

Introduction à OpenRefine

Cette introduction à OpenRefine a été créée par Owen Stephens (owen@ostephens.com), pour le compte de la *British Library*, en juillet 2014.

Elle a été traduite en français et amendée par Hervé Piedcoq (hpiedcoq@15cpp.fr) en mars 2015.

Ce travail est placé sous licence Creative Commons Attribution 4.0 International Licence <http://creativecommons.org/licences/by/4.0/> .

Il est suggéré que si vous utilisez travail, vous incluiez la phrase suivante « Développé par Owen Stephens pour le compte de la *British Library* ».

Convention : OpenRefine est un logiciel initialement écrit en anglais. Une traduction en français est en cours d'élaboration. Toutefois, afin de clarifier au mieux le fonctionnement de cet outil, nous avons pris le parti de franciser toutes les fonctions mais également de maintenir dans le texte, le nom de la fonction en anglais en italique et entre parenthèse, afin de pouvoir coller à la nombreuse documentation en anglais disponible en ligne.

1 – Pour commencer

Qu'est-ce qu'OpenRefine ?

On décrit souvent OpenRefine comme un outil pour travailler ou retravailler des « données sales », mais qu'est-ce que cela veut dire ?

Il est sans doute plus facile de décrire et présenter le type de données avec lesquelles OpenRefine est à l'aise pour travailler, et le type de problématiques qu'il peut vous aider à résoudre.

OpenRefine est particulièrement utile là où vous avez des données présentées dans un format dit « tabulaire » (en lignes et en colonnes), mais qui présentent des incohérences, ou des inconsistances dans leurs données, ou dans la terminologie utilisée. On peut citer à titre d'exemple :

- Avoir un aperçu rapide d'un tableau de données
- Résoudre les problèmes de cohérence, ou d'inconsistance de ce tableau de données
- Vous aider à séparer et à affiner ce tableau
- Faire correspondre des données avec celle contenues dans d'autres tableaux
- Enrichir ces données avec d'autres informations en provenance d'autres sources

On peut par exemple envisager de travailler sur des scénarii de ce type :

1. Vous voulez connaître combien de fois une certaine valeur apparaît dans un tableau.
2. Vous voulez savoir comment les valeurs sont réparties au sein de votre groupe de données
3. Vous voulez modifier le formatage d'une liste de dates hétérogènes, en utilisant un seul format standard

Vos données	Ce que vous voulez obtenir
1 ^{er} janvier 2015	01/01/2015
01/01/2015	01/01/2015
2015-01-01	01/01/2015
Jan 1 2015	01/01/2015

4. Vous avez une liste de noms ou de termes, tous différents mais qui se réfèrent tous à la même entité (Personne, lieu ou encore concept...)

Vos données	Ce que vous voulez obtenir
Paris	Paris
Paris]	Paris
(Paris,	Paris
paris	Paris

5. Vous avez une multitude d'informations contenue dans une seule colonne et vous souhaitez

séparer ces données en de multiples colonnes

Vos données	Ce que vous voulez obtenir							
	Institution	Service	Adresse1	Adresse2	Ville	Région	Pays	Code Postal
Université de Paris 4, département de sociologie, 4 rue des fossés, 75012 Paris	Université de Paris 4	Département de sociologie	4 rue des fossés		Paris		France	75012
Ecole Polytechnique, Informatique, Résidence des fleurs, 16 rue Montmartre, 45000 Orléans, France	Ecole Polytechnique	Informatique	16 rue Montmartre	Résidence des fleurs	Orléans		France	45000

6. Vous voulez enrichir vos données à l'aide d'une source d'informations

Vos données	Date de naissance (issue de VIAF, base de donnée internet)	Date de décès (issue de VIAF, base de donnée internet)
Braddon, M. E. (Mary Elizabeth)	1835	1915
Rossetti, Wiliam Michael	1829	1919
Prest, Thomas Peckett	1810	1879

Télécharger OpenRefine

Vous pouvez télécharger OpenRefine sur le site du projet : <http://openrefine.org/download.html>
A la date de rédaction de ce document, la version officielle et stable d'OpenRefine est appelée « *Google Refine 2.5* », mais on général, on préférera utiliser la version « *OpenRefine 2.6 – Development Version* » que nous recommandons et qui malgré son nom de *Beta* est très stable depuis longtemps.

Il existe des versions pour Windows, Mac OS et Linux

Installer et utiliser OpenRefine

Quand vous téléchargez OpenRefine pour Windows ou Linux depuis l'adresse internet ci-dessus, vous récupérerez un fichier compressé .zip. Pour installer OpenRefine il suffit de décompresser l'archive dans le répertoire où vous souhaitez installer le programme. Il peut s'agir d'un répertoire personnel ou dans Applications, ou encore un répertoire programme ; OpenRefine peut fonctionner quelque soit l'endroit où il est installé.

Sur MacOS, vous récupérez une archive .dmg, qu'il suffit de copier dans le répertoire Applications

(ou ailleurs).

OpenRefine est une application java, il est donc nécessaire d'installer au préalable un environnement java sur sa machine (JRE). Si vous ne l'avez pas installé, vous pouvez récupérer le JRE sur le site <http://java.com>.

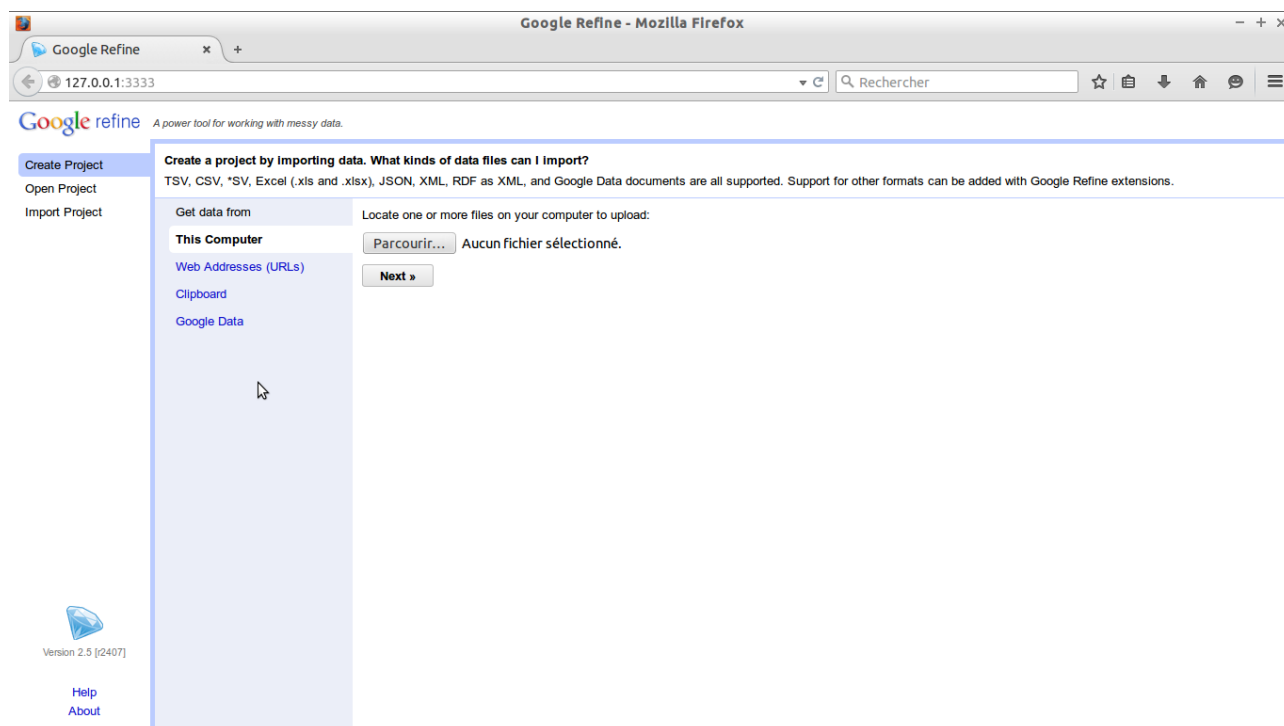
Pour lancer Refine :

Sous Windows : Naviguez dans le répertoire où est installé OpenRefine, et double-cliquez sur « *google-refine.exe* » ou « *openrefine.exe* » ou bien encore « *refine.bat* » selon la version que vous avez installé.

Sous Linux : Naviguez dans le répertoire où est installé OpenRefine dans un terminal et tapez « *./refine* »

Sous Mac : Double-cliquez sur l'icône OpenRefine dans le répertoire d'installation.

