Nettoyer ses données avec OpenRefine

Exercice adapté de la leçon de Seth van Hooland, Ruben Verborgh, and Max De Wilde sur Programming Historian

Objectifs

Ne faites pas confiance à vos données : c'est le message-clé de ce tutoriel qui se concentre sur la façon dont les chercheurs et chercheuses peuvent évaluer et agir sur l'exactitude des données. Dans cet exercice, vous apprendrez les principes et la pratique du nettoyage de données, ainsi que la façon dont OpenRefine peut être utilisé pour effectuer quatre tâches essentielles qui vous aideront à nettoyer ces données :

- supprimer les doublons ;
- séparer plusieurs valeurs contenues dans un même champ ;
- analyser la distribution des valeurs dans un ensemble de données ;
- regrouper différentes représentations de la même réalité.

Ces étapes sont illustrées à l'aide d'une série d'exercices basés sur une collection de métadonnées du Powerhouse museum, démontrant comment les méthodes (semi-) automatisées peuvent vous aider à corriger les erreurs dans vos données.

Pourquoi se préoccuper de la qualité de nos données ?

Les doublons, les valeurs vides et les formats incohérents sont des phénomènes auxquels nous devrions être prêt(e)s à faire face lorsque nous utilisons des ensembles de données historiques. Cet exercice vous apprendra à détecter les incohérences dans les données contenues dans une feuille de calcul ou une base de données. L'accroissement des pratiques de partage, de regroupement et de réutilisation de données, par ailleurs de plus en plus nombreuses sur le Web, obligent les historiens et les historiennes à agir face aux problèmes inévitables de qualité des données. Grâce à l'utilisation d'OpenRefine, vous pourrez facilement identifier les erreurs récurrentes telles que des cellules vides, des doublons, des incohérences orthographiques, etc. OpenRefine vous permet non seulement d'avoir un diagnostic rapide sur la qualité de vos données, mais aussi d'agir sur certaines erreurs de manière automatisée.

Description de l'outil OpenRefine

Dans le passé, les historiens et les historiennes devaient compter sur des spécialistes des technologies de l'information pour avoir un diagnostic sur la qualité des données et exécuter des tâches de nettoyage. Cela nécessitait la création de programmes informatiques sur mesure lorsque vous travailliez avec des ensembles de données importants. Heureusement, l'avènement des outils de transformation interactive de données permet désormais de réaliser des opérations rapides et peu coûteuses sur de grandes quantités de données, même effectuées par des professionnel(le)s sans compétences informatiques approfondies.

Les outils de transformation interactive de données ressemblent aux tableurs de bureau que nous connaissons tous, et partagent avec eux certaines fonctionnalités. Vous pouvez par exemple utiliser une application telle que Microsoft Excel pour trier vos données en fonction de filtres numériques, alphabétiques ou personnalisés, ce qui vous permet de détecter plus facilement les erreurs. La configuration de ces filtres dans une feuille de calcul peut être fastidieuse, car il s'agit d'une fonctionnalité secondaire. Sur un plan plus général, nous pourrions dire que les tableurs sont conçus pour fonctionner sur des lignes et des cellules individuelles, alors que les outils de transformation interactive de données fonctionnent sur de

grandes plages de données à la fois. Ces « tableurs surpuissants » offrent une interface intégrée et conviviale à travers laquelle les utilisateurs finaux peuvent détecter et corriger les erreurs.

Plusieurs outils généralistes pour la transformation interactive des données ont été développés ces dernières années, tels que Potter's Wheel ABC et Wrangler. Ici, nous voulons nous concentrer spécifiquement sur OpenRefine (anciennement Freebase Gridworks et Google Refine) car c'est un outil convivial, gratuit et open-source pour traiter et nettoyer efficacement de grandes quantités de données au travers d'une interface basée sur un navigateur.

En plus des opérations de profilage et de nettoyage des données, des extensions OpenRefine permettent aux utilisateurs et utilisatrices d'identifier des objets textuels dans du texte non structuré, un processus appelé reconnaissance d'entités nommées, et peuvent également lier leurs propres données avec des bases de connaissances existantes. Ce faisant, OpenRefine peut être un outil pratique pour lier des données avec des objets et des autorités qui ont déjà été déclarés sur le Web par des organisations telles que la Bibliothèque du Congrès ou l'OCLC. Le nettoyage des données est une condition préalable à ces étapes ; le taux de réussite de la reconnaissance d'entités nommées et un processus de croisement productif de vos données avec les autorités externes dépendent de votre capacité à rendre vos données aussi cohérentes que possible.

