

 http://www.openrefine.org/	OpenRefine (ex-Google Refine)	
	Outils d'analyse et de raffinement de données	
	System Requirements : Windows, Mac, Linux	
	Développé par :	Google & communauté « Open Source »
Licence BSD (gratuit)	Personne à contacter :	Vandy.Berten@smals.be

Fonctions

OpenRefine a été historiquement développé par Metaweb, sous le nom de Freebase Gridworks, société rachetée par Google en 2010, qui l'a renommée Google Refine. En 2012, Google ayant arrêté son développement, l'application a été reprise par la communauté « Open Source » sous le nom d'OpenRefine.

Le but de ce logiciel est de pouvoir finement explorer une base de données (ou plus généralement, un ensemble de données), pour mettre en évidence sa structure, des cas de redondance, d'incohérence, des erreurs, des imperfections, ..., pour ensuite pouvoir en améliorer la qualité, par des regroupements, des restructurations, des corrections en série voire même des connexions avec d'autres bases de données (géographique, dictionnaires, ...)

OpenRefine permet d'importer des données venant de multiples formats, tels que, parmi d'autres, des fichiers Excel ou OpenOffice (ou LibreOffice), du XML, du CSV (Comma Separated Values) ou du texte brut.

En ce qui concerne l'analyse, outre les possibilités classiques de tri, de filtre (avec expressions régulières), OpenRefine offre deux outils puissants d'analyse : le « faceting » et le « clustering ».

Le « faceting » permet de résumer, pour un champ, l'ensemble des valeurs présentes et le nombre d'occurrences. Cela n'a bien évidemment d'intérêt que dans le cas de champs à un nombre restreint de valeurs, comme une catégorie, ou un pays. Cela permet la visualisation de la répartition des valeurs, mais également d'identifier des problèmes de cohérence. Par exemple, on pourrait identifier que « Belgique », « belgique » ou « Belg. » devraient en fait être tous orthographiés de la même façon. Les « numeric facets » permettent, eux, de visualiser la répartition d'une valeur numérique, mais aussi de mettre en évidence des valeurs non numériques, alors qu'elles devraient l'être, ou des « trous » dans la numérotation. D'autres types de faceting, paramétrables, permettent de nombreux regroupements, basés sur la taille du champ, le nombre de lettres, la présence de blancs ou d'erreurs, ou de duplicatas.

Le clustering permet, lui, de regrouper des valeurs semblables, dans le but de les fusionner par la suite. Par exemple, dans une bibliothèque, des catégories « Livre d'Histoire », « Livre d'histoire », ou « Livres d'Histoire » peuvent être identifiées et être regroupées d'un seul clic.

Conclusions et Recommandations

OpenRefine est une très bonne alternative gratuite aux outils professionnels d'analyse de qualité de données, souvent hors de prix pour des petites organisations. Il est par ailleurs beaucoup plus simple à utiliser, et permet une prise en main rapide, même sans connaissances poussées en informatique. Il n'en a cependant pas toute la puissance, tant en terme de fonctionnalités qu'en terme de capacité à gérer des grands volumes de données.

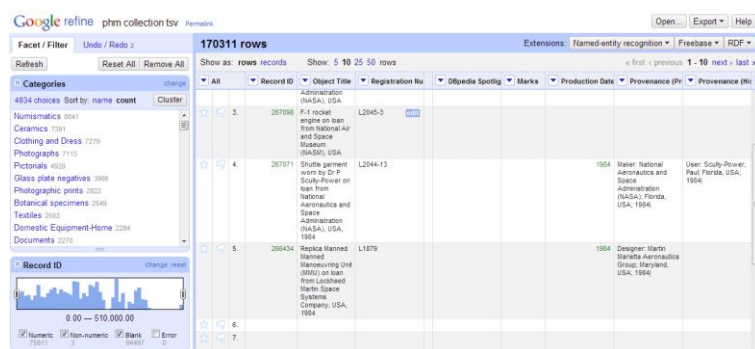
Il faut néanmoins souligner qu'une analyse fine de données ne se limite pas à l'usage d'un logiciel, aussi puissant soit-il, ni même à des connaissances purement techniques. Il faut non seulement une connaissance poussée du domaine d'application, mais également une grande expertise en matière de qualité de données, de *profiling*, de standardisation ou de *matching*.

