# Gestion des Données Ouvertes dans L'Agriculture et la Nutrition

*Ce cours en ligne est le fruit d'une collaboration entre les partenaires de GODAN Action, y compris Wageningen Environmental Research (WUR), AgroKnow, AidData, l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO), le Forum Mondial sur la Recherche Agricole (GFAR), l'Institut des Etudes du Développement (IDS), le Land Portal, l'Open Data Institute (ODI) et le Centre Technique de Coopération Agricole et Rurale (CTA).*



*GODAN Action est un projet de trois ans du Département pour le Développement International du Royaume-Uni pour permettre aux utilisateurs, producteurs et intermédiaires de données de s'engager efficacement avec les données ouvertes et maximiser leur potentiel d'impact dans les secteurs de l'agriculture et de l'alimentation. Nous travaillons en particulier à renforcer les capacités, à promouvoir des normes communes et les meilleures pratiques et à améliorer la manière dont nous mesurons l'impact. [www.godan.info]*

# UNITE 2 : UTILISATION DES DONNES OUVERTES

# LECON 2.2 : QUALITE ET PROVENANCE

Photo par Département de l'agriculture des États-Unis Sous licence CC BY 2.0

## Objectifs et résultats d'apprentissage

Les données ouvertes deviennent utilisables lorsqu'un humain peut les comprendre et qu'une machine peut les manipuler. Les utilisateurs de données ouvertes ont besoin de l'autorisation de leur éditeur, accordée par une licence ouverte. Mais la licence ouverte ne suffit pas à elle seule pour garantir la qualité et la maniabilité des données. Il y a peu de chance que les gens puissent traiter des données pour tirer des enseignements et démontrer des avantages si elles ne sont pas utilisables.

À la fin de cette leçon, vous devrez être en mesure :

- De répertorier, décrire et appliquer les facteurs qui affectent l'utilisation des données ouvertes
- D'utiliser des outils qui vous aident à évaluer la facilité d'utilisation des données ouvertes
- De vérifiez vos données par rapport à d'autres sources de données
- D'identifier la provenance d'un ensemble de données ouvert (lisible par machine / lisible par l'homme)

- De décrire les données utilisables selon différents points de vue
- Sommaire



- Liste des figures
-

# 1. Introduction

La qualité et la provenance sont deux aspects importants qui déterminent l'utilisabilité d'un ensemble de données. Cette leçon porte sur les différents aspects qui constituent un ensemble de données de qualité et sur un certain nombre de lignes directrices au sujet des meilleures pratiques qui facilitent la publication de données utilisables de haute qualité.

Une partie de la qualité d'un ensemble de données est dictée par l'historique, ou la provenance, de cet ensemble de données. Il est souvent plus important de savoir que les données proviennent d'une source fiable et qu'elles ont été recueillies grâce à des méthodes fiables (ou au moyen de méthodes comportant des contraintes connues) que d'avoir un ensemble de données avec un vocabulaire ou un schéma bien contrôlé.

Bien que les données soient de nature technique, toutes les mesures de qualité ne sont pas techniques ; s'assurer qu'un ensemble de données est lisible par machine ne signifie pas toujours qu'il est utilisable ; les exigences de qualité ne sont pas toutes techniques. L'évolution des lignes directrices sur les meilleures pratiques en est le reflet et, dans cette leçon, nous examinerons les aspects techniques et non techniques qui font un ensemble de données utilisable de haute qualité.

# 2. Marques et mesures de qualité

L'évaluation de la qualité des données ouvertes ne peut pas être effectuée rapidement. Il existe un certain nombre de normes et de labels de qualité, basés sur la communauté, qui peuvent vous aider à évaluer la qualité et la maniabilité des données.

L'une des premières marques de qualité à émerger pour les données ouvertes est les « 5 étoiles des données ouvertes liées ». L'attribution de chaque étoile est séquentielle et commence par l'obligation d'appliquer une licence ouverte aux données. Les étoiles restantes sont divisées en deux et se concentrent sur les données ouvertes disponibles sur le Web à télécharger et deux qui se concentrent sur les données étant « sur le web » à utiliser via une API qui peut instantanément récupérer des ressources. En dehors de l'aspect concernant les licences, la ligne directrice 5 étoiles est axée sur la disponibilité technique des données et ne comporte pas les aspects non techniques qui constituent des données utilisables de qualité.

