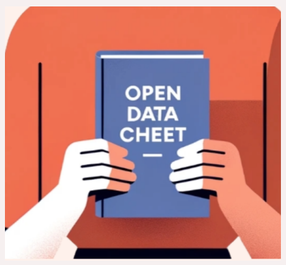


Open Data *Cheat Sheet*

Par DATACTIVIST

Les fondamentaux

Bases légales

Loi pour une République Numérique
2016, oblige les établissements publics à ouvrir leurs données par défaut

Article 15 de la Déclaration des Droits de l'Homme et du Citoyen
La société a le droit de demander compte à tout agent public de son administration.

Loi CADA
Elle est chargée de veiller à la liberté d'accès aux documents administratifs

Textes fondateurs

2005 : Open Definition
Une définition juridique des droits de l'usager d'un savoir ouvert

2007 : la rencontre de Sebastopol
Une déclaration définissant les grands principes de l'Open Government Data

2011 : la déclaration pour un gouvernement ouvert
75 pays se sont rassemblés pour définir les principes d'un gouvernement ouvert

Sites utiles

data.gouv.fr
Les données ouvertes de l'Etat et des administrations françaises

teamopendata.org
Le forum de la communauté open data francophone

open.datactivist
Des contenus en accès libre sur l'open data

Ressources d'Open Data France
Outils et guides en accès libre

Réutilisations de données

Yuka
S'informer sur les produits alimentaires grâce aux données d'Open Food Facts

Les demandes de valeurs foncières
Estimer et suivre les prix de l'immobilier grâce aux données ouvertes par l'Etat

Les applications de mobilité
Trouver des itinéraires et se déplacer plus facilement grâce aux données ouvertes par les exploitants de transports en commun

Intérêt de l'open data

Améliorer les politiques publiques

- Valorisation du travail des agents et de l'administration
- Servir l'intérêt général (pour tous et toutes)
- Assurer la continuité de service
- Gain de temps
- Levier d'innovation et changement interne

Encapaciter les citoyens

- Transparence
- Concertation
- Co-production
- S'assurer qu'on est pris en compte

Des opportunités économiques

- Création de services innovants
- Alimenter la recherche et développement

Check list qualité

- **Le contenu des données est satisfaisant**
- **Les métadonnées sont consultables**
- **Le jeu de données ne contient pas de données à caractère personnel**
- **Les données répondent aux besoins des utilisateurs internes et externes**
- **Le format des données facilite leur usage**
- **Le standard utilisé facilite le croisement de données**

Outils

R STUDIO
R Studio est un excellent choix pour un usage en data science ou data analyse qui requiert des traitements de jeux de données volumineux. Sa prise en main nécessite plusieurs jours.

AIRTABLE
Airtable permet d'éditer des tableurs et de connecter des objets entre eux, de manière intuitive et collaborative. Cependant, de nombreuses fonctionnalités ne sont accessibles qu'avec un abonnement.

GRIST
À l'instar d'Airtable, Grist permet d'éditer et partager des tables de données que l'on peut lier entre-elles.

OPEN DATA EDITOR
Open Data Editor a été déployé par l'Open Knowledge Foundation. L'outil permet de préparer des jeux de données et réaliser des traitements, le tout en ligne et gratuitement.

DATAWRAPPER
Permet de réaliser très facilement des graphiques et des cartes en important des données

OPEN REFINE
OpenRefine est un outil gratuit et open source pour travailler avec des données à nettoyer ou à transformer d'un format vers un autre.

Lexique

DCP - Données à caractère personnel
"Toute information relative à une personne physique susceptible d'être identifiée, directement ou indirectement. Par exemple : un nom, une photo, une empreinte, une adresse postale, une adresse mail, un numéro de téléphone, un numéro de sécurité sociale, un matricule interne, une adresse IP, un identifiant de connexion informatique, un enregistrement vocal, etc."

Anonymisation
Traitement des données personnelles qui vise à rendre impossible toute réidentification de la personne. Elle est irréversible (technique destructive d'informations). Elle permet de conserver les données sans limitation.

Pseudonymisation
Traitement des données personnelles qui vise à ne plus attribuer les données à une personne physique identifiée. Elle consiste à remplacer les données directement identifiantes (ex : nom, prénom) par des données indirectement identifiantes (un alias, un numéro). Elle est réversible.

Schéma de données
Les schémas de données sont des gabarits qui précisent les différents champs au sein d'un jeu de données, leur ordre, les valeurs possibles, etc. Ils permettent une montée en qualité des données et facilitent les usages en simplifiant la compilation entre données de différents producteurs. Source : schema.data.gouv.fr

Format de données
La manière dont les données sont structurées et mises à disposition des personnes et des machines. Il s'agit autant du format du fichier de données (csv, xlsx, geojson...) que du format du contenu des données. Les formats ouverts et adaptés aux traitements par machines favorisent les usages.