Description de l'exercice : Powerhouse Museum

Le Powerhouse Museum de Sydney a mis à disposition en 2013 sur son site web un ensemble de métadonnées sur sa collection librement utilisable (le jeu de données n'est plus accessible directement sur le site du powerhouse museum, mais il a été sauvegardé). Le musée est l'un des plus grands musées de science et de technologie au monde, offrant un accès à près de 90 000 objets, des machines à vapeur à la verrerie fine en passant par la haute couture et les puces informatiques.

Le Powerhouse Museum a rendu sa collection largement disponible en ligne et la plupart de ses données étaient librement accessibles. Sur le site web du musée, il était possible de télécharger un fichier texte séparé par des tabulations appelé `phm-collection.tsv` sous la forme d'une feuille de calcul. Le fichier décompressé (58 Mo) contient des métadonnées de base (17 champs) pour 75 823 objets, publiées sous licence Creative Commons Attribution - Partage dans les Mêmes Conditions. Dans ce tutoriel, nous utiliserons une copie des données que archivées pour vous permettre de les télécharger (dans la suite de la leçon).

Tout au long du processus de profilage et de nettoyage des données, l'étude de cas se concentrera spécifiquement sur le champ `Categories` (Catégories), qui contient les termes du thésaurus des noms d'objets du Powerhouse Museum. Ce thésaurus reconnaît l'usage des mots et l'orthographe australiens et reflète de manière très directe les points forts de la collection. Dans celle-ci, vous trouverez davantage de représentations de l'histoire sociale et des arts décoratifs que de noms d'objets relatifs aux beaux-arts et à l'histoire naturelle.

Les termes du champ `Categories` comprennent ce que nous appelons un vocabulaire contrôlé. Un vocabulaire contrôlé consiste en des mots-clés décrivant le contenu d'une collection en utilisant un nombre limité de termes, et constitue souvent un point d'entrée crucial dans les ensembles de données utilisés par les historiens et les historiennes dans les bibliothèques, les archives et les musées. C'est pourquoi nous accorderons une attention particulière à ce champ. Une fois les données nettoyées, il devrait être possible de réutiliser les

termes du vocabulaire contrôlé pour trouver des informations supplémentaires sur les termes utilisés ailleurs en ligne, ce qui s'appelle « créer des données liées ».

Pour commencer : installer OpenRefine et importer des données

Téléchargez OpenRefine et suivez les instructions d'installation. OpenRefine fonctionne sur toutes les plateformes : Windows, Mac et Linux. OpenRefine s'ouvrira dans votre navigateur, mais il est important de réaliser que l'application s'exécute localement et que vos données ne seront pas stockées en ligne. Les fichiers de données sont disponibles sur la page web du cours (htptps://recherche.mael-lenoc.fr/uvsq-2026/).[1] Veuillez télécharger le fichier phm-collection.tsv qui sera utilisé tout au long de cet exercice avant de continuer.

Sur la page de démarrage d'OpenRefine, créez un nouveau projet en utilisant le fichier de données téléchargé et cliquez sur Suivant. Par défaut, la première ligne sera correctement analysée comme étant l'en-tête des colonnes, mais vous devez décocher la case `Utiliser le caractère “ pour fermer les cellules contenant les séparateurs de colonnes` si elle est cochée, car les citations dans le fichier n'ont aucune signification pour OpenRefine. En outre, cochez la case `Analyser le texte des cellules comme nombres` pour permettre à OpenRefine de détecter automatiquement les nombres. Maintenant, cliquez sur Créer un projet. Si tout se passe bien, vous verrez 75 814 lignes.

L'ensemble de données du Powerhouse Museum comprend des métadonnées détaillées sur tous les objets de la collection, y compris le titre, la description, plusieurs catégories auxquelles l'objet appartient, des informations sur la provenance et un lien pérenne (`persistent link`) vers l'objet sur le site Web du musée. Le lien perenne n'est malheureusement plus fonctionnel (il ne s'agissait donc pas d'un lien perenne...) mais vous pouvez voir un exemple sur l'image ci dessous.