L'interface d'OpenRefine est accessible dans votre navigateur web (Safari, Firefox ou Chrome). A l'exécution du programme celui-ci doit se lancer automatiquement et vous rediriger vers l'adresse : <http://127.0.0.1:3333>



Obtenir de l'aide

Il est possible d'obtenir de l'aide, de la documentation et des tutoriels sur internet :

- [Le wiki OpenRefine](#)
- [Free your Metadata](#)
- La [mailing-list](#) et le forum OpenRefine

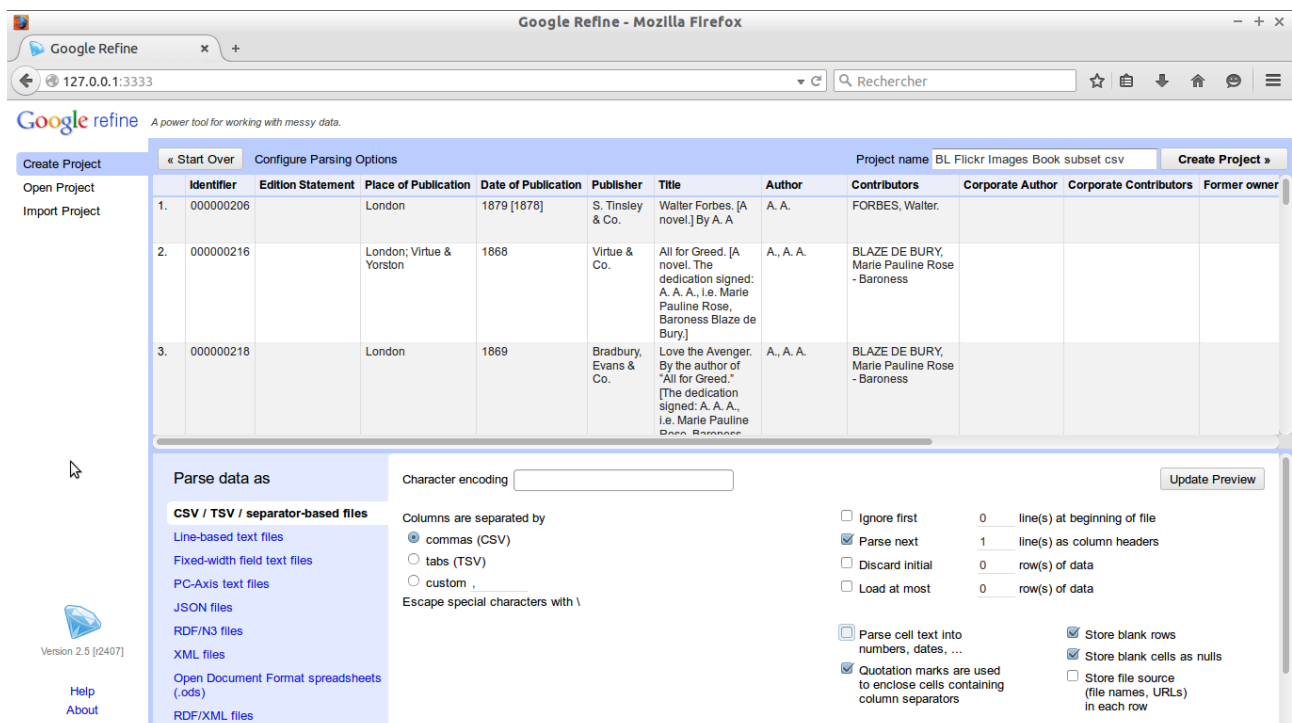
Exercice n°1 : Créer son premier projet OpenRefine (à l'aide du jeu de données fourni)

Il y a plusieurs manières d'entrer vos données dans OpenRefine. On peut les y télécharger dans une grande variété de formats tels que :

- TSV (données séparées par une tabulation)
- CSV (données séparées par une virgule)
- Excel
- JSON (Javascript object notation)
- XML
- Google Spreadsheet

Pour importer vos données, lancer OpenRefine puis :

- Cliquez sur Créer un projet (« *Create Project* »)
- Choisir Télécharger des données depuis cet ordinateur (« *Get data from this computer* »)
- Cliquez sur Choisir un ou des fichiers (« *Choose files* »)
- Choisir le fichier « *BL-Flickr-Images-Book-subset.csv* » (vous pouvez le télécharger depuis http://www.meanboyfriend.com/overdue_ideas/wp-content/uploads/2015/02/BL-Flickr-Images-Book-subset.csv)
- Cliquez sur Suivant (« *Next* »)



Ce premier écran vous présente quelques options qui vous assurent que le jeu de données sera importé dans OpenRefine correctement. Ces options varient en fonction du format de données que

vous importez.

Au cas présent, assurez-vous que :

- Vous choisissez le format de caractères UTF-8
- La première ligne des données sert d'en-tête à vos données (« *First row as column heading* »)
- OpenRefine ne cherche pas à détecter les nombres et les dates automatiquement.

Une fois ceci fait, cliquez sur Créer le projet (« *Create Project* »). Cette action va créer le projet et l'ouvrir pour vous.

Les projets sont sauvegardés tout au long de vos actions, il n'est nul besoin de procéder à des sauvegardes manuelles. Il vous est possible d'ouvrir un ancien projet en cliquant sur Ouvrir un projet (« *Open project* ») sur la page d'accueil d'OpenRefine, à gauche.

Pour aller plus loin...

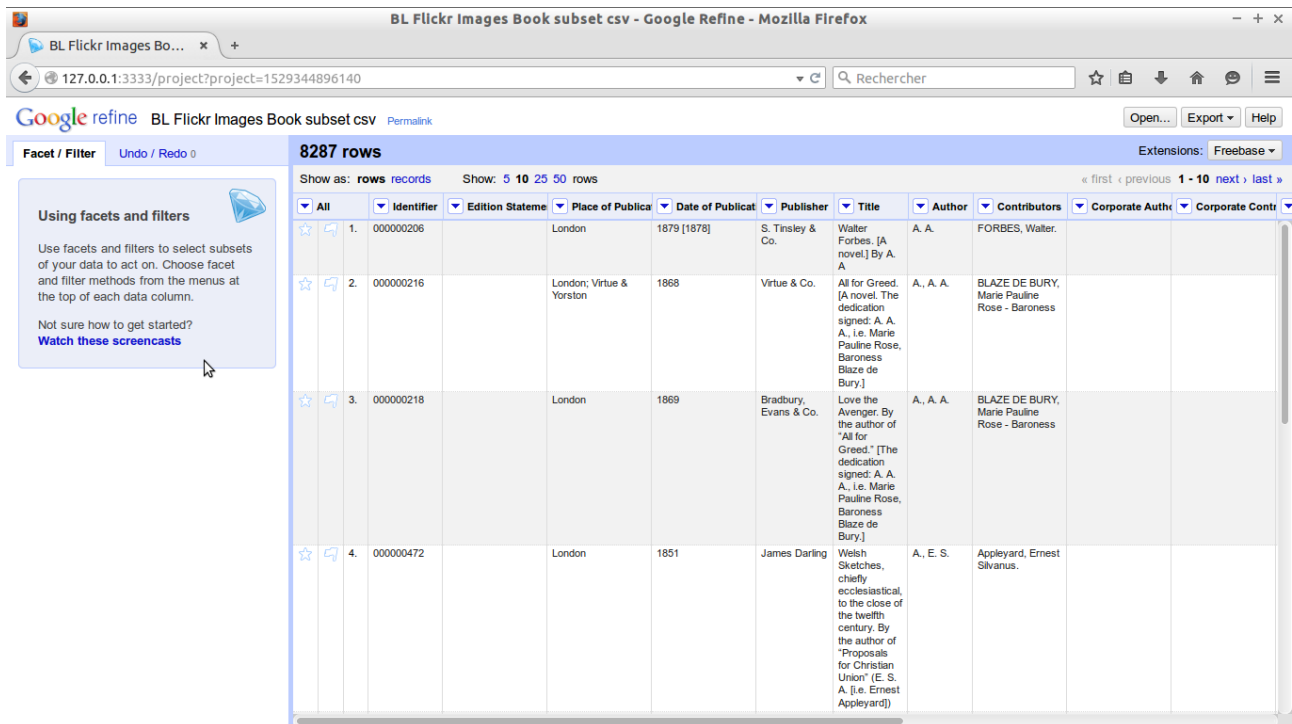
Regardez les autres options disponibles à la création de projets. Essayez de changer ces options, et de valider l'aperçu en cliquant sur Mettre à jour l'affichage (« *Update preview* »). Voyez comment vos données apparaissent alors.

Avez-vous accès à des données JSON, XML ? Dans ce cas, OpenRefine vous demandera un répertoire d'enregistrement, qui comprendra les morceaux de fichiers composant les lignes de votre tableau dans le projet OpenRefine.

2 - Les fonctions basiques d'OpenRefine

L'apparence d'OpenRefine

OpenRefine affiche les données sous forme tabulaire. Chaque ligne représente un « enregistrement » d'une donnée et chaque colonne représente un type d'information. Il s'agit là d'une vue qui s'apparente beaucoup à celle que vous pourriez avoir dans une feuille de calcul sous LibreOffice, ou dans une base de données.




Réordonner les colonnes, et les renommer

Beaucoup d'opérations dans OpenRefine sont accessibles via des menus déroulants situés en haut de chaque colonne.

Il est possible de modifier l'ordre des colonnes de manière globale, en cliquant sur le menu de la première colonne intitulée Tout (« All »), puis Editier les colonnes (« Edit columns ») et ré-ordonner (« Re-order / remove columns »).

Facet / Filter
Undo / Redo 0

Using facets and filters



Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? [Watch these screencasts](#)

8287 rows

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

▼ All	▼ Identifier	▼ Edition Stateme	▼ Place of Publica	▼ Date of Publicat
Facet ▶	06		London	1879 [1878]
Edit rows ▶				
Edit columns ▶	Re-order / remove columns...			
View ▶			Yorston	1868
☆	3.	000000218	London	1869

Il est alors possible de modifier cet ordre, ou de supprimer des colonnes par simple glisser/déposer.

book subset csv [Permalink](#)

8287 rows

Show as: rows records

▼ All	▼ Identifier	▼ Edition Stateme	▼ Place of Publica	▼ Date of Publicat
☆	1.			
☆	2.			
☆	3.			
☆	4.			

