



Open Refine

LOGICIEL POUR LA QUALIFICATION DES DONNÉES HISTORIQUES

Présentation

- 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 janvier 2022, version 3.5.2
- Le site pour l'installation et la documentation : <https://openrefine.org/>

Installation

- Sur le site d'*Open Refine*, allez dans la rubrique **Downloads** et téléchargez le fichier ZIP
- Décompressez le fichier ZIP
- Ouvrez le fichier ZIP et sélectionnez l'application **Open Refine**
- Une fenêtre invite de commande s'ouvre, laissez l'invite de commande ouverte sinon *Open Refine* ne fonctionne pas.
- Une fenêtre du navigateur s'ouvre automatiquement sur la page d'accueil du logiciel

Ouvrir un fichier

- Créer un projet
- Ouvrir un projet
- Importer un projet
- Langue
- Formats comme le CSV, TSV, Excel JSON sont acceptées pour créer un projet.
- On peut également connecter une base de donnée à *Open Refine*

The screenshot displays the OpenRefine web application interface. At the top, the OpenRefine logo is followed by the tagline "Un outil puissant pour travailler avec des données désordonnées." Below this is a navigation menu with four items: "Créer un projet" (highlighted in blue), "Ouvrir un projet", "Importer un projet", and "Langue". The main content area is titled "Créer un projet en important des données. Quelles sorte" and "Les documents de type TSV, CSV, *SV, Excel (.xls and .xlsx)". It features a section "Récupérer les données à partir de" with a list of options: "Cet ordinateur" (highlighted), "Adresses web (URLs)", "Presse-papier", "Database", and "Google Data". To the right of this list is a button "Choisir des fichiers" and a "Suivant »" button.

OpenRefine Un outil puissant pour travailler avec des données désordonnées.

Créer un projet
Ouvrir un projet
Importer un projet
Langue

Créer un projet en important des données. Quelles sorte
Les documents de type TSV, CSV, *SV, Excel (.xls and .xlsx)

Récupérer les données à partir de

Cet ordinateur
Adresses web (URLs)
Presse-papier
Database
Google Data

Chercher un ou plusieurs
Choisir des fichiers A
Suivant »

Ouvrir un fichier (exemple)

- Choisissez Adresses web (URLs)
- Ecrire l'URL pointant vers les données :
`https://tinyurl.com/n5ktws4n`
- Vérifiez que les données s'affichent correctement
- Créez votre projet

OpenRefine Un outil puissant pour travailler avec des données désordonnées. Nouvelle version [Télécharger OpenRefine v3.5.2 maintenant.](#)

Créer un projet
Ouvrir un projet
Importer un projet
Langue

Créer un projet en important des données. Quelles sortes de données puis-je importer ?
Les documents de type TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, OpenDocument (.ods) et Google Data ajoutés via des extensions OpenRefine.

Récupérer les données à partir de
[Cet ordinateur](#)
Adresses web (URLs)
[Presse-papier](#)
[Base de données](#)

Indiquer une ou plusieurs adresses web (URLs) pointant vers les données à télécharger :

OpenRefine Un outil puissant pour travailler avec des données désordonnées. Nouvelle version [Télécharger OpenRefine v3.5.2 maintenant.](#)

Créer un projet
Ouvrir un projet
Importer un projet
Langue

« Recommencer Configurer les options pour l'analyse syntaxique Nom du projet Étiquettes

