## Comment nettoyer les données

Un des plus grands défis que l’on rencontre lorsque l’on travaille avec des données est les erreurs. Ceux qui publient les données ne remarquent pas toujours les erreurs parce que les données changent au fil des années. Dans d’autres cas, les erreurs sont humaines et résultent de la mauvaise saisie de données (fautes de frappe ou fausses abréviations par exemple).

Lorsque l’on travaille avec des données, il est important de comprendre comment trouver les erreurs et les corriger afin de rendre les données plus utiles.

##### Dans ce module, nous allons explorer les points suivants :

* Les erreurs de données les plus fréquentes
* Les outils utiles pour nettoyer les données
* Pourquoi nettoyer les données ?

## Révéler la valeur : qu’entend-on par données ‘propres’ ?

Il est important d’avoir des données propres car celles-ci assurent un point de départ fiable à partir duquel vous pouvez travailler les données et générer de la nouvelle valeur.

Des données propres sont des données cohérentes et uniformes, sans doublons et prêtes à être consommées par des machines.

Les données propres facilitent la combinaison de différentes séries de données et permettent une compréhension plus approfondie.

## Les erreurs de données les plus fréquentes

Soyez vigilants face à certaines erreurs de données fréquentes dans chaque ensemble de données avec lequel vous travaillez.

### Le mauvais format pour les dates

On peut écrire les dates de plusieurs manières différentes. Un mélange de différents formats de dates est parfois simple à identifier :

8-sep-2013 vs 8/9/2013

et parfois moins évidents :

8/9/2013 vs 9/8/2013.

Un défi supplémentaire est que certaines applications comme Excel ‘devinent’ le format lorsque l’on ouvre un ensemble de données, ce qui peut masquer le vrai format des dates dans le fichier.

L’erreur la plus commune est de mélanger le format américain (MM/JJ/AA) et européen (JJ/MM/AA).

### Les représentations multiples

Pour gagner du temps lors de la saisie de données, certains termes sont souvent abrégés. Si ces abréviations ne sont pas compatibles, des erreurs peuvent apparaître dans la série de données. Des différences de majuscules, d’espaces, des erreurs d’accord des adjectifs sont autant de facteurs à l’origine d’erreurs.

### Les enregistrements en double

Les doublons apparaissent quand une même donnée est saisie plus d’une fois. Ils se produisent souvent quand des séries de données sont combinées ou parce qu’on ignorait que les données avaient déjà été entrées.

### Les données redondantes

Les données redondantes sont des données qui ne sont pas pertinentes pour votre travail avec un ensemble de données. Un ensemble de données est souvent crée avec un objectif spécifique et comprend ainsi de nombreux détails qui ne vous sont pas nécessaires.

Un exemple très commun de données redondantes est celui des rangées représentant les montants totaux.

D’autres fois, des colonnes des données ont été combinées ou encore reproduites pour améliorer la lisibilité pour l’utilisateur.

### Les échelles numériques cassées

On utilise souvent différentes échelles numériques pour rendre les valeurs numériques plus faciles à lire pour l’utilisateur. Ainsi, pour une série de données au sujet d’un budget, les chiffres sont souvent exprimés en millions. 1200000 s’écrit 1, 2 m. Cependant, les montants plus petits comme 800 000 sont écrits en entier. Une machine lira le plus grand chiffre comme ‘1,2’, ce qui causera des erreurs.

### Les étendues mixtes

Parfois, les données sont mesurées en échelles différentes, comme les tranches d’âge ou les échelles de salaire. Pour qu’une machine comprenne ces différentes échelles, il faut séparer les valeurs fortes des faibles.

### Les fautes d'orthographe

Les fautes d'orthographe sont souvent difficiles à identifier dans les données tabulaires.

N’oubliez pas d’utiliser votre correcteur orthographique pour déceler les erreurs.

## Des outils du nettoyage les données

Sans données propres, on ne peut pas construire d’information. Il existe un nombre d'outils utiles pour vous assister lors du processus de nettoyage des données.

### Des outils du nettoyage les données

Quand vous cherchez des erreurs dans les données, il peut être nécessaire de verifier et reverifier des séries de données entre plusieurs outils de nettoyage et de traitement.

Il est également important de conserver une trace des changements effectués pour les partager ouvertement avec d’autres de manière à ce que tout le monde puisse bénéficier de votre travail, particulièrement si les données que vous nettoyez sont déjà publiées et ouvertes.

Dans cette section, nous allons nous concentrer sur deux des outils principaux et leurs utilisations :

* Open Refine
* Excel & des programmes de feuille calcul semblables.

### Open Refine

Open Refine est un logiciel conçu pour traiter les données brutes. Il s’agit d’un programme, semblable à un navigateur qui analyse des colonnes de données et vous permet de corriger les erreurs d’un ensemble de données en une seule fois. Les erreurs potentielles qu’Open Refine peut corriger comprennent:

* Les formats des dates
* Les représentations multiples
* Les enregistrements en double
* Les données redondantes
* Les échelles numériques cassées
* Les étendues mixtes

### Les programmes de feuille calcul

Open Refine est un outil clé pour nettoyer les données. Cependant, il est parfois plus facile de corriger les erreurs suivantes à travers un programme de feuille calcul :

* Les fautes d'orthographe
* Les données redondantes
* La vérification numérique
* Corriger les données déphasées

## Pourquoi nettoyer les données ?

Au delà des outils qui peuvent être utilisés pour nettoyer l’open data, il est important de comprendre ce que le processus apporte?

