Source

Cette session a été conçue à l'aide de contenus de The Carpentries et de contenus provenant d'ateliers offerts par l'Université de Victoria. Vous pouvez accéder aux documents complets ici:

- Lien Data Carpentries : https://datacarpentr y .org/openrefine-socialsci/
- Lien Library Carpentry: htt p s://librarycarpentry.org/lc-open-refine/
- Lien vers l'Université de Victoria : https://docs.google.com/document/d/1sMO-KwmD3GSwrYuNyFGr0zV0zsxh_kkTJSAuEmmmMTA/edit

OpenRefine au cas où vous voudriez lire: https://docs.openrefine.org/

GREL pratique : https://guides.library.illinois.edu/openrefine/grel

Horaire et objectifs de la session

L'objectif principal de cet atelier est de vous guider à travers certaines des fonctionnalités clés d'OpenRefine les plus susceptibles d'être utilisées dans votre travail en bibliothèque. Dans cet atelier, vous apprendrez à :

- Démarrer un projet
- Utiliser OpenRefine pour nettoyer et organiser vos données en :
    - Utilisant les facettes et le clustering pour analyser les termes de normalisation des données
    - Utilisant des transformations courantes via GREL pour décaler, diviser et modifier vos données par lots
- Exporter votre projet dans le format souhaité

Trucs et astuces OpenRefine

Trucs et astuces dans OpenRefine

Indicateurs et étoiles

Les indicateurs (***flags***) et les étoiles (***stars***) sont d'excellents moyens de créer un sous-ensemble de données avec lequel travailler. Lorsque vous ajoutez un indicateur ou une étoile à vos données, OpenRefine traite les éléments comme une facette (***facet***), ce qui vous permet de travailler avec ces informations à l'aide des fonctions de facettes. Vous pouvez marquer manuellement n'importe quelle ligne en cliquant sur l'icône en

forme de drapeau ou d'étoile à gauche de l'écran. Vous pouvez également marquer plusieurs lignes d'un coup ou marquer de nouvelles lignes correspondant à vos lignes déjà marquées en utilisant le menu déroulant « **All** » et en sélectionnant **« Edit rows → Star rows »** ou « **Flag rows** ».

Pour plus d'informations (en anglais) : Facet by star or flag

Annuler et rétablir (Undo/Redo)

OpenRefine garde une trace de chaque opération que vous effectuez et vous permet facilement d'annuler ou de rétablir des étapes au fur et à mesure que vous travaillez sur vos données. De plus, OpenRefine enregistre vous permet de reprendre les étapes que vous avez effectuées sur un ensemble de données et de les appliquer à un autre ensemble de données par une opération de copier-coller. Vous pouvez accéder aux options Annuler (Undo) et Rétablir (Redo) dans le menu de gauche. Pour annuler ou rétablir une action, il vous suffit de cliquer sur l'étape à laquelle vous souhaitez revenir. Pour enregistrer l'historique de vos actions, vous pouvez cliquer sur le bouton **Extract**, où vous pouvez sélectionner les étapes que vous souhaitez enregistrer. Cet extrait sera au format JSON, qui peut être copié et enregistré dans un fichier texte. Lorsque vous souhaitez appliquer ces étapes, vous n'aurez qu'à coller le JSON et cliquer sur **Apply**. Les données d'annulation/rétablissement sont stockées avec le projet et sont enregistrées automatiquement au fur et à mesure que vous travaillez. Ainsi, la prochaine fois que vous ouvrirez le projet, vous pourrez accéder à l'historique complet des étapes que vous avez effectuées et annuler/rétablir exactement de la même manière.

Pour plus d'informations : Historique (Annuler/Rétablir)

Suppression des lignes vides

Parfois, vos données auront des lignes vides dispersées dans la feuille de calcul. Voici comment vous en débarrasser.

1. En regardant vos données dans la fenêtre d'affichage, vous voyez qu'il y a un certain nombre de lignes vides séparant les données.
2. Cherchons dans le fichier les lignes dont toutes les cellules sont vides. Cliquez sur la flèche déroulante à côté de « **All** » (à gauche). Passez votre curseur sur **Facet** et sélectionnez **Facet by blank (null or empty string)**.
3. Dans la facette qui apparaît, vous verrez « **true** » et « **false** » avec des chiffres. Les valeurs vides apparaîtront comme « **true** », puisqu'elles remplissent la condition du filtre (lignes vides). Passez votre curseur sur l'option « true » et sélectionnez « **include** ».