	id	surname	first_name	gender	title	profession_verbatim	occupation_group	legion_d_honneur	status	military_status	houseno	houseno_specification	name_old
1.	587	Mackau von	Unknown	M	Baron	Admiral, Senator	Militär	Grand-Croix	Active	Military	7	Unknown	De Cauma
2.	717	Schramm	Unknown	M	Graf	Divisionsgeneral, Senator	Militär	Grand-Croix	Active	Military	77	Unknown	Louis Le G
3.	4110	Kronowski	Unknown	M	Graf	Oberstlieutenant	Militär	Unknown	Active	Military	44	Unknown	Basse Du f
4.	12	Badens	Unknown	F	Gräfin von	Unknown	Adel	Unknown	Active	Civil	3	Unknown	Ferme
5.	97	Hamel von	Unknown	M	Graf	Unknown	Adel	Unknown	Active	Civil	89	Unknown	De Grenell
6.	112	Ordener	Unknown	M	General Graf	Senator	Adel	Grand Officier	Active	Civil	5	Unknown	De La Rév
7.	136	Bradel von	Unknown	M	Graf	Unknown	Adel	Officier	Active	Civil	Unknown	Unknown	Unknown

Considérer les données comme
Fichiers CSV / TSV / séparateur
[Fichiers texte à base de lignes](#)

Format des caractères

Les colonnes sont séparées par
☒ une virgule (CSV)
☐ une tabulation (TSV)
☐ personnalisé : ,

☐ Ignorer la ou les 0 première(s) ligne(s) du début du fichier
☒ Analyser la ou les 1 ligne(s) suivante(s) comme des entêtes de colonnes
☐ Ignorer la ou les 0 première(s) ligne(s) de données
☐ Charger au plus 0 première(s) ligne(s) de données


Les principaux onglets

- **Facette/Filtre** permet de voir les filtres que vous avez activés
- **Défaire/Refaire** permet d'annuler des actions ou d'enregistrer des actions
- **Ouvrir** pour ouvrir un autre fichier
- **Exporter** pour enregistrer votre fichier dans différents formats
- **Aide** vous redirige sur la page Github d'*Open Refine*

Ouvrir... Exporter ▼ Aide

Extensions: Wikidata ▼

« première < précédente 1 - 10 suivante > dernière »

 **OpenRefine** original_references.xlsx [Permalien](#)

Facette / Filtre

Défaire / Refaire 6 / 6

Extraire...

Appliquer...

Filtrer :

0. Create project

1. Text transform on 258 cells in column begin_page_no: grel:value.replace(".0","")

2. Text transform on 258 cells in column end_page_no: grel:value.replace(".0","")

3. Text transform on 258 cells in column

258 lignes

Voir en: lignes entrées Affic

▼ Toutes		▼ id	▼ scan_r
☆	🗑	1.	1
☆	🗑	2.	2
☆	🗑	3.	3
☆	🗑	4.	4
☆	🗑	5.	5
☆	🗑	6.	6
☆	🗑	7.	7
☆	🗑	8.	8

Exploration des données

- **Les Facettes** permettent de voir les différents choix dans une colonne de données.
- **Filtrer le texte**, permet de rechercher des données spécifiques dans votre fichier
- **Trier**, permet de choisir comment organiser ses données
- **Aperçu** pour cacher ou afficher des colonnes

« première < précédente 1 - 10 suivante > dernière »

▼ Beruf_Bezeichnung	▼ Beruf_Kategorie	▼ Adresse 1854	▼ Adr
Specerei- Stempelp Verkauf			
Rentner			
Wagenma			
Divisionsa des Straß			
Brückenb			
Rohstoff f			
Ebenisten			
Natürliche Blumen	Handel	6 u. 8 [passage] de l' Opéra	6 u. 8
Früher Advocat	Rentner	21 [rue d'] Antin	21
Ehemaliger Staatsmann	Beamte	6 [rue d'[Astorg	6
Kaufmann	Handel	4 [rue d'] Astorg	4

Transformations des données

- **Les Facettes** permettent de grouper les données qui sont similaires avec l'option *groupe*
- **Transformer** permet de nettoyer ses données avec le langage GREL qui est le langage du logiciel
- **Transformations courantes** permet d'effectuer un premier nettoyage des données avec des options déjà définies par *Open Refine*