De même, les principes FAIR [1] fournissent une ligne directrice similaire pour les éditeurs. Selon ces principes, les données devraient être : **trouvables, accessibles, interopérables et réutilisables (FAIR)**. Bien qu'ils se concentrent en

[1] Force11. Guiding Principles For Findable, Accessible, Interoperable And Reusable Data Publishing Version B1.0. https://www.force11.org/fairprinciples

grande partie sur des aspects techniques, les principes énoncent cependant certains termes relatifs à la provenance, indiquant que « les objets de données publiés doivent se référer à leurs sources avec des métadonnées et une provenance suffisamment riche pour permettre une **citation correcte** ».

Et le schéma 5 étoiles et les principes FAIR sont décrits plus en détail dans l'unité 4 : Échange de données ouvertes. Cette leçon va se pencher sur les autres aspects de la maniabilité et de la qualité tels que définis par l'Open Data Institutes Open Data Certificates.

# 3. Certificats de données ouvertes

L'Open Data Certificate (Certificat de données ouvertes) est un outil en ligne gratuit développé et maintenu par l'Open Data Institute, pour évaluer et reconnaître la publication durable de données ouvertes de qualité. Il aborde les aspects juridiques, pratiques, techniques et sociaux de la publication de données ouvertes en utilisant les meilleures pratiques.

Comme FAIR, le processus d'Open Data Certificates [2] prend une vision alternative mais complémentaire du schéma 5 étoiles. Un certificat mesure l'efficacité avec laquelle quelqu'un partage un ensemble de données pour en faciliter la réutilisation. La portée couvre plus que de simples problèmes techniques [3] y compris les droits et les licences, la documentation et les garanties concernant la disponibilité. Un certificat offre donc une évaluation plus arrondie de la qualité de publication d'un ensemble de données.

Pour les éditeurs de données, le processus d'évaluation d'un ensemble de données fournit un aperçu de la façon dont ils pourraient améliorer leur processus de publication. Le processus d'évaluation est donc précieux en lui-même, mais le certificat produit a également de la valeur pour les ré utilisateurs.

## 3.1. Être un ré utilisateur de données

Pour les utilisateurs de données, la qualité technique au point d'utilisation peut être suffisante pour leur utilisation particulière. Cependant, pour les ré utilisateurs, la durabilité et le soutien seront probablement plus dominants dans le processus de prise de décision que le format de fichier.

Les ré utilisateurs doivent avoir l'assurance que leur accès aux données sera cohérent et fiable. Un certificat de données ouvertes défie les éditeurs de penser au-delà des données afin de prendre en compte les principales considérations politiques dans la prise en charge des données. Ces considérations sont réparties en quatre catégories essentielles pour les ré utilisateurs.

[2] https://certificates.theodi.org
[3] https://certificates.theodi.org/about

## 3.2. Liste de vérification du ré utilisateur

À l'aide des certificats de données ouvertes, vous trouverez ci-dessous une liste de vérification des réutilisateurs pour les données ouvertes. La liste de vérification est divisée en quatre catégories reflétant les sections du certificat de données ouvertes.

### Légales

- Les données sont-elles sous licence ouverte et légalement utilisables ?
- Le modèle de données, le format ou la structure sont-ils également sous licence ouverte et légalement utilisables ?
- Les mentions de copyright sont-elles claires ?
- Y a-t-il des données ou des parties de données qui ne sont pas sous licence explicite ?
- Y a-t-il de la confidentialité et des contraintes éthiques potentielles à l'utilisation des données ?

### Practiques

- Les données sont-elles bien décrites ?
- La raison pour laquelle les données sont collectées est-elle claire ?
- L'utilisation des données par l'éditeur est-elle claire ?
- D'autres utilisations des données sont-elles décrites ?
- Les données sont-elles accessibles ?
- Les données sont-elles horodatées ou à jour ?
- Les données seront-elles disponibles pendant au moins un an ?
- Les données seront-elles mises à jour régulièrement ?
- Existe-t-il un processus de contrôle qualité ?

