Agence de Développement du Digital

Manuel opérationnel sur les standards et les formats de données ouvertes (Open Data)

Version 1.1

Décembre 2024

Table des Matières

Introduction :

Ce document est un manuel Open Data à destination des Gestionnaires de Données (GdD) au sein des structures publiques. L'objectif de ce manuel est de définir les standards pour le format et la structure des données ainsi que les critères de qualité à respecter pour la publication de données ouvertes afin de favoriser l'interopérabilité des données et de maximiser le potentiel de leur valorisation.

Ce manuel complémente les deux manuels déjà publiés, à savoir :

- Le « Manuel à destination des Gestionnaires de Données (GdD)[1] », publié en 2021, et qui présente les étapes à mettre en œuvre pour la publication de données sur le portail national de données ouvertes.
- Le « Guide méthodologique pour la mise en œuvre d'un inventaire de données au sein d'une structure publique[2] », publié en 2023, et qui présente notamment l'ensemble des métadonnées à associer à un jeu de données. Ces métadonnées référenceront les standards de données (métadonnées structurelles) et les formats de fichiers de données (métadonnées administratives).

Ce manuel est divisé en quatre partie. La première partie présente les concepts clés de l'Open Data, la deuxième partie présente le cycle de vie des données et l'importance de l'interopérabilité des données, la troisième partie introduit les standards pour la structure et les formats de données, et enfin la quatrième partie passe en revue les critères de qualité des données.

Ce manuel a pour objectif d'être un document vivant qui s'enrichit au fur et à mesure des retours d'expérience des gestionnaires de données, des défis qu'ils rencontrent et des mesures qu'ils entreprennent pour les résoudre. Les commentaires et les propositions d'évolution de ce document peuvent être envoyés à l'Agence de Développement du Digital via l'adresse suivante : opendata@add.gov.ma.

[1] https://data.gov.ma/sites/default/files/docs/Open_data_manuel_GgD%20_vf_Avril%202021.pdf
[2] https://www.data.gov.ma/sites/default/files/2023-03/guide_inventaire_1.3_Fev.%202023.pdf

1. Les fondamentaux de l'Open Data :

Les données ouvertes en tant que concept trouvent leur origine dans la conviction que les données publiques devraient être librement disponibles pour être publiées, utilisées et réutilisées sans limitations. Cette idée fondamentale s'est répandue à travers le monde depuis maintenant plus d'une décennie et accroît la demande pour des données exploitables informatiquement et libres de droits, sans limitations d'utilisation et de réutilisation. Alors que les pays développés comme le Royaume-Uni et les États-Unis ont été les premiers à s'engager dans cette voie, de plus en plus de pays en développement rejoignent aujourd'hui le mouvement et développent leur initiative nationale d'ouverture des données.

Plusieurs pays en développement ont soit déjà lancé leur propre initiative de publication de données en ligne, soit travaillent à la mise en œuvre d'initiatives de données ouvertes dans un proche avenir. En même temps, des organisations nationales et internationales, des organisations de la société civile, des institutions académiques et des entreprises du secteur privé se sont également jointes au mouvement Open Data et développent leurs capacités à exploiter ces données publiées.

Du côté de la société civile, la demande de données est généralement liée aux problématiques de la transparence et de la responsabilité des gouvernements au niveau central et local, tandis que le secteur privé recherche de nouveaux produits et des moyens pour renforcer la valeur économique. Ce mouvement en est peut-être encore à ses balbutiements dans le contexte des pays en voie de développement, mais les opportunités de développer le secteur sont immenses. Il est également utile d'explorer les points communs, mais aussi les différences entre les pays développés et ceux en voie de développement sur ce sujet. Les défis et les opportunités sont en effet légèrement différents. Des éléments tels que la (non) disponibilité de données sous format électronique, la collecte et la diffusion de données dans des pays ayant des siècles de tradition de secret, etc. doivent être soigneusement pris en compte. D'un autre côté, l'amélioration des services publics, l'emploi des jeunes, le développement social et économique sont des opportunités majeures dans le domaine de l'Open Data.

En effet, une initiative de données publiques ouvertes est une excellente occasion pour les citoyens d'un pays donné d'améliorer leur niveau de vie. Cependant, pour atteindre cet impact, il est essentiel que les données soient publiées à temps en utilisant des formats standard et documentés pour faciliter leur réutilisation et leur exploitation par les réutilisateurs. Dans ce cadre, les GdD sont des maillons essentiels de la réussite d'une initiative Open Data puisqu'ils sont responsables du traitement, de la publication et de la maintenance des données. Ce chapitre présente en détail les concepts de l'Open Data, les bases réglementaires qui gouvernent l'Open Data, les principes de la réutilisation des données publiques et les acteurs impliqués par leur ouverture.

1.1. Les concepts de l'Open Data

Une donnée est dite « ouverte » si elle répond à deux principaux critères :

- **La donnée est techniquement ouverte** : une donnée est techniquement ouverte si :
    - Elle est exploitable informatiquement ;

- Elle est disponible sous un format ouvert, c'est-à-dire non-propriétaire et documentée. Par exemple le format de fichier csv pour les données numériques. A l'inverse des formats comme les documents PDF, MS-Word (fichier en .doc) ou MS-Excel (fichier .xls) sont des formats propriétaires non ouverts.

L'objectif du présent manuel est de définir l'ensemble des standards techniques à mettre en œuvre lors de la publication de jeux de données.

- **La donnée est légalement ouverte** : une donnée est légalement ouverte si elle est associée à des droits de réutilisation (appelé « licence de réutilisation ») large qui autorise la réutilisation, le croisement et la redistribution de ces données, dans un contexte commercial ou pas, et requiert uniquement la mention de la source des données. Dans le cadre du chantier Open Data, le Maroc a opté pour la licence ODbL (Open Database License)[3] sur le portail Open Data[4] depuis la création dudit portail en 2011.

A ces deux critères intrinsèques aux données, d'autres éléments sont essentiels à la réussite d'une initiative Open Data et à l'émergence des impacts attendus :

- **Les données sont publiées sous le format le plus désagrégé possible** : Le principe de l'Open Data est de publier des données dites « brutes », c'est-à-dire telles qu'elles ont été collectées au plus haut niveau de désagrégation possible afin de maximiser leur réutilisation et leur valorisation.

  Les données brutes sont gratuites : Un des principes fondateurs de l'Open Data est l'accès gratuit aux données. Il est possible que l'administration, productrice de données, tire des revenus de ses données en offrant des services à forte valeur ajoutée sur ses propres données comme la création d'indicateurs composites ou la réalisation d'analyses spécifiques.
- **Les données doivent être facilement identifiables et accessibles** : Les données ouvertes et publiées par l'administration sont utiles et peuvent générer un impact économique et social si les réutilisateurs peuvent les trouver et y accéder facilement. Pour ce faire, la pierre angulaire d'une initiative Open Data est la mise en place d'un portail national de données ouvertes, qui est le point central d'accès à l'ensemble des données ouvertes de l'administration. Pour le cas du Maroc, le portail national a été lancé en 2011 ([5]). Ce portail héberge des jeux de données de certaines administrations, ou référence des jeux de données hébergés sur des portails open data sectoriels. Toutefois, la structure publique peut mettre en place son propre portail open data et le connecter au portail national pour que les jeux de données soient référencés, soit directement publier ses jeux de données sur le portail national.

[3] http://opendefinition.org/licenses/odc-odbl/
[4] http://www.data.gov.ma/data/fr/dataset?license_id=odc-odbl&_license_id_limit=0
[5] http://www.data.gov.ma

En termes d'accès, le téléchargement des données est libre et sans authentification afin de limiter au maximum les barrières pour les réutilisateurs.

- **Les données doivent être documentées** : Il est essentiel que les données publiées soient documentées afin de permettre leur réutilisation. La documentation des données se fait au travers de métadonnées (données sur les données). Trois types de métadonnées doivent être associées aux données :
    - **Les métadonnées descriptives** qui sont des métadonnées utiles pour la découverte et l'identification d'un jeu de données. Au minimum, ces métadonnées incluent le titre, la structure responsable, les coordonnées de la personne responsable au sein de la structure, la description du jeu ou la couverture géographique et spatiale ou les mots-clés. La liste des métadonnées descriptives importantes est fournie dans le « Guide méthodologique pour la mise en œuvre d'un inventaire de données au sein d'une structure publique de l'Administration[6] » ;
    - **Les métadonnées administratives** qui fournissent des informations sur les données comme la méthode de collecte, la périodicité de mise à jour, le format ou la licence de réutilisation. La liste des métadonnées descriptives importantes est fournie dans le « Guide méthodologique pour la mise en œuvre d'un inventaire de données au sein d'une structure publique de l'Administration » ;
    - **Les métadonnées structurelles** correspondent aux métadonnées internes, c'est-à-dire les métadonnées qui concernent la structure des objets du fichier tels que, pour un fichier de tableur par exemple, le contenu des tables et des colonnes, les clés ou les index. Les métadonnées structurelles peuvent inclure (bonne pratique) un schéma de validation[7]. Le schéma de validation décrit les types et contraintes des données et permet de faire une validation formelle des données. Ce type de validation a pour but de vérifier qu'aucune donnée ne soit incohérente[8].

