## Choisir le bon format pour l’open data

Le format d’un ensemble de données ouvertes est la manière dont les données sont structurées et mises à disposition des personnes et des machines.

Le choix du bon format permet une gestion et une réutilisation plus simples. Pour maximiser la réutilisation des données, il peut être nécessaire que ceux qui publient les données utilisent divers formats et structures sur différentes plateformes qui conviennent aux besoins de l’utilisateur.

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

* Pourquoi les formats sont si importants pour l’open data
* Choisir la bonne structure
* Accéder aux formats d’open data
* Choisir la simplicité avec CSV

## Exploitabilité, gestion et accès

L’open data, ce sont des données auxquelles tout le monde peut accéder et que tout le monde peut utiliser et partager. Dans cette section, nous verrons quels formats maximisent l’exploitabilité des données, facilitent l’accès pour tout le monde et simplifient la gestion pour ceux qui publient les données.

### Exploitabilité

Le format de données le plus exploitable est souvent celui qui a été utilisé lors de la création de l’ensemble de données. Dans de nombreux cas, le format est celui d’un logiciel propriétaire tel que Microsoft Excel. Bien que ces programmes permettent d’exporter les données sous d’autres formats, leur popularité peut signifier que les publier sous un autre format devienne un obstacle potentiel pour les utilisateurs.

D’autre part, les logiciels tels qu’Excel ont un contenu plus riche, comme les différents styles de tableaux ou de graphiques. Ces éléments peuvent aider à donner du contexte pour quelqu’un qui essayerait de comprendre les données.

### Gestion

Pour certaines organisations qui publient des données, l’open data est maintenant la source principale de leurs données, ainsi que pour le personnel au sein de l’organisation. L’utilisation de l’open data pour permettre au personnel d’accéder à l’information au sein de l’organisation a beaucoup d’avantages, y compris :

* L’accroissement des connaissances de l’équipe au sujet de l’open data.
* L’augmentation de la durabilité et de la confiance accordée aux données en en faisant un élément clé des opérations de l'organisation.
* La création d’opportunités claires pour l’innovation.

### Accès

Pour que l’open data soit aussi accessible que possible, on doit le publier sous:

* une forme que l’utilisateur comprend.
* une forme qu’une machine lit.
* une forme qui soutien la réutilisation facile des données.
* une forme accessible sans les outils chérs.

Le format approprié peut être différent pour chaque ensemble de données. Même si pour être lisible par un humain, un document (comme un fichier PDF) facilite l’accès aux données, ce n’est pas le format le plus lisible par une machine. Pour cette raison, on atteint souvent une accessibilité maximale en publiant les mêmes données sous de nombreux formats différents.

## Les structures les plus communément utilisées pour les données

Les données ne peuvent pas toujours être exprimées sous forme de feuille de calcul et il faut alors considérer d’autres formats. Il y a trois structures clés à connaître: tabulaire, hiérarchique et en réseau.

### La structure tabulaire

La structure la plus répandue est la structure tabulaire. On organise les données dans des colonnes et des rangées qui listent des valeurs séquentielles, comme les dépenses par exemple.

Si les données sont basées sur des entrées séparées qui ne sont pas liées, une structure tabulaire de données sous un format comme CSV est idéale. Ainsi, les chiffres de fréquentation des musées sur le portail de données italien constituent un excellent exemple d’une structure tabulaire de données.

[Lien vers le portail des données italien](http://www.datiopen.it/opendata/Visitatori_musei_pubblici_e_similari_titolo_d_accesso)

### La structure hiérachique

Les données hiérarchiques montrent les rapports entre des points de données, comme pour un arbre généalogique ou les régions d’un pays par exemple. Si l’ensemble de données dépend du rapport entre les points de données et suit une structure où les points sont liés sous forme d’ ‘arbres’ verticaux, une structure de données hiérarchique sous un format tel que JSON est idéale. Les données commerciales de gov.uk sont un bon exemple d’une structure de données hiérarchique appropriée.

[Lien vers json données du commerce de gov.uk](https://www.gov.uk/browse/business.json)

### La structure en réseau

Les données structurées en réseau permettent des rapports entre n’importe quelle combinaison d’éléments dans n’importe quelle direction.

Un bon exemple d’une structure des données en réseau est celui des réseaux sociaux. Pensez à votre réseau d’amis et à leurs amis sur Facebook et considérez les contacts de premier, deuxième et troisième degré sur LinkedIn.

Le Web est un autre exemple d’une structure de données en réseau, où les pages Web comportent des liens vers un nombre incalculable d’autres pages et ce dans n’importe quelle direction.

## Trouver l’open data sous le bon format

En tant qu’utilisateur d’open data, il est important de comprendre les différents formats de données utilisés par ceux qui les publient.

Le téléchargement des données n’est pas forcément la solution la plus appropriée car il se peut que le fichier soit trop large, que les données soient mises à jour trop régulièrement ou bien qu’elles soient trop complexes à exposer comme fichier statique. Certaines données doivent être divisées en ensembles de données plus petits. On préfèrera en publier d’autres sous forme de flux ‘live’ pour s’assurer qu’elles soient toujours mises à jour.

### L’open data téléchargeable

Les données tabulaires sont les plus souhaitables pour le téléchargement, ce qui explique pourquoi les portails d’open data gouvernementaux utilisent principalement les données tabulaires. Cependant, il faut prendre en compte les considérations suivantes lorsqu’il s’agit de gérer des millions de rangées de données :

* Serait-il souhaitable de diviser les données en ensembles de données plus petits ?
* À quelle fréquence devrait-on mettre à jour les données ?
* Comment les changements de méthode de publication affecteront ils les publications précédentes ?