Fonctions (suite)

Une fois les problèmes identifiés, OpenRefine propose de multiples outils pour corriger aisément les données : on peut ainsi renommer en blocs certains champs, supprimer des enregistrements vides, inutiles ou des doublons, séparer des cellules multi-valuées (plusieurs numéros de téléphones ou catégories séparés d'une virgule, par exemple) ... On peut également créer des nouvelles colonnes à partir d'une colonne existante, sur base soit d'expression régulière, soit d'un langage de programmation très puissant (GREL, pour General Refine Expression Language), proche de JavaScript. Cela permettra de vérifier, par exemple, que toutes les valeurs d'une colonne respectent bien une syntaxe spécifique (adresse e-mail, site web, numéro de téléphone...).

Il est par ailleurs possible à tout moment de revenir en arrière jusqu'à la création du projet, étape par étape, même après avoir redémarré le programme. Il ne faut donc pas avoir peur de procéder par essais et erreurs, comme c'est presque toujours nécessaire en analyse de qualité de données.

Tests et Résultats



Record ID	Object Title	Registration No.	Production Date	Provenance (Pi)	Provenance (Bi)
267096	F-1 rocket engine on loan from National Air and Space Museum (NASM), USA	L2045-3			
267071	Shuttle garment worn by Dr F. Sully-Prud'homme on loan from National Air and Space Administration (NASA), USA, 1964	L2044-13	1964	Water National Administration and Space Administration (NASA), Florida, USA, 1964	User: Sully-Prud'homme, Paul, Florida, USA, 1964
266434	Replica (Hanned) Mannequin Unit (MSU) on loan from Lockheed Martin Space Systems Company, USA, 1964	L1579	1964	Designer: Martin Marietta Astronautics Group, Maryland, USA, 1964	

Dans « Using OpenRefine » (Ruben Verborgh, Max De Wilde, Packt Publishing, 2013), les auteurs décrivent de façon très pédagogique et par l'exemple le cas de l'inventaire (réel) du « Powerhouse Museum » à Sidney, comprenant plus de 75 000 objets. Ils y identifient en quelques manipulations et avec succès de très nombreux cas d'incohérence, de manque de consistance, d'enregistrement vide ou presque, de doublons. Ils corrigent ensuite ces erreurs et regroupent ce qui doit l'être de façon largement automatisée.

Nous nous sommes également servi d'OpenRefine pour étudier la qualité d'un listing d'adresses e-mail. Nous avons pu aisément extraire les adresses syntaxiquement incorrectes à l'aide d'une expression régulière. De plus, après avoir extrait le nom de domaine sur lequel nous avons fait un « clustering », nous avons pu identifier de nombreux cas suspects de noms de domaines proches des noms les plus courants, comme par exemple hotmal ou hotmali au lieu de hotmail.

Il n'a néanmoins pas toute la puissance d'outils professionnels. Par exemple, ils permettent en général de générer une nouvelle table à partir d'une existante, là où OpenRefine permet simplement de créer de nouvelles colonnes à partir de colonnes existantes. De plus, il sera avec OpenRefine presque impossible de traiter des fichiers de plusieurs millions d'enregistrements. Par ailleurs, dans OpenRefine, retravailler une fonction de transformation demande des manipulations parfois laborieuses (via l'exportation en JSON puis une édition manuelle), là où les outils professionnels permettent aisément de le faire.

Cependant, dans de nombreuses situations, surtout pour des analyses occasionnelles ou une première phase d'approche, les possibilités d'OpenRefine sont largement satisfaisantes.

Une alternative intermédiaire consiste à s'adresser à Smals, dont le « DQ competence center », propose de réaliser des « data quality run sets » pour un prix réduit avec en plus l'expérience nécessaire (<https://www.smals.be/fr/content/data-quality>).

Budget

OpenRefine est un logiciel gratuit et Open Source.