Open Refine

Logiciel pour la qualification et le nettoyage des données

DHIP IHA

Qu'est ce qu'Open Refine ?

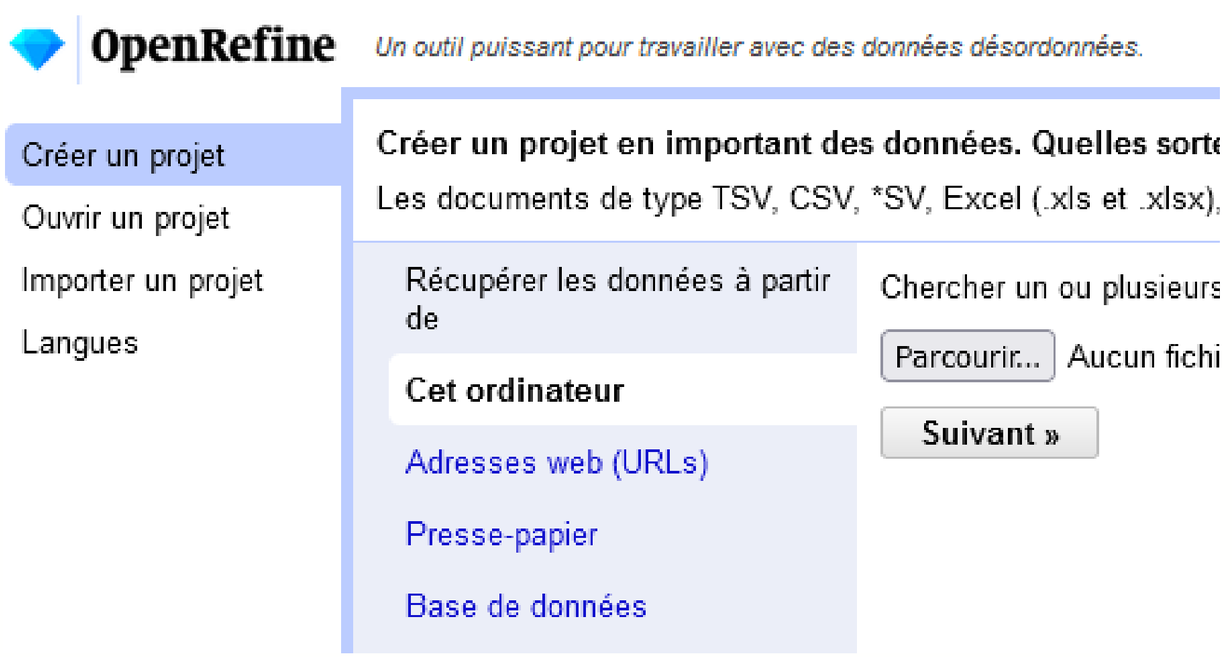
- Développé à l'origine sous le titre de Freebase Gridworks par Metaweb Technologies.
- Metaweb est racheté par Google en 2010, le logiciel est renommé Google Refine.
- En 2012, Google Refine devient Open Refine.
- Première version en 2010, dernière version stable 2022, version 3.6.2
- Le site pour l' installation et la documentation : https://openrefine.org/

Installation d'Open Refine

- Téléchargez la dernière version d'Open Refine sur https://openrefine.org
- Choisissez "windows-kit with embeded java" pour ne pas installer java sur l'ordinateur
- Décompressez le dossier et ouvrez-le
- Cliquez sur le fichier openrefine.exe, une fenêtre de commande s'ouvre, laissez-la ouverte pour le fonctionnement d'Open Refine

L'interface

- Créer un projet : pour tout nouveau projet
- Ouvrir un projet : reprendre un projet en cours
- Importer un projet : prendre un projet d'une autre plateforme Open Refine
- Langues : modifier la langue de l'application



