



Open Refine

Logiciel pour la qualification et le nettoyage des données





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

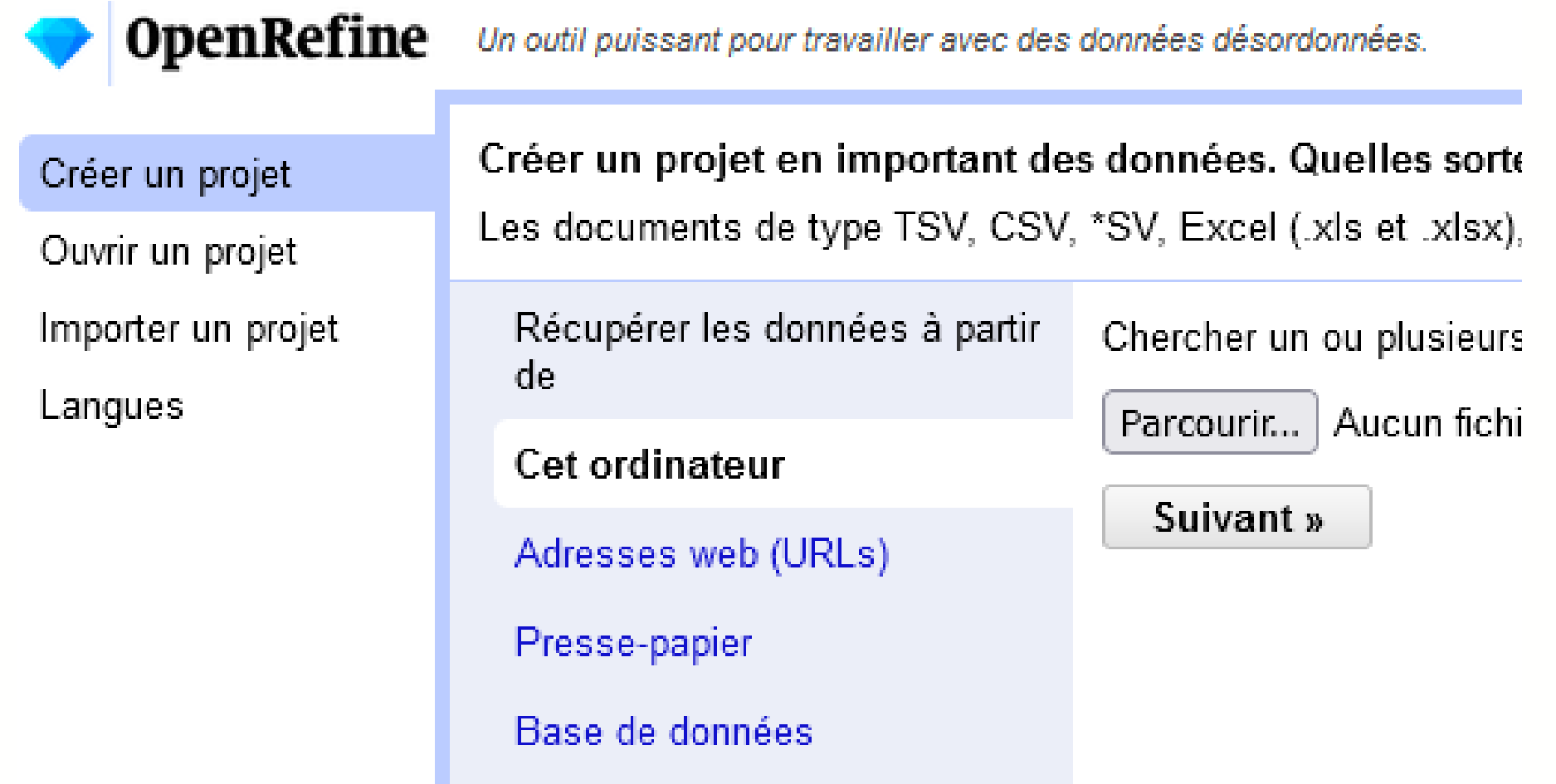


- Téléchargez la dernière version d'Open Refine sur <https://openrefine.org>
- Choisissez "windows-kit with embeded java" pour ne pas installer java sur l'ordinateur
- Décompressez le dossier et ouvrez-le
- Cliquez sur le fichier openrefine.exe, une fenêtre de commande s'ouvre, laissez-la ouverte pour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

