, il faut *transformer* une cellule.

Par exemple : - Modifier une chaîne de caractères XXX > XAXAX - Faire des *chercher/remplacer* complexe - Séparer des éléments ...

La difficulté est que la sémantique des expressions est très difficile à maîtriser.

Un exemple très simple :

Dans la colonne *Né le* et *Décédé le* :

```
- value.replace("1er", "1")
```

Autres modifications

Modification sur les colonnes

Pour séparer, joindre renommer...

▼ Né le	▼ Décédé le
Facette	septembre 1889
Filter le texte	anvier 1919
Éditer les cellules	mai 1898
Éditer la colonne	septembre 1924
Transposer	
Trier...	
Aperçu	
Réconcilier	

Diviser en plusieurs colonnes...

Joindre des colonnes...

Ajouter une colonne en fonction de cette colonne...

Ajouter une colonne en moissonnant des URL...

Ajouter des colonnes à partir de valeurs réconciliées...

Renommer cette colonne...

Supprimer cette colonne

Déplacer la colonne en premier

Déplacer la colonne en dernier

Déplacer la colonne à gauche

Déplacer la colonne à droite

Modifications sur les cellules

A utiliser pour les transformations de type de cellule (date, casse...) et pour supprimer rapidement les espaces invisibles en début et fin de cellule.

▼ Né le	▼ Décédé le
Facette	septembre 1889
Filter le texte	anvier 1919
Éditer les cellules	mai 1898
Éditer la colonne	
Transposer	
Trier...	
Aperçu	
Réconcilier	

Transformer...

Transformations courantes

Supprimer les espaces de début et de fin

Rassembler les espaces consécutifs

Convertir les entités HTML

Remplacer les guillemets courbés par des guillemets droits

En initiales majuscules (en capitales)

En majuscules

En minuscules

En nombre

En date

En texte

En valeurs nulles

Transformer en chaîne vide

Recopier les valeurs dans les cellules vides consécutives

Vider les valeurs répétées dans des cellules consécutives

Diviser les cellules multivaluées...

Joindre les cellules multivaluées...

Grouper et éditer...

Remplacer...

Réconcilier les données

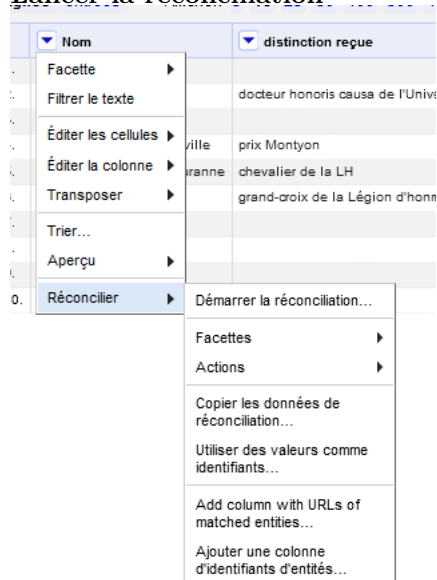
La réconciliation de données consiste à faire correspondre des données venant de différentes sources.

Dans notre exemple nous avons des hommes politiques, la plupart existe dans d'autres référentiels (par ex. sur Wikipedia), nous allons donc faire correspondre

notre liste avec une de ces sources et récupérer des éléments qui existent dans le référentiel que nous allons utiliser (par ex. le lieu de naissance).

Nous allons réconcilier notre fichier avec [Wikidata](#).

Lancer la réconciliation



Lors de votre première réconciliation, vous allez devoir *ajouter un service standard*, ici le service français de wikidata : <https://wikidata.reconci.link/fr/api>

Ajouter un service standard

Indiquer l'URL du service

<https://wikidata.reconci.link/fr/api>

Ajouter un service
Annuler

L'outil va repérer le type d'élément que chercher à faire correspondre (ici des êtres humains), vous pouvez *démarrer la réconciliation*.

Cela va prendre un peu de temps.

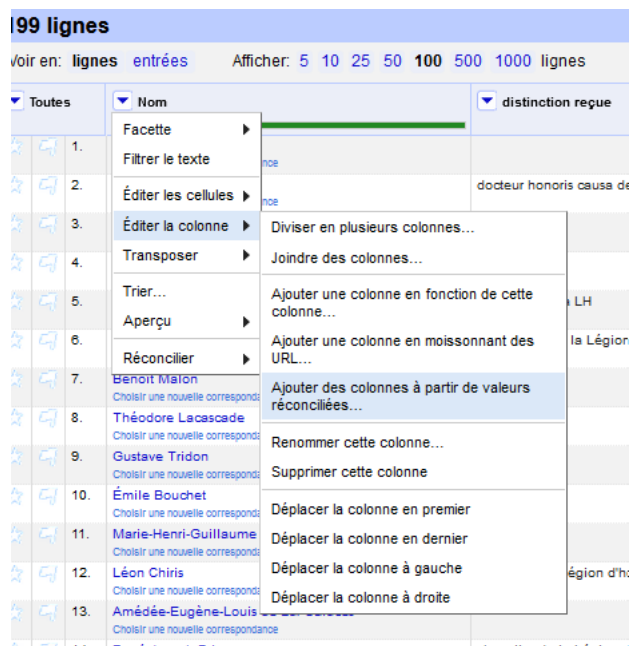
Validation de la réconciliation

Vous allez voir que vos noms sont devenus bleu (comme un lien hypertexte) pour la plupart. Ils sont donc reliés à la fiche wikidata correspondante.

Parfois l'outil hésite (homonymie), il va vous proposer de valider manuellement la bonne affectation. Au survol, vous voyez sa proposition, si c'est la bonne, vous pouvez *appairer cette cellule*.

Enrichir vos données

Désormais, vous allez pouvoir récupérer les éléments présents sur la fiche wikidata de ces personnes. Pour cela, nous allons ajouter des colonnes sur la base des cellules réconciliées.



L'outil va vous suggérer des propriétés à ajouter, vous allez pouvoir en ajouter autant que souhaité.



Et ainsi de suite...

Une fois votre fichier prêt vous pouvez le télécharger en l'exportant dans différents formats.

3. Alternative

Il n'existe pas réellement d'alternative à Open refine, vous pouvez utiliser Excel ou un tableur et les fonctions, mais cela nécessite une bonne maîtrise du tableur (et le risque d'erreur est souvent grand et le retour en arrière moins sécurisant que sur Open Refine).

Pour la réconciliation, en revanche, sauf à avoir de très bonnes connaissances en programmation, il n'y a pas de solution plus simple.

4. Ressources en ligne

- OpenRefine, "Excel aux hormones" pour nettoyage de données[<https://www.patrimoine-et-numerique.fr/tutoriels/52-36-openrefine-excel-aux-hormones-pour-nettoyage-de-donnees>]
- Tutoriel OpenRefine 3.4 : nettoyer, préparer et transformer des données - 06/11/2020[<http://bit.ly/tutoOpenRefine>]
- Cas concret d'utilisation d'OpenRefine pour les archives[<https://medium.com/@seeksanusername/cas-concret-dutilisation-d-openrefine-pour-les-archives-442726996b74>]
- Cas pratique à partir d'un inventaire en pdf <https://patrimoine-et-numerique.fr/tutoriels/69-38-d-un-inventaire-pdf-a-un-fichier-xml-cas-pratique-openrefine>