  En résumé, les métadonnées descriptives et administratives permettent la découverte de l'objet. Les métadonnées structurelles permettent d'appliquer, d'interpréter, d'analyser, de restructurer les données et de les relier à d'autres ensembles de données similaires.
- **Les données doivent être maintenues** : La qualité et donc l'impact d'un jeu de données résident, pour une partie importante, dans sa maintenance qui couvre plusieurs aspects :

[6] https://www.data.gov.ma/sites/default/files/2023-03/guide_inventaire_1.3_Fev.%202023.pdf
[7] Se reporter au « Manuel à destination des Gestionnaires de Données Open Data (GdD) » pour plus de détail sur les schémas de validation https://data.gov.ma/sites/default/files/docs/Open_data_manuel_GgD%20_vf_Avril%202021.pdf
[8] A noter que la validation formelle ne permet pas de vérifier la précision ni la fiabilité des données

- **La mise à jour** : Les jeux de données ouvertes publiés sur le portail Open Data doivent être actualisés aussi souvent que sont mises à jour les données de références à partir desquels ils ont été construits.
- **Les corrections et les gestions de versions :** La réutilisation d'un jeu de données et son analyse par des personnes extérieures à la structure qui en est responsable permettent généralement de mettre en lumière des problèmes spécifiques. Quand ces problèmes sont remontés par les réutilisateurs, via notamment le portail national des données ouvertes, il est essentiel pour le Gestionnaire de ces données de les corriger et publier des versions correctives pour s'assurer que les réutilisateurs continuent leur utilisation.
- **Le support** : Les réutilisateurs de données sont parfois confrontés à des défis, ou ont des questions sur la structure ou le contenu des données. Il est essentiel pour les gestionnaires de données d'assurer le suivi et le support de leurs jeux de données pour maximiser la réutilisation. Le portail national de données ouvertes permet aux réutilisateurs de poser des questions aux gestionnaires de données, et il est donc essentiel pour ces gestionnaires d'y répondre.

- **Les données doivent être standardisées** : Enfin, la valeur intrinsèque d'un jeu de données est difficilement mesurable. L'impact émerge généralement du croisement de plusieurs jeux de données qui font émerger de nouvelles connaissances importantes. Il est donc absolument nécessaire de faciliter le travail des réutilisateurs et faciliter le croisement de jeux en adoptant des standards communs pour des données partagées entre plusieurs administrations. L'objet du présent manuel est de définir ces standards de données.

L'ensemble de ces éléments sont documentés et consolidés dans la « Charte internationale sur les données ouvertes[9] » adoptée par de nombreux pays. Cependant, ces principes couvrent presque exclusivement les méthodes de publication de données et non le contenu a proprement dit, à savoir les données concernées. Les données couvertes par l'Open Data sont définies par le cadre légal de l'Open Data qui est présenté dans la section suivante. A noter que, dans le cadre de l'Open Data, le concept de données est un concept large qui couvre tout type de données depuis les données numériques jusqu'à des textes (textes de lois par exemple) ou des données géospatiales. De la même manière, l'ouverture des données concerne une grande variété de données comme les statistiques d'un secteur, mais également des données administratives comme par exemple la liste des communes d'un département, les résultats d'une élection, les variations quotidiennes du cours d'un produit alimentaire, les arrêts et horaires de passage d'un bus ou les faits constatés par les services de police par département.

En dehors de ces aspects techniques, un des principes clés de l'Open Data, élément de la Gouvernance Ouverte[10], est la collaboration entre l'administration et les acteurs non-

[9] https://opendatacharter.org/principles/
[10] https://fr.wikipedia.org/wiki/Gouvernement_ouvert

gouvernementaux (société civile, monde académique, secteur privé, etc.). L'émergence d'impacts sociaux et économiques est presque exclusivement due à la réutilisation des données ouvertes par les acteurs économiques et la société civile. Ces deux aspects, publication et réutilisation, sont très fortement liés entre eux, et l'interaction entre ces acteurs est essentielle pour la réussite d'une initiative Open Data.

1.2. Le cadre légal de l'Open Data

L'émergence d'une initiative Open Data robuste nécessite un cadre légal clair qui définit sans ambiguïté les données qui peuvent être publiées, celles qui ne peuvent pas l'être et celles qui nécessitent un traitement avant publication. Le développement de l'Open Data au Maroc s'appuie actuellement sur le droit d'accès à l'information, instauré par l'article 27 de la Constitution 2011 et la loi 31-13 relative au droit d'accès à l'information qui matérialise ce droit, la loi 09-08 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et la loi 05-20 relative à la cybersécurité.

- **La Loi n°31-13 du 22 février 2018, relative au droit d'accès à l'information**[11] : Cette loi définit de façon précise les informations qui peuvent être rendues publiques, soit mises à la disposition de leur demandeur soit publiées de façon proactive.
- **La Loi n°09-08 du 18 février 2009**, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel[12] : Cette loi fournit un cadre juridique approprié pour la protection de la vie privée et identifie en particulier les données qui rentrent dans le cadre de cette loi, et les traitements qui doivent être appliqués avant publication.
- **La Loi n°05-20 du 25 juillet 2020 [13],** relative à la cybersécurité qui fournit une base de

Anonymisation des données personnelles

Les deux grandes catégories de techniques pour anonymiser des données sont les suivantes :

- **Les techniques de rédaction** : technique dont le principe est de supprimer des champs ou des informations dans des lignes de données tout en conservant une intégrité suffisante pour permettre une analyse sémantique. Plusieurs techniques de rédaction existent comme la suppression d'attributs, la pseudonymisation, la généralisation ou le brassage.
- **Les techniques d'agrégation** : Les données sont agrégées délibérément pour garantir l'anonymat des données aberrantes. L'idée est de perdre les relations directes entre les données en échange de résumés de ces données afin de gagner en sécurité pour les personnes concernées. Il existe plusieurs techniques d'agrégation comme le comptage, les totaux, les moyennes, ou les distributions.

Toutes ces techniques sont décrites dans le « Manuel Open Data à destination des Gestionnaires de Données ».

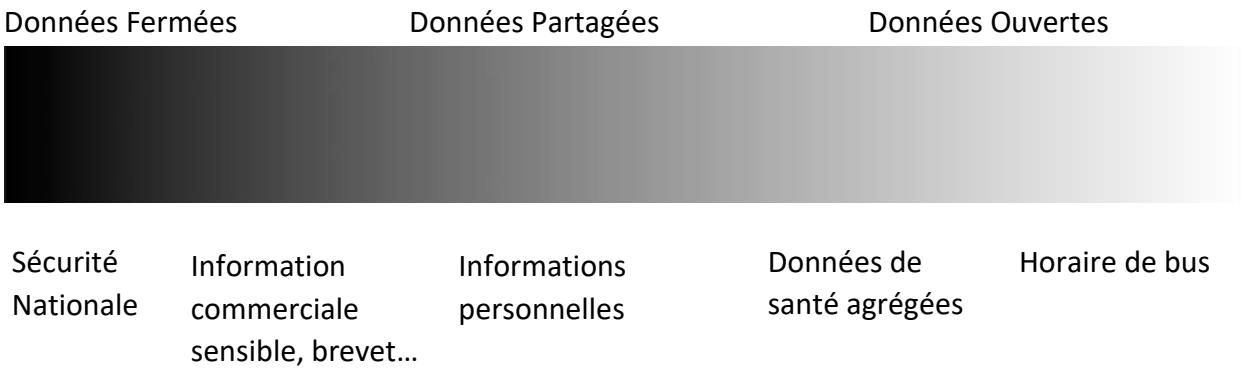
[11] https://www.cdai.ma/fr/3d-flip-book/1263/
[12] https://www.cndp.ma/wp-content/uploads/2023/11/Loi-09-08-Fr.pd
[13] https://www.dgssi.gov.ma/sites/default/files/legislative/brochure/2023-03/loi%2005-20.pdf

classification des actifs informationnels y compris le type de données en fonction de l'impact potentiel des incidents de cybersécurité.

Il est important de retenir les points clés de la Constitution marocaine et des lois précitées :

- L'Article 27 de la Constitution de 2011 dispose : « Les citoyennes et les citoyens ont le droit d'accéder à l'information détenue par l'administration publique, les institutions élues et les organismes investis d'une mission de service public. Le droit à l'information ne peut être limité que par la loi, dans le but d'assurer la protection de tout ce qui concerne la défense nationale, la sûreté intérieure et extérieure de l'État, ainsi que la vie privée des personnes, de prévenir l'atteinte aux droits et libertés énoncés dans la présente Constitution et de protéger des sources et des domaines expressément déterminés par la loi. »
- La Loi n°31-13 définit le champ d'application application du droit d'accès à l'information prévu par l'article 27 précité de la Constitution. Toutes les informations détenues par l'administration sont disponibles pour tout citoyen, quels que soient leur format et leur type (Article 2.a) dans les limites décrites dans l'Article 7 qui les exceptions à la loi.

Il est essentiel de noter que les lois n°31-13 et n°09-08 disposent clairement que lorsqu'un ensemble de données contient à la fois des données qui peuvent être publiées et des données qui ne peuvent pas être publiées, il est nécessaire de procéder au traitement des informations sensibles (anonymisation, oblitération… voir encadré page précédente) tout en publiant le reste des informations publiques. Le diagramme ci-dessous montre le continuum qui existe entre les données non-publiables et les données ouvertes.