[1] Si vous ne pouvez pas installer openrefine, il est possible d'utiliser openrefine en ligne en créant un compte sur https://usegalaxy.fr/ et en suivant les étapes décrites ici : https://training.galaxyproject.org/training-material/topics/digital-humanities/tutorials/open-refine-tutorial/tutorial.html#hands-on-get-the-data

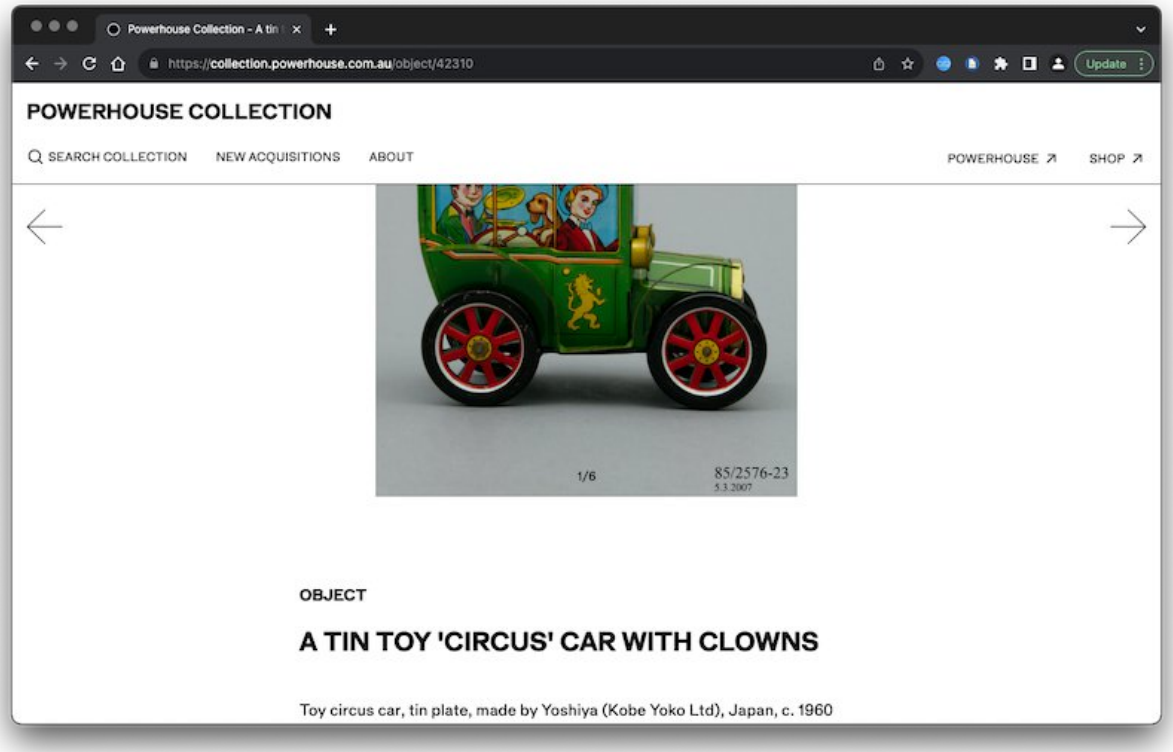

Figure 1: Capture d'écran d'un objet échantillon en provenance du site web du Powerhouse Museum

Apprendre à connaître ses données

La première chose à faire est de faire le tour de vos données et d'apprendre à les connaître. Prenez quelques minutes pour explorer les données et les différentes colonnes. Vous pouvez inspecter les différentes valeurs de données en les affichant dans des facettes. Vous pouvez considérer une facette comme un filtre à travers laquelle vous visualisez un sous-ensemble spécifique de données, en fonction d'un critère de votre choix. Cliquez sur le triangle situé devant le nom de la colonne, sélectionnez Facette et créez-en une. Par exemple, essayez une facette textuelle ou une facette numérique, selon la nature des valeurs contenues dans les champs (les valeurs numériques sont en vert). Sachez, cependant, que les facettes textuelles sont mieux utilisées sur des champs avec des valeurs récurrentes (`Categories` par exemple) ; si vous rencontrez une erreur « trop nombreux à afficher », vous pouvez choisir d'augmenter la limite du nombre de choix au-dessus des 2 000 par défaut, mais une limite trop élevée peut ralentir l'application (il vaut mieux éviter de dépasser 5 000). Les facettes numériques n'ont pas cette restriction. Pour plus d'options, sélectionnez Facettes personnalisées : la facette par valeur vide, par exemple, est pratique pour savoir combien de valeurs ont été remplies pour chaque champ. Nous les explorerons plus en détail dans les exercices suivants.

