Ouverture des données publiques au Maroc

Manuel à destination des Gestionnaires de l'Open Data (GdD)

Avril 2021

#ADD
Agence de Développement du Digital

WORLD BANK GROUP

Table des Matières

Table des Matières 2
Introduction 3
1. Les fondamentaux de l'Open Data 4
1.1. Les concepts de l'Open Data 4
1.2. Le cadre légal de l'Open Data 8
1.3. Les acteurs de l'Open Data 9
1.4. Les bénéficiaires d'une initiative Open Data 11
2. Cycle de vie de la donnée et Approche Open Data 13
2.1. Cycle de vie de la donnée 13
2.2. Les interlocuteurs du GdD 16
3. Publication de données 18
3.1. Structurer et organiser des données désordonnées à l'aide d'un tableur 19
3.2. Validation des données 30
3.3. Anonymisation des données 34
3.4. Génération d'un fichier CSV et publication sur le portail Open Data 36
Annexe A - Exemple complet de préparation de données 38

Introduction :

Ce document est un manuel à destination des Gestionnaires de Données Ouvertes (GdD) au sein des structures publiques. L'objectif de ce manuel est d'apporter à ces gestionnaires un support opérationnel dans la publication de données ouvertes sur le portail national de données ouvertes[1] ou sur les portails de leurs structures respectives. Le but de ce manuel est de documenter chaque étape du processus de publication à partir de la collecte et du traitement des données internes jusqu'à l'intégration sur le(s) portail(s).

Ce manuel est divisé en trois sections. La première partie présente les concepts clés de l'Open Data, la deuxième partie présente le cycle de vie des données et la troisième partie décrit et illustre en détail les étapes de la publication de données ouvertes.

Ce manuel a pour objectif d'être un document vivant qui s'enrichit au fur et à mesure des retours d'expérience des GdD, des défis qu'ils rencontrent et des stratégies qu'ils mettent en place pour les résoudre. Les commentaires et les propositions d'évolutions sont à envoyer à Opendata@add.gov.ma.

[1] http://www.data.gov.ma/

1. Les fondamentaux de l'Open Data :

Les données ouvertes en tant que concept trouvent leur origine dans la conviction que les données gouvernementales devraient être librement disponibles pour être publiées, utilisées et réutilisées sans limitations ni contrôle. Cette idée fondamentale s'est répandue à travers le monde depuis 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 lancé leur propre initiative de publication de jeux de données en ligne, soit travaillent à la mise en œuvre d'initiatives de données ouvertes dans un proche avenir. Dans le 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 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 notamment locaux et fédéraux, 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 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.

En quelques mots, une initiative de données publiques ouvertes est une excellente occasion pour les citoyens d'un pays donné d'améliorer leur vie. Cependant, pour obtenir un tel impact, il est essentiel que les données soient publiées à temps en utilisant des formats standards 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 de la publication et 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 critères :

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

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

- **La donnée est légalement ouverte** : une donnée est légalement ouverte si elle est associée à des droits de réutilisation (appelés « licence de réutilisation ») large qui autorisent la réutilisation, le croisement et la redistribution de ces données, dans un contexte commercial ou pas, et requièrent uniquement la mention de la source des données. Dans le cadre de l'Open Data, le Maroc a adopté une licence spécifique présentée sur le portail Open Data[2].

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. Ce principe s'appuie sur deux éléments :
  - Les données gouvernementales ont été collectées grâce aux impôts et taxes et ont donc été déjà payées par le contribuable.
  - La valorisation directe (vente des données) génère un revenu et un impact économique négligeable comparé à l'impact de la réutilisation de ces données en termes de création d'emploi, de création de richesse et de retombées directes (taxes, impôts sur la valeur ajoutée, etc.) et indirectes pour le pays.

  Il est possible que l'administration 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 ([3]). Ce portail héberge des jeux de données de certains organismes et institutions, et référence des jeux de données hébergés sur des portails open data sectoriels. Chaque structure publique a le choix, soit de mettre en place son propre

[2] http://www.data.gov.ma/fr/la-licence
[3] http://www.data.gov.ma

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. 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). Deux types de métadonnées doivent être associés aux données :
    - **Les métadonnées descriptives**, qui correspondent à des métadonnées externes, c'est-à-dire des métadonnées typiquement utilisées 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, la couverture géographique et spatiale, la périodicité de mise à jour, le format, la licence de réutilisation et les mots clés.
    - **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.

  En résumé, Les métadonnées descriptives 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 réside, pour une partie importante, dans sa maintenance qui couvre plusieurs aspects :
    - **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 permet généralement de mettre en lumière des problèmes spécifiques. Quand ces problèmes sont remontés par les réutilisateurs, via notamment le portail national de données, 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. 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. Ces données de références communes à l'administration sont appelées des référentiels de données et il est essentiel de s'assurer que les jeux de données ouvertes publiés les utilisent. A la date de rédaction de ce document, il est prévu qu'un ensemble de référentiels soient élaborés par l'Agence de Développement du Digital (ADD), en concertation avec le Comité de Pilotage Open Data. Il est important de noter que la mise en place et l'utilisation de référentiels est un processus à long terme. Dans un premier temps, l'administration se concentrera sur les référentiels et les standards de données les plus importants qui sont utiles à toutes les structures publiques (par exemple, le référentiel des noms géographiques ou le standard des données pour les marchés publiques). Ensuite, au fur et à mesure, des référentiels sectoriels seront adoptés pour faciliter notamment le croisement de données entre les administrations au niveau national, régional et local. Enfin, la nécessité et la volonté de concevoir et d'utiliser des référentiels et des standards de données se développeront au sein des administrations au fur et à mesure de l'augmentation de l'utilisation des données et de leur exploitation par ces administrations. Plus les équipes mettront les données au cœur de leur activité, plus elles saisiront la nécessité de ces référentiels pour faciliter leur travail, et plus elles mettront de l'énergie à les concevoir, les adopter, les diffuser et les utiliser.

L'ensemble de ces éléments sont documentés et consolidés dans la « Charte Internationale sur les données ouvertes[4] » 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.

[4] https://opendatacharter.net/principles-fr/

En dehors de ces aspects techniques, un des principes clés de l'Open Data, élément de la Gouvernance Ouverte[5], est la collaboration entre l'administration et les acteurs non-gouvernementaux. 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. De façon générale, le développement de l'Open Data s'appuie sur deux piliers législatifs essentiels que sont les lois d'accès à l'information et les lois relatives à la protection des données personnelles.

Au Maroc, il s'agit de :

- **La loi n°31-13 du 22 Février 2018, relative au droit d'accès à l'information**[6] : La loi n°31-13 définit de façon précise les informations publiques qui peuvent être publiées sur demande ou 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**[7] : La loi n°09-08 fournit un cadre réglementaire pour la protection de la vie privée et identifie en particulier les données qui rentrent dans le cadre de cette loi, les traitements qui doivent être appliqués avant publication.