Enlever les espaces

Un autre problème courant avec les données sont les espaces (attention, c'est au féminin en informatique !) au début et à la fin des données d'une cellule. Nous voulons généralement les supprimer, non seulement parce que ces espaces peuvent donner l'impression que les données de la cellule sont dans le désordre, mais aussi parce qu'une machine traitera ces espaces supplémentaires dans vos données comme faisant partie de la chaîne de caractères. Vous pouvez supprimer les espaces de début ou de fin lors de l'importation, mais au cours de vos traitements des données, vous pouvez en créer, en particulier lorsque vous fractionnez ou fusionnez des colonnes.

Si vous avez décoché cette case lors de l'importation de données, ou si des espaces de début ou de fin ont été introduites lors du fractionnement de colonnes ou d'autres opérations, OpenRefine fournit un outil qui vous permet de supprimer les espaces au début et à la fin de toutes les entrées qui en contiennent.

1. Modifiez le village sur la première ligne pour introduire une espace à la fin : God
2. Créez une nouvelle facette de texte pour la colonne du village. Vous devriez maintenant voir deux entrées différentes pour God, et l'une d'elles a une espace à la fin.
3. Pour supprimer l'espace, choisissez **Edit cells > Common transformations > Trim leading and trailing whitespace.**
4. Vous ne devriez plus voir que quatre choix dans votre facette de texte (Text Facet).

Téléchargement d'OpenRefine

1. Téléchargez OpenRefine à partir du site Web suivant : http://openrefine.org/download.html

Veuillez utiliser 3.5.2, la dernière version stable d'OpenRefine 3.5, mise en ligne le 26 janvier 2022. Veuillez sauvegardez votre répertoire d'espace de travail avant l'installation et signalez tout problème. Un journal des modifications est fourni sur la page de publication .

- **Kit Windows** Ceci nécessite l'installation de Java sur votre ordinateur. Téléchargez, décompressez et double-cliquez sur *openrefine.exe* ou *raffiner.bat* si le premier ne fonctionne pas.
- **Kit Windows avec Java intégré** Comprend OpenJDK Java , disponible sous le nom Licence GPLv2+CE . Téléchargez, décompressez et double-cliquez sur *openrefine.exe* ou *raffiner.bat* si le premier ne fonctionne pas.
- **Kit Mac** : Télécharger, ouvrir, faire glisser l'icône dans le dossier Applications et double-cliquer dessus. Vous n'avez pas besoin d'installer Java séparément.
- **Kit Linux** : Téléchargez, extrayez, puis tapez *. /refine* pour commencer. Cela nécessite que Java soit installé sur votre ordinateur.

2. Téléchargez le jeu de données suivant : https://ndownload e r.figshare.com/files/11502815

Démarrer un projet dans OpenRefine

1. Exécutez l' application OpenRefine (ou double-cliquez sur le fichier openrefine.exe). NB *Bien que l' application OpenRefine s'ouvre dans la fenêtre de votre navigateur par défaut, elle est exécutée localement et les données ne sont ni téléchargées ni partagées en ligne.*
2. La première page que vous verrez est la page d'importation de données. Nous allons créer un nouveau projet en important un fichier de données. Assurez-vous que « **Create project** » est surligné dans le menu de gauche. OpenRefine vous permet de vous connecter à des sources locales ou en ligne. Nous allons téléverser un fichier *.txt local.

3. Dans le menu **"Get data from"** , sélectionnez "**This Computer** " et cliquez sur "**Choose Files**..."
4. Dans la fenêtre qui s'ouvre, naviguez jusqu'au fichier « **SAFI-openrefine- csv » ;** sélectionnez-le et cliquez sur « **Open** ».
5. Allez à nouveau dans la fenêtre OpenRefine et cliquez sur « **Next** »
6. Avant d'importer vos données, OpenRefine fournit un aperçu de la façon dont elles sont lues.
7. Notre fichier organise les données contenues en valeurs séparées par des tabulations (TSV ou *tab-separated values*). Dans le menu « **Parse data as** », sélectionnez « **CSV / TSV / separator-based files** ». Remarquez tous les différents types de fichiers offerts !
8. Vous voyez les différentes options dont vous disposez pour analyser les données. Vous pouvez, par exemple, ignorer les lignes au début du fichier ou analyser certaines lignes comme en-têtes.
9. Maintenant que vos paramètres d'importation sont configurés, vous pouvez cliquer sur « **Create a project** » dans le coin supérieur droit.

