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.

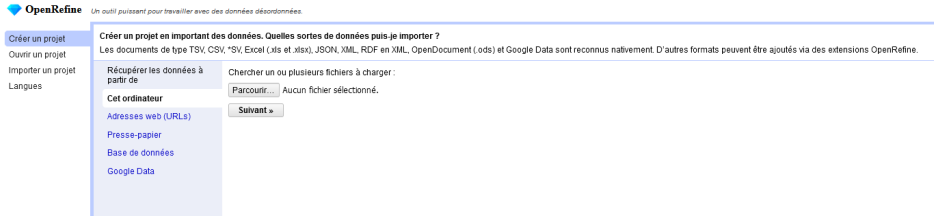
2. Découverte

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

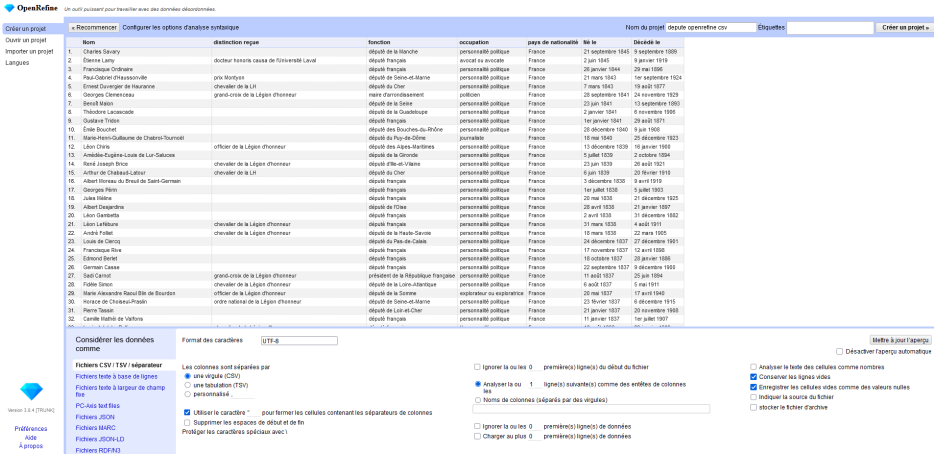
Charger le fichier, plusieurs format sont acceptés.



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 appaaître 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.

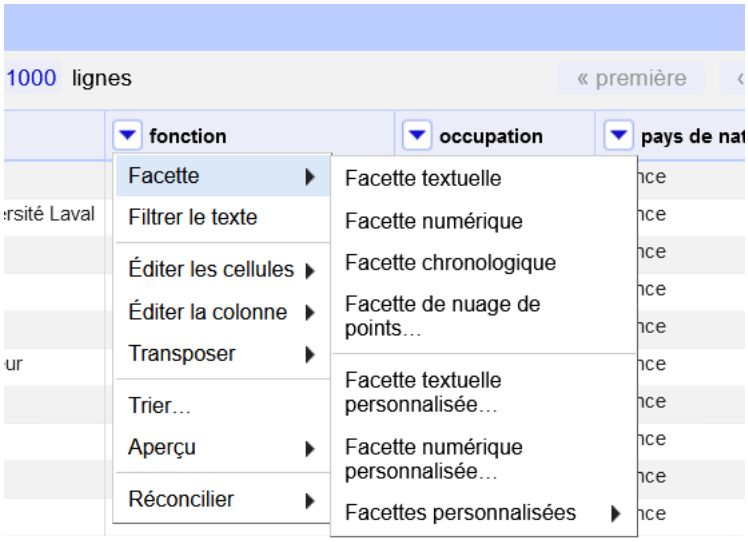


Bon à savoir : Open Refine ne modifie pas directement votre fichier (comme Excel par exemple), il créé un projet et toutes vos modifications sont enregistrées. Vos projets sont accessibles à droite *Ouvrir un projet* (ou en cliquant sur le logo) et une fois ouvert, l'onglet *Défaire/refaire* vous montre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

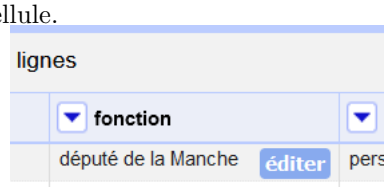
Les facettes (et la plupart des possibilités de traitement sur une colonne) sont accessibles en haut de chaque colonne avec le triangle.