Re-order / Remove Columns

Drag columns to re-order

Drop columns here to remove

Identifier
Edition Statement
Place of Publication
Date of Publication
Publisher
Title
Author
Contributors
Corporate Author
Corporate Contributors
Former owner
Engraver
Issuance type
Flickr URL
Shelfmarks

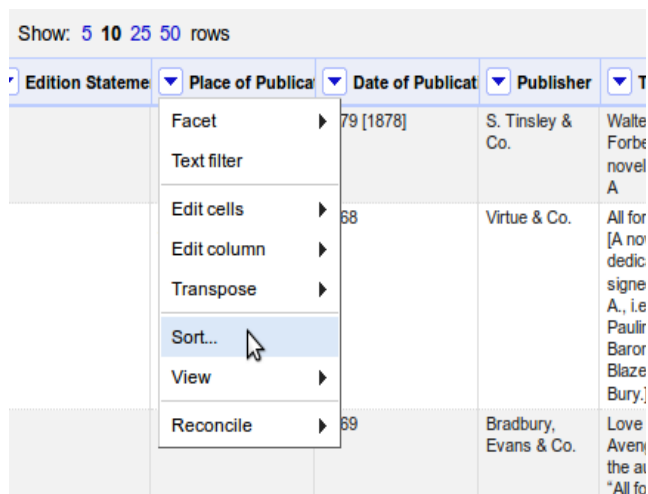
OK Cancel

Astuce :

Les performances d'OpenRefine sont directement liées au nombre de colonnes de votre projet. Il ne faut jamais hésiter à supprimer des colonnes devenues inutiles.

Trier les données

Pour trier vos données avec OpenRefine, il suffit de cliquer sur le menu déroulant de la colonne que vous désirez trier, et de choisir l'option Trier (« Sort »).



Une fois les données triées, un nouveau menu « Sort » apparaît alors à l'écran.



Contrairement à ce qui peut se passer avec Excel ou LibreOffice, le tri d'OpenRefine est temporaire et ne modifie pas vos données. Si vous désactivez ce tri, elle reviennent à leur état initial. Le bouton « Sort » qui vient d'apparaître sert alors à rendre ce tri permanent, ou à l'inverser, ou bien

encore à l'annuler.

Il est bien entendu possible de trier le tableau sur plusieurs colonnes à la fois.

Exercice 2 : Réordonner les colonnes et trier les données

- Trouvez la colonne « *Date of publication* » et triez-la par date de publication
- Placez la colonne de titre en seconde position, juste après la colonne « *Identifier* ».

Facettes

Les facettes (« *Facets* » en anglais) sont une des fonctions les plus utiles d'OpenRefine et peuvent à la fois vous donner une meilleure vue de vos données mais également vous aider à améliorer la consistance de celles-ci.

Une facette regroupe toutes les valeurs qui apparaissent dans une colonne et vous permet de trier les données par valeur et éditer celles-ci sur une grande quantité d'enregistrements (ou lignes) d'un seul coup.

La facette la plus simple est la facette de texte (« *Text Facet* »). Cette facette regroupe toutes les valeurs de texte situées dans une colonne, et les affiche avec le nombre d'occurrences correspondantes. Cette information apparaît toujours dans le panneau d'affichage situé à gauche de l'interface du logiciel.

The screenshot shows the Google Refine web interface in a Mozilla Firefox browser. The address bar shows the URL: 127.0.0.1:3333/project?project=1529344896140. The page title is "BL Flickr Images Book subset csv - Google Refine - Mozilla Firefox". The interface includes a search bar with the text "Rechercher". Below the search bar, there are tabs for "Facet / Filter" and "Undo / Redo 0". The main area displays a table with 8287 rows. The table has columns: Identifier, Edition Statement, Place of Publication, Date of Publication, Publisher, Title, Author, Contributors, Corporate Author, and Corporate Contributor. The first four rows of data are visible. A sidebar on the left contains a "Facet / Filter" panel with a "Using facets and filters" section. The sidebar also includes a "Watch these screencasts" link.

	Identifier	Edition Statement	Place of Publication	Date of Publication	Publisher	Title	Author	Contributors	Corporate Author	Corporate Contributor
1.	000000206		London	1879 [1878]	S. Tinsley & Co.	Walter Forbes. [A novel.] By A. A.	A. A.	FORBES, Walter.		
2.	000000216		London; Virtue & Yorston	1868	Virtue & Co.	All for Greed. [A novel. The dedication signed: A. A. A. i.e. Marie Pauline Rose, Baroness Blaze de Bury.]	A., A. A.	BLAZE DE BURY, Marie Pauline Rose - Baroness		
3.	000000218		London	1869	Bradbury, Evans & Co.	Love the Avenger. By the author of "All for Greed." [The dedication signed: A. A. A. i.e. Marie Pauline Rose, Baroness Blaze de Bury.]	A., A. A.	BLAZE DE BURY, Marie Pauline Rose - Baroness		
4.	000000472		London	1851	James Darling	Welsh Sketches, chiefly ecclesiastical, to the close of the twelfth century. By the author of "Proposals for Christian Union" (E. S. A. [i.e. Ernest Appleyard])	A., E. S.	Appleyard, Ernest Silvanus.		

Pour créer une facette dans une colonne, cliquez sur le menu situé en haut de celle-ci, et choisissez Facette (« *Facet* ») puis Facette de texte (« *Text Facet* »). La facette va apparaître à gauche.

A titre d'exemple, la copie d'écran ci-après montre une facette réalisée sur la colonne « *Issuance type* », qui contient, on le voit, deux valeurs « *Continuing* » (19 enregistrements) et « *Monographic* » (8268 enregistrements).

The screenshot shows the OpenRefine web interface in a Mozilla Firefox browser. The title bar indicates the project is 'BL Flickr Images Book subset csv'. The address bar shows a local URL. The interface includes a search bar, a list of recent projects, and a 'Google refine' logo. The main area displays a table with 8287 rows. A facet is applied to the 'Issuance type' column, showing two choices: 'continuing' (19) and 'monographic' (8268). The table columns are: Title, Author, Contributors, Corporate Author, Corporate Contributor, Former owner, Engraver, Issuance type, Flickr URL, and Shelfmarks. The first row shows a title 'Brieven uit Engeland aan mijn vriend Z. Door J. W. d. C. (J. W. del Campo.) (1e brief. De landbouw in Norfolk.-Tweede brief. De hoofdstad Londen en het Kristallen Paleis te Sydenham.-Derde brief. De Engelsche landbouw.)', author 'CAMPO, J. W. d.', contributors 'C., J. W. d.', and issuance type 'monographic'. The second row shows a title 'Het Jaar 1566. Eine historische proeve uit den Nederlandschen vrijheidsoorlog. Meerdereels naar onuitgegeven bescheiden bewerkt door M. L. van Deventer. Met een voorwoord van Dr. R. C. Bakhuizen van den Brink', author 'DEVENTER, Marinus Lodewijk van.', contributors 'BRINK, Reinier Cornelis Bakhuizen van den.', and issuance type 'monographic'. The third row shows a title 'Journaal van A. D., ... 1591-1602 (van 't gene doegdelijkx gepasseert is in den oorlog der Staeten Generael tegens de Spangarden, etc.) Uitgegeven ... met inleiding en ...', author 'DUYCK, Anthonis - Advokaat-', contributors 'Fiskaal van den Raad van State', and issuance type 'monographic'.

Il est possible d'afficher plusieurs valeurs de la facette en une seule fois, en cliquant sur l'option « *include* » qui apparaît au survol de la souris sur une des valeurs de la facette.

Il est également possible d'inverser le filtre pour afficher les valeurs qui ne correspondent pas à la valeur souhaitée. Cette option apparaît en haut du panneau de la facette, lorsque vous sélectionnez une valeur de filtre.

Filtres

Si on peut filtrer ses données en utilisant les facettes d'OpenRefine, il est également possible de visualiser ses données en utilisant la fonction de filtres de texte en cliquant sur « *Text filters* », en haut d'une colonne. Dans ce cas il sera possible d'afficher les enregistrements comportant dans cette colonne, la valeur qui vous intéresse. Il vous suffit d'entrer cette valeur dans la boîte de dialogue qui apparaît alors à gauche dans OpenRefine.

Google Refine interface showing 4219 matching rows (8287 total). The table is filtered by 'Place of Publication' set to 'London'. The table columns are: All, Identifier, Edition Statement, Place of Publication, Date of Publication, Publisher, Title, Author, Contributors, Corporate Author, and Corporate Contributor.

All	Identifier	Edition Statement	Place of Publication	Date of Publication	Publisher	Title	Author	Contributors	Corporate Author	Corporate Contributor
	2051. 000906365] London	[1872.]	[Printed for Private Circulation]	The Exile of Calauria; or, the Last days of Demosthenes. [A drama, in verse, by Stratford Canning, Viscount Stratford de Redcliffe.] MS. corrections [by the author:] General Appendix		Stratford de Redcliffe, Stratford Canning - Viscount		
	4931. 0024372 edit] London	1874	[Privately printed]	Mont Blanc. A comedy, in three acts [and in prose] ... Part of the plot ... derived from "Le Voyage de M. Perrichon," by ... E. Labiche et E. Martin	Mayhew, Henry	MAYHEW, Athol.		
	3910. 001907138		1. pp. 32. London	1700		A Journey to Hell; or, a visit paid to the Devil: a poem. [By Edward Ward.]		WARD, Edward - Author of the "London Spy."		
	4026. 001965006		200. London	1828		Three days at Killarney, with other Poems. (Notes.) [By C. Hoyle.]		HOYLE, Charles.		