Attention : OpenRefine n'aime pas Firefox et ne fonctionnera pas du tout avec Internet Explorer.

Numérisation et standardisation de vos données avec les facettes et le clustering

Facettes

L'utilisation de la fonction **Facet** regroupe toutes les valeurs similaires qui apparaissent dans une colonne et vous permet de filtrer les données en fonction de ces valeurs ou encore de modifier les valeurs de plusieurs cellules en même temps. Les facettes vous permettent d'analyser et de regrouper vos données, ainsi que de repérer les erreurs et les incohérences en vous montrant une vue d'ensemble de vos données, en plus de vous donner la possibilité de filtrer les sous-ensembles que vous souhaitez modifier en lots. Il existe plusieurs types de facettes : textuelle/text, chiffrée/numeric, chronologique/timeline, nuage de points/scatterplot et personnalisé/custom. Nous allons nous concentrer sur les facettes textuelles aujourd'hui.

Pour en savoir plus : Explorer les facettes

Ici, nous utiliserons les facettes pour rechercher les erreurs potentielles dans la saisie des données de la colonne **« village » .**

1. Trouvez la colonne **"village" .**
2. Cliquez sur la flèche vers le bas et choisissez **Facet > Text** .

3. Dans le panneau de gauche, vous verrez une boîte contenant chaque valeur unique de la colonne **"village"** avec un chiffre représentant le nombre d'occurrences de cette valeur dans la colonne.
4. Essayez de trier cette facette par **nom/name** et par **nombre/count**. Voyez-vous des problèmes dans les données ? Si oui, lesquels ?
5. Passez votre curseur sur l'un des noms de la liste **Facet**. Vous devriez voir qu'une fonction d'édition est disponible.
    1. NB À cette étape, nous pourrions nettoyer les erreurs que nous voyons, mais nous allons vous montrer comment les nettoyer avec le clustering dans la section suivante.
6. Vous disposez également d'une option d'inclusion qui vous permet d'inclure autant de facettes que vous le souhaitez.

Exercice pratique

1. À l'aide des facettes, déterminer le nombre de valeurs « **interview_date »** différentes dans les résultats de l'enquête.
2. La colonne est-elle au format texte ou date ?
3. Utilisez Facet pour produire un affichage chronologique, **«Timeline facet»** pour « **interview_date** ». Vous devrez utiliser **Edit cells > Common transformations > To date** pour convertir cette colonne en dates.
4. Quelle est la période durant laquelle la plupart des entretiens ont-ils été faits ?

Groupement/Clustering

Le groupement (*clustering*) est un excellent moyen de trouver des fautes de frappe, des abréviations, des différences de casse et des variantes orthographiques. Le *clustering* cherche dans les données d'une colonne et rassemble les différentes facettes qui pourraient être des variations de la même chose. Ceci est crucial lorsqu'il s'agit de facettes textuelles, car OpenRefine traite une facette de texte comme une chaîne de caractères, ce qui signifie que toute variation (y compris la casse des lettres) sera lue comme une facette distincte. Par exemple, grâce au groupement, vous pouvez rechercher street, Street et St. et les standardiser en choisissant le terme souhaité. Dans l'exemple suivant, nous vous guiderons à travers le groupement. Pour l'instant, il n'est pas crucial de comprendre l'algorithme exact de chaque méthode de groupement. Ce qui est important, c'est de garder une trace de ce que vous utilisez, d'expérimenter pour voir ce qui fonctionne et de faire attention à ne pas fusionner des termes qui ne sont pas identiques !

Prenons un exemple.