### Techniques

- Les données sont-elles disponibles dans un format adapté au contenu ?
- Les données sont-elles disponibles à partir d'un emplacement cohérent ?
- Les données sont-elles bien structurées et lisibles par machine ?
- Les termes et les acronymes complexes sont-ils définis dans les données ?
- Les données utilisent-elles un schéma ou une norme de données ?
- Existe-t-il une API disponible pour accéder aux données ?

### Social

- Existe-t-il une communauté d'utilisateurs de données ?
- Les données sont-elles déjà utilisées par un grand nombre de personnes ?
- Les données sont-elles officiellement prises en charge ?
- Des accords sur les niveaux de service sont-ils disponibles pour les données ?

- Sait-on clairement qui maintient et qui peut être contacté à propos des données ?

Si une publication a rempli un certificat de données ouvertes et appliqué la marque de qualité à ses données, un ré utilisateur peut rapidement trouver les réponses à toutes les questions de la liste de vérification de ré utilisateur ci-dessus. Une liste complète de tous les certificats est également disponible dans le registre d'ensembles de données certifiés [4].

## 3.3. Guide de provenance à l'intention des ré utilisateurs

Toutes les lignes directrices sur les meilleures pratiques présentées jusqu'à présent sont quelque peu axées sur les producteurs de données, qui possèdent et gèrent déjà les données sources. Cependant, toutes les données ouvertes ne proviennent pas de l'éditeur. Une grande quantité de données est dérivée d'autres données. Par exemple, une prévision météorologique est dérivée en appliquant des modèles complexes à des données météorologiques. Avec de nombreuses sources de données météorologiques et de nombreuses organisations qui utilisent différents modèles pour créer une prévision, cela peut conduire à des situations où même les prévisions basées sur les mêmes données d'entrée peuvent être très différentes, avec des conséquences potentiellement dévastatrices[5].

### *3.3.1. Liste de vérification de la provenance*

La liste de vérification ci-dessous aidera à établir la provenance d'un ensemble de données et aidera à établir le niveau de confiance dans cet ensemble de données.

- Les données sont-elles entièrement détenues et produites par le fournisseur de données ?
- Est-ce que quelqu'un d'autre produit des données comparables pour une vérification croisée ?
- Sait-on clairement si les données proviennent d'autres sources de données ?
- Les autres sources de données sont-elles claires ?
- Les autres sources de données sont-elles fiables et comparables aux autres fournisseurs de données ?
- Sait-on clairement si et comment les données ont changé (de n'importe quelle source) avant d'être disponibles en tant que données ouvertes à votre point d'accès ?

[4] https://certificates.theodi.org/en/datasets
[5] http://www.bbc.co.uk/news/av/world-24713504/michael-fish-s-denial-of-ahurricane-in-1987

Suivre ces points aidera à établir la fiabilité des différentes sources de données ouvertes. Cela peut même révéler un potentiel pour contourner la source de données actuelle et suivre la trace de la provenance vers une source originale qui peut être plus fiable ou offrir un service de données plus complet et / ou prise en charge.

# 4. Vérification de la qualité après l'accès

L'établissement de la confiance à un fournisseur de données, l'obtention éventuelle d'un contrat de service et l'accès aux données ne sont pas la fin du processus de vérification de la qualité. Une fois que l'accès aux données a été obtenu, il est essentiel de vérifier les données, surtout avec votre propre usage à l'esprit.

Cette étape de vérification et de préparation des données pour qu'elles soient prêtes à l'emploi comporte de nombreux aspects, la plupart basés sur des processus techniques. Il est plus que probable que des problèmes ou des incohérences seront trouvés dans les données à ce stade. C'est à ce moment qu'il est impérieux d'être connecté à la communauté ou de fournir et/ou d'avoir accès à des procédures de contrôle de la qualité pour vous aider à comprendre les données ou pour résoudre ces problèmes.

## 4.1. Validation des données

L'une des premières choses à faire avec les données est de les vérifier par rapport à tout schéma de données disponible et à la description de la structure de données. Cela révélera des omissions potentielles de la documentation pour les termes apparus dans les données. La vérification des données par rapport au schéma aidera également à vérifier que les données correctes existent dans l'ensemble de données.

Si vos données n'ont pas de schéma, il peut être nécessaire de transformer les données dans un format où vous pouvez facilement les appliquer ou créer votre propre schéma pour l'ensemble de données.