.LSX Permalien

258 lignes

Voir en: lignes entrées Afficher: 5 10 25 50 lignes

▼ Toutes	▼ id	▼ scan_no	▼ begin_page_no	▼ end_page_no	▼ id_300dpi	▼ id_72dpi	▼ id_preview
☆ ↗	1.	1	BHVP_703983_001	Facette ▶	1	1	1
☆ ↗	2.	2	BHVP_703983_002	Filtrer le texte	2	2	2
☆ ↗	3.	3	BHVP_703983_003	Éditer les cellules ▶	3	3	3
☆ ↗	4.	4	BHVP_703983_004	Éditer la colonne ▶	4	4	
☆ ↗	5.	5	BHVP_703983_005	Transposer ▶			
☆ ↗	6.	6	BHVP_703983_006	Trier...			
☆ ↗	7.	7	BHVP_703983_007	Aperçu ▶			
☆ ↗	8.	8	BHVP_703983_008	Réconcilier ▶			
☆ ↗	9.	9	BHVP_703983_009				
☆ ↗	10.	10	BHVP_703983_010				

Transformer...
Transformations courantes ▶
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

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 majuscules
En minuscules
En nombre
En date
En texte
En valeurs nulles
Transformer en chaîne vide

Transformations des données (*suite*)

- **Aperçu** permet de voir le résultat de la fonction GREL avant de l'exécuter
- **Historique** sont les fonctions qui ont déjà été utilisées
- **Étoilée** permet de voir les fonctions favorites
- **Aide** donne des éléments de syntaxe pour écrire des fonctions
- **value.replace** permet de remplacer ou de supprimer des termes

Transformation textuelle personnalisée sur la colonne scan_no

Expression

```
value.replace("BHVP","test")
```

Aperçu

Historique

Étoilée

Aide

row	value	value.replace("BHVP","test")
1.	BHVP_703983_001	test_703983_001
2.	BHVP_703983_002	test_703983_002
3.	BHVP_703983_003	test_703983_003
4.	BHVP_703983_004	test_703983_004
5.	BHVP_703983_005	test_703983_005
6.	BHVP_703983_006	test_703983_006
7.	BHVP_703983_007	test_703983_007

En cas d'erreur

- ☒ conserver l'original
☐ vider la cellule
☐ conserver l'erreur

☐ Retransformer fois maximum, tant que le

OK

Annuler

Transformations des données (*suite*)

- Pour transformer les données, on peut éditer la colonne
- Diviser en plusieurs colonnes
- Joindre des colonnes
- Ajouter des colonnes à partir de valeurs réconciliées

Rebellen Permalien

Ouvrir... Exporter ▼

Extensions: Wiki

46 lignes

Voir en: lignes entrées Afficher: 5 10 25 50 lignes « première < précédente 1 - 46 suivante > dernière »

Toutes	ID	Name	Vorname	Status	Profession/Position	Konfession	Ort	Ehepartner	Status Ehepartn
☆	1.	1	Facette						
☆	2.	2	Filtrer le texte						
☆	3.	3	Éditer les cellules						
☆	4.	4	Éditer la colonne						
☆	5.	5	Transposer						
☆	6.	6	Trier...						
☆	7.	7	Aperçu						
☆	8.	8	Réconcilier						
☆	9.	9							
☆	10.	10	Bálogh	Caspar					
☆	11.	11	Szuháy	Matthias					
☆	12.	12	Almássy	Franz					
☆	13.	13	Petróczy	Stephar					
☆	14.	14	Fejérpataki	Adam					
☆	15.	15	Szente	Valentin			Eperjes		
☆	16.	16	Bónis	(Franz)			Katholik	Preßburg	Szentbenedeki Elisabeth
☆	17.	17	Palóczáy	Adam			Katholik	Preßburg	
☆	18.	18	Keczer	Samuel			Katholik	Preßburg	?
☆	19.	19	Kökényesdy	Georg	Adel	Praefectus Nádasdy a Füzér	Katholik	Füzér	
☆	20.	20	Szepessy	Pál	Adel		Katholik	Somlyo	Csuda Susanna
☆	21.	21							
☆	22.	22							
☆	23.	23							
☆	24.	24							
☆	25.	25							
☆	26.	26							
☆	27.	27							
☆	28.	28							
☆	29.	29							
☆	30.	30							
☆	31.	31							
☆	32.	32							
☆	33.	33							
☆	34.	34							
☆	35.	35							
☆	36.	36							
☆	37.	37							
☆	38.	38							
☆	39.	39							
☆	40.	40							
☆	41.	41							
☆	42.	42							
☆	43.	43							
☆	44.	44							
☆	45.	45							
☆	46.	46							