A noter qu'une analyse du cadre juridique marocain lié à l'Open Data a été réalisée en 2019-2020 dans le cadre de l'élaboration de l'étude sur la préparation du Maroc à l'ouverture des données publiques (« Open Data Readiness Assesment - ODRA »).

1.3. Les acteurs de l'Open Data

La réussite d'une initiative Open Data requiert la contribution et la mise en œuvre de l'initiative au sein de chaque structure publique. La mise en place d'une dynamique positive au sein de ces structures publiques requiert la mise en place d'une structure organisationnelle spécifique propre avec des rôles définis pour certains agents :

- **Le Responsable de la structure publique** : Le responsable de la structure publique est l'acteur clé qui insuffle la vision, définit la stratégie et mobilise l'ensemble de sa structure pour la mise en œuvre d'une démarche Open Data. Sans un leadership fort qui entraîne la mobilisation des personnels, l'impact de l'ouverture des données serait faible.
- **Le Responsable Open Data (ROD)** : Le ROD est le deuxième acteur clé de la réussite de l'ouverture des données. Le ROD est l'acteur qui transforme la vision du responsable en plan d'action, coordonne les activités, met en œuvre les différents outils et assure le suivi et évaluation afin de remonter au responsable les défis et les actions correctives à mettre en place. Le ROD joue également un rôle de collaboration et coordination avec l'Agence de Développement du Digital (ADD), en tant que coordonnateur du chantier Open Data et responsable du portail national Open Data, afin de partager ses expériences, bénéficier du retour d'expérience des autres structures, et mettre en œuvre les directives globales adoptées. Le rôle et les fonctions du ROD sont détaillés dans le manuel open data à destination des Responsables Open Data[14].
- **Le Chargé d'accès à l'information** : Le chargé d'accès à l'information est un acteur indirect de l'Open Data. En tant que personne responsable de la mise en œuvre de la loi n°31-13 au sein de sa structure, il participe à l'identification des données qui peuvent ou doivent être publiées par la structure.
- **L'entité en charge des archives :** L'entité en charge des archives est aussi un acteur indirect de l'Open Data. En tant qu'entité en charge de la mise en œuvre de la loi n°69-99 relative aux archives, elle participe à l'identification des documents, informations et données qui peuvent être publiées en relation avec cette loi.
- **Le Responsable des Données Personnelles (RDP) :** Le RDP est un acteur indirect de l'Open Data. En tant que personne responsable de la mise en œuvre de la loi n°09-08, le RDP participe à l'identification des données qui relèvent de cette loi, et identifie les traitements qui doivent être appliqués aux données avant publication afin de respecter les clauses de cette loi.
- **Les Gestionnaires de Données (GdD)** : Les GdD sont les acteurs techniques clés de la publication de données. Ils sont en charge de la préparation et de la publication des données qu'ils gèrent et qui peuvent être publiées dans le cadre de l'ouverture des données publiques. Les GdD sont aussi en lien direct avec les réutilisateurs et

[14] https://data.gov.ma/sites/default/files/docs/Open_data_manuel_ROD_Avril%202021.pdf

interagissent avec eux sur les analyses de données ou les corrections d'erreurs potentielles. Les détails du processus de publication sont présentés dans le manuel open data à destination des Gestionnaires de l'Open Data (GdD)[15].

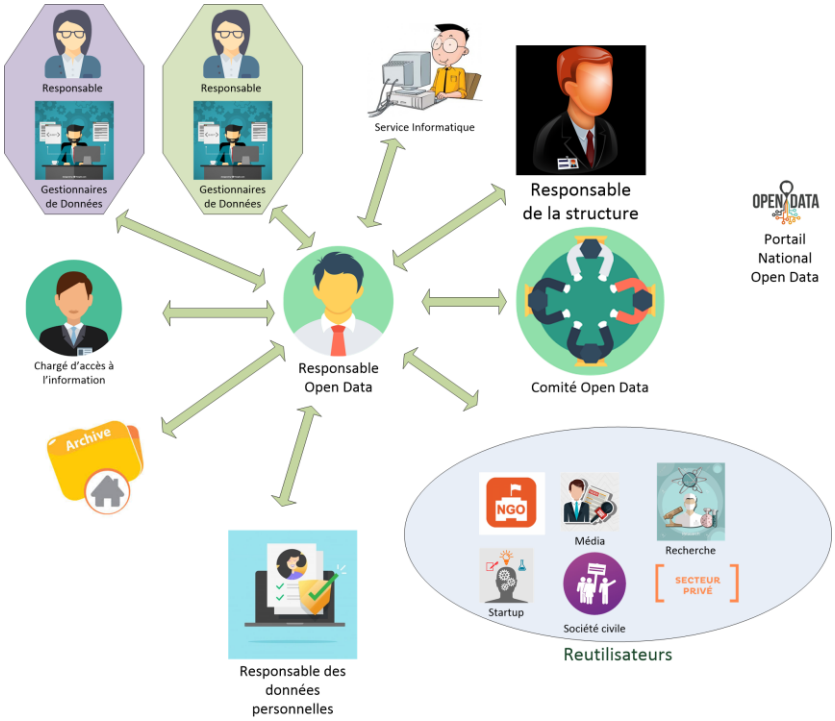
- **Les services informatiques (SI)** : Les SI sont des acteurs au service des GdD et au travers de leur expertise informatique permettent au GdD d'automatiser la publication de données depuis la préparation jusqu'à la publication sur le portail national et/ou le portail de la structure publique concernée en passant par le transcodage et le formatage.
- **Les réutilisateurs** : Les réutilisateurs sont les acteurs clés de la valorisation des données. Par leur réutilisation, l'analyse et le croisement de données, ou l'innovation, ils permettent, d'une part, de générer de la valeur sociale et économique et financière à partir de la donnée, et d'autre part, d'améliorer la qualité des données en mettant à jour des incohérences potentielles. Enfin, ils aident les structures à prioriser les publications de données en fonction de leurs besoins.

Le tableau ci-dessous présente la matrice des acteurs précités et leurs rôles.

Acteur	Stratégie open data	Publication de données et maintenance	Réutilisation de données
Responsable de la structure	X		
Responsable Open Data	X		
Chargé d'accès à l'information		X	
Responsable archives		X	
Responsable des données personnelles		X	
Gestion de Données		X	X
Services informatiques		X	
Réutilisateurs externes à la structure			X

[15] https://data.gov.ma/sites/default/files/docs/Open_data_manuel_GgD%20_vf_Avril%202021.pdf

Le diagramme ci-dessous présente l'écosystème Open Data au sein d'une structure publique.



1.4. Les bénéficiaires d'une initiative Open Data

L'objectif principal d'une initiative Open Data est de contribuer dans la génération d'un impact économique et social important pour le pays. Les bénéficiaires directs et indirects de l'ouverture des données sont nombreux :

- **Les structures publiques** : Les premiers bénéficiaires de l'ouverture des données publiques sont les structures publiques elles-mêmes pour deux raisons principales :
    - Elles peuvent accéder aux données des autres structures de façon simple et sans barrière administrative, ce qui facilite l'exploitation de ces données et la coordination de l'action publique entre différentes administrations. Pour illustrer ce phénomène, il est intéressant de noter par exemple que le Gouvernement de la province de Colombie Britannique au Canada rapporte que depuis la mise en place de son portail Open Data, à peu près un tiers des requêtes de jeux de données sur ce portail viennent des structures publiques.
    - La mise en place de l'Open Data les pousse à instaurer des processus robustes de gestion de données qui, d'une part, améliore la qualité des données, et d'autre part, facilite leur exploitation pour la prise de décision et la mesure de l'efficacité des politiques publiques.
- **Les acteurs de l'innovation** : L'accès à des données publiques, notamment des données géospatiales permet aux acteurs de l'innovation de concevoir et proposer des services innovants. Pour illustrer cette opportunité, il est intéressant de noter par exemple que

NESTA[16], un organisme non-gouvernemental anglais spécialisé dans l'innovation, a investi £1.2M pour stimuler l'innovation autour des données publiques du Royaume-Uni et a mesuré au bout de 3 ans que cet investissement avait généré un retour sur investissement pour chaque £1 entre £5 et £10 pour l'économie anglaise soit une génération de £5.3M à £10.8M en termes de création d'emplois et de richesse.

- **Le secteur privé** : Les données publiques ouvertes sont une opportunité majeure pour le secteur privé, que ce soit pour améliorer l'efficacité de leurs activités comme les études de marché, et les prospections, ou pour créer de nouveaux services. A titre d'exemple, l'ODI (Open Data Institute, institut indépendant dédié à l'Open Data[17]) a mené une étude en Angleterre qui a montré que 270 sociétés utilisaient les données publiques ouvertes, représentant un chiffre d'affaires de £92bn et plus de 500.000 emplois.
- **La société civile** : Un des enjeux de l'Open Data est l'amélioration de la transparence de l'Administration. Grâce aux données ouvertes, les organisations de la société civile, en particulier celles qui travaillent dans ce domaine, peuvent facilement mener leurs investigations. Pour celles qui travaillent dans des secteurs spécifiques (par exemple l'emploi, la santé…), elles peuvent trouver des données qui permettent d'orienter leurs activités.
- **Le secteur académique et universitaire** : L'ouverture des données publiques est un moteur essentiel pour la recherche universitaire. La mise à disposition de données permet aux chercheurs de conduire des recherches très précises (études sociales, économiques…) grâce aux données publiées.
- **Les médias** : L'émergence de l'Open Data dans le monde a permis également la création d'un nouveau type de journalisme, appelé Data journalisme, ou journalisme des données, où des journalistes exploitent et analyses des données ouvertes pour développer leurs articles et informer le public. De façon générale, l'Open Data est un outil essentiel pour les journalistes d'investigations, et une opportunité de développer et améliorer la technicité des journalistes, et les contenus média.
- **Les citoyens** : Les citoyens sont généralement des bénéficiaires indirects de l'ouverture des données. L'Open Data leur permet d'accéder à des services innovants grâce au travail des acteurs de l'innovation, leur permet d'être mieux informés de la gestion des affaires publiques grâce aux organisations de la société civile et aux médias, et leur permet plus généralement de participer plus activement dans la vie démocratique de leur pays.