Code INSEE ou COG
La manière dont les données sont structurées et mises à disposition des personnes et des machines. Il s'agit autant du format du fichier de données (csv, xlsx, geojson...) que du format du contenu des données. Les formats ouverts et adaptés aux traitements par machines favorisent les usages.

Science ouverte
La science ouverte vise à la diffusion sans entrave des données, méthodes et résultats de la recherche scientifique. Ce mouvement qui vise à démocratiser les savoirs scientifiques repose sur plusieurs piliers dont l'ouverture la plus large des données de la recherche selon le principe "aussi ouvert que possible, aussi fermé que nécessaire."

Données publiques ouvertes par défaut
Le premier principe de la charte internationale de l'open data demande que l'ouverture devienne la norme : les données publiques devraient être ouvertes, sauf exception légale, sans que le public ait à les demander. En France, ce principe s'est traduit dans la Loi pour une République Numérique de 2016 pour toutes les administrations et collectivités de plus de 3500 habitants et 50 agents.

Données complètes et à jour
Le deuxième principe de la charte internationale de l'open data réclame que les données soient publiées sous leur forme complète, dans le plus fort niveau de détail possible. Elles doivent être mises à disposition dès que possible et, lorsque c'est pertinent, en temps réel.

Formats de données

CSV — Plus facilement lisible par les humains (tableau)

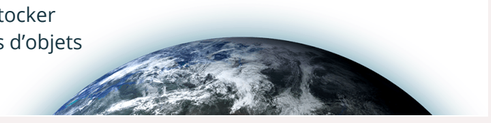
JSON — Polyvalent, moins lisible pour les humains

Formats géographiques

GeoJSON — Pour les formes géographiques et leurs attributs

SHP — Pour stocker des données vectorielles, mais peut être volumineux

KML — Permet de stocker des données d'objets en 3D



Métadonnées

Datasheets for Datasets est un modèle de documentation qui pose une série de questions pour guider la rédaction des métadonnées. En voici une traduction réalisée par Samuel Goëta et Laure Huguenin :

Motivations pour la création du jeu de données
- Pourquoi le jeu de données a-t-il été initialement créé ?
- Quelles ont été les utilisations non prévues du jeu de données ?
- Pour quelles autres tâches le jeu de données pourrait-il être utilisé ?
- Quelles sont les utilisations trompeuses du jeu de données ?
- Qui a financé ou soutenu la création du jeu de données ?

Composition du jeu de données
- Que contient le jeu de données principalement ? Les enregistrements représentent-ils principalement des documents, des personnes, des territoires, des entreprises... ?
- Dispose-t-on d'un schéma décrivant les variables du jeu de données ?
- Que contient chaque champ du jeu de données ?
- Est-ce que le contenu du jeu de données dépend de ressources externes ? De quelles garanties dispose-t-on concernant leur pérennité ?

Processus de collecte des données
- Comment les données ont été collectées ?
- Qui a assuré le processus de collecte de données ?
- Quelle a été la période de collecte des données ?
- Les données ont-elles été collectées directement ou à partir d'autres données ?
- Les données ont-elles été collectées sur un échantillon ? Quelle est la population complète ? Selon quelles méthodes ?
- Quelles sont les erreurs connues, les limites, les sources de bruit ou de redondances associées à ces données ?

Pré-traitement des données
- Comment les données ont-elles nettoyées ou préparées ?
- Les données « brutes » ont-elles été conservées ? Sont-elles diffusées ?
- L'outil de pré-traitement des données est-il disponible ?

Diffusion du jeu de données
- Les données sont-elles diffusées en ligne ? Selon quelles modalités (sur un portail open data, un site web, une API...) ?
- Si non, les données sont-elles diffusées au cas par cas ? à la demande ?
- Selon quelle licence les données sont-elles diffusées ?
- Des redevances ou des restrictions sont-elles appliquées ?

Maintenance du jeu de données
- Qui assure la maintenance du jeu de données ? Comment peut-on contacter cette personne ? Quel est le service responsable du jeu de données ?
- Est-ce que les rôles sont distincts entre la production des données, leur éditorialisation et leur diffusion ?
- Le jeu de données sera-t-il mis à jour ? Si oui, à quelle fréquence ?
- Si les données deviennent obsolètes, comment cette information sera-t-elle communiquée ?
- Est-il possible de contribuer à l'amélioration des données ? Comment ?

Considérations légales et éthiques
- Si le jeu de données concerne des individus, ont-ils exprimé leur consentement de manière claire ?
- Les individus ont-ils été informés sur la finalité du traitement de données ?
- Le jeu de données peut-il exposer de manière directe ou indirecte des individus ?
- Ces données sont-elles conformes au RGPD ?
- Les données peut-elles avantager ou désavantager des groupes sociaux ?
- Le jeu de données contient-il des informations pouvant être considérées comme inappropriées ou offensantes ?