Supprimer les lignes vides

Une chose que vous remarquez lors de la création d'une facette numérique pour la colonne `Record ID`, est que trois lignes sont vides. Vous pouvez les trouver en désélectionnant la case `Numérique`, en ne laissant que les valeurs non numériques. En fait, ces valeurs ne sont pas

vraiment vides mais contiennent un seul caractère d'espace, ce qui peut être constaté en déplaçant votre curseur à l'endroit où la valeur aurait dû être, et en cliquant sur le bouton Éditer qui apparaît. Pour supprimer ces lignes, cliquez sur le triangle situé devant la première colonne appelée Toutes, sélectionnez `Éditer les lignes`, puis Supprimer toutes les lignes correspondantes. Fermez la facette numérique pour voir les 75 811 lignes restantes.

Supprimer les doublons

Une deuxième étape consiste à détecter et supprimer les doublons. Ceux-ci peuvent être repérés en les triant par une valeur unique, telle que l'ID d'enregistrement (`Record ID`). Dans ce cas, nous supposons que l'ID d'enregistrement est unique pour chaque entrée. L'opération peut être effectuée en cliquant sur le triangle à gauche de Record ID, puis en choisissant `Trier…` et en sélectionnant la puce `Nombres`. Dans OpenRefine, le tri n'est qu'une aide visuelle, à moins que vous ne rendiez le réordonnancement permanent. Pour ce faire, cliquez sur le menu `Trier` qui vient d'apparaître en haut et choisissez `Retrier les lignes de façon permanente.` Si vous oubliez de le faire, vous obtiendrez des résultats imprévisibles plus tard dans ce tutoriel.

Les lignes identiques sont maintenant adjacentes les unes aux autres. Pour voir uniquement les facettes en double, , cliquez sur le triangle Record ID, choisissez `Facette`, `Facette personnalisée`, `Facette Doublon` Ensuite, effacez l'ID d'enregistrement des lignes ayant le même ID d'enregistrement que la ligne située au-dessus d'elles, en les marquant en double. Pour ce faire, cliquez sur le triangle Record ID, choisissez `Éditer les cellules` > `Vider les valeurs répétées dans les cellules consécutives` . Le statut vous indique que 84 colonnes ont été affectées (si vous avez oublié de réordonner les rangées de façon permanente, vous n'en obtiendrez que 19, si c'est le cas, annulez l'opération dans l'onglet Annuler / Rétablir et revenez au paragraphe précédent ; assurez-vous que les lignes sont réorganisées et non simplement triées). Éliminez ces lignes en créant une facette sur les cellules vides dans la colonne Record ID (`Facette`> `Facettes personnalisées`> `Facette par valeur vide (valeurs nulles ou chaîne vide)`), puis sélectionnez les 84 lignes vides en cliquant sur true, et supprimez-les en utilisant le triangle `Toutes` (`Éditer les lignes` > `Supprimer toutes les lignes correspondantes`). À la fermeture de la facette, vous voyez 75 727 lignes uniques.

Une vigilance particulière est nécessaire lors de l'élimination des doublons. Dans l'étape mentionnée ci-dessus, nous supposons que le jeu de données comporte un champ avec des valeurs uniques, indiquant que la ligne entière représente un doublon. Ce n'est pas nécessairement le cas, et il faut être très prudent pour vérifier manuellement si la ligne entière représente un doublon ou non.

Fractionner les données

Une fois les doublons supprimés, nous pouvons regarder de plus près le champ `Categories`. En moyenne, chaque objet a reçu 2,25 catégories. Ces catégories sont contenues dans le même champ, séparées par le caractère pipe |. Le dossier 9, par exemple, en contient trois : Mineral Samples | Specimens | Mineral Samples-Geological. Afin d'analyser en détail l'utilisation des mots-clés, les valeurs du champ `Categories` doivent être fractionnées en cellules individuelles sur la base du caractère pipe, en développant les 75 727 entrées en 170 167 lignes. Choisissez `Éditer les cellules, Diviser les cellules multivaluées`, et en entrant | comme séparateur de valeurs. OpenRefine vous informe que vous avez maintenant 170 167 lignes.