Exercice 1 : créer un projet

https://shorturl.at/kqstN

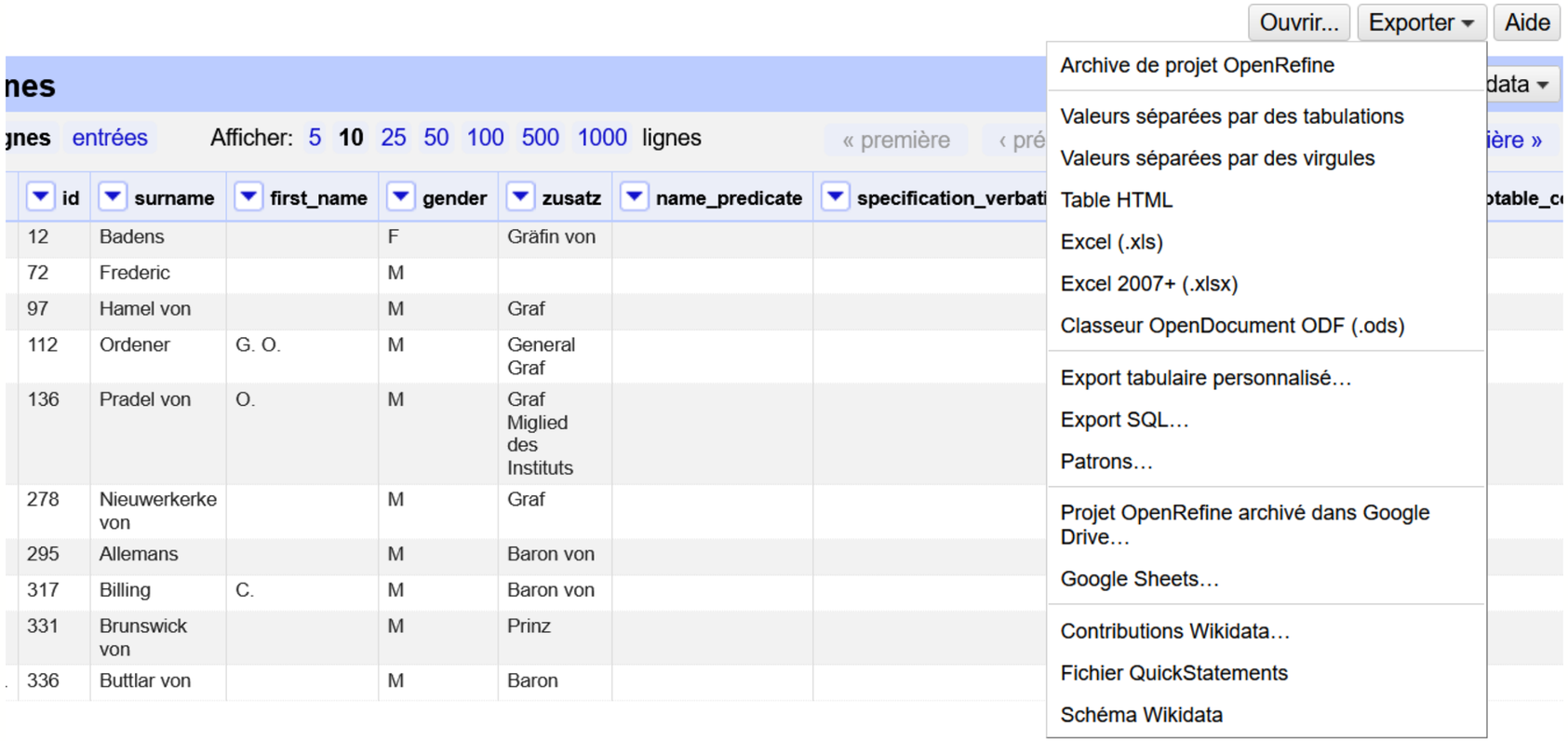
- Sélectionner "Créer un projet"
- Choisissez "À partir d'une adresse web (URLs)"
- Entrez l'adresse URL
- Cliquez sur suivant
- Ajuster les paramètres au fichier si nécessaire
- Donner un nom à votre projet et faites "suivant"

L'interface (2)

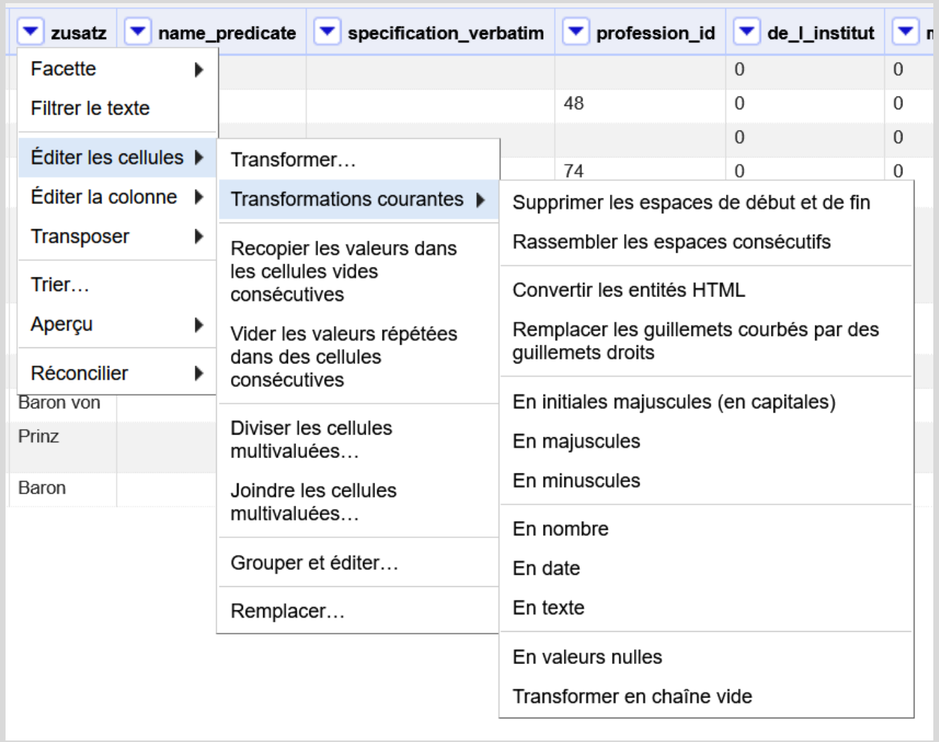
- Facette : Textuelle, Numérique, Chronologique
- Éditer les cellules : Transformer, Transformations courantes, grouper et éditer
- Éditer les colonnes : Diviser, joindre, renommer, supprimer
- Réconcilier : Démarrer la réconciliation, facettes, actions