### *4.1.1. Conception de schémas*

Un schéma est un plan de données qui définit un ensemble de contraintes et de règles d'intégrité relatives à la structure et au contenu d'une ressource de données. Un schéma définit trois éléments clés répertoriés ci-dessous.

*Colonne / Titres clés dans les données* : La définition d'un ensemble cohérent de titres de colonnes (ou de clés) pour un ensemble de données est essentielle pour garantir que les ensembles de données du même type peuvent être fusionnés et analysés facilement. Souvent, les titres de colonnes changent ou sont abrégés pour gagner du temps, mais cela entraîne de nombreux problèmes lors de l'analyse de données sur de longues périodes. L'ajout de titres

de colonnes est moins problématique mais doit être pris en compte lors de l'analyse des données.

*Types de valeurs* : Avec le titre de la colonne / les clés définis, il est important de définir le type de données valides pour les valeurs, par exemple : nombre, texte, date, coordonnée, etc.

*Contraintes de valeur* : Avec le type de valeur défini, des contraintes valides telles que la nécessité d'être unique, d'être dans une certaine unité (par exemple, gallons (UK)) ou d'être dans une certaine plage doivent être définies. Par exemple, une colonne pourrait être intitulée « Coût (£ m) » ; donc toutes les valeurs devraient être des nombres (sans virgules). Définir une plage valide permet également d'éviter et d'expliquer toute erreur dans les données. Par exemple, définir une plage de 0,001 à 100 sur le 'Coût (£ m)' (si l'on sait que le coût ne peut pas dépasser 100 millions de livres sterling). La validation de la portée empêche les utilisateurs de lire accidentellement le titre / les unités de la colonne et d'entrer 100000000 au lieu de 100 pour £ 100m.

### *4.1.2. Utilisation des schémas*

Tous les tableurs vous permettent de créer des règles de validation pour les données relatives à chaque titre de colonne. Cependant, très peu de progiciels permettent l'exportation de schémas à côté des données pour les autres utilisateurs sans ajouter d'extensions de développeurs complexes.

L'une des principales raisons à cela est la connexion entre les schémas et les bases de données hiérarchiques datant de la conception de la base de données précédente par E F Codd (1970). Les schémas sont un élément clé d'une base de données dans laquelle plusieurs tables sont liées à des relations prédéfinies. Depuis ce point, des implémentations ont été menées par ce modèle théorique et appliquées techniquement dans des progiciels de bases de données relationnelles comme MySQL.

Le développement de l'eXtensible Markup Language comme mécanisme de partage des données a émergé beaucoup plus tard en 1996. Cinq ans plus tard, en 2001, la spécification Schéma XML, désormais populaire, a été publiée pour faciliter le partage de données cohérentes et vérifiables. Du milieu des années 90 à la fin des années 2000 et encore aujourd'hui, le format XML était la norme de choix pour la représentation des données échangeables ; il a été conçu pour être lisible à la fois par machine et par l'homme.

Les données d'exemple suivantes (présentées ici sous forme de tableau) sont un exemple de données extraites d'un ensemble de données.

*Tableau1 Exemple de données*

| **Prénom** | **Nom** | **Instrument** | **Date de naissance** |
|---|---|---|---|
| John | Lennon | Vocal | 09 -10-1940 |
| Paul | McCartney | Guitare basse | 18 -06-1942 |
| George | Harrison | Guitare | 25 -02-1943 |
| Ringo | Starr | Batterie | 07 -07-1940 |

Les mêmes données sont montrées ci-dessous (Figure 1) en XML avec des extraits associés du schéma.