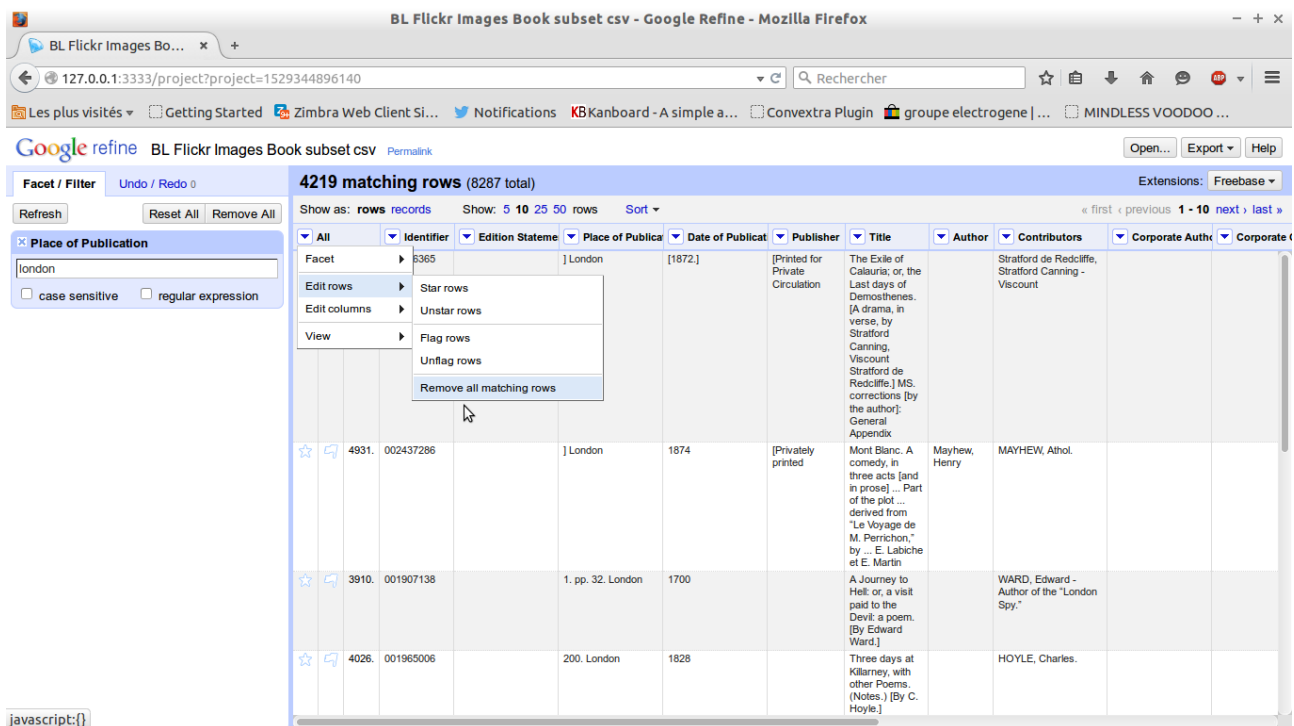
Travailler sur les données filtrées

Il est très important de comprendre ici qu'une fois que vous aurez filtré des données, TOUTES les opérations de modification que vous vous effectuerez ne s'appliqueront QUE sur les enregistrements filtrés, c'est à dire ceux qui sont affichés sur l'écran.

A titre d'exemple, si vous souhaitez éliminer des enregistrements qui correspondent à un filtre de tri spécifique, vous effectuerez les opérations suivantes :

- Filtrer les données à l'aide d'une facette ou d'un filtre
- Cliquer sur le menu déroulant de la colonne Tous (« All ») (qui est toujours la première)
- Choisir Edit rows (« Edit rows ») puis Eliminer toutes les lignes correspondantes (« Remove all matching rows »)

Ceci aura effectivement pour effet de supprimer toutes les lignes répondant à la condition du filtre, et uniquement celles-ci.



Exercice 3 : Supprimer toutes les publications répondant au critère « *Continuing* » du jeu de données

- Créez une facette sur la colonne « *Issuance type* »
- Filtrer les données pour n'afficher que celles répondant au critère « *continuing* »
- Supprimez ces enregistrements

Pour aller plus loin...

Créez une facette texte sur la colonne « *Date of publication* ». Quel est le problème avec cette facette ?

En utilisant cette facette, comment est-il possible d'identifier la valeur la plus fréquente dans cette colonne.

Qu'arrive-t-il à cette facette lorsque qu'on applique un filtre de texte sur la colonne « *Place of publication* » ?

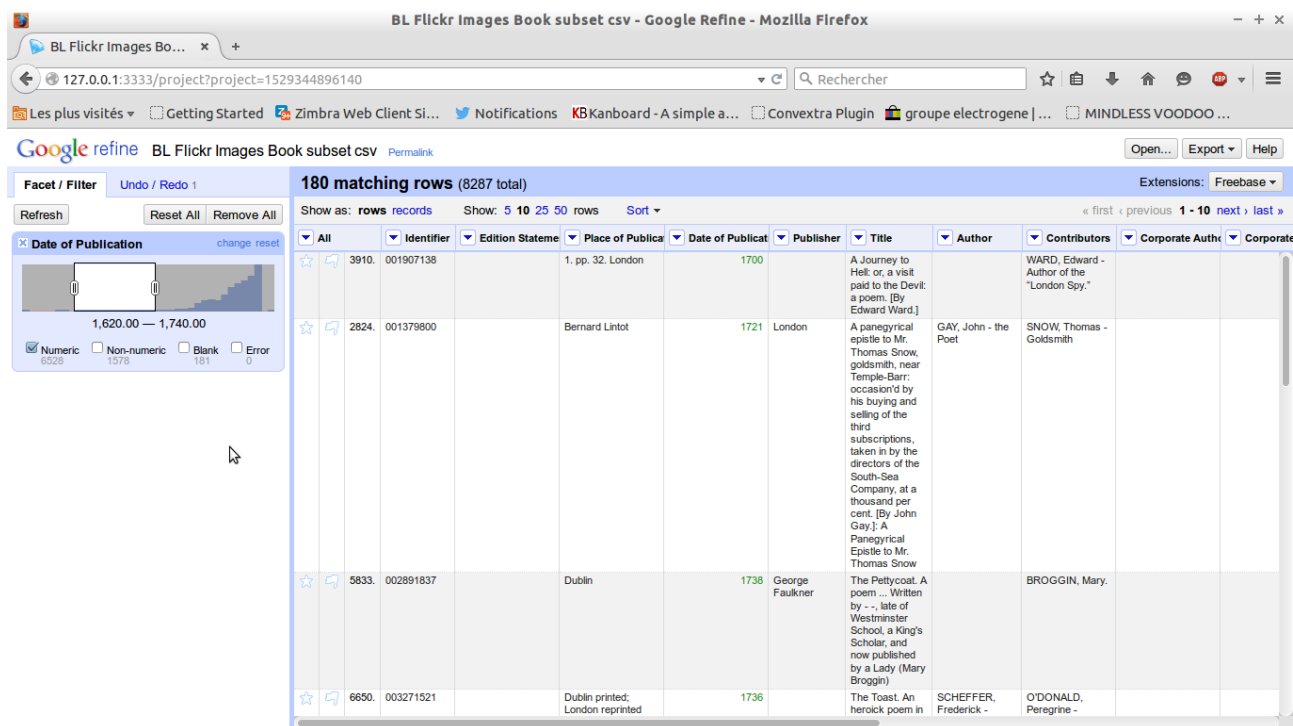
D'autres types de facettes

OpenRefine, en plus des facettes de texte, propose d'autres types de facettes :

- Facettes numériques (« *Numeric facets* »)
- Facettes temporelles (ou par date) (« *Timeline facets* »)

- Facettes personnalisées (« *Custom facets* »)
- Facette par nuage de points (« *Scatterplot facets* »)

Les facettes numériques et temporelles affichent à gauche des graphes de valeurs en lieu et place de listes de valeurs. Ces graphes utilisent des curseurs de type glisser/déposer que vous pouvez utiliser pour déterminer et affiner précisément l'écart de valeurs qui vous intéresse.



On utilise beaucoup plus rarement la facette par nuage de points et pour plus d'informations, nous vous recommandons la lecture de ce tutoriel :

http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial#Exploring_the_data_with_scatter_plots

Les facettes personnalisées sont une variété plus large de facettes incluant des facettes dont VOUS écrivez la structure.

Par défaut on trouve dans cette catégorie les facettes suivantes :

- Facette par mots (« *word facet* ») : cette facette découpe le texte contenu dans un enregistrement en mots et compte le nombre de ces mots.
- Facette par doublons (« *duplicate facets* ») : le résultat de cette facette est de type binaire vrai ou faux (« *true* » ou « *false* »). Les lignes sont marquées « vrai » (« *true* ») si la valeur de la colonne sur cette ligne est le doublon exact dans la même colonne sur une autre ligne.
- Facette par longueur de texte (« *text length facet* ») : crée une facette numérique basée sur la longueur du texte contenu dans chaque ligne d'une colonne donnée. Cette facette est notamment très utile pour marquer une donnée incorrecte ou inattendue, là où une longueur spécifique est attendue (ex : un code postal à 6 chiffres au lieu de 5 en France)
- Facette par vide (« *facet by blank* ») : le résultat de cette facette est de type binaire vrai ou faux (« *true* » ou « *false* »). Les lignes sont marquées « vrai » (« *true* ») si la valeur de la colonne sur cette ligne est vide. C'est très utile pour repérer des lignes avec des valeurs

manquantes.