L'interface (3)

- Ouvrir
- Exporter
- Aide

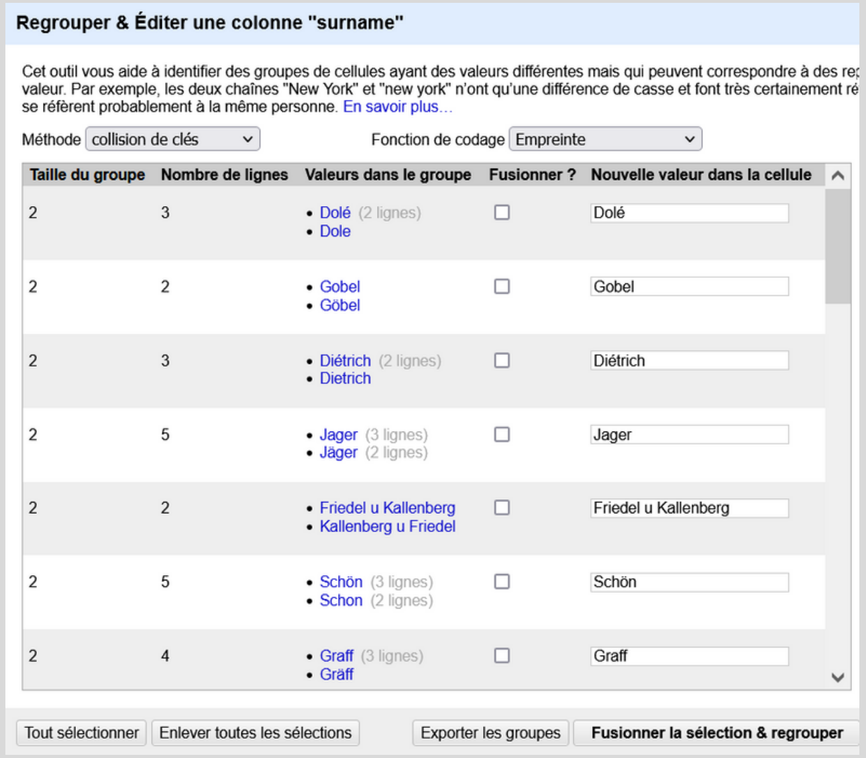


Transformer et nettoyer les données



- Transformations courantes :
  - Corriger les espaces
  - Valeurs en majuscules et miniscules
  - Trasnformer en nombre ou texte
- Transformer…
  - Utilisation du langage GREL d'Open Refine
- Grouper et éditer
  - Rassembler les données similaires

Transformer et nettoyer les données (2)