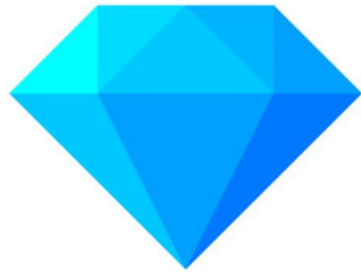




Exercice 1 : créer un projet

<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

OpenRefine persons [Permalien](#)

Facette / Filtre Défaire / Refaire 41 / 41

4772 lignes

Voir en: **lignes** entrées Afficher: 5 10 25 50 100 500 1000 lignes

			id	surname	first_name	gender	zusatz	name_p
☆	🗨	1.	12	Badens	Facette		Gräfin von	
☆	🗨	2.	72	Frederic	Filtrer le texte			
☆	🗨	3.	97	Hamel von	Éditer les cellules		Graf	
☆	🗨	4.	112	Ordener	Éditer la colonne		General Graf	
☆	🗨	5.	136	Pradel von	Transposer		Graf Miglied des Instituts	
☆	🗨	6.	278	Nieuwerkerke von	Trier...		Graf	
☆	🗨	7.	295	Allemans	Aperçu		Baron von	
☆	🗨	8.	317	Billing	Réconcilier	C.	Baron von	
☆	🗨	9.	331	Brunswick von		M	Prinz	
☆	🗨	10.	336	Buttlar von		M	Baron	



L'interface (3)

- Ouvrir
- Exporter
- Aide

nes

gnes entrées Afficher: 5 10 25 50 100 500 1000 lignes « première < pré

id	surname	first_name	gender	zusatz	name_predicate	specification_verbat
12	Badens		F	Gräfin von		
72	Frederic		M			
97	Hamel von		M	Graf		
112	Ordener	G. O.	M	General Graf		
136	Pradel von	O.	M	Graf Mitglied des Instituts		
278	Nieuwerkerke von		M	Graf		
295	Allemans		M	Baron von		
317	Billing	C.	M	Baron von		
331	Brunswick von		M	Prinz		
336	Buttlar von		M	Baron		

Ouvrir... Exporter ▼ Aide

- Archive de projet OpenRefine
- Valeurs séparées par des tabulations
- Valeurs séparées par des virgules
- Table HTML
- Excel (.xls)
- Excel 2007+ (.xlsx)
- Classeur OpenDocument ODF (.ods)
- Export tabulaire personnalisé...
- Export SQL...
- Patrons...
- Projet OpenRefine archivé dans Google Drive...
- Google Sheets...
- Contributions Wikidata...
- Fichier QuickStatements
- Schéma Wikidata



Transformer et nettoyer les données

▼ zusatz	▼ name_predicate	▼ specification_verbatim	▼ profession_id	▼ de_l_institut	▼ n
Facette				0	0
Filtrer le texte			48	0	0
Éditer les cellules	Transformer...		74	0	0
Éditer la colonne	Transformations courantes	Supprimer les espaces de début et de fin			
Transposer	Recopier les valeurs dans les cellules vides consécutives	Rassembler les espaces consécutifs			
Trier...	Vider les valeurs répétées dans des cellules consécutives	Convertir les entités HTML			
Aperçu		Remplacer les guillemets courbés par des guillemets droits			
Réconcilier		En initiales majuscules (en capitales)			
Baron von		En majuscules			
Prinz		En minuscules			
Baron		En nombre			
		En date			
		En texte			
		En valeurs nulles			
		Transformer en chaîne vide			

- Transformations courantes :
 - Corriger les espaces
 - Valeurs en majuscules et minuscules
 - Transformer 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)

Regrouper & Éditer une colonne "surname"

Cet outil vous aide à identifier des groupes de cellules ayant des valeurs différentes mais qui peuvent correspondre à des personnes ayant la même valeur. Par exemple, les deux chaînes "New York" et "new york" n'ont qu'une différence de casse et font très certainement référence à la même personne. [En savoir plus...](#)