Les facettes sont utilisées pour grouper des enregistrements par valeur commune et OpenRefine limite le nombre de valeurs autorisées par facette pour ne pas affecter le fonctionnement de l'outil ou éviter un dépassement de mémoire. Par exemple, si vous créez une facette avec trop de valeurs uniques (comme une facette sur la colonne «Book title » dans le fichier de démonstration, qui ne contient que des titres uniques), celle-ci sera très grande et pourra le cas échéant ralentir OpenRefine qui refusera alors de la créer.

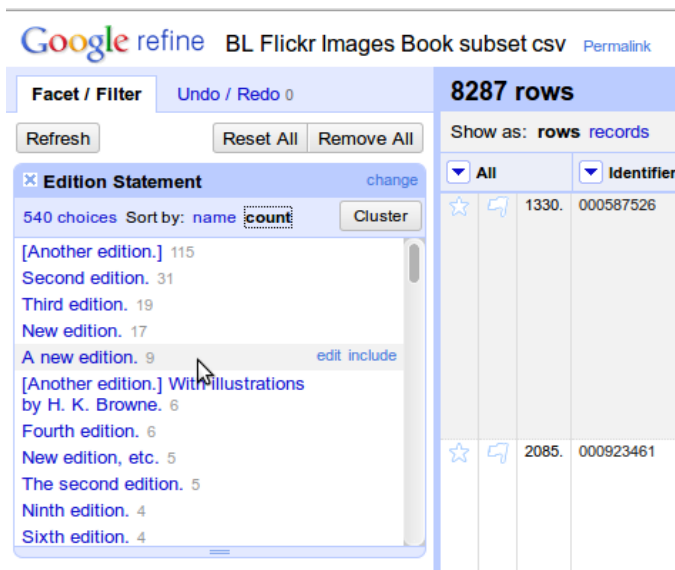
Exercice 4 : repérer toutes les publications qui n'ont pas de date de publication

- Utilisez la fonction Facette par vide (« *facet by blank* ») pour repérer toutes les publications qui n'ont pas de date de publication.
- Filtrez les données pour n'afficher que ces publications. Avez-vous remarqué quelque chose sur ces données ? (*astuce* : regardez sur la colonne « *Place of publication* »)

Modifier les données via les facettes

Si vous créez une facette de texte, vous pouvez alors éditer les valeurs d'enregistrement à l'intérieur de celle-ci et modifier ainsi un grand nombre de valeurs en une seule opération.

Pour faire ceci, survolez la valeur que vous souhaitez modifier avec la souris et cliquez sur le bouton Editer (« *edit* ») qui apparaît alors.



Cette approche, intéressante, est valable et pratique sur des facettes de petite taille, où vous rencontrerez des variations faibles (ponctuation, erreurs de typo, etc.). On citera par exemple une colonne devant contenir des termes d'une liste précise (jours de la semaine, mois...).

Dès que vous ferez une modification sur une valeur, la facette se mettra automatiquement à jour.

Exercice 5 : Éditer la colonne « *edition statement* » via une facette texte.

- Créez une facette de texte sur la colonne « *Edition statement* »
- Triez cette facette par décompte de valeur (« *sort by count* »), pour repérer les plus fréquentes
- Choisissez les valeurs se référant à la notion de « *second edition* » et modifiez les en utilisant un terme homogène.

Utiliser l'outil d'agrégation (« *clustering* ») pour repérer des valeurs identiques

Une autre fonction très pratique est proposée par le biais des facettes : la fonction d'agrégation (« *cluster* »). Cette fonction recherche des valeurs semblables dans la facette et vous permet de les fusionner à volonté en une seule valeur.

Cette fonction est redoutablement efficace lorsque vous avez des données avec des variations mineures, représentant des valeurs qui sont sûrement identiques. Exemple : « LYON SARL », « SARL LYON », « S.A.R.L. LYON » et S.A.R.L LY ON », sont certainement identiques à « SARL LYON ».

Cluster & Edit column "Publisher"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method **key collision** Keying Function **fingerprint** 38 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
5	32	<ul style="list-style-type: none">Printed for private circulation (25 rows)[Printed for private circulation] (3 rows)[Printed for private circulation] (2 rows)[Printed for Private Circulation] (1 rows)printed for private circulation (1 rows)	<input type="checkbox"/>	Printed for private circulation
4	63	<ul style="list-style-type: none">Privately printed (53 rows)[Privately printed] (4 rows)[Privately printed] (4 rows)Privately printed. (2 rows)	<input type="checkbox"/>	Privately printed
4	50	<ul style="list-style-type: none">Smith, Elder & Co. (47 rows)Smith Elder & Co. (1 rows)Smith, Elder Co. (1 rows)Smith, Elder, & Co. (1 rows)	<input type="checkbox"/>	Smith, Elder & Co.
3	88	<ul style="list-style-type: none">Macmillan & Co. (85 rows)Macmillan Co. (2 rows)MacMillan & Co. (1 rows)	<input type="checkbox"/>	Macmillan & Co.

Choices In Cluster
2 — 5

Rows In Cluster
2 — 88

Average Length of Choices
6 — 33

Length Variance of Choices
0 — 6

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

Les agrégats (« *clusters* ») sont créés automatiquement grâce à des algorithmes mathématiques. Plusieurs de ces algorithmes sont intégrés dans OpenRefine. Vous constaterez par expérience que certains algorithmes sont plus ou moins bien adaptés à votre jeu de données et donneront de plus ou moins bon résultats.

Pour obtenir plus d'informations sur ces algorithmes, et la méthode employée pour les créer, vous pouvez vous référer à cette page :

<https://www.github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth>

Pour chacun de ces agrégats, il vous est proposé l'option de fusion (« *merge* ») des valeurs, ce qui revient à remplacer l'ensemble de celles-ci par une seule valeur cohérente. Par défaut, OpenRefine utilise la valeur avec le plus grand nombre d'occurrences comme valeur de référence, mais il vous est également possible d'entrer vous-même la valeur désirée dans le champs Nouvelle valeur de cellule (« *New Cell Value* »).

Enfin, il est également possible d'accéder à l'outil d'agrégation directement depuis le menu déroulant de chaque colonne, en cliquant sur Editer les cellules (« *Edit cells* ») puis Agréger et éditer (« *Cluster and edit...* »).

Exercice 6 : Utiliser l'outil d'agrégation pour nettoyer les données de la colonne « *Publisher* »

- Créez une facette de texte sur la colonne « *Publisher* »
- Cliquez sur Agrégat (« *cluster* »)
- En utilisant la méthode « *key collision* » couplée à la fonction de clef « *fingerprint* », retravaillez les agrégats de valeurs en les fusionnant et en leur attribuant une valeur unique et cohérente.

Pour aller plus loin...

Expérimentez d'autres méthodes d'agrégation. Quelle est selon vous la plus efficace pour notre exemple ? Quel est l'inconvénient d'utiliser la méthode « *nearest neighbour* » avec notre jeu de données ?

Essayez d'utiliser l'outil d'agrégation avec d'autres colonnes comme « *Place of publication* » ou « *Date of Publication* ».

Introduction aux transformations

Par le biais des facettes, des filtres et des agrégats, OpenRefine offre une manière directe et simple pour visualiser rapidement vos données et effectuer des modifications dessus en utilisant des valeurs standardisées.

Toutefois, il est des cas fréquents où ces outils trouveront leurs limites et vous ne pourrez pas aller au bout de votre démarche à l'aide de ces seules options. On peut citer par exemple :

- Séparer des données contenues dans une colonne en plusieurs colonnes (ex : découper des adresses en plusieurs entités....)

- Standardiser le format de vos données sans en changer la valeur (en retirant la ponctuation, ou en modifiant le format d'une date, par exemple)
- Extraire une donnée spécifique d'une chaîne de caractères plus longue (ex : un numéro ISBN, un numéro de compte bancaire...)

Pour réussir cela, OpenRefine offre des fonctionnalités appelées « *Transformations* », qui ne sont rien d'autres que des manières de manipuler des données dans des colonnes. Ces transformations sont le plus souvent écrites dans un langage appelé « *GREL* » (« *Google Refine Expression Language* »). Ce langage est assez proche de celui que l'on peut trouver dans Excel (*Excel formula*), même si GREL est un peu plus orienté vers la manipulation de chaînes de texte, et moins de valeurs numériques.

La documentation complète de GREL est disponible en ligne à cette adresse :

<https://github.com/OpenRefine/OpenRefine/wiki/Google-refine-expression-language>

Le présent tutoriel ne couvre qu'une toute petite partie des innombrables fonctions de GREL.

Pour commencer à écrire une transformation, sélectionnez une colonne et cliquez sur *Edit cells* (« *Edit cells* ») puis *Transformer* (« *Transform ...*») et la fenêtre suivante va apparaître :

Custom text transform on column Date of Publication

Expression Language Google Refine Expression Language (GREL)

No syntax error.