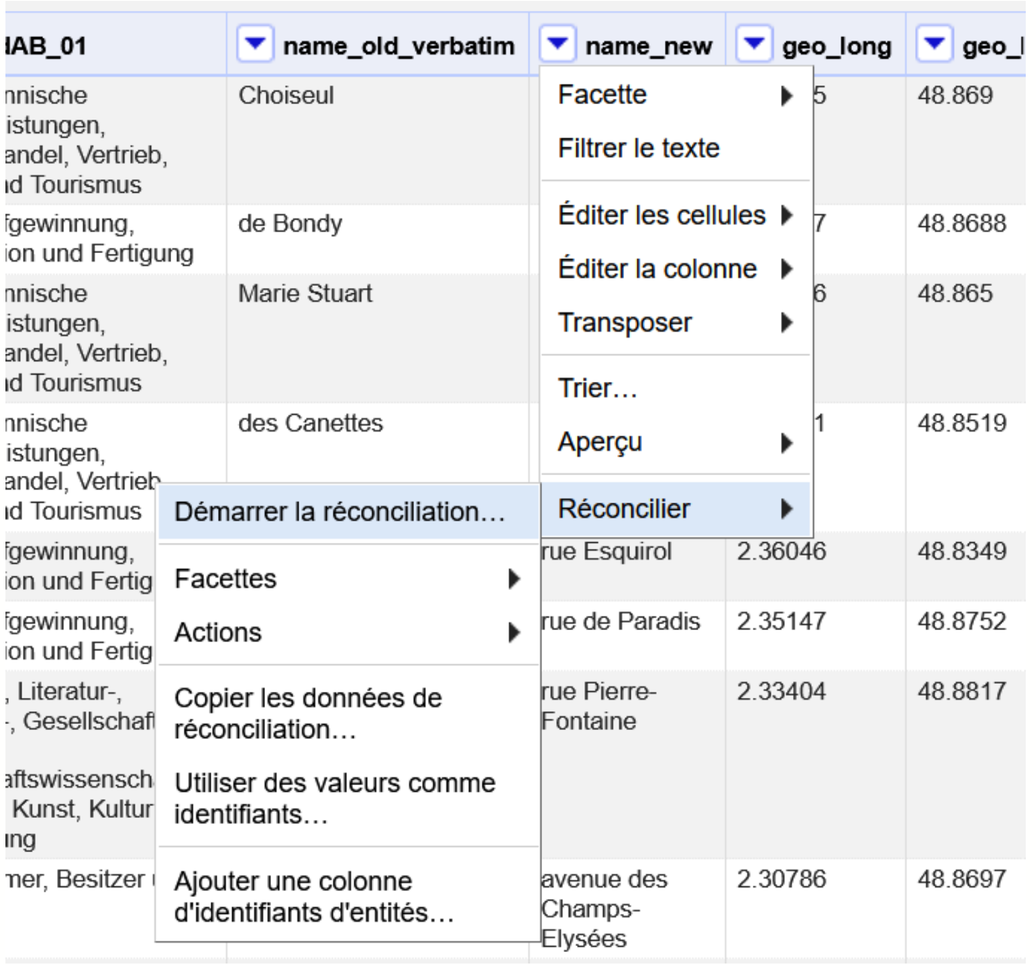
- Grouper et éditer
  - Valeurs dans le groupe : les différentes variantes dans le groupe
  - Nouvelle valeur : celle que nous allons choisir pour uniformiser le groupe
  - Fusionner la sélection et regrouper : uniformiser les données et vérifier s'il n'y pas d'autres variantes

Exercice 2 : nettoyer nos données

- À partir de notre de projet :
  - Convertir en majuscules les données de la colonne **surname**
  - Grouper et éditer les valeurs dans la colonne **profession_verbatim**
  - Transformer en nombre la colonne **id**