Il est important de bien comprendre le concept des lignes / entrées. Rendez la colonne Record ID visible pour voir ce qui se passe (pour cela, il suffit de cliquer sur le triangle puis sur `Masquer /Afficher` et ensuite sur `Masquer les autres colonnes`). Vous pouvez basculer entre les vues lignes et entrées en cliquant sur les liens ainsi étiquetés juste au-dessus des en-têtes de colonnes. Dans la vue des lignes, chaque ligne représente deux ID d'enregistrement (Record ID) et une seule catégorie, ce qui permet de les manipuler individuellement. La vue des entrées a une entrée pour chaque ID d'enregistrement (Record ID), qui peut avoir différentes catégories sur des lignes différentes (regroupées en gris ou en blanc), mais chaque enregistrement est manipulé dans son ensemble. Concrètement, il y a maintenant 170 167 affectations de catégories (lignes) réparties sur 75 736 éléments de collection (entrées). Vous avez peut-être remarqué que nous avons 9 entrées de plus que les 75 727 d'origine, mais ne vous inquiétez pas pour le moment, nous reviendrons à cette petite différence plus tard.

Appliquer une facette et regrouper des valeurs proches

Une fois que le contenu d'un champ a été correctement fractionné, les filtres, les facettes et les regroupements de valeurs proches peuvent être appliqués pour donner un aperçu rapide et simple des problèmes de métadonnées classiques. En appliquant la facette personnalisée `Facette par valeur vide`, on peut immédiatement identifier les 461 entrées qui n'ont pas de catégorie, et qui représentent 0,6% de la collection. L'application d'une facette textuelle au champ Categories permet d'avoir un aperçu des 4 934 catégories différentes utilisées dans la collection (la limite par défaut étant de 2 000, vous pouvez cliquer sur Définir la limite du nombre de choix pour la porter à 5 000). Les titres peuvent être triés par ordre alphabétique ou par fréquence (compte), donnant une liste des termes les plus utilisés pour indexer la collection. Les trois principales rubriques sont Numismatics (Numismatique) (8 041), Ceramics (Céramique) (7 390) et Clothing and dress (Vêtements et habillement) (7 279).

Après l'application d'une facette, OpenRefine propose de regrouper les choix de facettes en fonction de diverses méthodes de similarité avec le bouton `Groupe`. Comme l'illustre la figure 2, le regroupement des valeurs proches vous permet de résoudre des problèmes concernant les incohérences d'occurrences, l'utilisation incohérente de la forme singulière ou plurielle et les fautes d'orthographe simples. OpenRefine présente les valeurs proches et propose une fusion dans la valeur la plus utilisée. Sélectionnez les valeurs que vous souhaitez regrouper en sélectionnant leurs cases individuellement ou en cliquant sur `Tout sélectionner` en bas, puis choisissez `Fusionner la sélection et regrouper`.