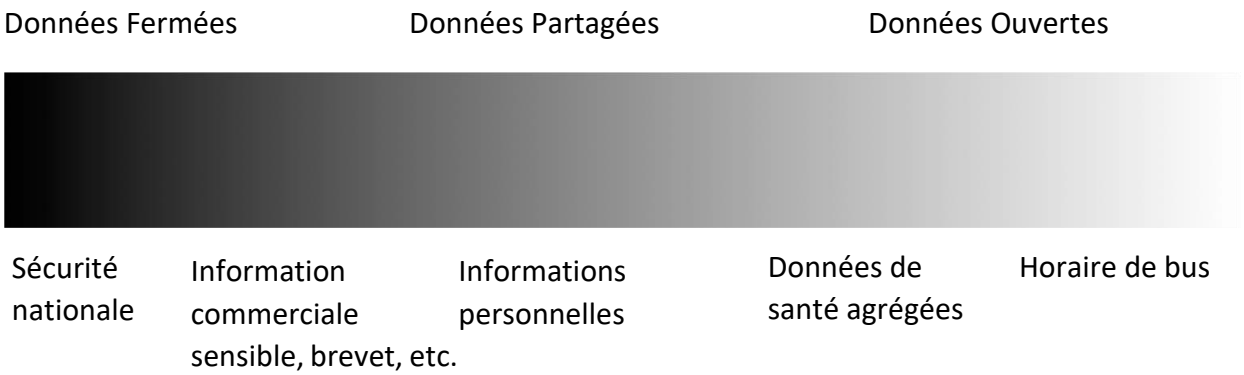
Il est important de retenir les points clés de ces législations ainsi que de la Constitution marocaine :

- L'Article 27 de la Constitution de 2011 stipule : « 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 met en application cet article. Toutes les informations détenues par l'administration sont disponibles pour tout citoyen quel que soit leur format et leur type (Article 2.a) dans les limites décrites dans l'Article 7 qui interdit la publication des données suivantes :

[5] https://fr.wikipedia.org/wiki/Gouvernement_ouvert
[6] https://www.ilo.org/dyn/natlex/docs/ELECTRONIC/107094/131706/F-985611108/MAR-107094.pdf
[7] http://adala.justice.gov.ma/production/html/Fr/156780.htm

- Informations relatives de la défense nationale et à la sécurité intérieure et extérieure de l'Etat.
- Informations relatives à la vie privée.
- Informations relatives à la protection des sources d'information.
- Informations couvertes par un secret (propriété intellectuelle, brevet, secret commerciaux, délibération du Conseil des Ministres et du Conseil du Gouvernement, etc.).
- Informations impliquées dans les procédures juridiques en cours.

- Il est essentiel de noter que les lois n°31-13 et n°09-08 stipulent clairement que lorsque qu'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 retrait des informations sensibles (anonymisation, oblitération, etc.) 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 détaillée du cadre juridique marocain dans le cadre de l'Open Data a été réalisée en 2019-2020 par la Banque Mondiale dans le cadre de la réalisation de l'étude sur la préparation du Maroc à l'ouverture des données publiques (« ODRA »[8]).

1.3. Les acteurs de l'Open Data :

La réussite d'une initiative Open Data et la mise en place d'une dynamique positive au sein des différentes structures publiques requiert la collaboration et la coordination de plusieurs acteurs présentés ci-dessous :

- **Le Responsable de la structure** : 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 entraine la mobilisation des personnels, l'impact de l'ouverture des données sera faible.

[8] Voir le Chapitre 2 du rapport de l'étude ODRA

- **Le Responsables 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'actions, coordonne les activités, mets 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 a également un rôle de collaboration et coordination avec l'ADD 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[9].
- **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, 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 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 directives de la 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 interagissent avec eux sur les analyses de données ou les corrections d'erreurs potentielles. Les détails du processus de publication sont présentés dans le chapitre 3 de ce présent document.
- **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. D'autre part, ils permettent également d'améliorer la qualité

[9] Cf. Manuel à destination des Responsables Open Data

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.

1.4. Les bénéficiaires d'une initiative Open Data :

L'objectif principal d'une initiative Open Data est de générer 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'autres 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 permettent 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[10], un organisme non-gouvernemental anglais spécialisé dans l'innovation, a investi £1.2M pour stimuler l'innovation autour des données gouvernementales du Royaume-Uni et a mesuré au bout de 3ans 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 création de richesses.
- **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 des nouveaux services. A titre d'exemple, l'ODI (Open Data Institute, Institut indépendant dédié à l'Open Data[11]) a mené une étude en Angleterre qui a montré que 270 sociétés utilisaient les données gouvernementales ouvertes, représentant un chiffre d'affaires de £92bn et plus de 500.000 emplois.

[10] https://www.nesta.org.uk/
[11] https://theodi.org/

- **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 emploi, 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, etc.) grâce aux données publiées.
- **Les médias** : L'émergence de l'Open Data dans le monde a permet 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 informer 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.

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

2.1. Cycle de vie de la donnée :

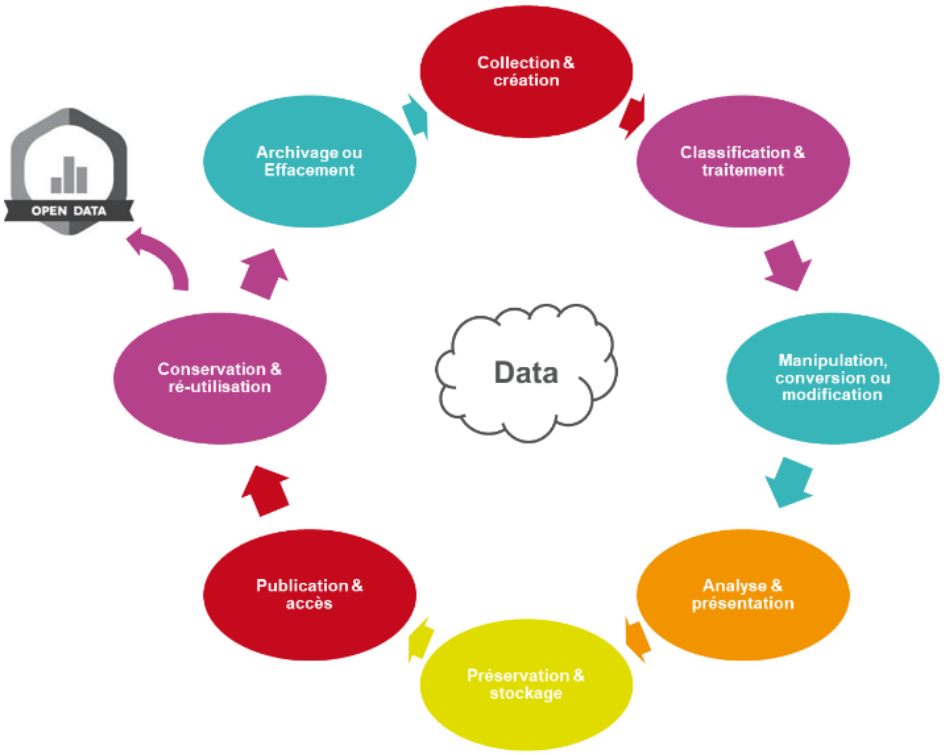
Le succès d'une initiative Open Data requiert à la fois la publication d'un nombre important de jeux de données de qualité, mis à jour régulièrement et également la réutilisation de ces données par les réutilisateurs afin de générer de la valeur économique et sociale à partir de ces données. Cependant, cette définition présente à première vue deux faiblesses importantes :

- **Une focalisation sur la publication de données** : Cette définition tend à suggérer que les structures publiques doivent principalement se focaliser sur la publication de données comme objectif et non comme le résultat d'un processus de gestion des données pérennes et robustes.
- **Un bénéfice et une génération de valeur pour les acteurs extérieurs (les réutilisateurs)** : Cette définition tend à suggérer qu'il est demandé aux structures publiques un effort et un travail (la publication de données) qui bénéficie in fine à d'autres acteurs externes à la structure (les réutilisateurs).