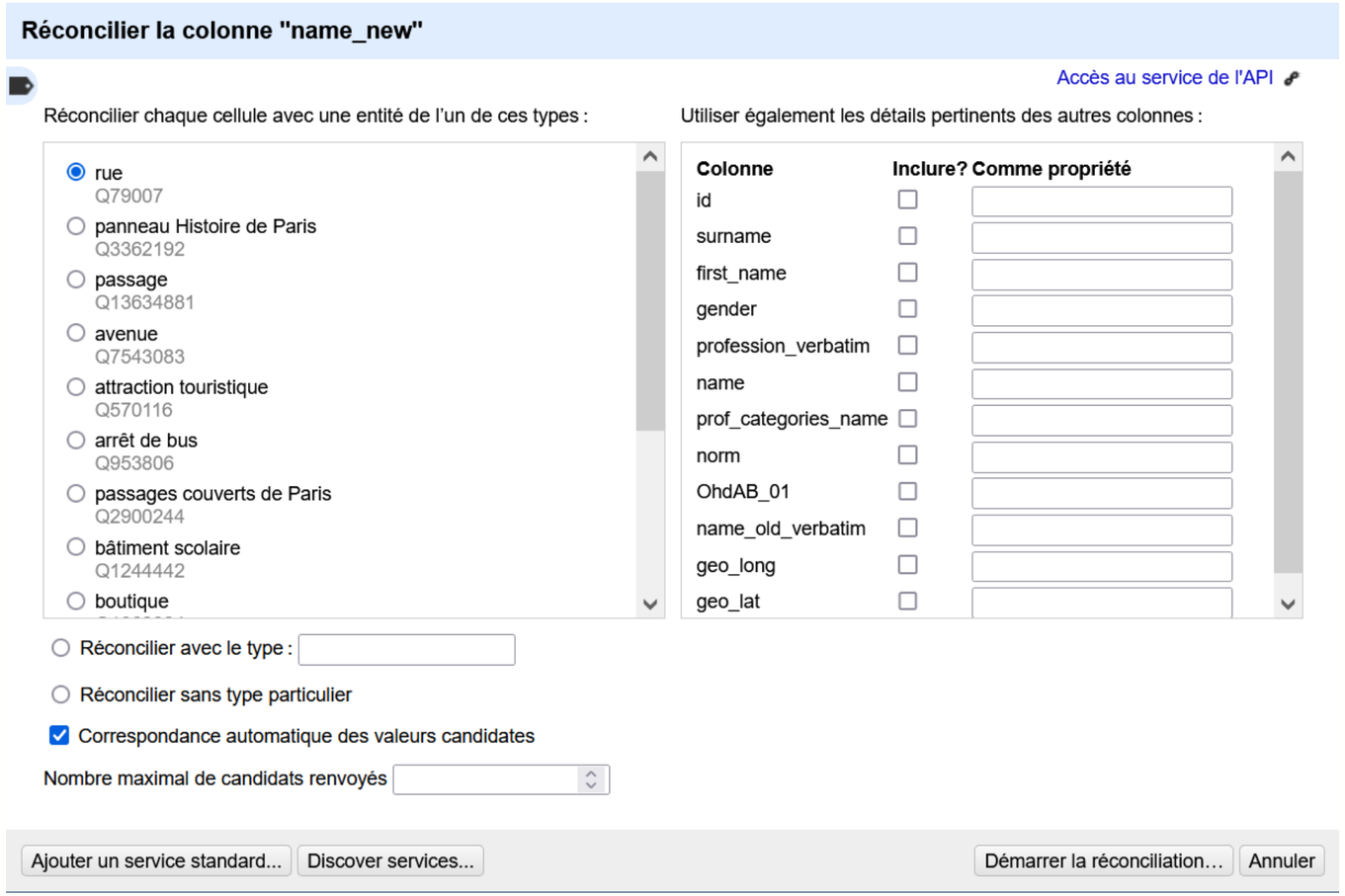
Enrichir ses données avec Open Refine

- Réconcilier : permet d'enrichir les données
  - Démarrer la réconciliation : permet de lancer l'enrichissement des données
  - Facettes : filtrer la réconciliation des données
  - Actions : Rejeter ou effacer la réconciliation