Réaliser des modifications sur le contenu des cellules

De façon unitaire

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



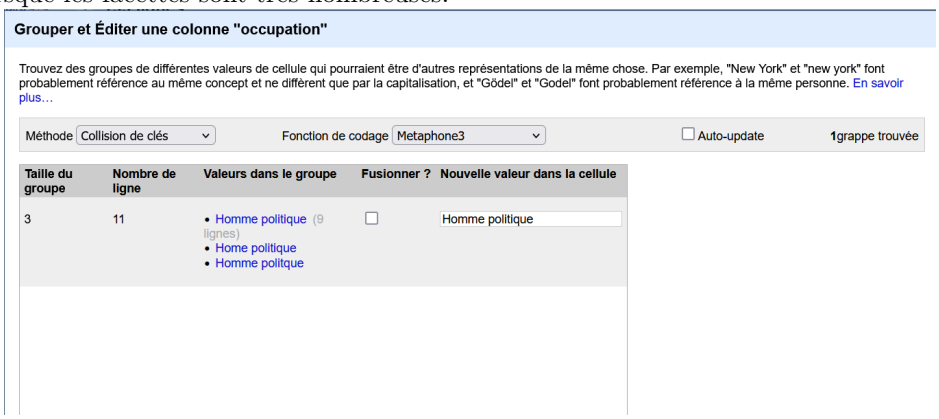
Par facettes

Vous pouvez également modifier l'ensemble des occurences 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 :

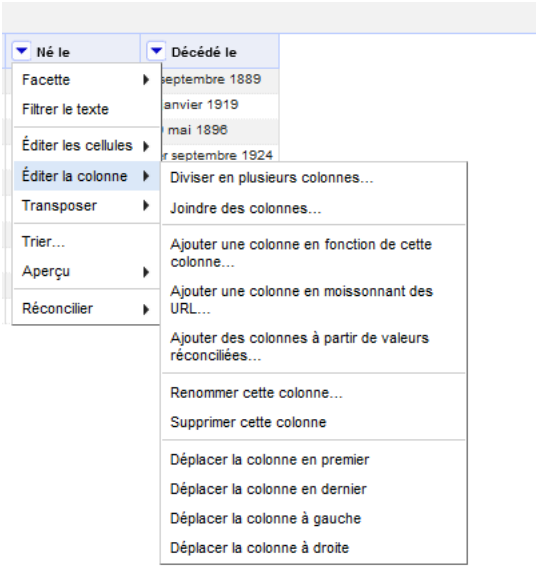
Dans la colonne *Né le* et *Décédé le* :

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

Autres modifications

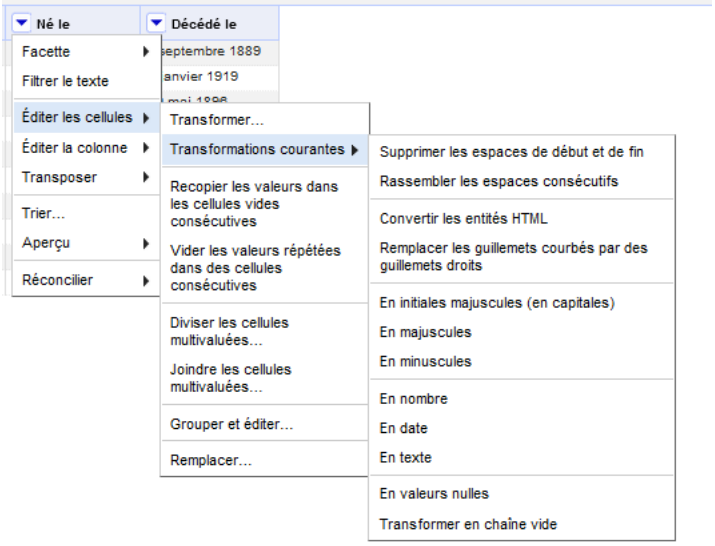
Modification sur les colonnes

Pour séparer, joindre renommer…



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.