### Mettre le nettoyage en priorité

Si les données ne sont pas propres, les décisions prises selon ces données peuvent être erronées. Cela peut affecter les produits ou les idées que vous développez à partir des données, et dans certains cas avoir un impact sur des milliers, voire des millions de personnes.

Une étude récente a démontré que le processus de préparation des données pour l’analyse peut prendre 60 à 80% du temps total nécessaire pour un projet centré sur des données

### Améliorer l’exploitabilité

Partager une série de données ouvertes et propres améliore l’exploitabilité des données. Les outils comme Open Refine vous permettent de partager l’historique des changements, ce qui permet aux utilisateurs de savoir ce qui a été nettoyé et comment. Le fait de rendre non seulement les données mais aussi le processus ouverts augmente la durabilité ainsi que la confiance que l’utilisateur peut avoir en votre ensemble de données.

On assiste aussi à l’émergence de communautés de nettoyage des données comme la communauté OpenStreetMap, communautés qu’il serait bon de soutenir plus activement.

[Lien vers le wiki communautaire danois OSM](http://wiki.openstreetmap.org/wiki/WikiProject_Denmark)

### Découvrir de nouvelles informations

Plusieurs outils de nettoyage de données comme Open Refine permettent de combiner et d’enrichir des ensembles de données ceci peut permettre de découvrir de précieuses nouvelles informations.

Ainsi, enrichir un ensemble de données faisant référence à des entreprises (en les nommant, par exemple) peut vous permettre de savoir si cette entreprise existe toujours et d’en connaitre la position financière actuelle.

[commencer enrichir un ensemble de données](http://training.theodi.org/resources/odt/Enrichingdata.pdf)

## Êtes-vous prêt(e) pour l’open data de qualité ?

Saurez-vous lister les raison pour lesquelles les données doivent être nettoyées, quels outils utiliser et quelles sont les erreurs les plus courantes dans des ensembles de données?

## Pourquoi nettoyer les données ?

Les données ouvertes brutes, ou ‘non nettoyées’, posent un problème parce qu’elles...

* ne peuvent pas être reportées sur un graphique
* ne peuvent pas être publiées
* peuvent donner lieu à des conclusions erronées

##### Vous avez raison !

L’open data brut pose un problème parce qu’il peut donner lieu à des conclusions erronées. Cela ne devrait pas empêcher la publication ou l’utilisation des données, cependant il faut les nettoyer.

##### Êtes vous sûr(e) ?

L’open data brut pose un problème parce qu’il peut donner lieu à des conclusions erronées. Cela ne devrait pas empêcher la publication ou l’utilisation des données, cependant il faut les nettoyer.

## Quels sont deux des outils les plus utiles pour nettoyer l’open data ?

Quels sont les deux outils les plus utiles pour nettoyer l’open data ?

* Microsoft Access et Adobe Photoshop
* Apple Notes et Microsoft Word
* Microsoft Excel et Open Refine

##### Vous avez raison !

Open refine et Excel (ou tout programme de feuille de calcul similaire) sont les plus utiles pour nettoyer l’open data.

##### Êtes vous sûr(e) ?

Open refine et Excel (ou tout programme de feuille de calcul similaire) sont les plus utiles pour nettoyer l’open data.

## Reconnaître des erreurs

Lesquels des points suivants sont des erreurs des données qu’il faut nettoyer ?

* Des échelles numériques mixtes
* Une manque de granularité
* Des données redondantes
* Aucune licence
* Des représentations multiples

##### Vous avez raison !

Des échelles numériques mixtes, des données redondantes et des représentations multiples des mêmes données ne sont que quelques erreurs de données non-propres qu’il faut corrigir. Bien que la manque de granularité et une licence soit un problème, elle n’est pas directement liée au fait que les données ne sont pas propres.

##### Êtes vous sûr(e) ?

Des échelles numériques mixtes, des données redondantes et des représentations multiples des mêmes données ne sont que quelques erreurs de données non-propres qu’il faut corrigir. Bien que la manque de granularité et une licence soit un problème, elle n’est pas directement liée au fait que les données ne sont pas propres.

## Combien de temps devrait-on consacrer au nettoyage et à la préparation des données ?

Combien de temps devrait-on consacrer au nettoyage et à la préparation des données pour tout projet centré sur des données?

* Entre 20% et 40%
* Entre 40% et 60%
* Entre 60% et 80%

##### Vous avez raison !

Tout porte à croire que le processus de nettoyage, transformation et préparation des données pour l’analyse prend entre 60 et 80% du temps total d’un projet centré sur des données.

##### Êtes vous sûr(e) ?

Tout porte à croire que le processus de nettoyage, transformation et préparation des données pour l’analyse prend entre 60 et 80% du temps total d’un projet centré sur des données.

## Le nettoyage des données

Le nettoyage des données est essentiel pour garantir que les bonnes conclusions sont tirées à partir des données.

Ceux qui publient les données ne remarquent pas toujours les erreurs parce que les données changent au fil des années. Dans d’autres cas, les erreurs sont humaines et résultent de la mauvaise saisie de données (fautes de frappe ou fausses abréviations par exemple).

Notez que le nettoyage des données risque de prendre plus de temps que vous ne pensez. Un processus de nettoyage minutieux vous fera gagner du temps sur le long terme.

Complétez notre exercice de nettoyage de données pour découvrir Open Refine et constatez par vous-même à quel point ce programme est facile à utiliser.