[16] https://www.nesta.org.uk/
[17] https://theodi.org/

2. Cycle de vie de la donnée & Approche Open Data

2.1. Cycle de vie de la donnée

Le succès d'une initiative Open Data requiert à la fois la publication d'un nombre important de jeux de données de qualité, mis à jour régulièrement et également la réutilisation de ces données par les réutilisateurs afin de générer de la valeur économique et sociale à partir de ces données. Cependant, cette définition sous-entend deux éléments essentiels explicités ci-dessous :

- **La publication des données est une étape dans leur cycle de vie :** La publication d'un jeu de données est le résultat d'un processus de gestion des données pérennes et robustes. En l'absence de tels processus au sein des structures publiques, l'ouverture des données ne saurait être pérenne et durable.
- **Un bénéfice et une génération de valeur pour tous les réutilisateurs et en particulier la structure publique qui gère les données publiées** : La publication de données ouvertes demande aux structures publiques un effort et un travail (la publication de données) spécifiques. Cependant, il est essentiel de noter que ce travail doit bénéficier à tous les réutilisateurs et en premier lieu la structure publique qui publie les données.

De nombreux pays ont focalisé leur action sur la mise en ligne de données et sur la réutilisation de données publiées au travers d'activités ponctuelles comme par exemple des hackathons. Les résultats d'une telle stratégie ont toujours été extrêmement limités et peu pérennes. D'une part, bien que la publication ponctuelle de données ouvertes soit possible, si elle n'est pas intégrée dans un processus complet de production, les données ne seront pas maintenues et mises à jour, perdant rapidement toute valeur potentielle. D'autre part, si les structures publiques qui publient des données n'y trouvent aucun bénéfice direct, leur motivation se diluera au cours du temps.

Le défi majeur de la mise en œuvre de l'Open Data est la transformation des structures publiques afin de mettre les données au centre de l'action publique. L'objectif est de mettre en place plusieurs éléments au sein des structures publiques :

- **Des processus robustes de gestion de données** où la publication de données ouvertes devient un effet de bord intégré dans la production et l'exploitation des données depuis leur collecte jusqu'à leur archivage. Le schéma ci-dessous présente la place de la publication Open Data dans le cycle de vie de la donnée.

Il est important de noter que la publication de données sous format open data doit également mettre en œuvre le stockage de ces données, qui sont mises à disposition sur un portail de données ouvertes. Ces portails peuvent soit héberger les données localement (logiciel de portail hébergé par l'administration), soit les héberger en ligne (dans le cas des logiciels de portail géré par des prestataires sous la forme de service SaaS[18]).

Il est également important de noter que les licences de réutilisation des données ouvertes imposent la non-révocabilité du droit de réutilisation, ce qui implique que les données publiées ne peuvent pas être retirées du portail. Elles peuvent être corrigées en cas d'erreur, mais l'accès à un jeu de données et à toutes les séries publiées doivent être géré en tenant compte de la réglementation en vigueur.

Enfin, un des éléments clé de la gestion de données concerne l'interopérabilité de ces données. L'objectifs des données est d'appuyer les analyses et les croisements qui facilitent la prise de décision. Ces analyses requièrent que les données soient facilement exploitables et interopérables. Cette dimension est détaillée dans la section suivante.

- **Des nouvelles capacités** : La mise en place de processus robustes de gestion de données, l'exploitation, le croisement, l'analyse, la visualisation et la publication de données nécessitent des compétences spécifiques qui doivent être développées au sein de la structure.

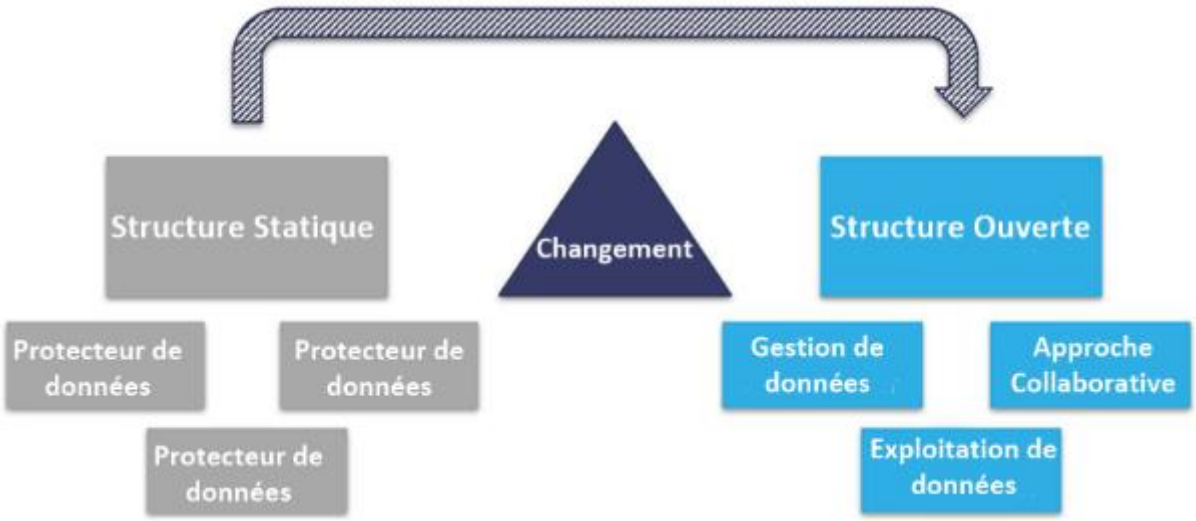
[18] https://fr.wikipedia.org/wiki/Software_as_a_service.

- **Des nouvelles approches collaboratives** : Le principe fondateur de l'Open Data est le rapprochement entre les structures publiques et les acteurs non-gouvernementaux. Un des objectifs de l'Open Data et de faciliter les collaborations entre les structures publiques et ces acteurs.

La mise en place concomitante de ces trois éléments, supportés par une législation incitative, est la clé de la réussite d'une initiative Open Data et apporte un ensemble de bénéfices mesurables pour les structures publiques notamment :

- **Economie et efficacité budgétaire** : Un des premiers impacts de la mise en place de ces approches au sein des structures concerne les économies potentielles. Par exemple, la centralisation et le partage vont faire clairement apparaître les duplications de collecte de données qui pourraient exister au sein d'une même structure ou entre les structures.
- **Efficacité** : L'utilité principale des approches basées sur les données concerne l'amélioration de l'efficacité de l'action publique. D'une part, les politiques publiques bénéficient de prise de décision scientifique basée sur les données réelles de terrain. Ensuite, les collaborations entre les acteurs gouvernementaux et non-gouvernementaux permettent non seulement de construire un consensus partagé sur le diagnostic d'un problème donné, mais permettent également une action collective concertée sur les grands défis qui touchent le pays que ce soit lors de crises graves comme celle du Covid-19 ou sur des sujets à long terme comme la modernisation et le développement global du pays. Enfin, l'échange facilité et automatique de données entre les structures publiques peut, par exemple, identifier et éviter les dépenses inutiles, et permettre d'améliorer l'efficacité des processus et la prestation des services publics.
- **Innovation** : Une des idées fondatrices de l'Open Data est liée à la collaboration entre les structures publiques et les acteurs non-gouvernementaux qui permet l'exploitation de l'innovation citoyenne dans la résolution des problèmes de l'administration. Les approches collaboratives Open Data favorisent naturellement l'émergence de solutions innovantes issues des échanges et de la cocréation entre des acteurs ayant des points de vue complémentaires.
- **Amélioration des services publics** : Un des effets les plus visibles de la publication de données notamment par les acteurs de l'innovation (startup…) est l'émergence de nouveaux services pour les citoyens qui viennent compléter les services publics de l'administration. D'autre part, les approches collaboratives notamment et l'organisation de compétitions par exemple permet à l'Administration de facilement codévelopper avec ces acteurs de l'innovation de nouvelles applications innovantes et à faible coût comparé aux appels d'offres classiques.

En résumé, l'émergence d'une initiative Open Data qui produit des résultats économiques et sociaux probants requiert la transformation interne des structures publiques comme présentée dans le diagramme ci-dessous.