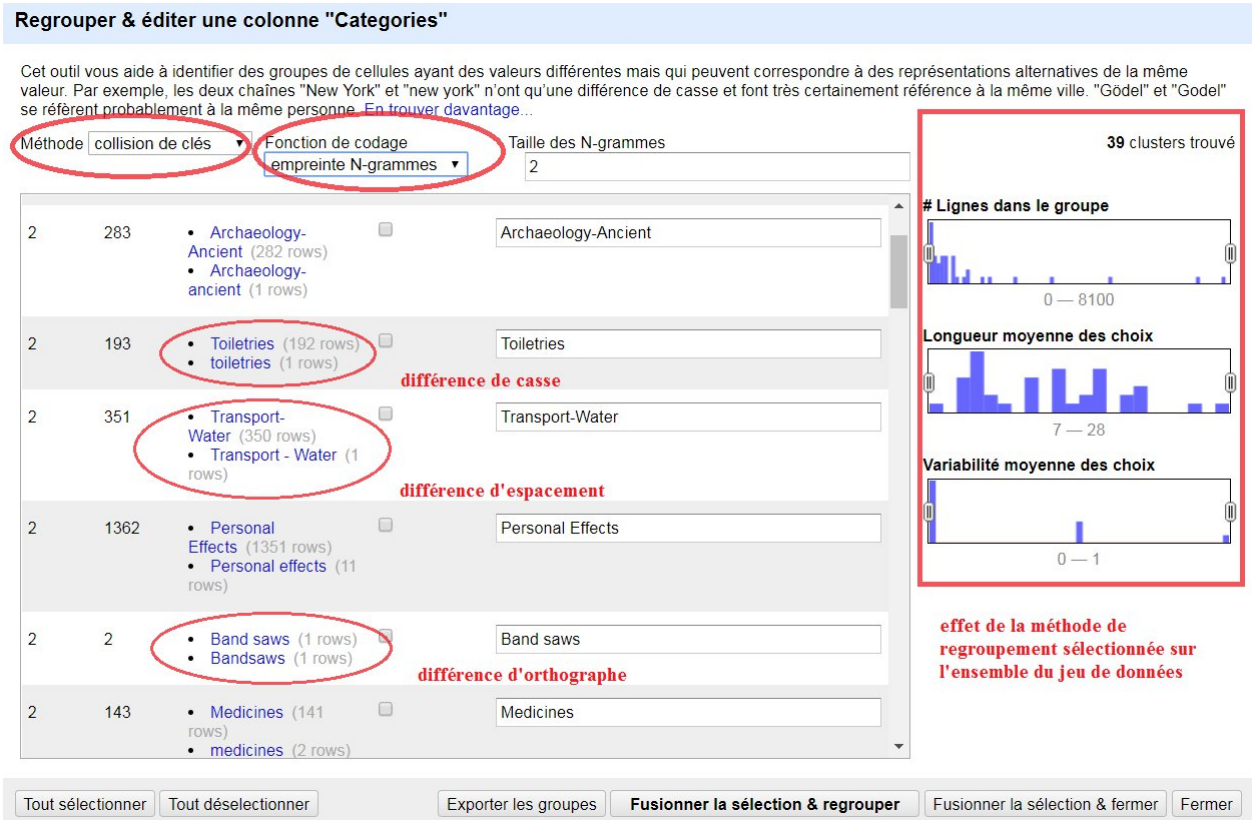

Figure 2 : Vue de valeurs regroupées

La méthode de regroupement par défaut est assez basique, elle ne trouve donc pas encore tous les regroupements possibles. Expérimentez différentes méthodes pour voir quels résultats elles donnent. Soyez prudent cependant : certaines méthodes sont trop poussées, vous pourriez donc finir par regrouper des valeurs qui ne sont pas équivalentes. Maintenant que les valeurs ont été regroupées individuellement, nous pouvons les regrouper dans une seule cellule. Cliquez sur le triangle `Categories` et choisissez `Éditer les cellules> Joindre les cellules multivaluées> OK`. Choisissez le caractère pipe (|) comme séparateur. Les lignes ressemblent maintenant à ce que nous avions avant, avec un champ Categories à plusieurs valeurs.

Appliquer des transformations ad hoc à l'aide d'expressions GREL

Vous vous souvenez peut-être qu'il y a eu une augmentation du nombre d'entrées après le processus de fractionnement : neuf entrées sont soudainement apparues. Afin de trouver la cause de cette disparité, nous devons remonter dans le temps, avant l'opération de fractionnement des catégories en lignes séparées. Pour ce faire, activez l'onglet Défaire / Refaire à droite de l'onglet Facette / Filtre, et vous obtiendrez un historique de toutes les actions que vous avez effectuées depuis la création du projet. Sélectionnez l'étape juste avant Split multi-valued cells in column Categories (Fractionner les cellules à valeurs multiples dans la colonne Categories). Si vous avez suivi notre exemple, cela devrait être Remove 84 rows (Supprimer 84 lignes). Puis revenez à l'onglet Facette / Filtre.

Le problème est survenu lors de l'opération de séparation sur le caractère pipe, il y a donc de fortes chances que tout ce qui s'est mal passé soit lié à ce caractère. Appliquons un filtre dans la colonne `Categories` en sélectionnant `Filtrer le texte dans le menu`. Entrez tout

d'abord un seul | dans le champ de gauche : OpenRefine vous informe qu'il y a 71 064 entrées correspondantes (c'est-à-dire des entrées contenant un pipe) sur un total de 75 727. Les cellules qui ne contiennent pas de pipe peuvent être des cellules vides, mais aussi des cellules contenant une seule catégorie sans séparateur, comme l'enregistrement 29 qui n'a que des "Scientific instruments" (« instruments scientifiques»). Maintenant, entrez un second | après le premier de sorte à obtenir || (double pipe) : vous pouvez voir que 9 entrées correspondent à ce modèle. Ce sont probablement les 9 entrées responsables de notre discordance : lorsque OpenRefine les sépare, le double pipe est interprété comme une rupture entre deux entrées au lieu d'un double séparateur sans signification. Maintenant, comment pouvons-nous corriger ces valeurs ? Allez dans le menu du champ `Categories` et choisissez `Éditer les cellules> Transformer….` Bienvenue dans l'interface de conversion de texte personnalisée, une fonctionnalité puissante d'OpenRefine utilisant le langage GREL (spécifique d'OpenRefine).