### Données en temps réel (‘Live Data’) et flux de données

Pour certaines données, la mise à disposition sous forme de fichiers téléchargeables n’est pas adaptée.

Ces données sont souvent mises à jour tellement régulièrement que les fichiers à télécharger seraient trop lourds pour la plupart des utilisateurs. Ce type de structure d’open data peut être mis à disposition par l’intermédiaire d’une machine, communément appelée API.

Il y a beaucoup de services de mise à disposition de ces interfaces disponibles sur le Web. On peut directement intégrer ces services à d’autres applications sur le Web. Pour voir un exemple d’une API d’open data, cliquez sur ce lien vers la compagnie ferroviaire belge.

[Lien vers l’API de iRail](https://hello.irail.be/api/1-0/)

## Choisir le bon format pour l’open data

Quand il s’agit de format d’open data, commencez par utiliser CSV.

Un fichier CSV (en anglais, ‘comma separated values’, soit valeurs séparées par des virgules), est tout simplement constitué de lignes de donnés, où chaque point de données est séparé du prochain par une virgule. Le format CSV est parfaitement adapté pour les données tabulaires et peut être facilement téléchargé ou enregistré sous des applications comme Excel, ce qui le rend accessible aux utilisateurs.

Même si le format CSV ne maintient pas le formatage ou les graphiques comme Excel, il s’agit d’un format ouvert et lisible par des machines. Le format CSV est le format le plus simple capable de supporter une vaste réutilisation de l’open data. En d’autres termes, CSV est ‘le plus petit dénominateur commun’ de l’open data et devrait être utilisé en priorité autant que possible.

## Formats de données géospatiales

Les données géospatiales sont souvent plus complexes que de simples ensembles de données tabulaires.

Elles existent sous forme d’ensembles de données hierarchisées, détaillant les départements / états, ou sous forme d’ensemble de données en réseau, donnant des détails sur les routes.

Lors de la publication de ce type de données, il faut considérer des formats tels que [geoJSON](https://en.wikipedia.org/wiki/GeoJSON) (basés sur JavaScript Object Notation - JSON) et [KML](https://en.wikipedia.org/wiki/Keyhole_Markup_Language) (basés sur Extensible MArkup Knowledge - XML). Ces formats sont conçus spécifiquement pour être largement exploitable et peuvent facilement être importés ou exportés dans des outils de cartographie tels que [Open Street Map](http://www.openstreetmap.org/) et [CartoDB](https://cartodb.com).

## Vos formats sont-ils réglementaires?

Vérifiez vos connaissances sur les formats d’open data grâce aux questions suivantes. Vous souvenez-vous des points importants ?

## Pourquoi est-il important de choisir le bon format pour l’open data ?

Pourquoi est-il important de choisir le bon format pour l’open data ?

* Assurer que la vie privée est protégée.
* Faire baisser les coûts.
* Faciliter l'exploitabilité, la gestion et l’accès.

##### Vous avez raison !

Choisir la bonne forme permet de la gestion et la réutilisation plus simples. Pour maximiser la réutilisation des données, il peut être nécessaire que la maison d’éditions utilise une variété des formes et structures disponibles à travers des différentes plates-formes qui conviennent aux besoins de l’utilisateur.

##### Êtes vous sûr(e) ?

Choisir la bonne forme permet de la gestion et la réutilisation plus simples. Pour maximiser la réutilisation des données, il peut être nécessaire que la maison d’éditions utilise une variété des formes et structures disponibles à travers des différentes plates-formes qui conviennent aux besoins de l’utilisateur.

## Quelle structure de données utilisez vous pour les données dans l’image ?

Quelle structure de données utilisez vous pour les données dans l’image ?

* Tabulaire
* Hiérarchique
* En réseau

##### Vous avez raison !

L’image montre que les rapports entre les points de données sont importants. En plus, les liens sont tous ‘parents-enfants,’ ce qui signifie qu’une structure hiérarchique est idéale.

##### Êtes vous sûr(e) ?

L’image montre que les rapports entre les points de données sont importants. En plus, les liens sont tous ‘parents-enfants,’ ce qui signifie qu’une structure hiérarchique est idéale.

## Sous quel format l’open data devrait-il être disponible de préférence ?

Sous quel format l’open data devrait-il être disponible de préférence ?

* CSV
* PDF
* DOC

##### Vous avez raison !

CSV (en anglais, comma separated values), est un format facile à comprendre et hautement réutilisable. Il est ouvert pour que tout le monde puisse développer de nouveaux outils pour l’intérpreter. Les autres formats sont utiles, mais ne sont pas aussi largement adoptés pour les données.

##### Êtes vous sûr(e) ?

CSV (en anglais, comma separated values), est un format facile à comprendre et hautement réutilisable. Il est ouvert pour que tout le monde puisse développer de nouveaux outils pour l’intérpreter. Les autres formats sont utiles, mais ne sont pas aussi largement adoptés pour les données.

## Les formats d’open data

Le format d’un ensemble de données ouvertes est la manière dont les données sont structurées et mises à disposition des personnes et des machines.

Le choix du bon format permet une gestion et une réutilisation plus simples. Pour maximiser la réutilisation des données, il peut être nécessaire que ceux qui publient les données utilisent divers formats et structures sur différentes plateformes qui conviennent aux besoins de l’utilisateur.

On privilégiera l’utilisation du format CVS pour le partage de données ouvertes aussi souvent que possible. Le format CSV est un format facile à comprendre, hautement réutilisable et lisible par des machines.