Réconcilier les données

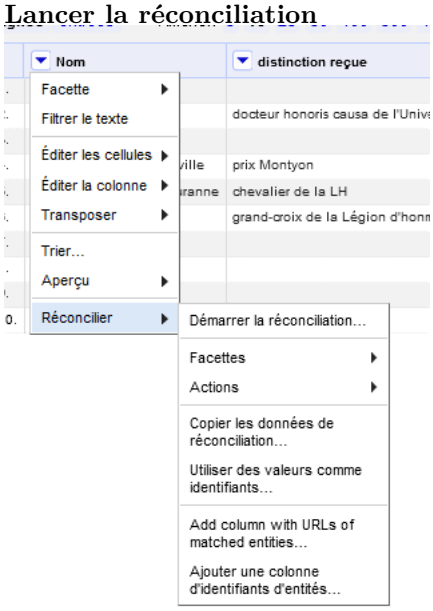
La réconciliation de données consiste à faire correspondre des données venant de différentes sources.

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

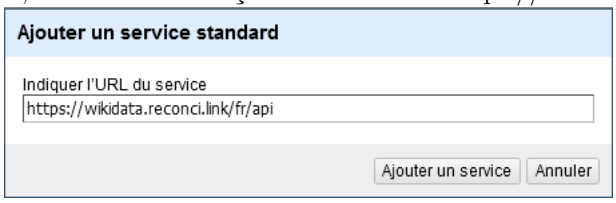
notre liste avec une de ces sources et récupérer des éléments qui existent dans le référentiel que nous allons utiliser (par ex. le lieu de naissance).

Nous allons réconcilier notre fichier avec Wikidata.

Lancer la réconciliation



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



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

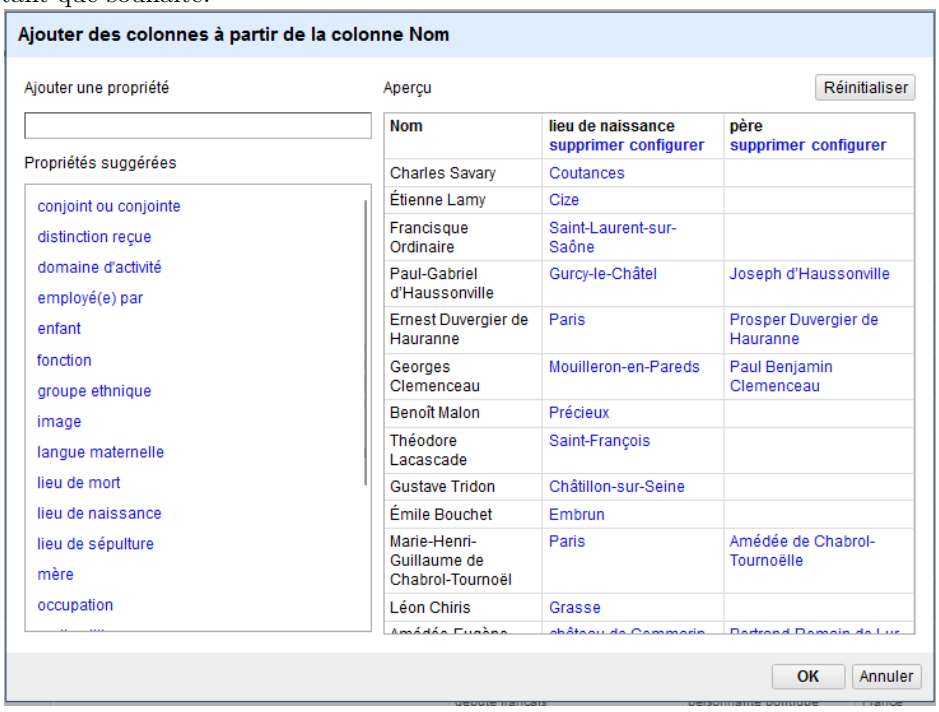
Vous allez voir que vos noms sont devenus bleu (comme un lien hypertexte) pour la plupart. Ils sont donc reliés à la fiche wikidata correspondante.

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 *apparier 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