Réconciliation des données

- Dans la colonne souhaitée, sélectionnez **réconciliation** puis **démarrer la réconciliation**
- Choisissez un **service de réconciliation**
- Ensuite, choisissez un **type d'entité**
- **Démarrer la réconciliation** pour faire correspondre vos données

<https://reconciliation-api.github.io/testbench/>

The screenshot shows the 'Réconcilier la colonne "Adresse_Straße"' interface. On the left, a list of services is available for selection, including Wikidata (en), Getty Vocabularies Reconciliation Service, OpenLibrary, Geonames Reconciliation Service, VIAF, GND reconciliation for OpenRefine, Wikibase Registry Test (en), Wikidata reconcil.link (de), and Wikidata reconcil.link (fr). The main area displays a list of entity types for reconciliation, with 'street' (Q79007) selected. Below this, there are options to reconcile with a specific type, without a specific type, or with automatic correspondence of candidate values (checked). A field for the maximum number of candidates returned is also present. On the right, a table lists columns to include in the reconciliation process, such as Person_ID, Person_Name_Vorname, Person_Name_Nachname, Person_Name_Namenszusatz, Person_Geschlecht, Beruf_Bezeichnung, Beruf_Kategorie, Adresse_1854, Adresse_Hausnummer, Adresse_Zusatz, Adresse_Straße 2, and Adresse_Straße 3. At the bottom, there are buttons for 'Ajouter un service standard...', 'Démarrer la réconciliation', and 'Annuler'.

Réconcilier la colonne "Adresse_Straße" » Accès Service API

Réconcilier chaque cellule avec une entité de l'un de ces types: Utiliser également les détails pertinents des autres colonnes:

☒ street Q79007
☐ Dead end street Q12731
☐ avenue Q7543083
☐ walkway Q13634881
☐ watercolor painting Q18761202
☐ tourist attraction Q570116
☐ bus stop Q953806
☐ thoroughfare Q83620
☐ covered passages of Paris

☐ Réconcilier avec le type:
☐ Réconcilier sans type particulier
☒ Correspondance automatique des valeurs candidates
Nombre maximal de candidats renvoyés

Ajouter un service standard...

Colonne	Inclure?	Comme propriété
Person_ID	<input type="checkbox"/>	<input type="text"/>
Person_Name_Vorname	<input type="checkbox"/>	<input type="text"/>
Person_Name_Nachname	<input type="checkbox"/>	<input type="text"/>
Person_Name_Namenszusatz	<input type="checkbox"/>	<input type="text"/>
Person_Geschlecht	<input type="checkbox"/>	<input type="text"/>
Beruf_Bezeichnung	<input type="checkbox"/>	<input type="text"/>
Beruf_Kategorie	<input type="checkbox"/>	<input type="text"/>
Adresse_1854	<input type="checkbox"/>	<input type="text"/>
Adresse_Hausnummer	<input type="checkbox"/>	<input type="text"/>
Adresse_Zusatz	<input type="checkbox"/>	<input type="text"/>
Adresse_Straße 2	<input type="checkbox"/>	<input type="text"/>
Adresse_Straße 3	<input type="checkbox"/>	<input type="text"/>