Méthode: collision de clés Fonction de codage: Empreinte

Taille du groupe	Nombre de lignes	Valeurs dans le groupe	Fusionner ?	Nouvelle valeur dans la cellule
2	3	<ul style="list-style-type: none">Dolé (2 lignes)Dole	<input type="checkbox"/>	<input type="text" value="Dolé"/>
2	2	<ul style="list-style-type: none">GobelGöbel	<input type="checkbox"/>	<input type="text" value="Gobel"/>
2	3	<ul style="list-style-type: none">Diétrich (2 lignes)Dietrich	<input type="checkbox"/>	<input type="text" value="Diétrich"/>
2	5	<ul style="list-style-type: none">Jäger (3 lignes)Jäger (2 lignes)	<input type="checkbox"/>	<input type="text" value="Jäger"/>
2	2	<ul style="list-style-type: none">Friedel u KallenbergKallenberg u Friedel	<input type="checkbox"/>	<input type="text" value="Friedel u Kallenberg"/>
2	5	<ul style="list-style-type: none">Schön (3 lignes)Schon (2 lignes)	<input type="checkbox"/>	<input type="text" value="Schön"/>
2	4	<ul style="list-style-type: none">Graff (3 lignes)Gräff	<input type="checkbox"/>	<input type="text" value="Graff"/>

Tout sélectionner Enlever toutes les sélections Exporter les groupes Fusionner la sélection & regrouper

- 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

IAB_01	▼ name_old_verbatim	▼ name_new	▼ geo_long	▼ geo_l
nnische istungen, andel, Vertrieb, id Tourismus	Choiseul	Facette ▶	5	48.869
fgewinnung, ion und Fertigung	de Bondy	Filtrer le texte		
nnische istungen, andel, Vertrieb, id Tourismus	Marie Stuart	Éditer les cellules ▶	7	48.8688
nnische istungen, andel, Vertrieb, id Tourismus	des Canettes	Éditer la colonne ▶	6	48.865
nnische istungen, andel, Vertrieb, id Tourismus		Transposer ▶		
fgewinnung, ion und Fertig		Trier...		
fgewinnung, ion und Fertig		Aperçu ▶	1	48.8519
, Literatur-, , Gesellschaft		Démarrer la réconciliation...		
aftswissensch Kunst, Kultur ing		Facettes ▶	rue Esquirol	2.36046 48.8349
ner, Besitzer		Actions ▶	rue de Paradis	2.35147 48.8752
		Copier les données de réconciliation...	rue Pierre- Fontaine	2.33404 48.8817
		Utiliser des valeurs comme identifiants...		
		Ajouter une colonne d'identifiants d'entités...	avenue des Champs- Elysées	2.30786 48.8697



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

Réconcilier la colonne "name_new"

[Accès au service de l'API](#)

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

- ☒ rue
Q79007
- ☐ panneau Histoire de Paris
Q3362192
- ☐ passage
Q13634881
- ☐ avenue
Q7543083
- ☐ attraction touristique
Q570116
- ☐ arrêt de bus
Q953806
- ☐ passages couverts de Paris
Q2900244
- ☐ bâtiment scolaire
Q1244442
- ☐ boutique

Utiliser également les détails pertinents des autres colonnes :

Colonne	Inclure?	Comme propriété
id	<input type="checkbox"/>	<input type="text"/>
surname	<input type="checkbox"/>	<input type="text"/>
first_name	<input type="checkbox"/>	<input type="text"/>
gender	<input type="checkbox"/>	<input type="text"/>
profession_verbatim	<input type="checkbox"/>	<input type="text"/>
name	<input type="checkbox"/>	<input type="text"/>
prof_categories_name	<input type="checkbox"/>	<input type="text"/>
norm	<input type="checkbox"/>	<input type="text"/>
OhdAB_01	<input type="checkbox"/>	<input type="text"/>
name_old_verbatim	<input type="checkbox"/>	<input type="text"/>
geo_long	<input type="checkbox"/>	<input type="text"/>
geo_lat	<input type="checkbox"/>	<input type="text"/>

