

# Demo Open Refine

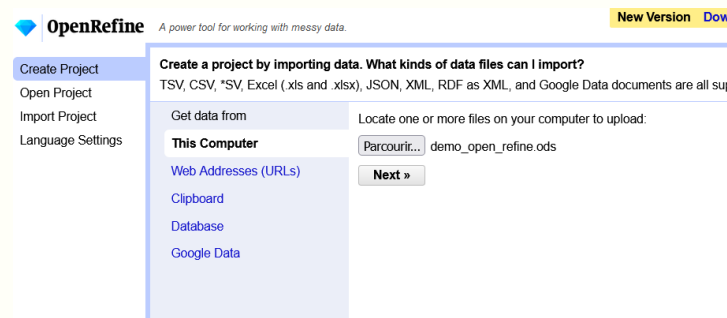
## *Cours open data - Ministère de la Culture*

Dataactivist – Anne-Laure Donzel

2023-11-16

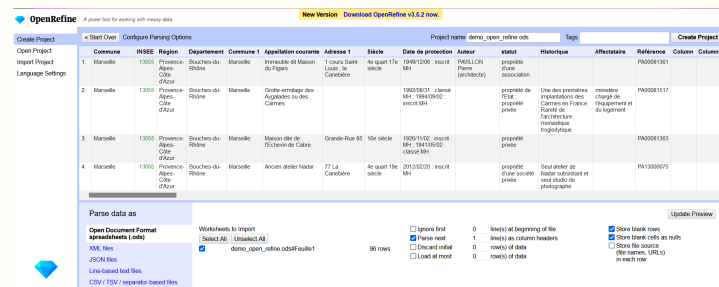
## *Charger un fichier*

Charger le fichier, plusieurs format sont acceptés.



## *Sélectionner les options et créer le projet*

Suivant les cas : choix du séparateur, de l'encodage, de l'en-tête... puis créer le projet.



## *Les facettes*

Le contenu de chaque colonne peut être visualiser sous la forme de facette. Test sur la facette *Commune* : changer la valeur blank

OpenRefine demo\_open\_refine ods Permalink

Facet / Filter Undo / Redo 0/0 96 rows

Show as: rows records Show: 5 10 25 50 rows

Commune 1 change

1 choices Sort by: name count Cluster

Marseille (0)

(blank) 1

Facet by choice counts

	Commune	INSEE	Région	Département	Appellation cour	Adresse 1	Siècle	Date de protect	Auteur
1	Marseille	13055	Provence-Alpes-Côte d'Azur	Bouches-du-Rhône	test filter	numéric facet	4e quart 19e siècle	1949/12/06 : insc MH	PROLLOP
2	Marseille	13055	Provence-Alpes-Côte d'Azur	Bouches-du-Rhône	test filter	numéric facet	4e quart 19e siècle	1949/12/06 : insc MH	PROLLOP
3	Marseille	13055	Provence-Alpes-Côte d'Azur	Bouches-du-Rhône	test filter	numéric facet	4e quart 19e siècle	1949/12/06 : insc MH	PROLLOP
4	Marseille	13055	Provence-Alpes-Côte d'Azur	Bouches-du-Rhône	test filter	numéric facet	4e quart 19e siècle	1949/12/06 : insc MH	PROLLOP

## Changer une valeur

Faire une facette sur *Affectataire* : changer la Direction Générale du Patrimoine en Ministère de la Culture

Méthode 1 : en modifiant la valeur dans la fenêtre des facettes (*edit*)

OpenRefine demo\_open\_refine ods Permalink

Facet / Filter Undo / Redo 2/2 95 rows

Show as: rows records Show: 5 10 25 50 rows

Affectataire change

8 choices Sort by: name count Cluster

direction générale des patrimoines 2

ministère chargé de l'équipement 2

ministère chargé de l'équipement et du logement 2

ministère chargé de l'intérieur 1

ministère chargé de la culture 2

ministère chargé de la défense 2

ministère chargé de la défense ; direction générale des patrimoines 1

ministère de l'écologie, du développement durable et de l'énergie 1

direction générale des patrimoines

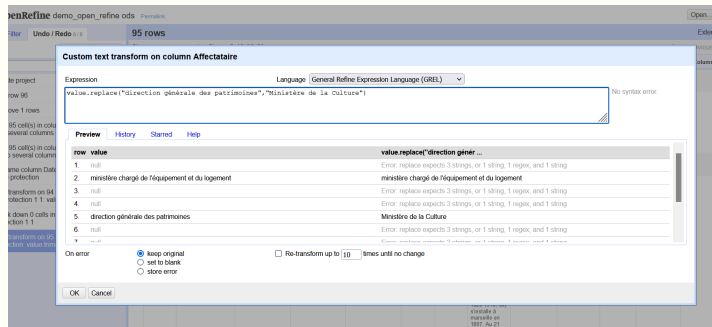
Apply Enter Cancel ESC

Affectataire	Commune 1	Appellation cour	Adresse 1	Siècle	Date de pro
direction générale des patrimoines	Marseille	Immeuble dit Maison du Figaro	1 cours Saint-Louis ; la	4e quart 17e siècle	1949/12/06 : insc MH
	Marseille	Maison dite de l'Echevin de Cabre	Grande-Rue 85	16e siècle	1926/08/31 : clas MH ; 1941/05/02 classé MH
	Marseille	Ancien atelier Nadar	77 La Canebière	4e quart 19e siècle	2012/02/20 : insc MH