Ajouter un service standard... Démarrer la réconciliation Annuler

Réconciliation des données (suite)

- **Type d'action de jugement** permet de voir les données qui ont été réconciliées
- **Auto** sont les données qui ont trouvé une correspondance
- **Similar**, *Open Refine* nous laisse le choix pour faire correspondre les données
- **None / Unconciliated**, les données ne sont pas réconciliées
- **Blank**, la cellule est vide

The screenshot shows the OpenRefine web interface for a dataset named 'adressbuch1854 (brouillon)'. The main table displays 4772 lines of data. The left sidebar shows a facet for 'Adresse_Straße heute Type d'action de jugement' with 2 choices: 'auto' (2164) and 'similar' (2450). The main table has columns for 'Adresse_Straße', 'Adresse_Zusatz', and 'Adresse_Straße'. A context menu is open over the 'Adresse_Straße' column, showing options like 'Facette', 'Filtrer le texte', 'Éditer les cellules', 'Éditer la colonne', 'Transposer', 'Trier...', 'Aperçu', and 'Réconcilier'. The 'Réconcilier' option is selected, and a sub-menu is open showing 'Démarrer la réconciliation...', 'Facettes', 'Actions', 'Copier les données de réconciliation...', 'Utiliser des valeurs comme identifiants', and 'Ajouter une colonne d'identifiants d'entités'. The 'Facettes' sub-menu is also open, showing options like 'Par avis', 'Type d'action de jugement', 'Date d'action de jugement', '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 candidat', and 'Types de meilleurs candidats'.

OpenRefine adressbuch1854 (brouillon) Permalien

Facette / Filtre Défaire / Refaire 649 / 649

Rafraîchir Tout réinitialiser Tout supprimer

4772 lignes

Voir en: lignes entrées Afficher: 5 10 25 50 lignes

Adresse_Straße heute Type d'action de jugement

2 choix Trier par: nom compte

auto 2164

similar 2450

(blank) 158

Facette par nombre de choix

Adresse_Straße Adresse_Zusatz Adresse_Straße Adresse_Straße Adresse_Straße

Facette

Filtrer le texte

Éditer les cellules

Éditer la colonne

Transposer

Trier...

Aperçu

Réconcilier

Démarrer la réconciliation...

Facettes

Actions

Copier les données de réconciliation...

Utiliser des valeurs comme identifiants

Ajouter une colonne d'identifiants d'entités

Par avis

Type d'action de jugement

Date d'action de jugement

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 candidat

Types de meilleurs candidats

Réconciliation des données (suite)

- **Apparier cette cellule**, permet de faire correspondre uniquement cette cellule avec cette source de données externe
- **Apparier toutes les cellules identiques**, permet de faire correspondre toutes les cellules qui ont le même nom avec cette source de données externe

Adresse_Straße	Paris city digital	Latitude	Longitude	Adresse_Arrond	Adresse_Arrond	instan
Avenue De La Motte-picquet	5230	48.855798	2.307759	10	7th arrondissement of Paris	Choisir une nouvelle correspondance
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> avenue de La Motte-Picquet (100)						
<input checked="" type="checkbox"/> Créer un nouveau sujet						
Chercher une correspondance						
Avenue des Champs-Élysées					8th arrondissement of Paris	Choisir une nouvelle correspondance
Avenue des Champs-Élysées					8th arrondissement of Paris	Choisir une nouvelle correspondance
Unknown				1	8th arrondissement of Paris	Choisir une nouvelle correspondance
rue Joseph-Sansboeuf	4898	48.8755	2.32294	1	8th arrondissement of Paris	Choisir une nouvelle correspondance
impasse Guéménée	4329	48.8541	2.36577	8	4th arrondissement of Paris	Choisir une nouvelle correspondance
boulevard des Italiens	4687	48.87141944	2.33699722	2	9th arrondissement of Paris	Choisir une nouvelle correspondance
rue d'Antin	0348	48.869	2.33351	2	2nd arrondissement of Paris	Choisir une nouvelle correspondance
rue d'Astorg	0474	48.8731	2.31949	1	8th arrondissement of Paris	Choisir une nouvelle correspondance
rue d'Astorg	0474	48.8731	2.31949	1	8th arrondissement of Paris	Choisir une nouvelle correspondance