☐ Réconcilier avec le type :

☐ Réconcilier sans type particulier

☒ Correspondance automatique des valeurs candidates

Nombre maximal de candidats renvoyés

Ajouter un service standard... Discover services... Démarrer la réconciliation... Annuler



- Matched
 - Les données sont réconciliées avec le service de données externe
- None
 - Les données ne sont pas réconciliées
- Unreconciled :
 - 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

Ajouter des colonnes à partir de la colonne name_new

Add Property Preview Reset

Suggested Properties

- destination du parcours
- forme géométrique
- hauteur
- identifiant Banque de noms de lieux du Québec
- identifiant BC Geographical Names
- identifiant Dharma Drum Buddhist College d'un lieu
- identifiant Geschichtliches Ortsverzeichnis
- identifiant GNS Unique Feature
- identifiant Vietvārdu datubāze
- interagit physiquement avec
- largeur
- localisation administrative
- longueur
- longueur

OK Cancel



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

Ouvrir... Exporter ▾

4772 lignes

Voir en: **lignes** entrées Afficher: 5 10 25 50 100 500 1000 lignes « première < précédente

<input type="checkbox"/> Toutes	<input type="checkbox"/> id	<input type="checkbox"/> surname	<input type="checkbox"/> first_name	<input type="checkbox"/> gender	<input type="checkbox"/> zusatz	<input type="checkbox"/> name_predicate	<input type="checkbox"/> specification_verbatim
☆	1.	12	Badens		F	Gräfin von	
☆	2.	72	Frederic		M		
☆	3.	97	Hamel von		M	Graf	
☆	4.	112	Ordener	G. O.	M	General Graf	
☆	5.	136	Pradel von	O.	M	Graf Mitglied des Instituts	
☆	6.	278	Nieuwerkerke von		M	Graf	
☆	7.	295	Allemans		M	Baron von	
☆	8.	317	Billing	C.	M	Baron von	
☆	9.	331	Brunswick von		M	Prinz	
☆	10.	336	Buttlar von		M	Baron	

Archive de projet OpenRefine

Valeurs séparées par des tabulations

Valeurs séparées par des virgules

Table HTML

Excel (.xls)

Excel 2007+ (.xlsx)

Classeur OpenDocument ODF (.ods)

Export tabulaire personnalisé...

Export SQL...

Patrons...

Projet OpenRefine archivé dans Google Drive...

Google Sheets...

Contributions Wikidata...

Fichier QuickStatements

Schéma Wikidata



Exporter son projet avec OpenRefine (2)

Exportateur tabulaire personnalisé

Contenu Télécharger Téléverser Code d'option

Choisir et trier les colonnes à exporter

- ☒ id
- ☒ surname
- ☒ first_name
- ☒ gender
- ☒ profession_verbatim
- ☒ name
- ☒ prof_categories_name
- ☒ norm
- ☒ OhdAB_01

Tout sélectionner Enlever toutes les sélections

Options pour **id**

pour réconcilier les cellules, sortie

- ☒ Nom d'entité correspondant
- ☐ ID de l'entité correspondante
- ☒ Lien vers la page de l'entité correspondante
- ☐ Contenu de la cellule
- ☐ Aucune sortie pour les cellules sans correspondance

☒ ISO 8601, par ex. 2011-08-24T18:36:10+08:00

- ☐ Format court
- ☐ Format long
- ☐ Personnalisé [Aide](#)
- ☐ Format moyen
- ☐ Format local complet

☐ Utiliser le fuseau horaire local ☐ Supprimer l'heure

☒ Écrire les entêtes de colonnes ☐ Écrire les lignes vides (dont toutes les cellules sont des valeurs nulles) ☐ Ignorer les facettes et les filtres et exporter toutes les lignes

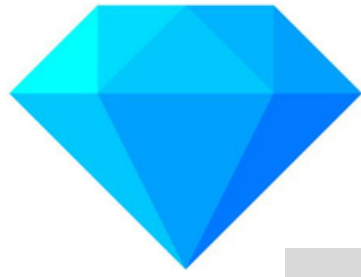
Annuler

- 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 ?