1. Dans la facette **"village"** que nous avons créée à l'étape ci-dessus, cliquez sur le bouton **Cluster** .

2. Dans la fenêtre contextuelle (*pop-up*) résultante, vous pouvez modifier la méthode (**Method)** et la fonction de saisie (**Keying function)**. Essayez différentes combinaisons pour voir quelles différentes fusions de valeurs sont suggérées.
3. Sélectionnez la méthode **key collision** et la fonction de saisie **metaphone3 .** Cela devrait identifier deux groupements.
1. Cliquez sur la case **Merge ?** à côté de chaque groupement, puis cliquez sur **Merge selected and Recluster** pour appliquer les corrections au jeu de données.
2. Essayez de sélectionner à nouveau différentes méthodes et fonctions de saisie pour voir quelles nouvelles fusions sont suggérées.
3. Vous devriez constater qu'en utilisant les paramètres par défaut, plus aucun groupement n'est trouvé, par exemple pour fusionner Ruaca-Nhamuenda avec Ruaca ou Chirdozo avec Chirodzo.
4. Pour fusionner ces valeurs, nous allons les survoler dans la facette textuelle de **« village »**, sélectionner **« edit »** et modifier manuellement les noms. Changer Chirdozo en Chirodzo et Ruaca-Nhamuenda en Ruaca. Vous devriez maintenant avoir quatre groupements : Chirodzo , God, Ruaca et 49.

Important : Si vous **fusionnez (« merge »)** en utilisant une méthode ou une fonction de saisie différente, ou plus de fois que décrit dans les instructions ci-dessus, vos solutions pour les exercices ultérieurs ne seront pas les mêmes que celles présentées dans les solutions.

Vous voulez en savoir plus ? Consultez la documentation : En savoir plus sur le clustering

Présentation des transformations

Les transformations sont des moyens de manipuler des données dans des colonnes au-delà du groupement et des filtres. Dans OpenRefine, ces transformations sont effectuées à l'aide d'un langage appelé GREL (General Refine Expression Language) que vous pouvez considérer comme similaire aux formules d'Excel. Avec les transformations, vous pouvez :

- Fractionner les données d'une seule colonne en plusieurs colonnes (par exemple, fractionner une adresse en plusieurs parties). Vous pouvez également fusionner des colonnes !
- Normaliser le format des données dans une colonne sans modifier les valeurs (par exemple, supprimer la ponctuation ou normaliser un format de date)
- Extraire un type particulier de données d'une chaîne de texte plus longue (par exemple , trouver des ISBN dans une citation bibliographique)

Nous n'avons pas le temps de couvrir tout ce que vous pouvez faire avec GREL aujourd'hui, mais nous allons couvrir certaines transformations populaires afin que vous puissiez avoir une idée de la puissance du programme et de comment vous pouvez l'utiliser en complément d'Excel pour nettoyer vos données.

En savoir plus sur ce que GREL peut faire : https://docs.openrefine.org/m a nual/grelfunctions .

Nettoyage des données dans les cellules avec des transformations

Les données de la colonne **" items_owned "** sont un ensemble d'éléments sous forme de liste. La liste est entre crochets et chaque élément est entre guillemets simples. Avant de diviser la liste en éléments individuels dans la section suivante, nous souhaitons d'abord supprimer les crochets et les guillemets.

1. Cliquez sur la flèche pointant vers le bas qui se trouve en haut de la colonne « **items_owned** ». Choisissez **Edit cells > Transform...**

2. Cela ouvrira une fenêtre dans laquelle vous pourrez saisir une expression GREL.

3. D'abord, nous allons supprimer tous les crochets de gauche ([). Dans la zone Expression, saisissez : value.replace ("[", "") et cliquez sur OK.
4. L'expression a la signification suivante : prenez la valeur de chaque cellule de la colonne sélectionnée et remplacez tous les "[" par "" (c'est-à-dire rien ! Donc, supprimer les [).
5. Essayez maintenant de supprimer le crochet fermant et les guillemets en utilisant la même méthode.
6. Cliquez sur OK. Vous devriez voir dans la colonne " **items_owned "** que tous les crochets et guillemets ont été supprimés.

Maintenant que nous avons nettoyé les caractères superflus de notre colonne « **items_owned »** , nous pouvons utiliser une facette textuelle pour voir quels éléments étaient généralement possédés ou rarement possédés par les personnes interrogées.