Ajouter des colonnes à partir de valeurs réconciliées

- Allez dans **éditer la colonne > ajouter des colonnes à partir de valeurs réconciliés**
- Cette option permet d'ajouter de nouvelles colonnes dans votre tableau de données avec la réconciliation faite précédemment
- Choisir la propriété qui correspond aux données que vous souhaitez afficher

Ajouter des colonnes à partir de la colonne Adresse_Straße heute

Add Property

Suggested Properties

country

country

destination point

Dharma Drum Institute of Liberal Arts place ID

GNS Unique Feature ID

highway system

Historical Gazetteer (GOV) ID

Latvian toponymic names database ID

length

located in the administrative territorial entity

occupant

OS grid reference

start point

Statistics Canada Geographic code

structure replaced by

Preview

Reset

Adresse_Straße heute	located in the administrative territorial entity
	supprimer configurer
avenue de La Motte-Picquet	7th arrondissement of Paris
Avenue des Champs-Élysées	Champs-Élysées
Avenue des Champs-Élysées	Champs-Élysées
<not reconciled>	
rue Joseph-Sansboeuf	Paris
impasse Guéménée	4th arrondissement of Paris
boulevard des Italiens	2nd arrondissement of Paris
rue d'Antin	Paris
rue d'Astorg	Paris
rue d'Astorg	Paris

OK

Cancel

Export des données

- **Archive de projet Open Refine** génère un projet Open Refine au format .tar
- **Valeurs séparées par des tabulations/virgules** génèrent des fichiers TSV et CSV
- **Table HTML** construit un tableau de données au format HTML qu'on peut insérer dans une page web
- **Excel et Excel 2007** génère des tableaux pour différentes versions d'Excel
- **ODF** pour un classeur manipulable sous **Libre Office** ou **Open Office Calc**
- **Patrons** pour un format JSON
- **Avec google drive** en archive ou google sheets.

rouillon) Permalien

Ouvrir... Exporter ▼ Aide

4772 lignes

Voir en: **lignes** entrées Afficher: 5 10 25 50 lignes

			Person_ID	Person_Name_V	Person_Name_N	Person_Name_N	Person_Geschle	Beruf	
☆	1.	3672		Keller		männlich	Specerei-, Stempelpa Verkauf		28
☆	2.	291		Ackermann		männlich	Rentner		12
☆	3.	502		Potoska von	Gräfin	weiblich			12
☆	4.	767		Röser		männlich	Wagenmal		64
☆	5.	513	C.	Reib II.		männlich	Divisionsau des Straße Brückenba		
☆	6.	3280		Koch		männlich	Rohstoff für Ebenisten	Handel	8 impasse Guéménée 8
☆	7.	809		Baron		männlich	Natürliche Blumen	Handel	6 u. 8 passage de l' Opéra 6 u.

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...

OpenRefine project archive to Google Drive...

Google Sheets

Contributions Wikidata...

Fichier QuickStatements

Schéma Wikidata

Liens utiles :

- Seth van Hooland, Ruben Verborgh, et Max De Wilde, «Nettoyer ses données avec OpenRefine,» traduction par Sybille Clochet, *The Programming Historian en français* 1 (2019), <https://doi.org/10.46430/phfr0004>.
- Mathieu Saby, *Nettoyer et préparer ses données avec Open Refine*, 6/11/2020, <https://msaby.gitlab.io/tutoriel-openrefine/index.html>
- Documentation Open Refine, <https://docs.openrefine.org/>