2.2. Interopérabilité

La section précédente montre l'importance de la réutilisation des données pour la génération de valeurs sociales et financières. Cependant, il est essentiel de souligner que le principe fondateur de l'Open Data établit que la génération de valeurs est issue du croisement de données issues de différents secteurs, départements ou ministères. Ce sont les analyses croisées qui vont permettre de faire émerger de nouvelles connaissances, de nouvelles analyses, qui vont faciliter la prise de décision, ou évaluer les impacts des politiques publiques. De ce fait, un élément clé et un des objectifs de la publication de jeux de données sont de faciliter ce croisement, et favoriser l'interopérabilité des données. Dans ce cadre, trois dimensions sont à considérer[19] :

1. **La facilité d'exploiter un jeu de données** : La première dimension qui facilite la réutilisation dépend de chaque jeu de données de façon individuelle. Pour faciliter sa réutilisation, un jeu de données ouvertes doit utiliser un format ouvert[20] et être exploitable informatiquement.
2. **L'utilisation de standards de données communs** : La deuxième dimension concerne les standards de données pour représenter les informations telles que les chiffres, les dates ou les coordonnées GPS. L'utilisation de standards différents entre les jeux complexifie le croisement de jeux de données.

[19] A noter que nous parlons, dans ce contexte, uniquement de la dimension technique intrinsèque des données. Comme mentionné dans le chapitre 1, d'autres éléments impactent la réutilisation comme les aspects légaux ou la facilité de chercher et trouver les jeux de données publiés par l'administration.

[20] Le principe des formats ouverts est de faciliter l'exploitation d'un jeu de données en le fournissant sous un format qui peut être manipulé sans logiciel propriétaire spécifique

3. **L'utilisation d'identifiants communs** : Différents secteurs et différentes structures publiques manipulent des entités communes. Par exemple les sociétés enregistrées dans un pays sont gérées par plusieurs ministères comme le ministère du Commerce pour tout ce qui est enregistrement, le Ministère de l'Économie et des Finances pour tout ce qui touche à la taxation, etc. De la même manière, la gestion des bâtiments publics relève généralement de plusieurs ministères. Il est donc important pour ces entités communes d'être référencées sous un même identifiant de façon globale et centrale au niveau de l'Administration. La liste des identifiants uniques et de leurs caractéristiques est communément appelée données de référence ou référentiels. Ces référentiels peuvent être soit globaux (par exemple le registre des sociétés) soit spécifiques à un secteur (par exemple la nomenclature des actes médicaux au niveau de la santé). La mise en place de référentiels communs est un élément critique pour l'interopérabilité des données et les croisements de données entre les différents secteurs.

L'objet du présent manuel est de couvrir les points 1 et 2 présentés ci-dessus.

Il est important de noter que ce manuel complémente le Cadre Général d'Interopérabilité (CGI) sur des éléments qui ne sont pas couverts par le CGI. Cependant, une partie des standards de données est commune avec le CGI et les standards qui sont communs avec la version actuelle du CGI[21] sont identifiés dans le présent document.

[21] Version 3.0 – Octobre 2023

3. Les standards de données

Cette section présente les différents standards de format et de codage des données et jeux de données qui sont couramment adoptés pour faciliter l'exploitation, l'analyse et le croisement de ces jeux de données. A noter que les standards présentés dans cette section répondent aux critères du Cadre général d'interopérabilité (CGI[22]) :

- **Ouvert** : La spécification fonctionnelle et technique de la norme doit être complète, publique et sans restriction
- **Pertinent** : La pertinence d'une norme fait référence à son utilité, sa nécessité et sa simplicité de mise en œuvre et cela, indépendamment du fait que la norme est reconnue ou non.
- **Mature** : Une norme est mature quand :
    - Elle est soutenue par les infrastructures technologiques ;
    - Elle a démontré sa fiabilité à la suite de son application ;
    - Elle présente une certaine stabilité dans son utilisation.
- **Indépendant** : Une norme devrait être indépendante de toute infrastructure technologique, logiciel ou bien matériel.
- **Facile à déployer** : Le déploiement de la norme ne doit pas être contraignant et générer des coûts de déploiement supplémentaires.
- **Soutenu par l'industrie** : la norme doit être bien établie dans l'industrie pour son périmètre d'usage. Elle doit avoir bâti une solide réputation dans le domaine auquel il se rattache.

Ils sont issus des organismes de standardisation reconnus et listés dans le CGI. Certains des standards présentés ci-dessous sont communs avec le CGI et sont identifiés comme tels.

Les formats de fichier

Le premier critère des données ouvertes et le premier facteur qui impacte la réutilisation d'un jeu de données sont liés au format du fichier de données. Pour permettre au plus grand nombre d'utiliser un fichier de données, il est nécessaire que ce fichier soit dans un format documenté et public, aussi appelé « format ouvert », manipulable par des outils standard et ne nécessitant pas l'acquisition de logiciels spécifiques. Chaque type de données peut toujours être fourni sous différents formats ouverts. Le tableau ci-dessous présente différents types de formats.

[22] La version de référence dans ce document est la version 3.0 d'Octobre 2023

Type de données	Formats ouverts	Formats propriétaires
Données tabulaires	CSV[23] (*)	Excel (.xls)
Image bitmap[24]	PNG(*), JPEG (*)	PSD (Adobe Photoshop), TIFF (*)
Image vectorielle[25]	SVG (*)	AI (Adobe Illustrator)
Texte	TXT(*), RTF (*), HTML(*)	PDF, Word (.doc)
Données géospatiales	Geojson, geotiff, GML(*, norme ISO 19136-1:2020), KML, WFS (*) ou shapefile Voir également les normes NM ISO 19131[26] et ISO 19119[27]	DXF (autocad)
Données structurées/hiérarchiques	XML(*), JSON (*), RDF, ODF, EPUB	AZW (Amazon Kindle)
Données multimédia	Vidéo : OGM Son : OGG(*), WEBM(*)	Vidéo : AVI (*), MPEG4(*) Son : MP3(*)
API	OAS(*), REST(*)	

Tableau 1 : Formats ouverts (*) format recommandé dans le CGI

Les formats mentionnés ci-dessus sont des formats de fichier informatique qui sont indépendants du contenu et de la thématique du jeu de données. Certains secteurs ont également standardisé des formats spécifiques de plus haut niveau avec une sémantique et une structure spécifique dédiée à des applications spécifiques. On peut citer comme exemple GTFS (Global Transit Feed Specification[28]) pour les données de transport public ou OCDS (Open Contracting Data Standard[29]) pour les données de marchés publics. L'utilisation de ces formats spécifiques présente plusieurs avantages, notamment :

- Une interopérabilité au niveau sémantique : ces standards permettent de croiser des données au niveau sémantique même si elles sont de nature différente. Par exemple GTFS permet de modéliser toutes les données de transport, quel que soit le mode de transport (autobus, train, avion, bateau…). Ce standard a permis l'émergence des applications de mobilité multimodale qui permet à un usager de déterminer son trajet en utilisant plusieurs modes de transport.
- Une interopérabilité au niveau international : L'utilisation de formats globaux permet de croiser des données au niveau international. Par exemple, GTFS permet à une application de définir un trajet pour un utilisateur au travers de plusieurs pays. Les

[23] https://fr.wikipedia.org/wiki/Comma-separated_values A noter que plusieurs formats de fichier CSV existent et le chapitre 4 du présent manuel fait des recommandations spécifiques pour un format particulier.
[24] Il s'agit ici d'images graphiques qui ne peuvent pas être représentées sous un autre format. Il est par exemple exclu d'utiliser un format graphique pour représenter un tableau de données
[25] Il s'agit ici d'images graphiques qui ne peuvent pas être représentées sous un autre format. Il est par exemple exclu d'utiliser un format graphique pour représenter un tableau de données
[26] https://www.imanor.gov.ma/wp-content/uploads/2024/02/17.8.431-ISO-19131.pdf
[27] https://www.iso.org/fr/standard/59221.html
[28] https://gtfs.org/fr/
[29] https://standard.open-contracting.org/latest/fr/

organismes publics ou privés qui proposent leurs données sur ce format peuvent de ce fait atteindre plus de clients.

- Une exploitation facilitée : La définition et l'adoption des standards sectoriels amènent la création de communautés motivées sur le sujet et, dans la très grande majorité des cas, l'émergence d'outils dédiés qui s'appuient sur ces standards. L'utilisation de ces standards permet de ce fait l'utilisation de ces outils. Par exemple, il existe des moteurs internationaux qui fonctionnent sur le standard OCDS. L'utilisation d'OCDS pour les marchés publics permet de facto l'intégration de ces jeux de données dans ces moteurs internationaux et donc un nombre accru de participants dans les appels d'offres avec potentiellement des offres plus intéressantes pour les entités contractantes.

La structure des données tabulaires

Cette section traite des bonnes pratiques pour la mise en forme des données tabulaires.

Format exploitable informatiquement[30]

Une des recommandations importantes pour la publication de données ouvertes concerne la structure d'un fichier de données tabulaires. Les bonnes pratiques internationalement reconnues incluent les directives suivantes :

- Il doit être possible de traiter les données directement, en y effectuant toutes les opérations appropriées telles que le tri des colonnes, le filtrage des lignes, l'exécution d'agrégats de valeurs. Ces opérations nécessitent des données bien structurées et notamment la mise en place de tables pures sans cellules fusionnées, sans en-têtes ni de notes de bas de page, de commentaires ou de données cachées dans du texte.
- Les éléments communs des données sont exprimés de manière uniforme - par exemple, les dates sont toujours dans le même format, les codes ou les noms sont toujours dans le même cas, et les nombres sont exprimés de manière cohérente et homogène.