Preview History Starred Help

row	value	value
1330.	1860-62	1860-62
2085.	1856	1856
2316.	1862-66	1862-66
3212.	1832	1832
3895.	1873	1873
5017.	1860	1860
5001.	1857-60	1857-60
5002.	1857-60	1857-60
5003.	1857-60	1857-60
5004.	1857-60	1857-60

On error ☒ keep original ☐ Re-transform up to times until no change
☐ set to blank
☐ store error

OK Cancel

Sur cet écran, vous constatez que vous disposez d'une zone de texte pour pouvoir écrire une transformation (la zone « *Expression* »), et d'une zone de prévisualisation des 10 premières lignes de votre jeu de données (la zone « *value* », à droite).

Votre transformation doit être une expression GREL valide. L'expression la plus simple de GREL est « *value* », qui recopie la valeur et la remplace par elle-même... sans aucune modification!

Les fonctions de GREL sont écrites en attribuant une valeur (qui peut être une chaîne de caractères, une date, un nombre, etc.) à une fonction GREL. Certaines de ces fonctions utilisent deux paramètres ou deux options pour fonctionner. GREL supporte deux types de syntaxes :

- *value.fonction(options)*
- *fonction(value, options)*

Leur résultat est identique et l'usage dépend uniquement de vos habitudes personnelles!

A côté de la fonction prévisualisation (« *Preview* »), on trouve deux outils intéressants :

- Historique (« *History* ») qui contient la liste de vos précédentes transformations et vous offre la possibilité de les réutiliser ou de les mettre en favori pour un accès encore plus pratique
- Favoris (« *Starred* ») qui reprend ainsi vos fonctions favorites
- Aide (« *Help* »), une liste des fonctions de GREL et une brève explication de comment elles fonctionnent

Quelques transformations simples

Expression GREL	Exemple	Action attendue
toUppercase(string)	toUppercase(value) value.toUppercase()	Convertir la chaîne de caractères en majuscule
toLowercase(string)	toLowercase(value) value.toLowercase()	Convertir la chaîne de caractères en minuscule
toTitlecase(string)	toTitlecase(value) value.toTitlecase()	Première lettre de la chaîne en majuscule, les autres en minuscule
trim(string)	trim(value) value.trim()	Supprimer les espaces et tabulations au début et à la fin d'une chaîne de caractères
substring(string, number from, optional number to)	substring(value, 0, 4) value.substring(0, 4)	Prendre les 4 premiers caractères d'une chaîne
Replace(string, string to find, replacement string)	replace(value, "a", "b") value.substring("a", "b")	Dans une valeur données, remplacer la lettre <i>a</i> par la lettre <i>b</i> .
+	"http://" + value	Concatène « http:// » et la chaîne de caractères

Exercice 7 : Nettoyage d'une date de publication par l'emploi de transformations simples

- Créez une facette sur la colonne « *Date of Publication* »
- Triez-la par nom (« *name* »)
- Quels sont les plus gros problèmes rencontrés dans cette liste ?
- Utilisez la fonction GREL remplacer (« *replace* ») pour supprimer les caractères [,] et ? dans la colonne « *Date of Publication* ».

Pour aller plus loin...

Quels sont les autres difficultés que vous pouvez repérer dans cette colonne ?

Quelle approche pourriez-vous utiliser pour traiter ces problèmes ?

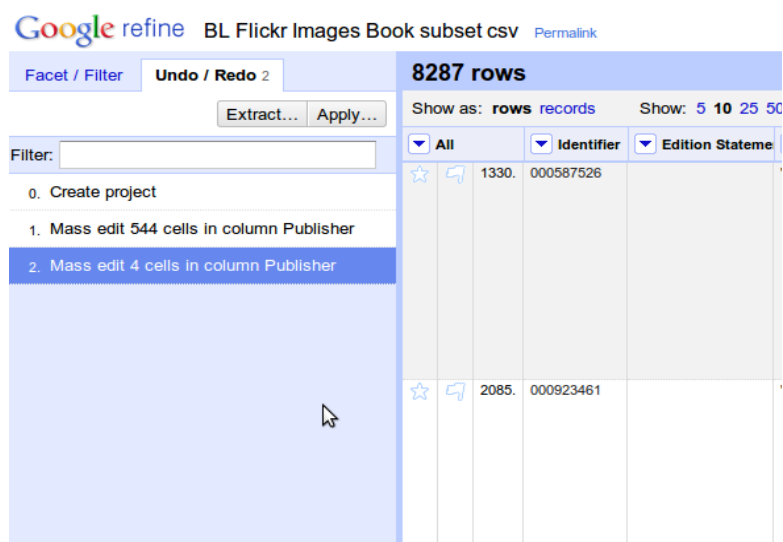
Pensez-vous qu'il est possible de réussir à faire une seule colonne avec une date au format 4 caractères représentant la date de publication ?

Faire et défaire...

OpenRefine vous permet de faire et défaire (« *redo* » et « *undo* ») vos modifications, et ce autant de fois que vous avez effectué d'opérations. Ce qui signifie que vous pouvez à tout moment tenter une transformation, et l'annuler si elle ne vous convient pas ou ne fonctionne pas.

La manière dont OpenRefine enregistre les étapes de transformation d'un de vos jeux de données vous permet même de les réutiliser par simple copier/coller sur un autre jeu de données.

Les options « *Undo* » et « *Redo* » sont accessibles via le panneau gauche.



L'écran *Undo/Redo* vous présente une liste des étapes que vous avez accompli. Pour annuler une étape, il vous suffit de cliquer sur le dernier état que vous souhaitez retrouver.

A ce stade, les étapes suivantes sont grisées mais sont toujours accessibles (on peut encore annuler l'annulation!).

Mais attention. Dès lors que vous opérez une nouvelle transformation, l'état modifié devient alors définitif.

Si vous désirez automatiser certaines transformations, les réutiliser sur un autre projet, cliquez sur le bouton Extraire (« *Extract* »), qui vous permet de sélectionner les étapes que vous souhaitez sauvegarder et de les copier au format texte JSON.

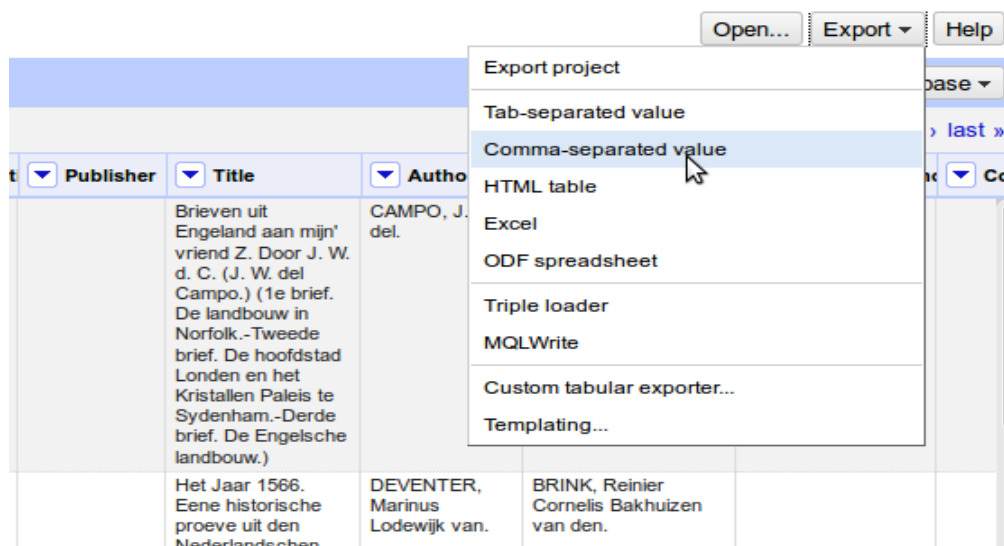
Pour réutiliser ces étapes, copiez-collez celles-ci, et utilisez le bouton Appliquer (« *Apply* »). Vous pourrez ainsi réutiliser votre savoir-faire dans tous vos projets semblables.

Les données contenues dans l'espace *Undo/Redo* (vos étapes) sont stockées dans votre projet OpenRefine et enregistrées automatiquement tout au long de votre travail, donc la prochaine fois que vous réutiliserez OpenRefine, vous pourrez accéder à l'ensemble de l'historique de vos étapes, et faire un *Undo/Redo* de la même manière!

Exporter vos données

Une fois que vous avez terminé votre nettoyage de données dans OpenRefine, vous désirez peut-être exporter ces données.

Cette option d'exportation se trouve en haut à droite de l'interface, via le bouton Exporter (« *Export* »).



Les formats d'export proposés par défaut sont HTML, Excel (.xls), CSV et TSV.

Mais vous pouvez également écrire votre propre format d'exportation, choisir les champs spécifiques à exporter, ajouter un en-tête (« *header* »), ou un pied-de-page (« *footer* ») et préciser le format exact du fichier.

3 – Types de données et expressions régulières

Comprendre les notions de données et d'expressions régulières vous aidera à écrire des transformations plus complexes mais aussi plus efficace avec GREL.

Types de données dans OpenRefine

Toute donnée dans OpenRefine est typée. Le type le plus commun et le plus souvent rencontré est la chaîne de caractère simple (« *string* »). Mais d'autres sont disponibles et permettent des transformation particulières, ou des transformations d'un type à l'autre :

- Chaîne (« *string* »)
- Nombre (« *number* »)
- Date
- Booléen (« *boolean* »)
- Liste de valeurs (« *array* »)

La signification des trois premiers types est assez évidente, mais les deux suivants méritent une explication complémentaire.