Le mot « `value` » dans le champ texte représente la valeur actuelle de chaque cellule, que vous pouvez voir ci-dessous. Nous pouvons modifier cette valeur en lui appliquant des fonctions (voir la documentation GREL pour une liste complète). Dans ce cas, nous voulons remplacer les doubles pipes par un seul pipe. Cela peut être réalisé en entrant l'expression GREL suivante (assurez-vous de ne pas oublier les guillemets) :

```
value.replace('||', '|')
```

Sous le champ texte Expression, vous obtenez un aperçu des valeurs modifiées, avec les doubles pipes supprimés. Cliquez sur OK et réessayez de fractionner les catégories avec `Modifier les cellules> Diviser les cellules multivaluées` : le nombre d'entrées restera à 75 727 (cliquez sur le lien entrées pour vérifier une nouvelle fois).

Un autre problème qui peut être résolu avec l'aide de GREL est celui des entrées pour lesquelles la même catégorie est listée deux fois. Prenez par exemple le cas 41, dont les catégories sont Models | Botanical specimens | Botanical Specimens | Didactic Displays | Models. La catégorie Models apparaît deux fois sans raison valable, nous voulons donc supprimer ce doublon. Cliquez sur le triangle `Categories` et choisissez `Éditer les cellules> Joindre les cellules multivaluées> OK`. Choisissez le caractère pipe comme séparateur. Maintenant, les catégories sont listées comme avant. Sélectionnez ensuite `Éditer les cellules> Transformer`, également dans la colonne des catégories. En utilisant GREL, nous pouvons diviser successivement les catégories sur le caractère pipe, rechercher des catégories uniques et les joindre à nouveau. Pour ce faire, tapez simplement l'expression suivante :

```
value.split('|').uniques().join('|')
```

Vous remarquerez que 33 008 cellules sont concernées, soit plus de la moitié de la collection.

Exporter vos données nettoyées

Depuis que vous avez chargé vos données dans OpenRefine, toutes les opérations de nettoyage ont été effectuées dans la mémoire du logiciel, laissant votre jeu de données d'origine intact. Si vous voulez enregistrer les données que vous avez nettoyées, vous devez les exporter en cliquant sur le menu Exporter en haut à droite de l'écran. OpenRefine prend en charge une grande variété de formats, tels que CSV, HTML ou Excel : sélectionnez ce qui vous convient le mieux ou ajoutez votre propre modèle d'exportation en cliquant sur Templating. Vous pouvez également exporter votre projet au format OpenRefine interne afin de le partager avec d'autres utilisateurs.

Pour aller plus loin avec vos données nettoyées

Une fois vos données nettoyées, vous pouvez passer à l'étape suivante et explorer d'autres fonctionnalités intéressantes d'OpenRefine. La communauté d'utilisateurs d'OpenRefine a développé deux extensions particulièrement utiles qui vous permettent de lier vos données à des données déjà publiées sur le Web. L'extension RDF Transform transforme les mots-clés en texte brut en URL. L'extension NER vous permet d'appliquer la reconnaissance des entités nommées, qui identifie les mots-clés dans un texte et leur donne une URL.

Conclusion

Si vous devez vous souvenir d'une seule chose après cet exercice, ce doit être celle-ci : aucun jeu de donnée n'est parfaitement « propre », mais vous pouvez y remédier. Comme nous l'avons montré ici, il y a déjà beaucoup de choses que vous pouvez faire par vous-mêmes pour améliorer la qualité de vos données. Vous avez ainsi appris comment avoir un rapide aperçu du nombre de valeurs vides que contient votre jeu de données et à quelle fréquence une valeur particulière (par exemple un mot-clé) est utilisée dans une collection. Cet exercice vous a également montré comment résoudre des problèmes récurrents tels que les doublons et les incohérences orthographiques de manière automatisée à l'aide d'OpenRefine.