L'objectif de ces recommandations est de faciliter l'exploitation du jeu de données par un programme informatique[31], et de faciliter son analyse et la réalisation de visualisations avec des outils comme des tableurs.

Les tableaux ci-dessous illustrent un cas de données qui ne sont pas exploitables informatiquement et un cas de données qui le sont.

[30] Appelé également format exploitable par une machine par traduction littérale du terme anglais « machine-readable format »

[31] En anglais cette caractéristique est appelée « machine-readable »

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		Table 4. Gender Development Index															
2											SDG 3				SDG 4.3		
3			Gender Development Index				Human Development Index (HDI)				Life expectancy at birth				Expected years of schooling		
4							Value				(years)				(years)		
5	HDI rank	Country	Value		Group	[b]	Female		Male		Female		Male		Female	[c]	Male
6			2017		2017		2017		2017		2017		2017		2017		2017
7		VERY HIGH HUMAN DEVELOPMENT															
8	1	Norway	0,991		1		0,945		0,953		84,2		80,5		18,6	[d]	17,2
9	2	Switzerland	0,987		1		0,937		0,949		85,3		81,5		16,1		16,3
10	3	Australia	0,975		2		0,926		0,950		85,0		81,2		23,3	[d]	22,5
11	4	Ireland	0,979		1		0,926		0,946		83,6		79,7		19,7	[d]	19,5
12	5	Germany	0,967		2		0,919		0,951		83,5		78,9		16,9		17,0
13	6	Iceland	0,966		2		0,920		0,952		84,4		81,5		20,5	[d]	18,2
14	7	Hong Kong, China (SAR)	0,965		2		0,916		0,949		87,1		81,2		16,3		16,4
15	7	Sweden	0,992		1		0,927		0,934		84,3		80,9		18,4	[d]	16,9
16	9	Singapore	0,982		1		0,922		0,939		85,2		81,1		16,4	[g]	16,0
17	10	Netherlands	0,966		2		0,913		0,944		83,7		80,3		18,3	[d]	17,8
18	11	Denmark	0.980		1		0.919		0.938		82.8		79.0		19.8	[d]	18.4

Tableau 2 : Données non exploitables informatiquement

	A	B	C	D	E	F	G	H	I	J	K	L	M
	HDI Rank (2017)	Country	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
	168	Afghanistan											
	68	Albania	0.645	0.626	0.610	0.613	0.619	0.632	0.641	0.641	0.652	0.662	0.669
	85	Algeria	0.577	0.581	0.587	0.591	0.595	0.600	0.608	0.617	0.627	0.636	0.644
	35	Andorra											0.759
	147	Angola										0.374	0.387
	70	Antigua and Barbuda											
	47	Argentina	0.704	0.713	0.720	0.725	0.728	0.731	0.738	0.746	0.753	0.764	0.771
	83	Armenia	0.631	0.628	0.580	0.588	0.599	0.605	0.612	0.623	0.636	0.642	0.647
0	3	Australia	0.866	0.867	0.868	0.872	0.875	0.883	0.886	0.889	0.892	0.895	0.898
1	20	Austria	0.795	0.800	0.805	0.807	0.813	0.817	0.820	0.824	0.835	0.834	0.838
2	80	Azerbaijan						0.612	0.612	0.617	0.626	0.633	0.640
3	54	Bahamas											0.776
4	43	Bahrain	0.746	0.752	0.757	0.765	0.769	0.775	0.778	0.779	0.783	0.786	0.792
5	136	Bangladesh	0.387	0.394	0.402	0.409	0.417	0.425	0.433	0.442	0.451	0.460	0.468
6	58	Barbados	0.716	0.718	0.718	0.721	0.727	0.731	0.735	0.740	0.736	0.743	0.752
7	53	Belarus						0.657	0.661	0.667	0.671	0.676	0.683
8	17	Belgium	0.806	0.810	0.825	0.839	0.845	0.852	0.857	0.862	0.866	0.868	0.873

Tableau 3 : données exploitables informatiquement

Données longues et données larges

Toute série de données est constituée de valeurs numériques (généralement) décrites par des métadonnées normalisées (temps, zone, description spécifique, etc.). Il y a deux façons principales de présenter ces données, soit au format appelé « données larges », soit au format appelé « données longues », détaillées ci-dessous :

- **Les données larges** présentent des données numériques en plusieurs colonnes. Soit sous forme de catégories (par exemple, chaque pays est présenté dans sa propre colonne), soit par date (par exemple, chaque mise à jour annuelle donne lieu à une nouvelle colonne). Les nouvelles données traversent l'écran de gauche à droite. L'exemple ci-dessous montre des données sous format large ou l'ajout d'une nouvelle année de données pour un pays donné entraînera l'ajout d'une colonne.

Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963	1964	1965
Aruba	ABW	Urban population	SP.URB.TOTL	27526	28141	28532	28761	28924	29082
Afghanistan	AFG	Urban population	SP.URB.TOTL	755836	796272	839385	885228	934135	986074
Angola	AGO	Urban population	SP.URB.TOTL	509222	597288	628381	660180	691532	721552
Albania	ALB	Urban population	SP.URB.TOTL	493982	513592	530766	547928	565248	582374
Andorra	AND	Urban population	SP.URB.TOTL	7839	8766	9754	10811	11915	13067
Arab World	ARB	Urban population	SP.URB.TOTL	28797177	30292822	31856717	33513046	35275337	37163923
United Arab Emirates	ARE	Urban population	SP.URB.TOTL	67927	74975	84367	95215	106178	116473
Argentina	ARG	Urban population	SP.URB.TOTL	15076842	15449950	15815502	16183085	16552517	16923103
Armenia	ARM	Urban population	SP.URB.TOTL	960956	1012430	1065431	1119586	1174560	1229980
American Samoa	ASM	Urban population	SP.URB.TOTL	13324	13729	14254	14871	15522	16176
Antigua and Barbuda	ATG	Urban population	SP.URB.TOTL	21466	21472	21458	21443	21449	21489

Tableau 4 : Données sous format large

Les données larges sont souvent utilisées pour la visualisation et le traitement des données, car les données peuvent facilement être regroupées dans les axes nécessaires. Cependant, il s'agit d'un format d'archivage difficile, car la mise à jour d'une telle série de données nécessite l'équivalent de la création d'un nouveau champ (par année dans l'exemple ci-dessus) et donc la modification de la structure de données. Ensuite chaque ligne doit être mise à jour avec les informations appropriées. Cette opération sera coûteuse pour une grande base de données, et signifie également que l'écriture d'un programme informatique pour interroger vos données est plus difficile et devra être mis à jour régulièrement à chaque changement de structure de données.

- **Les données longues** présentent les données numériques sur plusieurs lignes avec une seule colonne pour les valeurs. Les nouvelles données sont ajoutées en ajoutant des lignes. L'exemple ci-dessous montre des données sous format long.

Departme	Entity	Date	Expense T	Expense A	Supplier	Transactic	Amount
Departme	Departme	########	CASH FUN	FINANCE &	NOTTINGH	HAFS-796	45000000
Departme	Departme	########	CASH FUN	FINANCE &	NOTTINGH	HAFS-796	85000000
Departme	Departme	########	CASH FUN	FINANCE &	OLDHAM I	HAFS-796	28000000
Departme	Departme	########	CASH FUN	FINANCE &	OXFORDSI	HAFS-797	75000000
Departme	Departme	########	CASH FUN	FINANCE &	PETERBOR	HAFS-797	22000000
Departme	Departme	########	CASH FUN	FINANCE &	PLYMOUTI	HAFS-797	35000000
Departme	Departme	########	CASH FUN	FINANCE &	PORTSMO	HAFS-797	27000000
Departme	Departme	########	CASH FUN	FINANCE &	REDBRIDG	HAFS-797	31000000

Tableau 5 : Données sous format long

Le format long pour les données est le plus approprié pour l'archivage et pour représenter la structure que vous trouverez habituellement dans une base de données. Chaque ligne d'une longue série de données représente une ligne dans une base de données. L'ajout de nouvelles informations est relativement simple puisque vous ne devez mettre à jour qu'une seule ligne à la fois. La méthode recommandée pour la publication de données ouvertes est le format long.

L'encodage des données

Les caractères dans les chaînes de caractères sont codés dans les fichiers informatiques[32] (à chaque code correspond un caractère). La liste des caractères et leur code sont stockés dans un registre appelé « registre de caractères codés » (Voir la norme Unicode[33] ou la norme ISO/IEC 10646[34]). Il existe de nombreux registres comme ASCII[35], ou UTF-8[36] et ces registres contiennent plus ou moins d'entrées qui permettent de représenter plus ou moins de caractères et donc de couvrir plus ou moins de langues et leurs spécificités (par exemple la lettre latine á, l'idéogramme chinois 請 ou le caractère devanagari ह्र montrent la variété des caractères existants). Tous les encodages ne se valent pas. Par exemple l'encodage ASCII contient 128 entrées dédiées à l'anglais et l'encodage UTF-8 contient plus de deux millions d'entrées. L'utilisation d'un encodage inapproprié amène des représentations erronées. Par exemple, la chaîne « Où sont les caractères accentués ? » pourrait devenir « OÃ¹ sont les caractÃ¨res accentuÃ©s ? ».