De nombreux pays ont tenté d'appliquer à la lettre cette définition et ils 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.



- **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.
- **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.

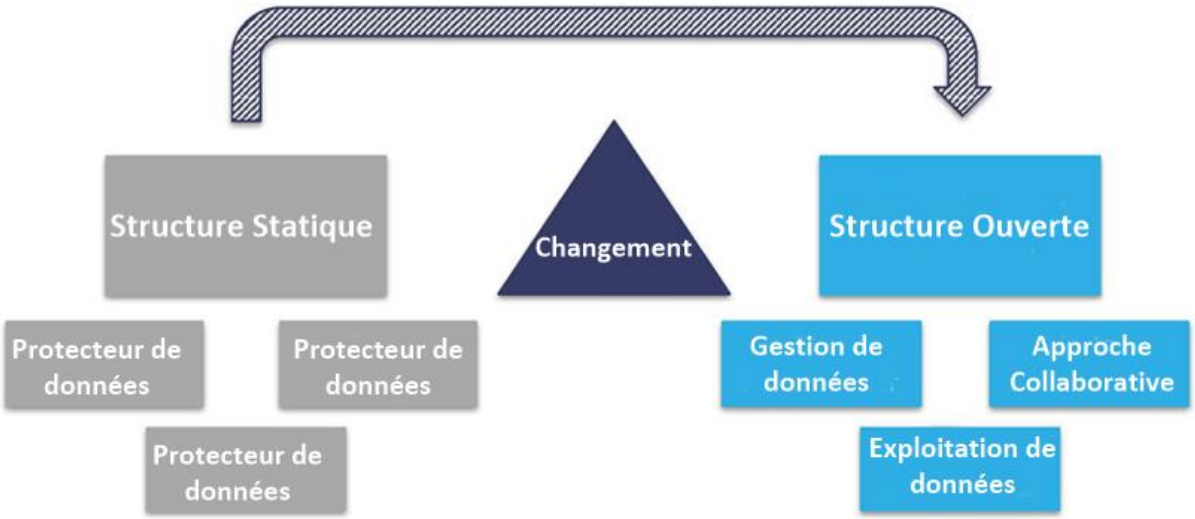
La mise en place concomitante ces trois éléments, supportée par une législation incitative, sont les clés de la réussite d'une initiative Open Data et apporte un ensemble de bénéfices mesurables pour les structures publiques notamment :

- **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 apparaitre 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 permettent à l'Administration de facilement codévelopper avec ces acteurs de l'innovation de nouvelles applications innovantes et à faible coût comparé aux appels d'offre classiques.

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



La figure ci-dessus met en évidence le rôle fondamental des gestionnaires de données. Il est important de noter que ces gestionnaires de données ont deux rôles principaux :

- **Producteur de données** : la première fonction du GdD consiste à préparer, publier sur le portail Open Data et maintenir les données qu'il gère. Dans le cadre de cette

fonction, il est amené à interagir avec les réutilisateurs pour fournir un support à l'exploitation des jeux publiés (réponse aux questions, correction d'erreurs…).

- **Analyste de données** : la deuxième fonction du GdD est d'exploiter et analyser les données qu'il gère afin d'en extraire de nouvelles connaissances, ou de faciliter la prise de décision. L'exploitation du plein potentiel de ces données nécessite le croisement avec d'autres jeux de données issus de sa propre structure ou d'autres structures. Dans ce cadre, le GdD est un réutilisateur des données de l'Administration à part entière. Enfin, le GdD est également en contact avec les autres réutilisateurs pour discuter de leur propre analyse, utiliser, le cas échéant, leurs compétences et favoriser le diagnostic ou l'analyse commune de ces données.

Ce manuel est dédié à la fonction de producteur de données.

2.2. Les interlocuteurs du Gestionnaire de Données :

Le gestionnaire de données est un des acteurs de l'Open Data comme présenté à la section 1.3, et dans le cadre de sa mission de producteurs de données publiques, il interagit avec plusieurs des autres acteurs comme présenté dans le diagramme ci-dessous :

Le GdD travaille en étroite collaboration avec :

- **Le Responsable Open Data (ROD) :** Le ROD est la personne au sein de la structure du GdD en charge du développement de l'Open Data. Il est au service des GdD pour faciliter leur travail de publication. En particulier le ROD :
    - **Aide à la résolution des problèmes administratifs :** Le GdD se focalise sur les problèmes techniques liés à la publication. Le ROD est en charge de la résolution des problèmes administratifs, notamment des droits d'accès et de la classification des données pour identifier celles qui peuvent être publiées.
    - **Communique les standards techniques adoptés au niveau national (par le Comité de Pilotage Open Data) :** La publication de données ouvertes est régie par des standards techniques (métadonnées obligatoires, format de fichiers, référentiels…) qui sont adoptés par les instances compétentes au niveau national, en l'occurrence le Comité de pilotage Open Data. Le ROD transmet et informe les GdD sur les éléments techniques à intégrer dans le processus de publication.
    - **Organise les formations nécessaires aux GdD** : la publication de données ouvertes, tout comme l'analyse et le croisement de jeux de données, nécessitent une expertise spécifique qui fait parfois défaut à certains GdD. Le rôle du ROD est d'identifier les besoins et de s'assurer que les GdD suivent les formations appropriées et acquièrent les capacités nécessaires à l'exécution de leurs tâches.
    - **Aide à la priorisation des publications** : Par ses interactions avec les réutilisateurs, et au travers des choix stratégiques de la structure identifiés

dans le plan annuel d'activité Open Data, le ROD facilite la priorisation des publications par les GdD.

- **Remonte les défis qui pourraient se présenter auprès du Comité de Pilotage Open Data et aide à leur résolution** : Enfin, il peut arriver que les GdD soient confrontés, pour des jeux de données particuliers, à des défis qu'ils ne savent pas résoudre. Le rôle du ROD en collaboration avec l'ADD et le Comité de Pilotage Open Data, est d'aider le GdD et de lui trouver le support technique adéquat.

Le ROD facilite également l'accès aux données des autres structures dans le cadre du rôle d'analyste de données du GdD.

- **Le service informatique** (SI) : Les GdD travaillent en collaboration avec le service informatique qui pourra les aider à l'automatisation des tâches de publication et l'intégration de la publication Open Data dans le cycle global de production des données.
- **Le chargé d'accès à l'information** (CAI) : Le CAI est un interlocuteur potentiel des GdD. En fonction des requêtes de demande d'informations reçues ou dans le cadre de la politique de publication proactive, le CAI en collaboration avec le ROD pourra demander la publication de jeux spécifiques aux GdD.
- **L'entité en charge des archives** : L'entité en charge des archives est un interlocuteur potentiel des GdD. En fonction des requêtes de demande d'accès aux archives, l'entité en charge des archives en collaboration avec le ROD pourra demander la publication de jeux spécifiques aux GdD ou l'informer sur les droits de publications de jeux de données spécifiques.
- **Le Responsable des Données Personnelles (RDP)** : Le RDP est un interlocuteur important des GdD qui leur permettra d'identifier les jeux de données qui contiennent des données personnelles et qui doivent donc être anonymisés avant publication.
- **Les réutilisateurs** : Dans le cadre de son rôle de producteur de données, le GdD apporte un support aux réutilisateurs, notamment en corrigeant les erreurs éventuelles remontées, en répondant aux questions, ou en mettant à jour les données en fonction de leur cycle de production.
- **Le portail Open Data** : Le portail national Open Data n'est pas un acteur de l'Open Data à proprement dit, mais un élément essentiel pour les GdD. D'une part, il est le point central de publication des données. D'autre part, il est également l'outil principal d'interaction avec les réutilisateurs au travers des questions et commentaires reçus.