XML (Télécharger)

```
<People>
   <Person>
      <FirstName>John</FirstName>
      <LastName>Lennon</LastName>
      <Instrument>Vocal</Instrument>
      <DateOfBirth>1940-10-09</DateOfBirth>
   </Person>
   <Person>
      <FirstName>Paul</FirstName>
      <LastName>McCartney</LastName>
      <Instrument>Bass Guitar</Instrument>
      <DateOfBirth>1942-06-18</DateOfBirth>
   </Person>
   <Person>
      <FirstName>George</FirstName>
      <LastName>Harrison</LastName>
      <Instrument>Guitar</Instrument>
      <DateOfBirth>1943-02-25</DateOfBirth>
   </Person>
   <Person>
      <FirstName>Ringo</FirstName>
      <LastName>Starr</LastName>
      <Instrument>Drums</Instrument>
      <DateOfBirth>1940-07-07</DateOfBirth>
   </Person>
</People>
```

Extrait de schéma XML (Télécharger)

```
...
<xs:simpleType name="birthsDate">
   <xs:restriction base="xs:date">
      <xs:minInclusive value="1800-01-01"/>
      <xs:maxInclusive value="2017-07-31"/>
      <xs:pattern value=".{10}"/>
   </xs:restriction>
</xs:simpleType>
<xs:simpleType name="instrument">
  <xs:restriction base="xs:token">
     <xs:enumeration value="Vocal"/>
     <xs:enumeration value="Guitar"/>
     <xs:enumeration value="Bass Guitar"/>
     <xs:enumeration value="Drums"/>
   </xs:restriction>
</xs:simpleType>
...
<xs:complexType>
   <xs:sequence>
      <xs:element name="FirstName" type="xs:string"/>
      <xs:element name="LastName" type="xs:string"/>
      <xs:element name="Instrument" type="instrument"/>
      <xs:element name="DateOfBirth" type="birthsDate"/>
   </xs:sequence>
</xs:complexType>
...
```

*Figure 1 Exemple de données en format XML avec les extraits associés du schéma*

Un tel schéma peut être utilisé par des machines pour valider automatiquement la structure et le contenu d'un ensemble de données; il y a aussi un outil de validation en ligne [6] qui peut être utilisé avec les deux exemples ci-dessus.

[6] http://www.utilities-online.info/xsdvalidation/#.WhK94rZ0cUs
[7] http://json-schema.org

Bien que le XML ait été largement adopté, 2005 a été une année importante pour les données sur le Web avec l'émergence d'Ajax (Garrett 2005). Permanent pour JavaScript et XML, Ajax est un ensemble de techniques de développement web qui utilise JavaScript pour charger dynamiquement des données dans des applications web. L'objectif initial était de permettre l'utilisation dynamique des données XML dans les applications Web et d'encourager davantage la diffusion des données. Dans la pratique, les applications modernes substituent couramment JSON (Notation d'objets JavaScript) à XML en raison des avantages de JSON provenant de JavaScript, 21% plus rapide à travailler avec comme résultat et beaucoup moins verbeux.

De même que pour XML, JSON fournit une spécification JSON Schema[7]; Cependant, ceci n'est pas encore entièrement développé à la spécification XML Schema et n'a pas la capacité de valider une plage d'entrées (par exemple les valeurs minimum et maximum permises). La figure 2.2.2 ci-dessous montre les mêmes données que l'exemple XML, cette fois en JSON avec le schéma JSON équivalent.

| JSON | Schéma JSON (Télécharger) |
|---|---|
| <pre>{<br>"FirstName": "John",<br>"LastName": "Lennon",<br>"Instrument": "Vocal",<br>"DateOfBirth": "1940-10-09"<br>},<br>{<br>"FirstName": "Paul",<br>"LastName": "McCartney",<br>"Instrument": "Bass Guitar",<br>"DateOfBirth": "1942-06-18"<br>},<br>{<br>"FirstName": "George",<br>"LastName": "Harrison",<br>"Instrument": "Guitar",<br>"DateOfBirth": "1943-02-25"<br>},<br>{<br>"FirstName": "Ringo",<br>"LastName": "Starr",<br>"Instrument": "Drums",<br>"DateOfBirth": "1940-07-07"<br>}<br>]</pre> | <pre>{<br>"fields":<br>[<br>{<br>},<br>{ "name": "FirstName",<br>"type": "string",<br>"constraints": {<br>"required": true<br>}<br>"name": "LastName",<br>},<br>{ "constraints": {<br>"required": true<br>}<br>"name": "Instrument",<br>"enum": ["Vocal", "Guitar", "Bass Guitar", "Drums"],<br>"constraints": {<br>"required": true<br>},<br>}<br>{<br>"name": "DateOfBirth",<br>"type": "string",</pre> |

```
            "format": "date-time"
        }
    ] }
```

*Figure 2 Exemple de données en format JSON avec les extraits associés du schéma*

Voici le schéma JSON pour notre fichier de données tabulaires. Essayez-le par vous-même en téléchargeant les données tabulaires [7] (dans CSV) et schéma[8][9] (dans JSON) et utiliser ensuite csvlint[10] (un validateur de données tabulaires) pour valider que les données sont conformes au schéma.