Pour éviter ces problèmes, UTF-8 est l'encodage défini au niveau international[37] qui permet de couvrir l'ensemble des caractères de tous les langages. C'est également l'encodage recommandé dans le CGI.

Les types de données

Dans un fichier de données, plusieurs types de données peuvent cohabiter (données décimales, date, numéro de téléphone, etc.). Certains de ces types peuvent poser des défis particuliers qui sont présentés dans les sous-sections suivantes. Dans tous les cas, il est critique de renseigner et documenter les standards, conventions et approches utilisés pour chaque colonne dans les métadonnées structurelles du jeu de données pour faciliter à la fois la validation formelle des données et leur interprétation. Les métadonnées doivent renseigner dans le détail le plus fin les informations sur chaque colonne. Cela inclut par exemple les limites supérieures et inférieures autorisées des valeurs (mesure, poids, coordonnées GPS…) ou les liens entre les colonnes.

Valeurs numériques

Les valeurs numériques, entières ou décimales, font l'objet de différentes conventions dans différents pays ou régions. Par exemple :

- Dans les pays anglo-saxons, le séparateur de décimale est « . » (on écrit « 1.5 » par exemple). Dans les pays francophones, ce séparateur est « , » (on écrit « 1,5 » par exemple)

[32] Pour plus d'information sur l'encodage des caractères, voir https://www.w3.org/International/questions/qa-what-is-encoding.fr.html
[33] https://www.unicode.org/standard/standard.html
[34] https://www.iso.org/fr/standard/76835.html
[35] https://fr.wikipedia.org/wiki/American_Standard_Code_for_Information_Interchange
[36] https://fr.wikipedia.org/wiki/UTF-8
[37] https://www.w3.org/International/questions/qa-choosing-encodings.fr.html

- Dans les pays anglo-saxons, un séparateur de milliers est utilisé et peut-être l'espace ou « , » (on écrit « 1 000 » ou « 1,000 »). Dans les pays francophones, il n'y a pas de séparateur de milliers (on écrit « 1000 »)

Ces différences génèrent des problématiques d'interopérabilité, notamment au niveau international. De ce fait, l'organisme de standardisation IEEE a défini une norme internationale (IEEE 754-2019[38]) qui impose le « . » comme séparateur de décimal et interdit l'utilisation de séparateur de milliers (par exemple « 1001.5 » respecte cette norme).

Date

La gestion des dates, soit courte (uniquement le jour, le mois et l'année) soit longue (jour, mois, année, heure, minute, seconde) pose des problématiques d'interopérabilité de plusieurs ordres :

- Dans les pays anglo-saxons, les dates courtes sont écrites selon le format MM-JJ-AA (mois-jour-année) alors que dans les pays francophones, le format est JJ-MM-AAAA (jour-mois-année)
- Le séparateur entre les jours, mois et années peut prendre plusieurs formes (« - », « / »…)
- Dans les dates longues, la connaissance du fuseau horaire est essentielle pour comparer différentes dates entre elles
- Plusieurs formats de dates longues existent. Par exemple « 12/05/2022 2:57pm », « 2002-05-12 12:27:58 » ou « 2022-05-12T14:57:00Z **»**

Pour résoudre ces défis, l'organisme de standardisation ISO a défini une norme (ISO 8601[39]), norme également adoptée par l'IMANOR (NM ISO 8601[40]) qui fixe un format de date universel (AAAA-MM-JJTHH:MM:SS(.sss)Z exemple : 2022-05-12T14:57:00Z pour le 12 mai 2022 à 14:57:00 UTC). Ce format est référencé dans le CGI.

Numéro de téléphone

La représentation des numéros de téléphone pose différents problèmes d'interopérabilité :

- Le groupage des chiffres et les séparateurs entre les groupes de chiffres varient. Par exemple, un numéro français ou marocain pourrait s'écrire 05 37 54 56 57 et un numéro américain 555-555-1234.
- Certaines notations incluent le code pays, voire le signe « + » et parfois des parenthèses pour indiquer un indicatif à ne pas utiliser à l'international. Par exemple les numéros ci-dessus pourraient être écrits : +33 5 37 54 56 57, +212 (0)5 37 54 56 57 ou 1-555-555-1234.

[38] https://ieeexplore.ieee.org/document/8766229
[39] https://www.iso.org/fr/iso-8601-date-and-time-format.html
[40] https://www.imanor.gov.ma/Norme/nm-iso-8601/

Afin de résoudre ces défis, l'Union Internationale des Télécommunications (UIT) a adopté un standard (ITU E.164[41]) qui impose une représentation unique des numéros de téléphone avec les caractéristiques suivantes

- Le numéro commence par le code international du pays sans le signe +
- Puis le numéro complet sans groupage et sans séparateur directement après le code pays et sans les chiffres optionnels locaux

Les exemples précédents s'écriraient donc : 33537545657, 212537545657 et 15555551234.

Coordonnées géographiques

Les coordonnées géographiques sont des types de données couramment utilisés et qui présentent des défis spécifiques d'interopérabilité qui sont détaillés ci-dessous :

- Il existe deux notations possibles pour les coordonnées géographiques : soit sous format décimal, soit sous format degrés/minutes/secondes
- Les coordonnées géographiques correspondent aux coordonnées d'un point dans un référentiel donné, appelé système géodésique[42]. Il existe plusieurs systèmes géodésiques comme WGS 84, NAD 83, PZ-90, GCJ-02 ou BD-09. Le premier est de loin le plus utilisé parce qu'il est associé au système GPS.
- Les coordonnées géographiques sont constituées de deux parties d'information (longitude et latitude). Certaines notations et certains logiciels, par exemple Google Maps, utilisent la notation « latitude, longitude ». D'autres, comme Open Street Map, utilisent la notation « longitude, latitude ».

Il n'y a pas de standard établi qui fournisse une référence pour résoudre ces défis. Pour maximiser l'interopérabilité des données GPS, les bonnes pratiques recommandent :

- La mention du système géodésique dans les métadonnées structurelles
- L'utilisation de la notation décimale pour faciliter la vérification formelle du contenu
- La séparation de la latitude et la longitude en deux colonnes séparées

Données atomiques

Comme mentionné au paragraphe précédent, certains types de données comme les coordonnées géographiques regroupent plusieurs informations ensemble. Dans de tels cas, les bonnes pratiques recommandent de séparer ces données en plusieurs colonnes pour faciliter la validation formelle de chaque colonne.

Données textuelles

Les données textuelles, par exemple une adresse ou un nom, sont les données les plus difficiles à normaliser et à comparer. En effet, différents éléments impactent les comparaisons : la casse du texte, la langue utilisée, les espaces, etc. Il n'y a pas de standards

[41] https://www.itu.int/rec/T-REC-E.164/
[42] https://fr.wikipedia.org/wiki/Syst%C3%A8me_g%C3%A9od%C3%A9sique

internationaux ou de conventions qui permettent de résoudre ces défis. Deux approches complémentaires sont à considérer :

- Une documentation exhaustive du champ dans les métadonnées pour préciser la langue et les conventions (casse…) utilisées
- L'utilisation de référentiels de données pour les informations clés comme les adresses, les noms (lieux, noms propres, noms d'espèce, etc.)

Seule l'utilisation de référentiels permet de résoudre toutes les ambiguïtés liées aux valeurs textuelles.

Valeur manquante

Il est courant que dans certaines lignes d'un tableau de données des valeurs soient manquantes, soit parce qu'elles ne sont pas applicables à la ligne donnée, soit parce que l'information n'est pas disponible. Suivant les conventions locales, plusieurs options pour marquer ces valeurs manquantes sont utilisées comme 0 ou « - » voire « . » ou d'autres caractères. Toutes ces options présentent plusieurs inconvénients comme :

- L'impossibilité de valider formellement les données (par exemple quand une chaîne de caractères remplace une donnée numérique)
- Un impact sur les formules qui seraient appliquées (moyenne…) si la valeur utilisée pour indiquer la valeur manquante est 0

De ce fait, les bonnes pratiques recommandent de laisser la cellule vide pour les valeurs manquantes.

Unités de mesure

Les données numériques sont généralement associées à une unité de mesure en fonction de ce qu'elles reflètent comme information : des longueurs, des poids, des volumes, des surfaces, etc. Plusieurs systèmes métriques existent comme le système international d'unités[43] ou le système impérial[44]. Ces systèmes utilisent différentes unités en fonction du contenu. Les bonnes pratiques internationales recommandent le système international d'unités qui est standardisé par l'organisme ISO (ISO 80000-1[45]). Dans tous les cas de figure, il est indispensable de mentionner dans les métadonnées structurelles pour chaque colonne le système métrique utilisé et l'unité de la valeur.

[43] https://fr.wikipedia.org/wiki/Syst%C3%A8me_international_d%27unit%C3%A9s
[44] https://fr.wikipedia.org/wiki/Unit%C3%A9s_de_mesure_anglo-saxonnes#:~:text=Le%20%C2%AB%20syst%C3%A8me%20imp%C3%A9rial%20d'unit%C3%A9s,en%20bronze%20date%20de%201845).
[45] https://www.iso.org/fr/standard/76921.html

Récapitulatif des standards recommandés :

Le tableau ci-dessous résume les standards et approches recommandées. Les standards avec une * sont les standards qui font également partie du cadre général d'interopérabilité.