3. Publication de données :

La tâche de publication de données ouvertes se compose de quatre étapes :

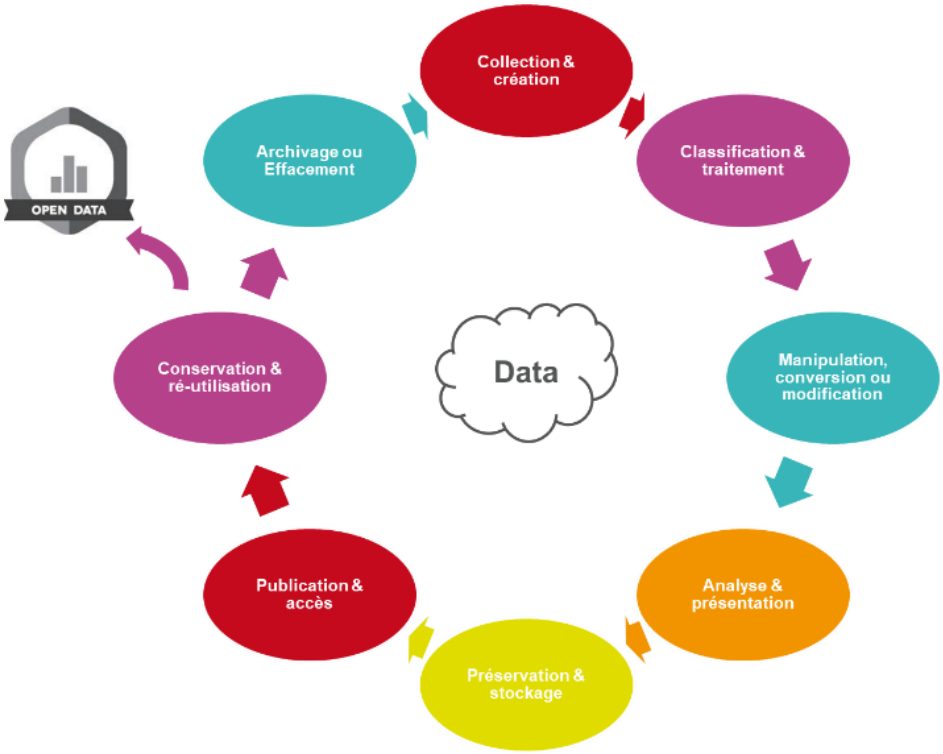
- **La sélection des données :** La première phase essentielle est pour le GdD d'identifier parmi les données, celles qui peuvent être publiées dans le cadre de l'Open Data. Généralement, un très grand nombre de jeux sont susceptibles d'être publiés. La difficulté majeure du GdD sera de prioriser les publications et sélectionner les jeux qui ont potentiellement plus d'impacts. La priorisation est un travail relativement complexe qui se conduit en collaboration avec le ROD en fonction des priorités de la structure, de l'importance des jeux de données pour d'autres administrations, et de la demande émanant des réutilisateurs non-gouvernementaux. A noter que la mise en place d'un inventaire de données au sein de la structure facilitera la sélection des données et permettra aux GdD d'identifier les jeux qui peuvent être publiés ainsi que les traitements à leur appliquer avant publication.
- **La préparation des données** : l'objectif de cette étape est de transformer des données destinées principalement pour un utilisateur humain en des données exploitables informatiquement, disponibles sous un format ouvert et qu'il est possible d'analyser et de croiser.
- **La validation des données** : L'objectif de cette étape est de détecter les incohérences ou les problèmes de format et de corriger les données afin de maximiser leur qualité avant publication.
- **L'anonymisation des données** : La dernière étape consiste à anonymiser les données qui contiennent des informations rentrant dans le cadre de la loi n°09-08 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel.

Chacune de ces tâches est présentée en détail dans les trois sections suivantes de ce chapitre. Il est important de noter que ce manuel s'adresse à des gestionnaires de données maitrisant des outils de tableurs comme Excel ou Google Sheet. Les tâches présentées sont donc des tâches manuelles répétitives à accomplir à partir d'un tableur. Une partie de ces tâches peuvent être automatisées informatiquement par le développement de programmes spécifiques. Le développement de ce type de programmes nécessite des compétences de développeurs informatiques et ne sont pas couvertes dans ce manuel. La mise au point de ces programmes nécessite le plus souvent la collaboration entre les GdD et le service informatique où ce type de profils (développeur logiciel) est disponible. A noter qu'il sera le plus souvent nécessaire aux développeurs informatiques d'acquérir les compétences spécifiques liés à la manipulation de données avant de pouvoir assister les GdD pour la réalisation des programmes informatiques cités.

3.1. Structurer et organiser des données désordonnées à l'aide d'un tableur :

3.1.1. Le cycle de vie de la gestion des données :

Si le processus de création, de maintenance et d'exploitation de nouvelles données est souvent présenté comme un cycle, il s'agit plutôt d'une spirale. Chaque cycle monte en spirale, et donne lieu à plus d'informations. Cependant, l'efficacité de chaque étape est définie par les besoins de ses utilisateurs, et sa pertinence par rapport au processus ou aux événements qu'elle reflète.



3.1.1.1. Collecte et création :

Avant de pouvoir collecter des données, il faut connaître toute une série de choses :

- Pourquoi collectons-nous les données ?
- À quoi servent-elles ?
- Avons-nous le consentement et/ou l'autorisation légale de collecter ces données ?
- Existe-t-il une série de données existantes que ces données complètent ou élargissent ?
- Comment les données seront-elles collectées ?
- Qui sera responsable de la qualité des données et comment cette qualité sera-t-elle mesurée ?
- Qui aura accès aux données et quel est le degré de sensibilité de ces données (par exemple, identification personnelle) ?
- Utilisons-nous une classification ou un format de métadonnées standardisé et adopté?

3.1.1.2. Classification et traitement :

Le créateur des données est le mieux placé pour savoir de quoi il s'agit et devrait attribuer des mots-clés comme descripteurs. Ces données sur les données sont appelées métadonnées. Ce terme est utilisé pour deux concepts fondamentalement différents :

- **Les métadonnées structurelles** correspondent aux métadonnées internes, c'est-à-dire les métadonnées concernant la structure des objets de la base de données tels que les tables, les colonnes, les clés et les indices ;
- **Les métadonnées descriptives**, qui correspondent à des métadonnées externes, c'est-à-dire des métadonnées typiquement utilisées pour la découverte et l'identification, comme des informations telles que le titre, l'auteur, les sujets, les mots clés ou l'éditeur utilisées pour rechercher et localiser un objet;

Les métadonnées descriptives 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 métadonnées peuvent permettre l'interopérabilité entre différents systèmes. Une structure approuvée pour interroger la "fiabilité" d'une série de données peut permettre à des systèmes logiciels indépendants de trouver et d'utiliser des données à distance.

Au-delà des métadonnées, il existe également des mécanismes permettant de structurer les relations entre les hiérarchies de mots clés. Ces mécanismes sont connus sous le nom d'ontologies et, avec les métadonnées, peuvent être utilisés pour définir avec précision et permettre la découverte de données.