Méthode 2 : en sélectionnant la colonne, *edit cells* et *replace*, permet de faire le changement grâce à un chercher-remplacer

iteur	statut	Historique	Affectataire	Référence	Column	Column 16
ON	propriété d'une association		Facet	81361		
cte)	propriété de l'Etat ; propriété privée	Une des premières implantations des Carmes en France. Rareté de l'architecture monastique troglodytique.	Text filter			
	propriété privée		Edit cells			
	propriété d'une société privée	Seul atelier de Nadar subsistant et seul studio de photographe professionnel du 19e siècle	Edit column			
			Transpose			
			Sort...			
			View			
			Reconcile			
				PA130		

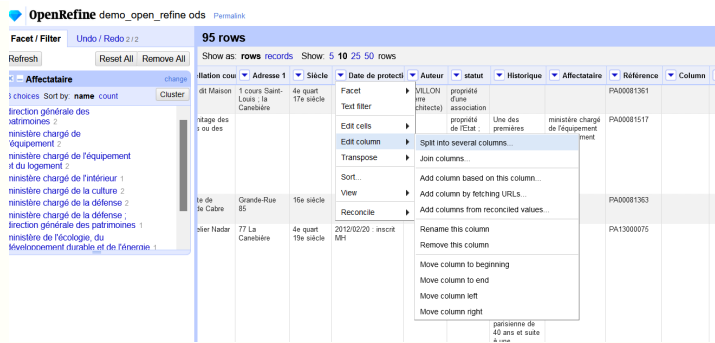
**Méthode 3** : par une expression régulière, *edit cells*, *transform* puis saisir `> value.replace("direction générale des patrimoines", "Ministère de la Culture")`



**Méthode 4** : par du rapprochement sémantique. Pour cela il faut choisir une autre colonne, par exemple *commune*

*Séparer des colonnes*

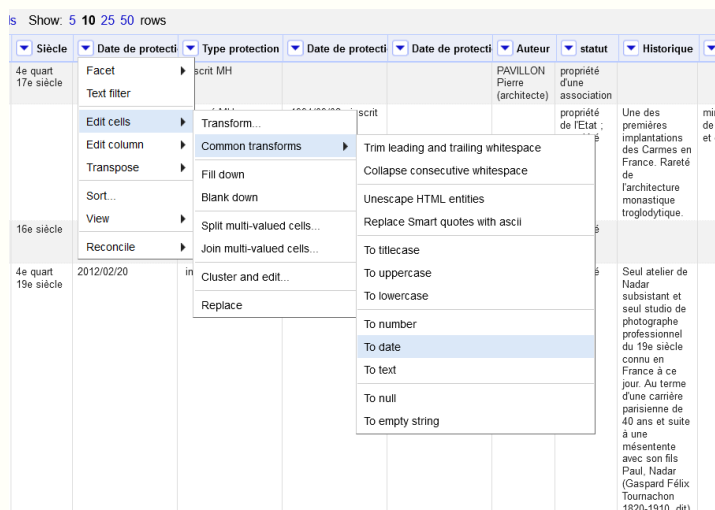
Choisir la colonne *Date de protection*, *split into several columns* et choisir les paramètres. Ne garder que la date dans la colonne.



## Autres changements

### Changer le type de champ

La colonne date peut être transformée en une véritable colonne de date : choisir la colonne, *edit cells*, *commons transforms*, *to date*



## Supprimer des espaces

Lors des changements précédents des espaces sont apparus en tête de champ, ils peuvent être supprimés : *edit cells*, *Trim leading and trailing whitespace*

## Aller plus loin avec Open Refine

Open Refine peut appeler l'API de Wikidata, la base de données en web sémantique de Wikimedia. Il est ainsi possible de récupérer des éléments de Wikidata.

Par exemple, notre fichier comprend, pour certains monuments, un auteur. Si ces auteurs existent sur Wikidata, il est possible de rapatrier des informations, par exemple leur lieu de naissance.

Cette opération s'appelle une *réconciliation* de données.