Un booléen dans OpenRefine est une valeur binaire qui peut prendre deux états, vrai ou faux (« *true* » ou « *false* »). On peut utiliser ces valeurs booléennes directement dans une cellule d'OpenRefine, mais cela ne présente qu'un intérêt très relatif et on préférera les utiliser pour des transformations et en tant que composantes d'expressions GREL.

Ainsi, `value.contains("test")` générera un booléen « *true* » ou « *false* » suivant le cas où la valeur *test* se trouve ou non dans la cellule.

Ce type peut donc être utilisé par exemple pour appliquer des transformations si une condition est ou non remplie : « si la valeur est « *true* » alors.... Sinon »

Une liste de valeur (« *array* »), est représentée dans OpenRefine par l'usage de crochets droits contenant une liste de valeurs, entourées par des "" et séparés par des virgules. Ainsi, une liste complète des jours de la semaine s'écrit :

```
["Lundi","Mardi","Mercredi","Jeudi","Vendredi","Samedi","Dimanche"]
```

Ces listes peuvent être triées, dédoublonnées, et manipulées de bien d'autres façons dans OpenRefine mais ne peuvent apparaître tel que dans une cellule d'OpenRefine. Ces listes sont donc le plus souvent le résultat de transformations.

Par exemple, la fonction Séparer (« *split* ») découpe une chaîne et la transforme en liste de valeurs séparées par un séparateur.

Pour mieux comprendre, imaginons la chaîne suivante contenue dans une cellule :

```
"Lundi,Mardi,Mercredi,Jeudi,Vendredi,Samedi,Dimanche"
```

Si on applique la fonction ***value.split(",")*** sur cette cellule, on obtiendra le résultat suivant :

```
["Lundi","Mardi","Mercredi","Jeudi","Vendredi","Samedi","Dimanche"]
```

Ceci peut bien évidemment être combiné avec d'autres transformations d'OpenRefine comme le tri par exemple.

Si vous réalisez un **`value.split(",").sort()`** sur la même cellule, vous obtiendrez alors le résultat suivant :

```
["Dimanche","Jeudi","Lundi","Mardi","Mercredi","Samedi","Vendredi"]
```

Si vous souhaitez extraire une valeur particulière de cette liste, il suffit de préciser sa position dans la liste (en commençant par 0) :

`value.split(",")[0]`

pour l'exemple "Lundi,Mardi,Mercredi,Jeudi,Vendredi,Samedi,Dimanche", donnera le résultat :

"Lundi"

Vous pouvez même imagine vouloir concaténer et trier des éléments d'une liste de valeur pour en faire une chaîne.

L'expression GREL sera alors :

`value.split(",").sort().join(",")`

et donnera une liste des jours de la semaine, classés par ordre alphabétique et séparés par une virgule :

"Dimanche,Jeudi,Lundi,Mardi,Mercredi,Samedi,Vendredi"

Expressions régulières

Une expression régulière (« *regex* ») est une manière de représenter des schémas types dans une chaîne de caractères. Elles sont utilisées pour rechercher du texte qui correspond spécifiquement à ce schéma type décrit dans la « *regex* ». Ces « *regex* » sont précédées et suivies du signe « / ». Pour écrire ces « *regex* », vous devez connaître la syntaxe spécifique utilisée pour représenter les différents types de caractères qu'on rencontre dans une chaîne. Dans le tableau ci-dessous, vous trouverez les plus souvent utilisées.

Type de caractères	Syntaxe de la « <i>regex</i> »	Explication
N'importe quel caractère	.	Une <i>wildcard</i> .
Une liste ou un écart de caractères	[<liste/écart de caractère>]	<p>Vous pouvez mettre entre crochets une liste ou un écart de caractères pour retrouver n'importe lequel de ceux-ci :</p> <p>[ABC] retrouve toutes les valeurs contenant A, B ou C (notez qu'il faut respecter la casse).</p> <p>[A-Z] retrouve toutes les valeurs contenant une lettre en majuscule</p> <p>[A-Za-z0-9] retrouve tous les valeurs contenant une minuscule, ou une majuscule ou un chiffre.</p>
N'importe quel chiffre	\d	\d est l'équivalent de [0-9] et retrouve toutes les valeurs contenant un chiffre.
N'importe quel caractère partie d'un mot	\w	\w est l'équivalent de [A-Za-z0-9_] et retrouve tous les valeurs contenant une minuscule, ou une majuscule ou un chiffre ou un <i>underscore</i> .
Tout caractère invisible	\s	\s retrouve toutes les valeurs contenant un espace, une tabulation ou un retour chariot par exemple.
Qui commence par	^	^ retrouve les valeurs qui commencent par un schéma particulier
Qui finit par	\$	\$ retrouve les valeurs qui finissent par un schéma particulier

Ces caractères spéciaux peuvent se combiner avec n'importe quel caractère et forment alors une expression régulière. Disons que nous souhaitons trouver les valeurs avec la graphie s ou z de « *organise* » et « *organize* » nous pourrions utiliser :

/organi.e/

Ici le point peut représenter n'importe quel caractère, donc nous retrouverions « *organise* » et « *organize* » mais aussi « *organite* » et « *organile* ».

Nous devons donc être un peu plus précis en écrivant :

/organi[sz]e/

On peut également ajouter à ces caractères spéciaux des opérateurs de répétition qui précisent à la commande combien de fois le schéma est susceptible de se répéter. Ces opérateurs de répétition s'appliquent au caractère ou à l'expression qui les précèdent immédiatement.

Ces opérateurs sont les suivants :

Opérateur	Signification	Explication/Exemple
*	Le caractère ou l'expression qui précèdent peuvent être répétés X fois (ou aucune)	L'expression régulière <code>./</code> représente donc n'importe quel chaîne de caractère
+	Le caractère ou l'expression qui précèdent peuvent être répétés une ou X fois	Contrairement à « <code>*</code> », utiliser « <code>+</code> » indique que le caractère ou l'expression doivent apparaître au moins une fois <code>/porte\s+/monnaie/</code> retrouverait « porte monnaie » (un espace), et « porte monnaie » (deux espaces), mais pas « porte-monnaie ».
?	Le caractère ou l'expression qui précèdent peuvent être répétés une ou aucune fois	<code>/colou?r/</code> retrouverait « color » et « coulour », par exemple.

Par ailleurs, il est également possible de spécifier le nombre exact de répétitions souhaité ou un intervalle de répétitions, en utilisant les accolades :

`/a{2}/`

retrouve les valeurs contenant deux fois la lettre « a » (c'est à dire « aa »)

`/a{2,4}/`

Retrouve les valeurs contenant « aa », « aaa » et « aaaa »

Pour aller plus loin...

Rédigez une « *regex* » qui retrouve un nombre à quatre chiffres dans une chaîne.

Rédigez une « *regex* » qui retrouve une date écrite sous la forme « dd-MM-yyyy ».

Rédigez une « *regex* » qui retrouve une date écrite sous la forme « dd-MM-yyyy » ou « dd-MM-yy »

Rédigez une « *regex* » qui retrouve un numéro ISBN à 13 chiffres.

Il existe plusieurs sources de documentation et de nombreux tutoriels sur les expressions régulières et notamment :

- <http://www.regular-expressions.info>
- <http://software-carpentry.org/v4/regex/index.html>
- <http://www.codeproject.com/Articles/939/An-Introduction-to-Regular-Expressions>

- <http://regex.bastardsbook.com>

Utiliser la transformation « *match* » avec les expressions régulières

Une autre fonctionnalité offerte par les expressions régulières est que vous pouvez capturer des parties de la chaîne de caractères correspondante et ensuite les travailler avec OpenRefine, grâce à la fonction Correspondre (« *match* »). « *match* » vous permet donc d'extraire certaines parties caractéristiques d'une valeur en utilisant une expression régulière.

Pour faire ceci, nous allons dire à la fonction quel morceau de chaîne nous intéresse en l'englobant dans des parenthèses au sein de l'expression régulière.

A titre d'illustration, prenons les valeurs suivantes :

pp. 40. G Bryan & Co : Oxford
pp. 64. W. Cann : Plymouth
pp. 92. Heath Cranton : London

Ces données représentent des nombres de pages, suivi d'éditeurs de littérature, et de lieux de publication. Pour triturer ces données, vous pouvez utiliser la fonction « *match* » de cette manière :

value.match(/pp. (\d*).*/)

Cette fonction va retrouver le nombre de pages (qui suit « pp ») sur chaque ligne et le mettre dans une liste de valeur dans OpenRefine ce qui donnera le résultat suivant :

```
["40"]  
["64"]  
["92"]
```

Dans la fonction Correspondre (« *match* »), l'expression régulière utilisée doit correspondre à la chaîne complète recherchée mais seules les parties entre parenthèses sont placées dans la liste de valeurs en sortie. Un exemple plus complexe vous permettra d'appréhender un peu mieux ce concept :

value.match(/pp. (\d*). (.*) : \s*(.*)/)

Cette expression possède trois zones de capture : le nombre de pages, l'éditeur et le lieu de publication, ce qui nous donne la sortie suivante :

```
["40","G Bryan & Co","Oxford"]  
["64","W. Cann","Plymouth"]  
["92","Heath Cranton","London"]
```

Exercice 8 : Extraction des dates de publication dans la colonne « *Place of publication* »

Dans l'exercice 4, vous avez pu noter que dans certains enregistrements, la valeur « *Date of publication* » était nulle et qu'une date se trouvait alors dans la colonne « *Place of publication* ». Cet exercice a pour but de vous faire appliquer tout ce que vous avez appris en manipulant les concepts de facettes, transformations, type de données, et expressions régulières, pour extraire les dates manquantes et les replacer dans la bonne colonne.