L'ajout de métadonnées aux ressources de données existantes peut être un processus coûteux et exigeant en main-d'œuvre. Cela peut devenir un obstacle à la mise en œuvre d'un système complet de gestion des connaissances.

Les types de métadonnées descriptives qu'il est recommandé d'inclure et les termes généralement utilisés pour les désigner sont présentés dans la table ci-dessous :

Obligatoire	Recommandé	Important
Titre	Mots-clés	Dernière mise à jour
Description	Conditions d'utilisation	Fréquence de mise à jour
Thème(s)	Courriel de contact	Couverture géographique
Organisme de publication		Couverture temporelle
		Validité
		Ressources connexes
		Règlements

3.1.1.3. Manipulation, conversion ou altération :

Cette partie du processus est celle où les données sont transcrites, traduites, vérifiées, validées, nettoyées et gérées.

C'est ce qui présente le plus grand risque pour la cohérence des données. Tout changement de format ou toute manipulation, ou même la copie d'un fichier d'un système à un autre, introduit un risque de corruption des données. De même, cela augmente également le risque que des données - qu'elles soient erronées ou non - soient accidentellement communiquées aux utilisateurs ou au public avant qu'elles ne soient prêtes.
C'est ce que l'on appelle **la préparation des données**.

3.1.1.4. Analyse et présentation :

L'analyse ne fait peut-être pas toujours partie du rôle du gestionnaire des données, mais c'est certainement la raison pour laquelle les données sont collectées.

C'est là que les données sont interprétées, combinées à d'autres ensembles de données pour produire une méta-analyse, et que l'analyse devient l'histoire que vous souhaitez raconter à partir des données.

La qualité de l'histoire dépend de celle de vos données et la recherche n'est considérée comme valable que si les données qui l'alimentent sont également publiées.

L'objectif de l'analyse est d'informer le comportement collectif ou individuel, d'influencer la politique ou de soutenir l'activité économique, parmi beaucoup d'autres. La confiance peut être obtenue en "montrant votre fonctionnement" (c'est à dire comment vous avez analysé les données), ce qui inclut la publication des données.

3.1.1.5. Préservation et stockage :

Les données doivent être préservées de la corruption et être disponibles pour une utilisation ultérieure une fois l'analyse initiale terminée. Il est également nécessaire de préserver les données en cas de questions concernant l'analyse.

Le stockage à long terme exige que les **métadonnées** soient bien définies et soient vraiment utiles pour garantir que la compréhension de ce que les données décrivent soit encore possible longtemps après leur collecte initiale.

Les données peuvent finir par être stockées dans plusieurs formats ou sur plusieurs systèmes. Il est essentiel que la primauté soit établie (c'est-à-dire quel ensemble de données est la version originale et prime sur les autres) et que les différents formats soient maintenus alignés.

3.1.1.6. Publication et accès :

Même lorsque les données ne sont diffusées qu'au sein du gouvernement ou d'une structure - et non pour le public - il y aura toujours d'autres personnes qui voudront utiliser vos données, ou qui en tireront profit si elles savent qu'elles existent. La plus grande inefficacité dans la gestion des données survient lorsque la collecte de données primaires est répétée parce qu'une autre administration/structure ou une autre équipe de la structure concernée a besoin des mêmes données sans savoir qu'elles existent déjà.

La diffusion ne consiste pas simplement à mettre les données à la disposition du public, mais aussi à créer un processus prévisible pour cette diffusion.

Les données régulièrement collectées (comme les taux d'inflation par exemple) ont besoin d'un cycle de diffusion prévisible, car de nombreuses entreprises fondent leurs décisions d'investissement sur la disponibilité de ces informations.

Publier un calendrier de diffusion pour vos consommateurs et réutilisateurs de données (et s'y tenir) leur permet de planifier leur propre analyse, ou de réagir à votre analyse.

3.1.1.7. Conservation et réutilisation :

Une fois les données publiées, la question qui se pose est de savoir pendant combien de temps elles seront disponibles, et ce, en tenant compte de la durée de conservation des données prévue par la réglementation en vigueur. Les données de recherche devraient, en règle générale, être disponibles à perpétuité. Les données chronologiques sont d'autant plus utiles qu'elles ont été collectées longtemps. La suppression soudaine de données peut provoquer d'énormes perturbations. Si des systèmes n'ont pas été mis en place, la conservation peut devenir un problème très coûteux. Il est important, pour favoriser la réutilisation, de définir clairement les droits d'auteur et les licences qui permettent de réutiliser librement les données à des fins légitimes.

3.1.1.8. Archivage :

Les jeux de données peuvent devenir très volumineux et peuvent n'être consultés que rarement. Cela peut devenir problématique pour le stockage à long terme. Un processus d'archivage - où les données peuvent être stockées à moindre coût tout en restant accessibles - doit être envisagé.

En résumé, le GdD doit :

- S'assurer qu'il y a un propriétaire des données qui est responsable de leur cycle de vie, y compris leur publication et la réponse aux commentaires ou aux questions ;
- Préparer un plan de collecte de données, en veillant à ce que les définitions et la structure des données soient conformes aux normes internationales ;
- S'assurer que les métadonnées sont définies consensuellement entre les parties prenantes et qu'elles sont utiles et normalisées ;
- S'assurer que les données font l'objet d'une licence appropriée pour garantir leur diffusion et leur réutilisation.

3.1.2. Préparation des données et de leur publication :

Il existe un certain nombre de formats communs pour la distribution des données. Certains sont considérés comme "ouverts" (tels que CSV, XML, texte et autres) et d'autres comme propriétaires (SAS, STATA, SPSS, etc.). XLS et XLSX, associés à Microsoft Excel, sont des formats relativement ouverts et un certain nombre de systèmes logiciels qui peuvent interpréter les données.

Les formats propriétaires sont légitimes, même dans le cadre d'une initiative de données ouvertes, car ce sont les systèmes logiciels utilisés par de nombreux utilisateurs professionnels de données. Cependant, comme ces formats ne sont souvent pas interopérables, le potentiel de réutilisation des données est limité, à moins que les formats

ouverts ne soient également pris en charge. La diffusion de données dans des formats propriétaires n'exclut pas la diffusion dans des formats ouverts et vice versa.

Les tableurs et les systèmes de données distribués ne disposent souvent pas d'une structure de données standardisée. Un chercheur qui souhaite combiner ces données avec d'autres données doit d'abord les normaliser, puis décider de termes normalisés pour définir les colonnes et les types de données dans ces colonnes.

La conversion de données tabulaires semi-structurées en un format exploitable par une machine donne le fichier CSV (« comma-separated-value » soit en français données séparées par des virgules). Ce sont des fichiers tabulaires avec une ligne d'en-tête qui définit chacune des données dans les colonnes et lignes ci-dessous. Sans tenir compte du respect d'autres normes, les fichiers CSV peuvent être organisés de telle sorte qu'ils soient "joints" sur une colonne commune. Par exemple, un ensemble de coordonnées géospatiales peut être utilisé pour relier plusieurs fichiers similaires couvrant différentes séries de données.