## Réconciliation des auteurs

Choisir la colonne puis *reconcile*, *start reconciling* ajouter le service de réconciliation français

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

de protectio	Type protection	Auteur	statut	Historique	Affectataire	Référence
6T00:00:00Z	inscrit MH	Facet				PA00081361
		Text filter				
1T00:00:00Z	classé MH	Edit cells		Une des premières implantations des Carmes en France. Rareté de l'architecture monastique troglodytique.	ministère chargé de l'équipement et du logement	PA00081517
		Edit column				
		Transpose				
		Sort...				
2T00:00:00Z	inscrit MH	View				PA00081363
		Reconcile		Start reconciling...		
0T00:00:00Z	inscrit MH		propriété d'une société privée	Facets		PA13000075
				Actic	Reconcile text in this column with it	
				Copy reconciliation data...		
				Use values as identifiers		
				Add entity identifiers column		
			d'une carrière parisienne de 40 ans et suite à une			

La réconciliation porte sur des être

humains, *start reconciling*

Reconcile column "Auteur"

Reconcile each cell to an entity of one of these types:

- ☒ être humain

Also use relevant details from other columns:

Column	Include? As Property
Commune	<input type="checkbox"/>
INSEE	<input type="checkbox"/>
Région	<input type="checkbox"/>
Département	<input type="checkbox"/>
Commune 1	<input type="checkbox"/>
Appellation courante	<input type="checkbox"/>
Adresse 1	<input type="checkbox"/>
Siècle	<input type="checkbox"/>
Date de protection 1	<input type="checkbox"/>
Type protection	<input type="checkbox"/>
statut	<input type="checkbox"/>
Historique	<input type="checkbox"/>

☐ Reconcile against type:

☐ Reconcile against no particular type

☒ Auto-match candidates with high confidence

Maximum number of candidates to return:

Le système a fait le lien entre notre fichier et Wikidata, il propose une correspondance (matching), si cela correspond bien il faut le valider.

on Auteur statut Historique Affectataire Référence Column

PAVILLON Pierre (architecte)  propriété PA00081361

☒ Pierre Pavillon (73)

☒ Create new item

Pierre Pavillon (Q15409315)

architecte français

l'architecture monastique troglodytique.

Alors ?

La qualité de la colonne ne permet pas de faire une bonne réconciliation.

*Réconciliation à partir d'un identifiant*

Tentons la réconciliation sur une autre colonne : *Référence*. Wikidata intègre de nombreux référentiel dont celui sur les identifiants des monuments historiques.

17 matching rows (0/05 total)

### Reconcile column "Référence"

Reconcile each cell to an entity of one of these types:

- ☐ maison Q3947
- ☐ chapelle Q108325
- ☐ château Q751878
- ☐ cathédrale catholique Q56242215
- ☐ immeuble Q11755959
- ☐ église Q16970
- ☐ grotte Q35509
- ☐ atelier Q656720
- ☐ forteresse Q423914

Also use relevant details from other columns:

Column	Include? As Property
Commune	<input type="checkbox"/>
INSEE	<input type="checkbox"/>
Région	<input type="checkbox"/>
Département	<input type="checkbox"/>
Commune 1	<input type="checkbox"/>
Appellation courante	<input type="checkbox"/>
Adresse 1	<input type="checkbox"/>
Siècle	<input type="checkbox"/>
Date de protection 1 1	<input type="checkbox"/>
Type protection	<input type="checkbox"/>
Auteur	<input type="checkbox"/>
statut	<input type="checkbox"/>

Reconcile against type:

☐ Reconcile against no particular type

☒ Auto-match candidates with high confidence

Maximum number of candidates to return

Add Standard Service... Start Reconciling Cancel

Cette fois cela fonctionne mieux, on peut ensuite récupérer les coordonnées géographiques des monuments, *add columns, add columns from reconciled values*

5 rows

### Add columns from reconciled column Référence

Add Property Preview Reset

coordonnées géographiques	Référence	coordonnées géographiques
coordonnées géographiques	P625	43.2964,5.37806
coordonnées géographiques de l'objet		43.3592,5.36194
identifiant GeoNames	P1566	43.2975,5.37111
identifiant dans la base de données géographique GeoNames		43.297675,5.380574
identifiant Géopatronyme	P3370	43.299444,5.364722
identifiant d'un nom de famille dans la base de données Géopatronyme		43.297916666667,5.368711111111
géométrie géographique	P3896	
données géographiques de Wikimedia Commons		43.29766667,5.36875
identifiant Geographical Names Board of New South Wales	P3517	43.274304,5.385904
Identifiant sur le site web du Geographical Names Board of New South Wales		43.2953,5.5025
identifiant du registre national géorgien des monuments	P4166	43.279691111111,5.325138888889
identifiant pour un monument dans le registre national des monuments		43.296455,5.371056
identifiant George Eastman Museum	P10018	43.2744049,5.3635565
identifiant d'une personne ou d'une organisation sur le site web du George Eastman Museum		43.29472,5.3625
		43.2992,5.46694
		43.300205,5.367774

OK Cancel