1. Cliquez sur la flèche vers le bas en haut de la colonne « **items_owned** ». Choisissez **Facet > Custom text facet...**
2. Dans la zone Expression, tapez value.split (";")
3. Cliquez sur OK
4. Vous devriez maintenant voir une nouvelle zone de facette textuelle dans le volet de gauche avec chaque valeur unique repérée dans les données qui se trouvent dans les cellules.

Exercice

Effectuez les mêmes étapes de nettoyage et de facettes textuelles personnalisées pour la colonne « **months_lack_food »** . Pendant quel(s) mois les agriculteur˙rice·s ont-ils et -elles été les plus susceptibles de manquer de nourriture ?

Fractionner les données en colonnes

Cellules

Parfois, vous devez séparer les informations contenues dans une même cellule en lignes distinctes. Dans la section précédente, nous avons travaillé dans la colonne " **items_owned "** , en nettoyant quelques caractères supplémentaires et nous avons vu comment gérer en facettes des cellules

contenant plusieurs valeurs. Mais si vous envisagez de faire une analyse dans un autre programme comme Excel, vous voudrez peut-être séparer ces données.

1. Cliquez sur la flèche déroulante à côté de l'en-tête de colonne « **items_owned** ». Survolez « **Edit cells** » et sélectionnez « **Split multi-valued cells**… »
2. Dans le menu qui apparaît, spécifiez que la colonne sera divisée par un séparateur et saisissez un point-virgule " **;** ” dans la zone de texte.

Voyez ce qui se passe : vous verrez que chaque valeur a été ajoutée dans une nouvelle ligne. N'est-ce pas génial comme affichage ? C'est une excellente option lorsque vous souhaitez analyser plusieurs réponses à la même question dans quelque chose comme un tableau croisé dynamique.

Exercice

Pouvez-vous diviser **« liv_owned »** de la même manière ? Pourquoi le feriez-vous? Comment le feriez-vous ?

Fusionner des colonnes

Vous pouvez également fusionner les informations de deux colonnes (ou plus). Lorsque vous exécutez cette fonction, toutes les chaînes sont ajoutées dans la cellule la plus haute de l'enregistrement dans l'ordre dans lequel elles apparaissent.

1. Pour commencer, sélectionnez la première colonne que vous souhaitez fusionner – ici, ce sera **« province ».** Dans le menu déroulant, sélectionnez **"Edit Column",** puis **"Join columns".**
2. Une fenêtre contextuelle s'ouvrira sur la gauche avec une liste de laquelle vous pourrez choisir les colonnes que vous souhaitez fusionner. Dans ce cas, nous aimerions fusionner la province et le district pour faire une phrase "[province], [district], [ward] ».

3. Ajoutez un séparateur. Pour cet exercice, nous allons ajouter les caractères suivants comme séparateur : **", "** (virgule espace). Vous devez généralement choisir un séparateur qui n'est pas couramment utilisé dans vos données afin de savoir où se trouve le changement et de le supprimer facilement par programmation. Une ligne verticale ("|" ou *pipe* en anglais) ou ";" (point-virgule) sont deux bons choix. Puisque nous fusionnons ceci pour créer une phrase lisible pour les humains, nous choisissons des caractères qui se lisent normalement.

Exemple
Essayez de joindre **"ward "** et **"village".** Y a-t-il d'autres instances dans cette feuille de calcul que vous aimeriez fusionner ? Quel type de séparateur utiliseriez-vous ?

Nettoyage final et exportation

Il est temps d'enregistrer et d'exporter. Quel type de fichier allez-vous choisir ? OpenRefine offre beaucoup d'options. De plus, OpenRefine enregistre automatiquement votre progression.

1. Si vous souhaitez exporter votre travail dans OpenRefine afin de pouvoir le reprendre sur un autre ordinateur, cliquez sur le bouton **Export** dans le coin supérieur droit, puis sélectionnez « **Export project** ». Vous aurez la possibilité d'enregistrer le fichier dans un répertoire local ou sur Google Drive.
2. Si vous avez terminé de nettoyer vos données, vous pouvez choisir d'exporter un fichier avec les paramètres actuels activés. Cliquez sur **Export,** dans le coin supérieur droit, puis sélectionnez votre format préféré (par exemple Excel (.xls), ou valeur séparée par des tabulations (tsv), etc.)