Au niveau international, il est généralement conseiller de respecter les directives suivantes, détaillées dans le reste de cette section, pour maximiser la lisibilité des documents et des données en ligne :

- Un demandeur peut ouvrir le fichier de données dans des progiciels gratuits et largement accessibles - cela signifie que des formats tels que CSV doivent être préférés ou proposés en plus de formats tels qu'Excel, et que les formats propriétaires qui ne peuvent être ouverts qu'avec des logiciels commerciaux ou spécialisés doivent être évités.
- Il est 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 - cela nécessite des données bien structurées. Dans la mesure du possible, la signification des données ne doit pas être contenue dans la présentation : table pure sans cellules fusionnées, pas d'en-têtes ni de notes de bas de page, pas de données cachées dans du texte, etc.
- Les éléments communs de l'ensemble de 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 (par exemple 1,000 ou 1000 mais pas un mélange des deux).
- La signification des champs et des valeurs est clairement documentée - soit par une désignation claire des champs, soit par des descriptions accompagnant les données.

En outre, dans la mesure du possible, l'exploitabilité par une machine est améliorée si le jeu de données utilise des normes communes lorsqu'elles existent- y compris des identificateurs et des noms de champs standards. Il peut s'agir de normes telles que le vocabulaire des dépenses publiques développé pour le gouvernement, ou de normes tierces telles que KML pour l'indication des "points d'intérêt".

Enfin, Il est également important de noter qu'il peut être possible de fournir des données dans divers formats exploitables par une machine et, dans la mesure du possible, le fournisseur de données doit échanger avec les réutilisateurs pour identifier le meilleur format. Par exemple, les points d'intérêt pourraient être fournis sous forme de feuille de calcul CSV ou KML. L'idéal serait que les deux soient fournis. Cependant, selon le contexte et le réutilisateur, l'un peut être plus approprié que l'autre.

3.1.2.1. Données longues et données larges :

Toute série de données est constituée de valeurs numériques (généralement) décrites par des termes de métadonnées normalisés (temps, zone, description spécifique, etc.). Il y a deux façons principales de présenter ces données lisibles par machine, qui peuvent être résumées en données larges ou données longues. Vous devez faire un choix délibéré quant au format que vous choisirez, et chacun a ses propres forces et faiblesses :

- **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 données entrainera l'ajout d'une colonne.

Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963	1964	1965
Aruba	ABW	Urban populatio	SP.URB.TOTL	27526	28141	28532	28761	28924	29082
Afghanistan	AFG	Urban populatio	SP.URB.TOTL	755836	796272	839385	885228	934135	986074
Angola	AGO	Urban populatio	SP.URB.TOTL	569222	597288	628381	660180	691532	721552
Albania	ALB	Urban populatio	SP.URB.TOTL	493982	513592	530766	547928	565248	582374
Andorra	AND	Urban populatio	SP.URB.TOTL	7839	8766	9754	10811	11915	13067
Arab World	ARB	Urban populatio	SP.URB.TOTL	28797177	30292822	31856717	33513046	35275337	37163923
United Arab Emirates	ARE	Urban populatio	SP.URB.TOTL	67927	74975	84367	95215	106178	116473
Argentina	ARG	Urban populatio	SP.URB.TOTL	15076842	15449950	15815502	16183085	16552517	16923103
Armenia	ARM	Urban populatio	SP.URB.TOTL	960956	1012430	1065431	1119586	1174560	1229980
American Samoa	ASM	Urban populatio	SP.URB.TOTL	13324	13729	14254	14871	15522	16176
Antigua and Barbuda	ATG	Urban populatio	SP.URB.TOTL	21466	21472	21458	21443	21449	21489
Australia	AUS	Urban populatio	SP.URB.TOTL	8378309	8589875	8832932	9034955	9245383	9459784
Austria	AUT	Urban populatio	SP.URB.TOTL	4561167	4592914	4624644	4658034	4692726	4726878
Azerbaijan	AZE	Urban populatio	SP.URB.TOTL	2051433	2110438	2171811	2233682	2293589	2349762
Burundi	BDI	Urban populatio	SP.URB.TOTL	58113	60329	62624	65010	67577	70985
Belgium	BEL	Urban populatio	SP.URB.TOTL	8463316	8500111	8545539	8624158	8720520	8814176
Benin	BEN	Urban populatio	SP.URB.TOTL	225533	243036	262053	282715	305170	329545
Burkina Faso	BFA	Urban populatio	SP.URB.TOTL	226977	234744	242709	251039	259788	268990
Bangladesh	BGD	Urban populatio	SP.URB.TOTL	2465493	2605371	2790354	2989609	3205156	3439969

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 ensuite la mise à jour de chaque ligne avec les informations appropriées. Cela peut être une opération coûteuse dans une grande base de données, et signifie également que l'écriture d'un programme informatique pour interroger vos données est plus difficile.

- **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-7967	45000000
Departme	Departme	########	CASH FUN	FINANCE &	NOTTINGH	HAFS-7968	85000000
Departme	Departme	########	CASH FUN	FINANCE &	OLDHAM I	HAFS-7969	28000000
Departme	Departme	########	CASH FUN	FINANCE &	OXFORDSI	HAFS-7970	75000000
Departme	Departme	########	CASH FUN	FINANCE &	PETERBOR	HAFS-7971	22000000
Departme	Departme	########	CASH FUN	FINANCE &	PLYMOUTI	HAFS-7972	35000000
Departme	Departme	########	CASH FUN	FINANCE &	PORTSMO	HAFS-7973	27000000
Departme	Departme	########	CASH FUN	FINANCE &	REDBRIDG	HAFS-7974	31000000
Departme	Departme	########	CASH FUN	FINANCE &	REDCAR A	HAFS-7975	18000000
Departme	Departme	########	CASH FUN	FINANCE &	RICHMON	HAFS-7976	21000000
Departme	Departme	########	CASH FUN	FINANCE &	ROTHERH/	HAFS-7977	29500000
Departme	Departme	########	CASH FUN	FINANCE &	SALFORD I	HAFS-7978	31000000
Departme	Departme	########	CASH FUN	FINANCE &	SANDWEL	HAFS-7979	43000000
Departme	Departme	########	CASH FUN	FINANCE &	SEFTON P	HAFS-7980	38000000
Departme	Departme	########	CASH FUN	FINANCE &	SHEFFIELD	HAFS-7981	70500000
Departme	Departme	########	CASH FUN	FINANCE &	SHROPSHI	HAFS-7982	32400000
Departme	Departme	########	CASH FUN	FINANCE &	SOLIHULL	HAFS-7983	26200000
Departme	Departme	########	CASH FUN	FINANCE &	SOMERSET	HAFS-7984	65000000
Departme	Departme	########	CASH FUN	FINANCE &	SOUTH BIF	HAFS-7985	49500000

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 préférence dans la publication de données ouvertes est pour le format long, et ce sera la méthode habituellement recommandée pour la publication. A noter que la conversion des données du format long au format large est relativement facile à faire, y compris de manière semi-automatique, alors que la conversion inverse nécessite une intervention humaine.

3.1.2.2. Métadonnées et schéma de destination :

La création d'un schéma est la première partie du processus de préparation des données. Votre schéma définit la structure cible de métadonnées pour votre processus de préparation. Ce n'est pas le format dans lequel vos données d'entrée arrivent, mais c'est ce à quoi vous voulez qu'elles ressemblent quand vous avez terminé.

Votre schéma définit les exigences, les contraintes et les valeurs par défaut raisonnables disponibles pour la saisie de données dans les champs définis par le schéma. Les colonnes

d'un fichier CSV ou Excel d'entrée seront restructurées en de nouvelles colonnes définies par les champs de votre schéma. Ces champs cibles sont susceptibles d'être ceux d'une base de données. Tant que vos données d'entrée ne sont pas conformes à cette structure, vos données ne doivent pas être publiées en tant que données ouvertes.