- Assurez-vous que vous ne travaillez QUE sur les enregistrements où la valeur « *Date of publication* » est nulle (*Astuce : voir pour cela l'exercice 4*).
- En travaillant sur la colonne « *Place of publication* », ajoutez une colonne basée sur cette colonne (« *Add column based on this column* »)
- Utilisez la fonction Correspondre (« *match* ») avec une expression régulière sur la colonne « *Place of Publication* », pour découvrir les valeurs qui se terminent par 4 chiffres.
 - Quand vous utilisez cette fonction, pensez à utiliser l'outil de capture de groupe, tel que vu ci-dessus
 - Le résultat de cette commande est une liste de valeurs, dont vous devrez extraire le contenu

4 – Utilisation avancée d'OpenRefine

OpenRefine est capable d'enrichir les données de vos projets à l'aide de données disponibles sur le web. A cette fin, on peut opérer de plusieurs manières : en utilisant un service distant par son URL ou en utilisant des données disponibles dans d'autres projets OpenRefine, par exemple.

Ainsi, il est possible de rechercher des noms dans la base de données VIAF (« *Virtual International Authority File* ») et de faire remonter des informations telles que les dates de naissance ou de décès de personnes célèbres. Typiquement, cette opération va se faire en deux étapes : retrouver ces données sur le service distant, puis extraire l'information pertinente qui vous intéresse.

Pour récupérer des données d'une source externe, sélectionnez la colonne qui vous intéresse, puis cliquez sur *Edit column* (« *Edit column* »), puis *Add column by fetching URLs* (« *Add column by fetching URLs* »). Ceci va faire apparaître une fenêtre, qui vous propose de créer une URL, destinée à lancer une requête sur le site qui vous intéresse.

Add column by fetching URLs based on column Author

New column name Throttle delay milliseconds

On error ☒ set to blank ☐ store error

Formulate the URLs to fetch:

Expression Language No syntax error.

Preview History Starred Help

row	value	value
1330.	CAMPO, J. W. del.	CAMPO, J. W. del.
2085.	DEVENTER, Marinus Lodewijk van.	DEVENTER, Marinus Lodewijk van.
2316.	DUYCK, Anthonis - Advokaat-Fiskaal van den Raad van State	DUYCK, Anthonis - Advokaat-Fiskaal van den Raad van State
3212.	null	null
3895.	JONGE, Johan Karel Jakob de.	JONGE, Johan Karel Jakob de.
5017.	MEYLINK, Antonius Alexis Josephus.	MEYLINK, Antonius Alexis Josephus.

OK Cancel

Par exemple si vous avez une liste d'auteurs et que vous souhaitez récupérer des informations issues du VIAF, vous allez donc créer une URL en ajoutant le paramètre de nom à la fin de celle-ci :

"http://viaf.org/viaf/AutoSuggest?query="+escape(value,'url')

(Nous admettons ici que vous travaillez sur une colonne qui contient les noms sur lesquels vous souhaitez travailler.)

Le service distant VIAF adresse en retour à cette requête, une liste possible des positifs au format JSON. Il existe dans GREL, une fonction spéciale pour extraire facilement des données de fichiers JSON, appelée « *parseJSON* ».

Ainsi, l'expression ci-dessous va extraire le prénom et l'identifiant VIAF :

forEach(value.parseJson().result, v, v.term+'|'+v.viafid)[0]

Explication : on attribue la valeur v au résultat, et dans ce résultat, on récupère la partie term (v.term) et la partie VIAFid (v.viafis), séparé par un « | ».

Ce processus de récupération de données sur des sources externes est assez lent et il est préférable de l'utiliser sur de petits jeux de données.

Exercice 9 : Collecter l'identifiant VIAF des auteurs

- Utilisez le menu déroulant sur la colonne « *Date of publication* » et sélectionnez le filtre de texte (« *text filter* »).
- Entrez la valeur 1800 pour limiter le nombre d'enregistrements à traiter à 26
- Utilisez le menu déroulant sur la colonne Auteurs (« *Authors* ») et sélectionnez Editer la colonne (« *Edit column* ») puis Ajouter une colonne liée à l'URL (« *Add column by fetching URLs* »).
- Nommez cette colonne « *VIAF JSON* »
- Dans la section « *Expression* », insérez l'expression suivante :

"http://viaf.org/viaf/AutoSuggest?query= "+escape(value,'url')

- Cliquez sur OK
- Laissez le temps à OpenRefine de récupérer les données pour vous, cela peut prendre quelques minutes.
- Utilisez le menu déroulant sur la colonne « *VIAF JSON* » que vous venez de créer et cliquez sur Cellule (« *Cells* »), puis transformer (« *transform* ») et dans la zone « *Expression* », entrez la transformation suivante :

forEach(value.parseJson().result, v, v.term+'|'+v.viafid)[0]

Vous devriez constater que les cellules ne gardent alors QUE le nom de l'auteur et sont ID, séparés par un |.

Services de réconciliation

Les services de réconciliation vous permettent de récupérer des informations sur vos données OpenRefine depuis des sources externes encore plus facilement qu'avec la méthode ci-dessus.

Mais ces sources nécessitent d'être optimisées pour OpenRefine, ce qui suppose de ne pouvoir utiliser que ce type de sources.

Extensions

Les nombreuses fonctionnalités d'OpenRefine peuvent être encore enrichies avec des extensions qui peuvent être téléchargées et ajoutées à votre installation d'OpenRefine.

Une liste des diverses extensions est disponible à cette adresse :

<https://github.com/OpenRefine/OpenRefine/wiki/Extensions>

Une extension en particulier essaye de combler les difficultés liées aux services de réconciliation évoqués ci-dessus en rendant possible l'interconnexion avec des sources de données utilisant le protocole SPARQL¹. Pour plus d'informations sur le sujet, référez-vous au chapitre « *RDF extension* » sur le site : <http://refine.deri.ie>, et notamment sur leur section « *showcases* ».

Enregistrements et lignes

Tous les exemples que nous avons étudié dans ce fascicule ont été traités en mode Ligne (« *row* »), en partant du principe que chaque ligne dans le tableau correspondait à un enregistrement. Mais OpenRefine est capable de traiter des modèles plus complexes d'enregistrements, qui autorisent notamment l'emploi de cellules à valeurs multiples pour un même enregistrement. A titre d'exemple, il est fréquent pour un livre d'avoir plusieurs auteurs.

Sur la capture d' écran ci-dessous, vous apercevez le sélecteur qui vous permet de passer facilement d'une vue à l'autre : Ligne ou Enregistrement (« *rows* » ou « *records* ») et vous pouvez constater que 11 de ces enregistrements ont par exemple, 2 contributeurs listés.

1 SPARQL est un langage utilisé pour faire des requêtes directement dans une adresse URL, destinées à lier des données locales à des données distantes.

8287 rows										
Show as: rows records Show: 5 10 25 50 rows « first < previous										
All	Identifier	Edition Statement	Place of Publication	Date of Publication	Publisher	Title	Author	Contributors	Corporate	
☆	10.	000001143		London	1679		A Satyr against Vertue. (A poem: supposed to be spoken by a Town-Hector. [By John Oldham. The preface signed: T. A.]	A., T.	OLDHAM, John.	
☆	11.	000001280		Coventry	1802	Printed by J. Turner	An Account of the many and great Loans, Benefactions and Charities, belonging to the City of Coventry ... A new edition. [The dedication signed: AB, CD, EF, GH, &c. By Edward Jackson and Samuel Carte.]		CARTE, Samuel.JACKSON, Edward - Rector of Southam, and CARTE (Samuel)	
☆	12.	000001808		Christiania	1859		Erindringer som Bidrag til Norges Historie fra 1800-1815. Anden Udgave ... Udgivet med nogle Rettelser og Tillæg af	AALL, Jacob.	AALL, J. C.LANGE, Christian Christoph Andreas.	

Cette approche « Enregistrement » est souvent assez utile et peut être utilisée quand deux lignes ou plus dans votre jeu de données représentent ou sont associées à la même entité, et que vous souhaitez fusionner ces lignes en une seule, tout en gardant la cohérence et l'intégrité de l'information.

Pour un complément d'information sur comment créer des enregistrements dans OpenRefine vous pouvez consulter la page suivante :

<http://googlerefine.blogspot.co.uk/2012/06/create-records-in-google-refine.html>

Utiliser la fonction de croisement (« cross »), pour rechercher des données dans un autre projet OpenRefine

De la même manière que nous pouvons rechercher des données sur des sources externes à OpenRefine, il est également possible de croiser des données avec celles issues d'un autre projet OpenRefine sur le même ordinateur. Pour ce faire, nous utiliserons la fonction de croisement (« cross »).

La fonction « cross » prend une valeur dans le projet OpenRefine sur lequel vous travaillez, et recherche cette même valeur dans la colonne d'un autre projet. Si elle trouve une telle valeur, elle récupère sous forme de liste de valeurs, celles associées à la valeur d'origine.

Cela peut servir à comparer des valeurs dans deux projets, mais également à enrichir un projet particulier.

Ex : un premier projet contient la liste des codes postaux d'Île-de-France.

Code postal	Ville
77114	Herme
77115	Blandy
77120	Amillis

Un second projet contient la ville mais pas le code postal.

En utilisant la fonction « cross », vous pourrez aller chercher cette information et compléter automatiquement cette information.