L'un des défis avec ces deux formats est l'exigence d'avoir deux fichiers : les données et le schéma. Cela conduit à la situation prévisible où les données sont conservées et partagées, mais le schéma est oublié ou perdu. L'utilisation d'espaces de noms et de données liées comme solution technique permet de fournir une solution, mais le principal problème réside toujours dans le manque d'intégration de ces normes dans les progiciels de gestion des données disponibles sur le marché.

Les tableurs comme Excel ont tendance à rendre les données de formatage et la mise en place de règles de validation trop complexes, comme en témoigne la facilité de les configurer dans des outils comme airtable[11]. Encore une fois, cependant, l'exportation du schéma n'est actuellement pas possible, pas plus que l'utilisation de schémas référencés et liés.

Il en est de même pour les logiciels de base de données et le passage aux structures de base de données noSQL qui ne sont pas contrôlées par un schéma relationnel étroitement défini. Alors que le passage à de nouvelles structures de base de données plates est bon pour la vitesse lors de la gestion de grands ensembles de données ; l'utilisation d'espaces de noms pour définir des schémas fait toujours défaut.

Les schémas ont clairement leurs avantages et devraient être adoptés là où les organisations dépendent fortement de données de haute qualité.

## 4.2. Nettoyage des données

L'un des plus grands défis lorsque vous travaillez avec des données est le traitement des erreurs. Souvent, les éditeurs de données ne remarquent même pas les erreurs, car les données peuvent changer au fil des années. Dans d'autres cas, les erreurs peuvent résulter d'erreurs humaines lors de la saisie de données, telles que des fautes de frappe ou des abréviations incorrectes.

---

[7] http://odedu.learndata.info/D5/course/en/assets/d466f522c9b280bdf8e0050e0899efe7cb184c9d.csv

[8] http://odedu.learndata.info/D5/course/en/assets/e367089a2d91deea74496387877

[9] d8b5a91e.json

[10] http://csvlint.io

[11] https://airtable.com

Même lorsqu'un schéma est disponible, des erreurs et des incohérences peuvent exister dans les données. Lorsque vous travaillez avec des données, il est important de savoir comment trouver les erreurs et les corriger pour rendre les données plus utiles.

Cette section présente un certain nombre d'exemples d'erreurs et d'incohérences différents dans les données, des contours qui peuvent être corrigés avec la validation du schéma et qui nécessitent un outil plus avancé comme Open Refine[12].

### *4.2.1. Formats de date incorrects*

Les dates peuvent être écrites de manière incohérente et selon différentes normes. L'une des plus grandes confusions existe entre les manières d'écrire des dates aux États-Unis et au Royaume-Uni. Aux États-Unis, le mois arrive en premier, puis le jour (par exemple, 12/30/2017), alors que c'est l'inverse au Royaume-Uni (30/12/2017). Ceci est facile à repérer lorsque le jour est plus grand que le 12, mais peut entraîner une confusion dans le cas contraire (par exemple, 6/7/2017 ?).

La norme ISO 8601 spécifie une série de règles pour écrire des dates et des heures pour résoudre ce problème et d'autres. L'ISO 8601 spécifie que les dates doivent être écrites avec l'année en première position (par exemple AAAA-MM-JJ HH : ii : ss). Non seulement est-il toujours facile à lire, mais il fonctionne également comme un moyen de trier dans l'ordre des dates, le facteur de tri le plus important étant le premier.