Nous utiliserons les définitions et le schéma de Frictionless Data[12] pour définir les champs de métadonnées structurelles.

3.1.2.2.1. Type et format :

Le type définit le type de données du champ, tandis que le format affine les propriétés spécifiques du type. Ce sont les types de base que vous êtes susceptible d'utiliser, avec des indentations pour les formats :

- string : toute chaîne basée sur du texte.
    - default : toute chaîne de caractères
    - email : une adresse électronique
    - uri : toute adresse web / URI
- number : toute valeur basée sur un nombre, y compris les nombres entiers et les nombres flottants.
- integer : toute valeur basée sur un nombre entier.
- boolean: une valeur booléenne [true, false]. Peut définir des contraintes de catégorie pour fixer le terme utilisé.
- object : toute donnée JSON valide.
- array : toute donnée valide basée sur un tableau.
- date : toute date sans heure. Doit être au format ISO8601, AAAA-MM-JJ.
- datetime : toute date avec une heure. Doit être au format ISO8601, avec l'heure UTC spécifiée (en option) comme AAAA-MM-JJ hh:mm:ss Zz.
- year : toute année, formatée sur quatre chiffres (AAAA).

3.1.2.2.2. Conception d'un schéma :

Si vous consultez par exemple le portail open Data français concernant les données hospitalières relatives à l'épidémie de COVID-19 en France[13], vous pourrez prévisualiser le fichier de schéma des métadonnées pour l'hospitalisation suivant :

[12] https://specs.frictionlessdata.io/table-schema/#field-descriptors
[13] https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/

Colonne	Type	Description_FR	Description_EN	Exemple
dep	integer	Département	Department	1
sexe	integer	Sexe	Sex	0
jour	string($date)	Date de notification	Date of notice	18/03/2020
hosp	integer	Nombre de personnes actuellement hospitalisées	Number of people currently hospitalized	2
rea	integer	Nombre de personnes actuellement en réanimation ou soins intensifs	Number of people currently in resuscitation or critical care	0
rad	integer	Nombre cumulé de personnes retournées à domicile	Total amount of patient that returned home	1
dc	integer	Nombre cumulé de personnes décédées à l'hôpital	Total amout of deaths at the hospital	0

Le premier terme de chaque ligne définit le nom de la colonne de destination (les termes des métadonnées structurelles), ainsi que son type de données, les descriptions en anglais et en français et un exemple de valeur attendue pour ce champ. Avant de commencer à restructurer des fichiers de données désordonnés, créez une telle définition de métadonnées structurelles pour vos données ; et la première chose à faire lorsque vous commencez est de regarder vos données et de comprendre comment elles sont actuellement structurées.

L'objectif que vous poursuivez en passant de données désordonnées à des données structurées est simple : Préserver tout ce qui est connu de vos données sources lorsque vous les restructurez.

En résumé, les exigences en matière de conception de schémas sont les suivantes :

- **Etablir une convention standard pour nommer les champs de colonnes**, et s'y tenir ; la convention se réfère à la structure physique du mot, comme startDate ou place_name. Ne les mélangez pas car c'est frustrant pour les utilisateurs.
- **Chaque valeur de chaque colonne de chaque table ne doit représenter qu'une seule chose** : si vous regardez par exemple la colonne région ci-dessous, vous verrez qu'elle doit être divisée en deux colonnes séparées, région et district, pour la Côte d'Ivoire :

Description	Région	Année	Valeur
Population	Gbôklé, Bas-Sassandra	2014	400,798
Population	Nawa, Bas-Sassandra	2014	1,053,084
Population	San-Pédro, Bas-Sassandra	2014	826,666
Population	Indénié-Djuablin, Comoé	2014	560,432
Population	Sud-Comoé, Comoé	2014	642,620

- **Définissez le type, le format et la définition pour chaque nom de colonne** : la définition du schéma ci-dessous est basée sur des données au format long. Si chaque année avait sa propre colonne, vous auriez alors - au lieu de « année » - une ligne pour 2014, alors la prochaine version pourrait inclure 2016 et ainsi de suite. Vous pouvez

voir que la mise à jour simultanée des données et du schéma représente plus de travail pour vous et vos utilisateurs que la simple mise à jour des données. Le schéma correspondant aux données ci-dessus (en supposant la colonne Région séparée en deux – Région, district - :

Colonne	**Type**	**Description**	**Exemple**
description	string	Définition de la série de données démographiques	Population
région	string	Nom d'une des 31 régions de Côte d'Ivoire	Sud-Comoé
district	string	Nom d'un des 14 districts de la Côte d'Ivoire	Comoé
année	date	L'année de la série de données	2014
valeur	integer	La valeur pour la série de donnée	560432

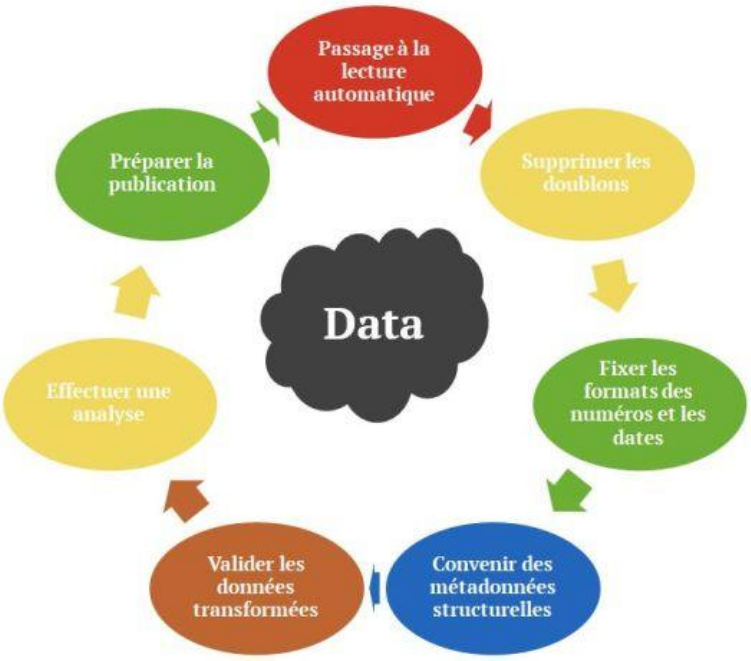
Il existe des exigences supplémentaires pour valider vos données - par exemple, noter que les virgules entre les chiffres doivent être supprimées - et nous allons les passer en revue ci-dessous.

3.1.2.3. Utiliser Excel pour nettoyer les données désordonnées :

Excel est l'un des logiciels les plus puissants utilisés par les analystes ordinaires (non-développeurs informatiques), et probablement l'outil de gestion et d'analyse de données le plus répandu dans le monde. Excel peut vous permettre de faire tout ce qui suit :

- Supprimer les enregistrements en double.
- Séparer les valeurs multiples contenues dans le même champ.
- Analyser la distribution des valeurs dans un ensemble de données.
- Regrouper les différentes représentations d'une même réalité.

La préparation et le nettoyage de données est un processus simple qui se fait étape par étape comme présenté dans le diagramme ci-dessous.