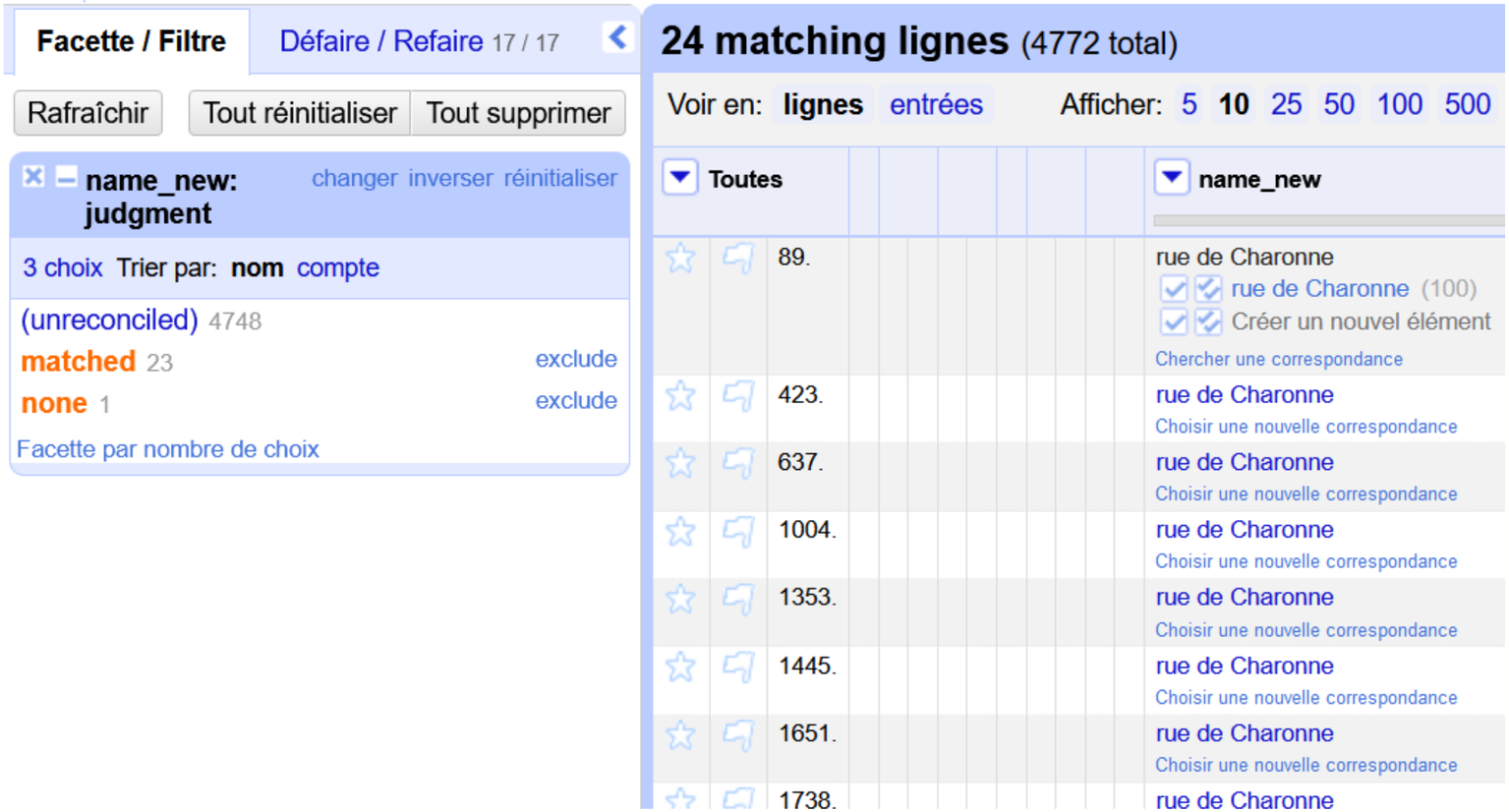


Enrichir ses données avec Open Refine (2)

- Réconcilier la colonne
  - Réconcilier chaque cellule avec une entité de l'un de ces types : choisir le type de données du côté du service de réconciliation
  - Démarrer la réconciliation : lancer l'enrichissement des données



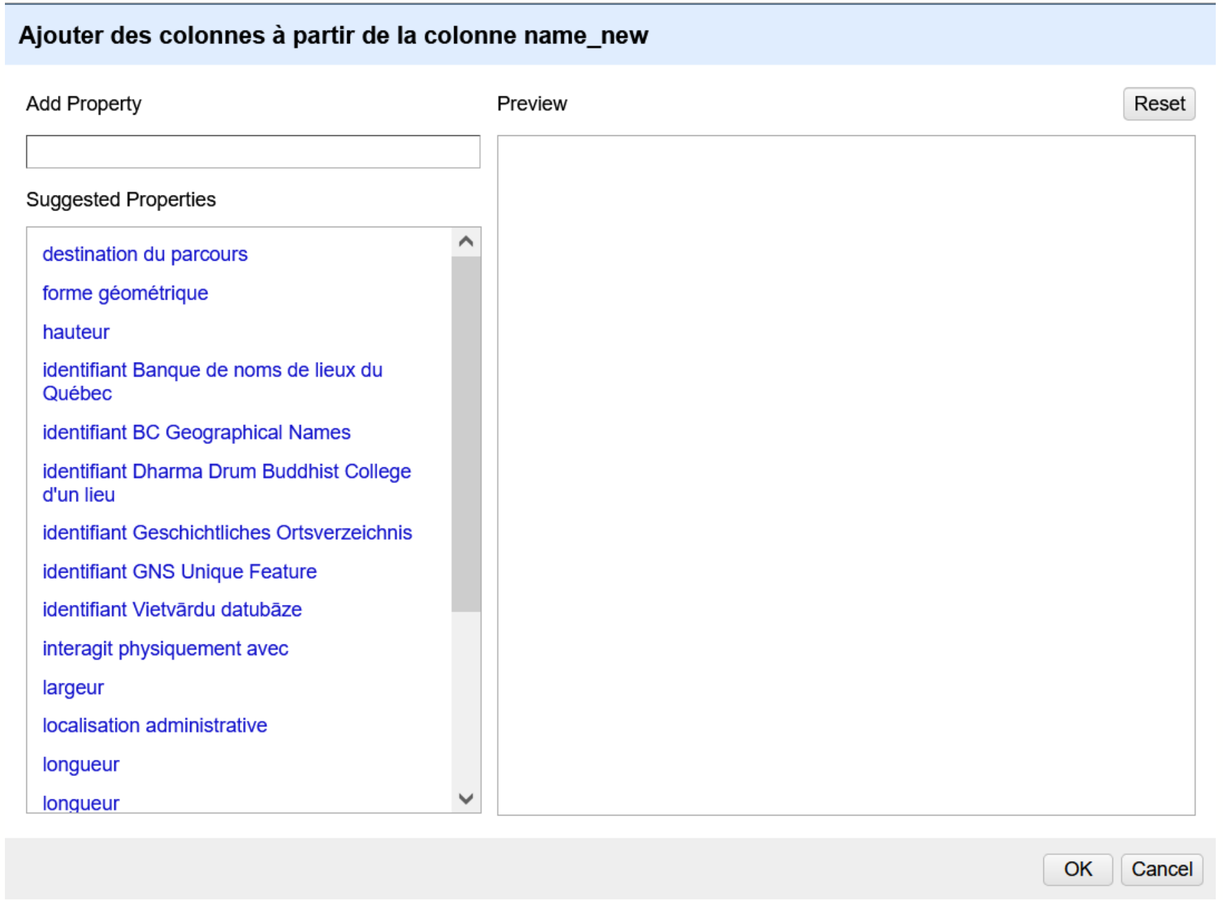
Enrichir ses données avec Open Refine (3)