Comme les dates sont compliquées, des efforts ont été faits au fil des années pour cacher la gestion des dates aux utilisateurs. Par exemple, si vous tapez " 8-7 " dans n'importe quelle cellule générale d'Excel, cela se traduira automatiquement par une date et enregistrera " 08-Jul' " dans un CSV, ce qui n'est pas une norme ISO. Le formatage des cellules comme date permet le formatage spécifique, mais le mélange des dates américaines et britanniques est toujours possible et visible uniquement lorsque le contenu est aligné à gauche (incorrect) ou à droite (correct) dans la cellule. Ce qui est encore plus inquiétant, c'est ce qui se passe lors de l'importation de données CSV dans Excel, par exemple une date ISO standard (AAAA-MM-JJ HH:ii:ss) sera traduite dans un format de date personnalisé (JJ/MM/AAAA HH:ii:ss) et ce format sera à nouveau enregistré dans votre fichier une fois la modification terminée dans Excel, à moins que le format soit modifié avant son enregistrement. La plupart des utilisateurs ignorent totalement qu'Excel a effectué cette traduction avant d'afficher les données.

Les dates sont difficiles à gérer, en particulier lorsque le logiciel fait des hypothèses au nom de l'utilisateur. Les schémas peuvent aider à condition que

[12] http://openrefine.org

les données correspondent au format requis et que les traductions n'aient pas lieu quelque part entre les deux.

#### *4.2.2. Représentations multiples*

Les gens essaient souvent de gagner du temps lorsqu'ils saisissent des données en abrégeant les termes. Si ces abréviations ne sont pas cohérentes, cela peut entraîner des erreurs dans l'ensemble de données. Les schémas qui utilisent des listes prédéfinies de termes acceptables peuvent aider ici, à condition que les utilisateurs ne soient pas en mesure d'ajouter facilement à la liste et de recréer ainsi la même erreur.

D'autres erreurs qui existent dans cette catégorie incluent des différences dans la capitalisation, l'espacement, le genre et la pluralisation des adjectifs (par exemple, councilman vs councilmen), qui peuvent tous causer des problèmes sérieux.

#### *4.2.3. Enregistrements en double*

Un enregistrement en double est l'endroit où la même donnée a été saisie plus d'une fois. Les enregistrements en double se produisent souvent lorsque les ensembles de données ont été combinés ou parce qu'il n'était pas connu qu'il y avait déjà une entrée. De plus, la duplication d'enregistrements peut se produire lorsqu'une personne peut être désignée par deux noms (par exemple Dave et David). Cela peut conduire à des cas où les enregistrements doivent être désambiguïsés pour découvrir si les auteurs des publications sont identiques ou différents. Ce type d'erreur ne peut pas être intercepté avec une validation de schéma.

#### *4.2.4. Données redondantes ou combinées*

Les données redondantes sont tout ce qui n'est pas pertinent pour votre travail avec l'ensemble de données. Souvent, un ensemble de données a été créé dans un but précis, ce qui nécessite des détails dont vous n'avez pas besoin. Les occurrences courantes de données redondantes incluent les lignes qui représentent les montants totaux. Ils apparaissent souvent lorsqu'un ensemble de données dans Excel a été exporté dans d'autres formats sans que la ligne 'Total' ne soit d'abord supprimée. À d'autres moments, des colonnes de données ont été combinées ou reproduites afin d'aider à la lisibilité humaine.

#### *4.2.5. Utilisation mixte des échelles numériques*

Les valeurs numériques dans les ensembles de données utilisent souvent des échelles différentes pour faciliter la lecture par l'homme. Dans les ensembles de données budgétaires, par exemple, les unités sont souvent en millions. 1 200 000 $ devient souvent 1,2 million de dollars. Cependant, de plus petits montants comme 800 000 $ sont encore écrits en entier. Pour une machine, cela signifie qu'ils lisent la plus grande valeur en $ 1,2, ce qui provoque des erreurs. Sinon, si la colonne est censée être en millions, alors le deuxième chiffre devient 800 000 000 000 $.

Malheureusement, les schémas ne permettent pas de repérer ce type d'erreur. Cela est dû au fait que toutes les valeurs pourraient être la valeur et des erreurs pourraient être causées à n'importe quel niveau. La définition de limites peut aider, mais pourrait ne pas résoudre le problème. Rendre les unités de mesure claires au point de collecte ou d'utilisation est essentiel ici pour garantir la cohérence des données et éviter les catastrophes [13].