Dimension		Standard recommandé	Approche recommandée
Format de fichier	**Données tabulaires**	CSV - RFC 4180 (*)	La première ligne contient les en-têtes de colonnes et le reste des lignes sont des lignes de données Les données sont sous format long
	Image bitmap	PNG(*), JPEG (*), TIFF (*)	
	Image vectorielle	SVG (*)	
	Texte	TXT(*), RTF (*), HTML(*)	
	Données géospatiales	Geojson, geotiff ou shapefile	
	Données structurées/hiérarchiques	XML(*), JSON (*), EPUB	
Encodage des fichiers		UTF-8 (*)	
Type de données	**Valeurs numériques**	IEEE 754-2019	
	Date	NM ISO 8601 (*)	
	Numéro de téléphone	ITU E.164	
	Coordonnées géographiques	Système géodésiques WGS 84 (GPS)	Séparation des informations de latitude et de longitude en 2 colonnes Utilisation de la notation décimale (et non de la notation en degré, minute et seconde)
	Données textuelles	Utilisation de référentiel	
	Données atomiques	Les données ayant plusieurs composantes comme les données géographiques sont séparées (une colonne par composante)	
	Valeur manquante	Cellule/champs vide	
Unité de mesure		Système international de données - ISO 80000-1	

4. La qualité des jeux de données

Cette section définit les critères de qualité auxquels doivent se conformer les jeux de données et les données qui les composent pour être publiés sur le portail national de données ouvertes. Le concept de « qualité des données » couvre à la fois la qualité intrinsèque des données[46] et le niveau d'interopérabilité de ces données. Deux niveaux de qualité sont définis :

- Un niveau de qualité minimale qui permet l'exploitation des données publiées
- Un niveau de qualité maximale qui maximise l'interopérabilité des données et donc leur potentiel de valorisation[47]

Il est important de noter que les critères définis ci-dessous de définissent pas entièrement l'ensemble des éléments nécessaires à la publication de données ouvertes. La liste complète des éléments à prendre en compte pour la publication d'un jeu de données est décrite dans la « Fiche d'évaluation de la qualité des jeux de données ouvertes » et dans le « Manuel à destination des Gestionnaires de Données Open Data (GdD) ». Les critères dans les sous-sections suivantes détaillent les éléments mentionnés dans les principes techniques du premier document cité (fiche d'évaluation). En dehors de ces critères, le jeu de données doit répondre à la définition des données ouvertes :

- Les données publiées doivent être sélectionnées conformément aux obligations légales décrites à la Section 1.2 du présent document. En particulier les jeux de données contenant des données personnelles qui relèvent de la Loi 09-08 doivent être anonymisés avant publication. Les techniques d'anonymisation sont présentées en détail dans le « Manuel à destination des Gestionnaires de Données Open Data (GdD) ».
- Les jeux publiés doivent être associés à une licence ouverte de réutilisation
- Afin de maximiser le potentiel de réutilisation, les choix des jeux de données à publier doivent être priorisés pour répondre aux demandes des différents types de réutilisateurs. L'approche de priorisation est également décrite à la fois dans le guide de l'inventaire et la fiche d'évaluation de la qualité des jeux de données

Niveau de qualité minimum

L'objectif de ce niveau est de viser à obtenir des jeux de données présentant les critères minimums au regard des standards de l'Open Data. Ces critères minimums sont les suivants :

- **Le Jeu de données est fourni sous un format ouvert** :

[46] La qualité intrinsèque des données correspond à la capacité des données à refléter le monde réel

[47] A noter que cette section ne se focalise que sur les aspects techniques (qualité intrinsèque et interopérabilité) d'un jeu de données. Le potentiel de valorisation d'un jeu de données est également en grande partie liée à la demande des réutilisateurs pour ce jeu de données et donc à la priorisation des publications. Cet aspect est couvert dans le « Guide méthodologique pour la mise en œuvre d'un inventaire de données au sein d'une structure publique de l'Administration »

- Le format CSV est utilisé pour les formats tabulaires. Le format CSV retenu est le format standard (RFC 4180[48]) international :
  - Le fichier utilise l'encodage UTF-8
  - La première ligne du fichier contient les en-têtes de colonne
  - Les données numériques répondent au format IEEE 754-2019 (séparateur décimal « . » et pas de séparateur de milliers. Exemple : 5320.87)
  - Les champs sont séparés par une virgule (« , »)
  - Les fins de lignes sont délimitées par le caractère « CRLF »
  - Le format Excel est toléré
- Les formats TXT, RTF ou HTML sont utilisés pour les fichiers textes. Les format Word, PPT et PDF sont tolérés.
- Les formats XML ou JSON sont utilisés pour les formats hiérarchiques
- Les formats PNG, TIFF ou JPEG sont utilisés pour les images bitmap
- Le format SVG est utilisé pour les images vectorielles
- Les formats GeoJSON, geotiff ou shapefile sont utilisés pour les formats de données spatiales
- Les standards mentionnés ci-dessus sont recommandés mais tout autre format ouvert et qui respecte un standard international est autorisé

- **Les données sont structurées et exploitables informatiquement** (machine-readable)
- **Les données doivent être des données brutes désagrégées**
- **Les données doivent être à jour au regard du cycle de collecte**
- **Le jeu de données doit être documenté (métadonnées)** en suivant le cadre de l'inventaire et l'ensemble des métadonnées retenues dans l'inventaire doivent être renseignées. Les métadonnées essentielles sont présentées dans le « Guide méthodologique pour la mise en œuvre d'un inventaire de données au sein d'une structure publique de l'Administration[49] ».

Tous les jeux de données publiés sur le portail national de données ouvertes doivent répondre à ces critères. Les jeux de données respectant ce cahier des charges seront identifiés avec un logo spécifique sur le portail (« Jeu de qualité »).

Niveau de qualité optimum

Le niveau de qualité optimum est un ensemble de critères supplémentaires par rapport au niveau précédent afin d'améliorer l'interopérabilité des données et de conduire des actions supplémentaires pour vérifier la qualité intrinsèque des données. En plus des contraintes du niveau précédent, le jeu de données devra répondre aux critères suivants :

- **Format des jeux de données** :
  - Les formats propriétaires Excel, Word et PPT ne sont plus tolérés

[48] https://datatracker.ietf.org/doc/html/rfc4180
[49] https://www.data.gov.ma/sites/default/files/2023-03/guide_inventaire_1.3_Fev.%202023.pdf

- Le jeu de données utilise des formats spécifiques au domaine quand ils existent. Par exemple
  - OCDS pour les données concernant les marchés publics
  - GTFS pour les données de transport[50]

Il existe de très nombreux standards spécifiques par secteur (SIRI, GDF, Lidar, NeTex…) qui augmentent le potentiel de réutilisation des données et qui peuvent également être utilisés.

- **Structure des données** : les données tabulaires utilisent un format long pour la présentation des données
- **Standard de données** : Les données sont représentées en utilisant des standards internationaux :
  - **Dates au format ISO 8601** (2022-05-12T14:57:00Z)
  - **Numéros de téléphone au format ITU E.164** (33688559955 – code pays et numéro local sans espace)
  - **Les champs textes doivent renseigner la langue utilisée** dans les métadonnées structurelles.
  - **Les valeurs nulles de champs doivent être vides**
  - **Les coordonnées géographiques**
    - Sont exprimées en degrés décimaux (33.9693414)
    - Sauf données spatiales spécialisées nécessitant un système géodésique particulier**,** les coordonnées géographiques font référence au système géodésique WGS84 (GPS)
    - La latitude et la longitude doivent être séparées en deux colonnes
  - Les unités des valeurs de chaque colonne doivent être exprimées dans les métadonnées structurelles
  - **Les champs qui correspondent à des référentiels doivent utiliser ces référentiels** et cette information doit apparaître dans les métadonnées structurelles.
- **Métadonnées**
  - Pour les données tabulaires, un schéma de validation de données doit être attaché aux données
  - Le producteur fournit toutes les informations adéquates sur la collecte de données pour maximiser la confiance des réutilisateurs et la crédibilité des données. Ces informations incluent le processus de collecte, les étapes de digitalisation, la méthode de collecte (enquêteurs, capteurs…) et la qualité associée (formation des enquêteurs…).
- **Les données doivent être vérifiées :** quatre composantes seront mises en œuvre :

[50] GTFS est le standard qui a permis par exemple l'émergence d'applications de mobilité et de transport multimodaux : des applications qui permettent à un usager qui désire se rendre d'un point A à un point B de connaitre toutes les options possibles et l'articulation des différents modes de transport. Voir par exemple l'application transit app https://play.google.com/store/apps/details?id=com.thetransitapp.droid&hl=en_US

- **La validation formelle des données :** les données doivent être valides au regard du schéma de validation. Les métadonnées doivent inclure un schéma de validation détaillé et l'application de ce schéma ne doit pas générer d'erreurs. En dehors de la validité formelle, le producteur applique les analyses courantes des données pour vérifier leur cohérence et leur précision avant de les publier, et documente ces tests.
- **La vérification de la complétude des données :** La complétude des données est une dimension importante de la qualité des données. Avant publication, le producteur de données devra s'assurer des points suivants
  - Toutes les séries historiques sont fournies.
  - Chaque série est complète

Les jeux de données respectant ce cahier des charges seront identifiés avec un logo spécifique sur le portail (« Jeu de qualité optimale »).