- Matched
  - Les données sont réconciliées avec le service de données externe
- None
  - Les données ne sont pas réconciliées
- Unreconcilied :
  - Nos données n'ont pas de correspondance avec le service de données externe

Enrichir ses données avec Open Refine (4)

- Ajouter des colonnes à partir de valeurs réconciliés
  - choisir les propriétés que nous souhaitons ajouter à notre dataset
  - une fois le choix confirmé les colonnes s'ajoutent à notre projet

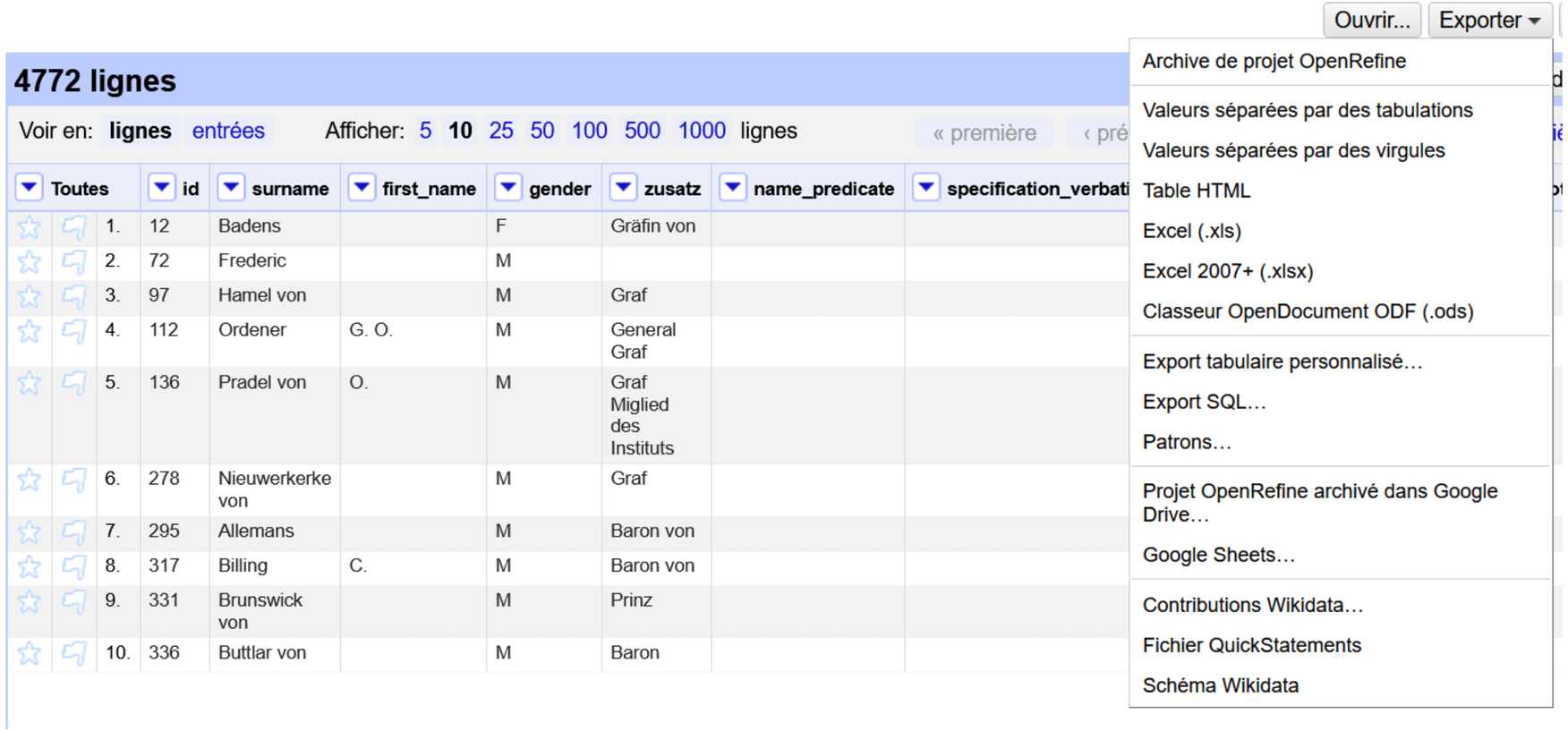


Exercice 3 : enrichir nos données

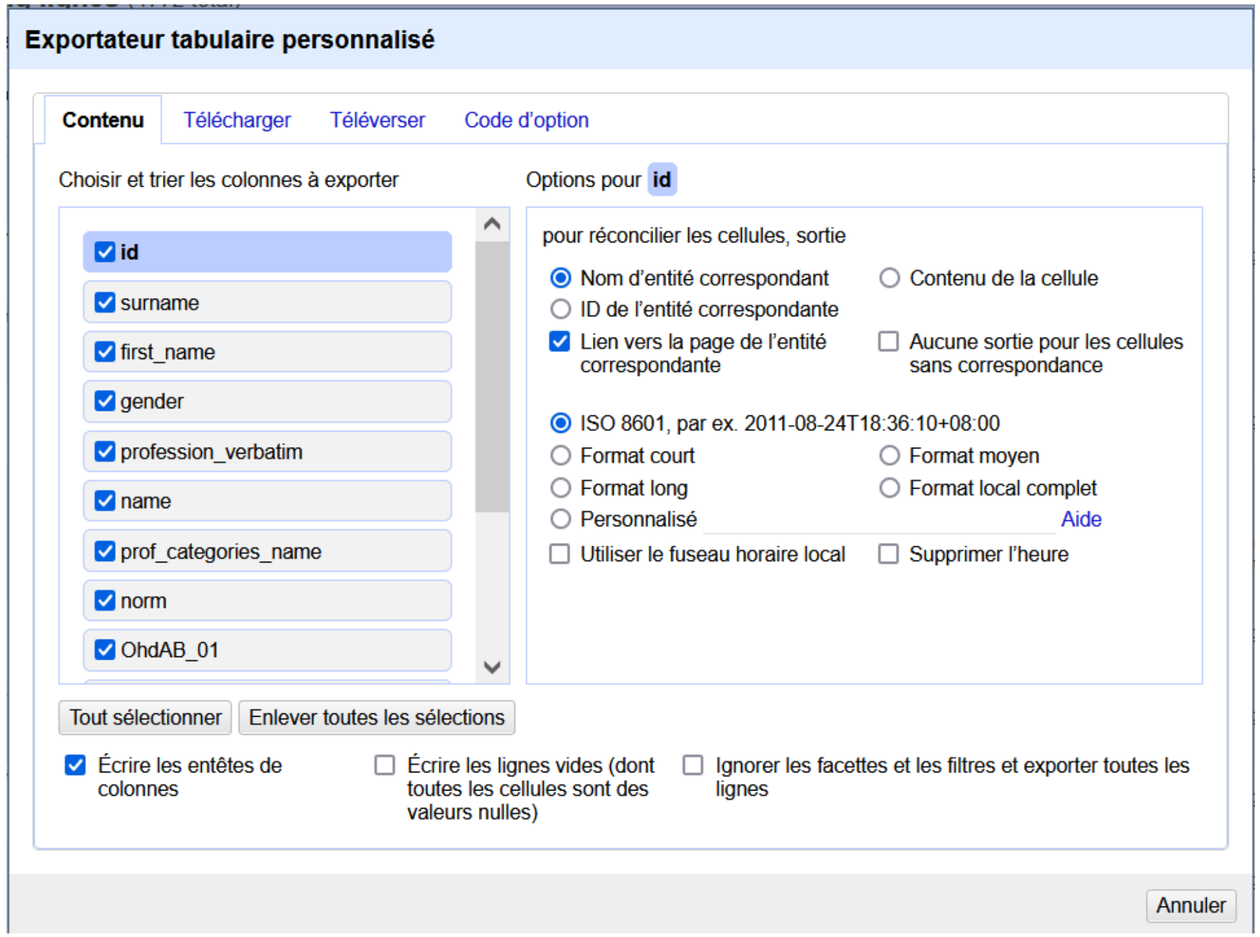
- À partir de notre de projet :
  - Enrichir la colonne **name_new** avec wikidata et le type "rue"
  - Associer les données à une rue du wikidata
  - Ajouter les coordonnées de chaque rue , en choisissant : ajouter des colonnes à partir de valeurs réconciliées et en sélectionnant la propriété "coordonnées géographiques"

Exporter son projet avec OpenRefine

- Exporter
  - CSV / TSV
  - Excel
  - LibreOffice
  - SQL
  - Patrons (JSON)
  - Google Sheets
  - Archive OpenRefine



Exporter son projet avec OpenRefine (2)



- Choisir export tabulaire personnalisée
  - choisir les colonnes qu'on souhaite exporter
  - choisir l'encodage du fichier (UTF8)
  - Télécharger le fichier

Exercice 3 : exporter notre projet

- À partir de notre de projet :
  - Choisir le format de votre choix
  - Exporter votre projet

Open Refine

Merci de votre attention !

Des questions ?