#### *4.2.6. Plages numériques*

Les données sont parfois mesurées en plages, telles que l'âge ou l'échelle salariale. Pour qu'une machine comprenne ces plages, il est important de séparer les valeurs haute et basse pour faciliter l'analyse. Il pourrait même être nécessaire de créer de nouvelles parenthèses si elles ont changé au fil des ans (par exemple, l'âge des personnes ou l'âge de la retraite a augmenté).

#### *4.2.7. Fautes d'orthographe*

Si vous avez beaucoup de textes libres dans les données, il est important de vérifier l'orthographe cohérente pour vous assurer que l'analyse peut être effectuée avec d'autres ensembles de données. L'orthographe peut ne pas être "correcte", par exemple colour par rapport à color, cependant, il devrait être cohérent de permettre aux ensembles de données d'être comparables et interopérables.

## 4.3. Boîte à outils pour le nettoyage des données

Lorsque vous recherchez des erreurs dans les données, il peut être nécessaire de télécharger et de charger des ensembles de données dans de nombreux outils de nettoyage et de traitement. Il est également important de garder une note sur les modifications apportées et de les partager ouvertement avec les autres afin que tout le monde puisse bénéficier de votre travail, en particulier si les données que vous nettoyez sont des données ouvertes existantes qui ont été publiées.

Nous avons déjà examiné un certain nombre d'outils de validation de schéma, mais il existe un certain nombre d'autres outils qui peuvent aider à effacer les données incohérentes.

**OpenRefine**

OpenRefine est un outil logiciel conçu pour gérer les données non nettoyées. L'outil est un navigateur basé sur des colonnes qui vous permet de corriger les erreurs sur l'ensemble d'un ensemble de données ouvertes en une seule action. Les erreurs pouvant être corrigées incluent:

- Les formats de date
- Les représentations multiples
- Les enregistrements en double

[13] http://www.nytimes.com/1983/07/30/us/jet-s-fuel-ran-out-after-metric-conversionerrors.html?mcubz=1

- Les données redondantes
- Les échelles numériques mixtes
- Les plages mixtes.

### Programmes de tableurs

OpenRefine est un outil clé pour le nettoyage des données. Cependant, il est parfois plus facile de corriger certaines erreurs dans un programme de tableur :

- Les fautes d'orthographe
- Les données redondantes
- La vérification numérique
- La correction des données décalées.

### Autres outils

- Drake[14]
- Data Wrangler[15] (outil développé conjointement par les universités de Stanford et Berkeley aux États-Unis)
- Data Cleaner[16]
- WinPure[17]

# Résumé

Pour que les données soient utilisables, il y a bien plus que de simples données techniques. De nombreuses directives essaient de simplifier les exigences pour que les données soient vraiment utilisables et chacune a ses mérites. La liste de vérification créée à partir du travail d'Open Data Certificates montre l'ampleur du défi que représente la production de données exploitables de haute qualité. L'Open Data Certificate est conçu pour évaluer les aspects juridiques, pratiques, techniques et sociaux de la publication de données ouvertes en utilisant les meilleures pratiques. Cependant, même cela ne tient pas compte de la vérification croisée des sources de données pour s'assurer que les bonnes données sont utilisées. La provenance des données est également essentielle, mais encore une fois difficile à suivre entièrement sur un ensemble de données.

Une fois que vous avez établi la confiance sur un ensemble de données, la vérification de son contenu est le prochain défi en attente. La création et l'adoption de schémas et de validateurs de données aident dans une certaine mesure, mais la nécessité de nettoyer et de valider les données (potentiellement à la main) ne disparaîtra pas de sitôt. En fait, comme nous le verrons dans la leçon suivante, le processus de préparation des données prêtes à l'analyse pourrait prendre jusqu'à 80% de votre temps !

[14] https://github.com/Factual/drake
[15] http://vis.stanford.edu/wrangler/
[16] http://datacleaner.org
[17] http://www.winpure.com/article-datacleaningtool.html

# Références


- Dodds, L. 2015. Comparing the 5-star scheme with Open Data Certificates. Available at: https://theodi.org/blog/5-star-open-data-certificates-timberners-lee
- Codd, E. F. 1970. A relational model of data for large shared data banks. *Communications of the ACM* **13 (**6), 377–387.
- Garrett, J. J. 2005. Ajax: A new approach to web applications. Available at: http://adaptivepath.org/ideas/ajax-new-approach-web-applications/