

Demo Open Refine

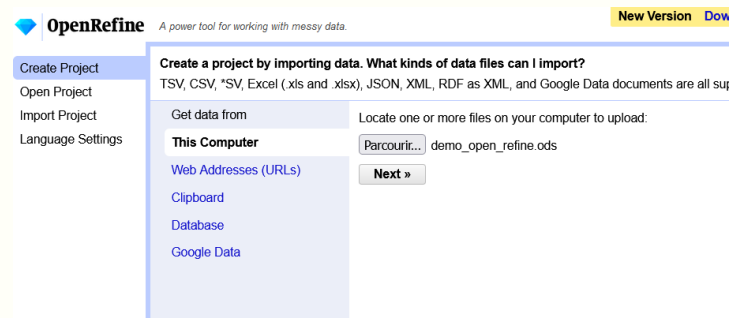
Cours open data - Ministère de la Culture

Dataactivist - Anne-Laure Donzel

2023-04-13

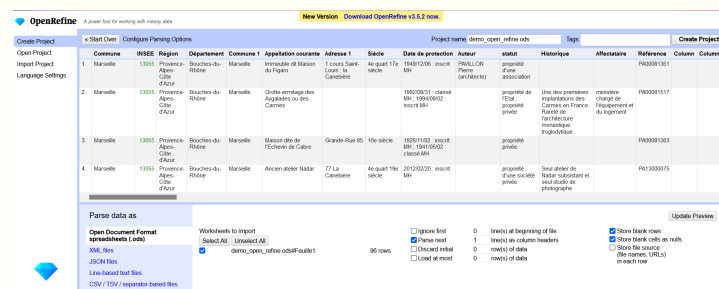
Charger un fichier

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 10 (blank) 1

Facet by choice counts

	Commune	INSEE	Région	Département	Commune 1	Appellation cour	Adresse 1	Siècle	Date de protect	Auteur	it
1	Marseille	13055	Provence-Alpes-Côte d'Azur	Bouches-du-Rhône							
2	Marseille	13055	Provence-Alpes-Côte d'Azur	Bouches-du-Rhône							
3	Marseille	13055	Provence-Alpes-Côte d'Azur	Bouches-du-Rhône							
4	Marseille	13055	Provence-Alpes-Côte d'Azur	Bouches-du-Rhône							

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

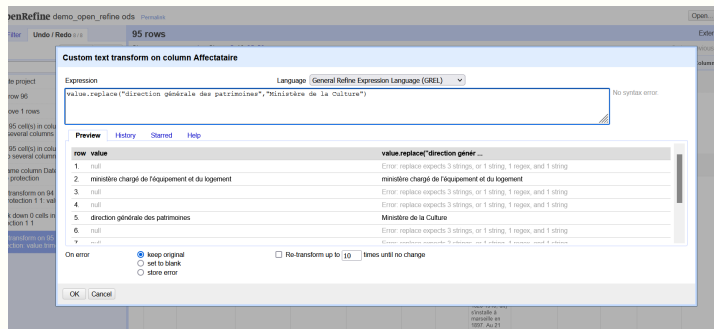
Apply Enter Cancel Esc

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	1926/11/02 : insc 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

acteur	statut	Historique	Affectataire	Référence	Column	Column 16
ON	propriété d'une association		Facet	81361		
acte)	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		

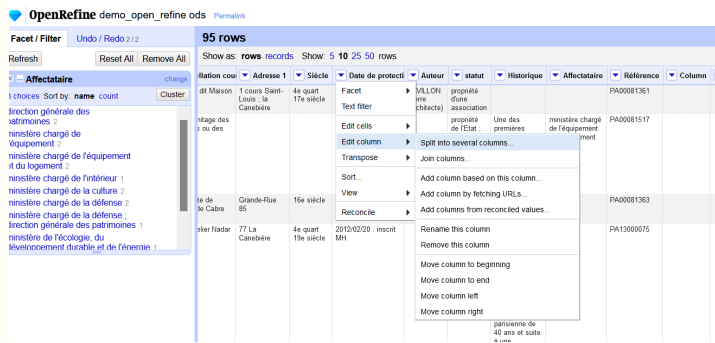
Méthode 3 : par une expression régulière, *edit cells, transform* puis saisir >
`value.replace("direction générale des patrimoines", "Ministère de la Culture")`



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

Séparer des colonnes

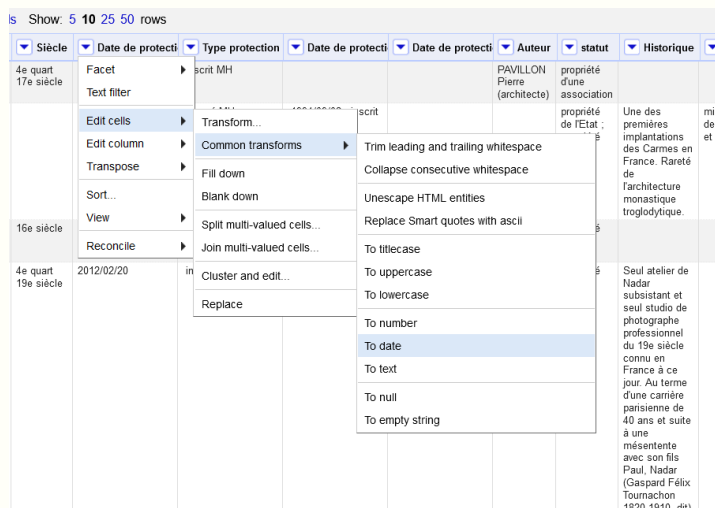
Choisir la colonne *Date de protection*, *split into several columns* et choisir les paramètres. Ne garder que la date dans la colonne.



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
1T00:00:00Z	classé MH	Text filter				
		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*

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"/>

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

Auteur	statut	Historique	Affectataire	Référence	Colonne
PAVILLON Pierre (architecte)	propriété			PA00081361	

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 matches row(s) / 0% total

Reconcile column "Référence"

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

- ☐ maison Q3347
- ☐ chapelle Q108325
- ☐ château Q751878
- ☐ cathédrale catholique Q56242215
- ☐ immeuble Q11755959
- ☐ église Q169970
- ☐ 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

1956-10-03T00:00:00Z inscrit MH REY Site (architecte) propriété d'une Édifice construit pour le PA13000010

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 remove configure

Searching...

coordonnées géographiques	P625	ro	des	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	le-Marie-		43.299444,5.364722
identifiant d'un nom de famille dans la base de données Géopatronymie				43.297916666667,5.368711111111
géométrie géographique	P3896	intents blancs		
données géographiques de Wikimedia Commons				43.29766667,5.36875
identifiant Geographical Names Board of New South Wales	P3517	me-des-		
Identifiant sur le site web du Geographical Names Board of New South Wales		dulevard		43.274304,5.385904
identifiant du registre national géorgien des monuments	P4166	uzine		43.2953,5.5025
identifiant pour un monument dans le registre national des monuments				43.279861111111,5.325138888889
identifiant George Eastman Museum	P10018	0 qui du Port		43.296455,5.371066
identifiant d'une personne ou d'une organisation sur le site web du G		vieux sans Lucia		43.2744049,5.3635565
		fort Saint-Jean		43.29472,5.3625
		oppidum des Bacu de Saint-Marcel		43.2992,5.48694
		ville de Chanté		43.300205,5.367774

OK Cancel

et réussit à en faire le rendez-