Les étapes sont les suivantes :

- **Créer le fichier de données avec de nouvelles feuilles de travail** pour chacune d'entre elles : Données originales | Données intermédiaires | Données finales.
- **Nettoyer les données** :
  - Créez une colonne ID reliant vos données originales à vos données intermédiaires. Il s'agit d'identifier chaque ligne de vos données parce que la transformation des données modifiera le nombre de lignes de votre fichier original et donc le numéro de ligne n'est pas un identifiant approprié.
  - Gérer les enregistrements en double en créant une clé de recherche qui doit identifier chaque ligne de manière unique ; Excel peut trier sur cette ligne et supprimer les doublons.
  - Supprimer les caractères indésirables (rechercher/remplacer).
  - Repérer les valeurs hors limites (c'est-à-dire les valeurs qui ne sont pas ce qu'elles devraient être).
  - Supprimer toute donnée non publiable (comme les informations personnelles ou toute donnée non approuvée pour la publication).
- **Traiter les données** :
  - Analyser les données (par exemple, en utilisant la méthode du texte en colonnes et en les séparant par des espaces ou des tabulations).
  - Recoder les données (pour convertir les termes en métadonnées normalisées).
  - Calculer de nouvelles valeurs (comme des totaux ou des moyennes)
  - Reformater les données dans un format standardisé.
- **Créer une copie des données prêtes pour l'analyse** en copiant les données intermédiaires (feuille InterimData) non formatées dans la feuille FinalData et en supprimant les colonnes inutiles ;

- **Documenter les données** :
    - Documentation au niveau du fichier, telle que le projet, la date d'achèvement, vérifiée par, etc. ;
    - Produire un fichier texte qui accompagne le fichier de données et qui comporte les métadonnées descriptives : description, date de production, propriétaire des données, etc.

Pour mettre en pratique les éléments présentés dans l'ensemble de la section 3.1.2, un exemple complet est présenté en Annexe A.

3.2. Validation des données :

L'objectif de cette section est d'utiliser un tableur pour s'assurer qu'un jeu de données correspond parfaitement au schéma qui le décrit (voir la section sur la conception du schéma à la section 3.1.2) et qu'il ne contient pas de valeur erronée syntaxiquement. Il est important de noter que la validation des données présentée ici consiste uniquement à vérifier que les données sont conformes au regard du schéma de données. Les données peuvent cependant être parfaitement conformes au schéma et être incohérentes ou erronées. Par exemple, si un jeu de données contient l'ensemble des produits et prix d'un producteur de lait, on pourrait trouver une ligne qui donne le prix d'un litre de lait à 12DH et une ligne qui donne le prix d'un carton de 5L à 92DH. Il y a clairement un problème sur l'un des deux prix, mais seule une analyse approfondie du jeu de données pourra la détecter. A l'inverse si une colonne de type date contient par exemple un chiffre seul ou une chaine de caractères, il sera possible dans le cadre d'une validation d'un schéma de détecter cette erreur.

3.2.1. Définition du schéma de validation dans Excel :

La suite de cette section est adaptée d'un tutoriel Microsoft Office en anglais[14]. Cette approche fonctionnera aussi bien dans OpenOffice que dans Google Sheets, bien que les étapes spécifiques soient différentes. Microsoft propose un fichier d'exemple que vous pouvez télécharger[15].

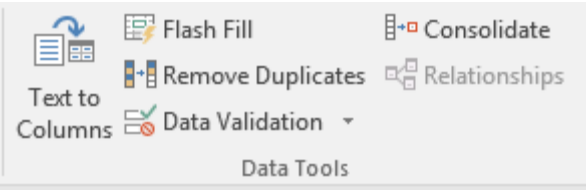
NB: dans ce paragraphe nous vous indiquons les instructions sous Excel à la fois en français dans le texte et en anglais en image. Vous devriez pouvoir trouver les options si vous avez une version française ou anglaise.

Le processus est simple :

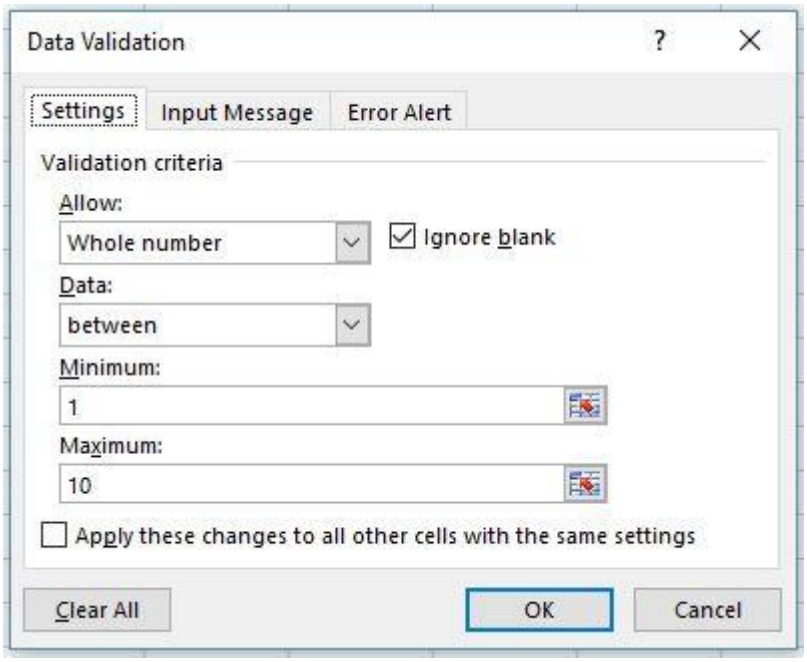
- Sélectionnez les cellules d'une colonne spécifique dont vous souhaitez vérifier le type
- Sélectionnez **Données > Outils de données > Validation des données**

[14] https://support.office.com/en-gb/article/apply-data-validation-to-cells-29fecbcc-d1b9-42c1-9d76-eff3ce5f7249

[15] https://download.microsoft.com/download/9/6/8/968A9140-2E13-4FDC-B62C-C1D98D2B0FE6/Data%20Validation%20Examples.xlsx



- Dans l'onglet **Option**, sous **Autoriser**, sélectionnez l'un des éléments suivants :



  - **Nombre Entier** : restreindre la cellule pour n'accepter que des nombres entiers.
  - **Décimal** : restreint la cellule pour n'accepter que les nombres décimaux ou entier.
  - **Liste** : choisir des données dans une liste déroulante, et limitée par les valeurs contraintes par enum.
  - **Date** : restreindre la cellule pour n'accepter que les dates.
  - **Heure** : restreindre la cellule pour n'accepter que les datetime.
  - **Longueur de Texte** : restreindre la longueur du texte, équivalent aux contraintes maxLength.
  - **Personnalisé** : pour une formule personnalisée.

- Sous **Données**, vous pouvez sélectionner une condition :
  - Comprise entre
  - Non comprise entre
  - Égale à
  - Différente de
  - Supérieure à
  - Inférieure à
  - Supérieure ou égale à

    - Inférieure ou égale à
- Définissez les autres valeurs requises, en fonction de ce que vous avez choisi pour **Autoriser** et **Données**. Par exemple, si vous faites un choix entre les valeurs **Minimum** : et **Maximum** : pour la ou les cellules.
- Cochez la case « **Ignorer si vide** » si vous souhaitez ignorer les espaces vides (c'est-à-dire pour les données manquantes). Notez, cependant, qu'Excel n'a pas de place pour les caractères spéciaux que vous pourriez utiliser comme marqueur pour les données manquantes (par exemple « .. » ou « NA » ou tout autre caractère), de sorte que ceux-ci seront mentionnés comme des erreurs.
- Sélectionnez **OK**.

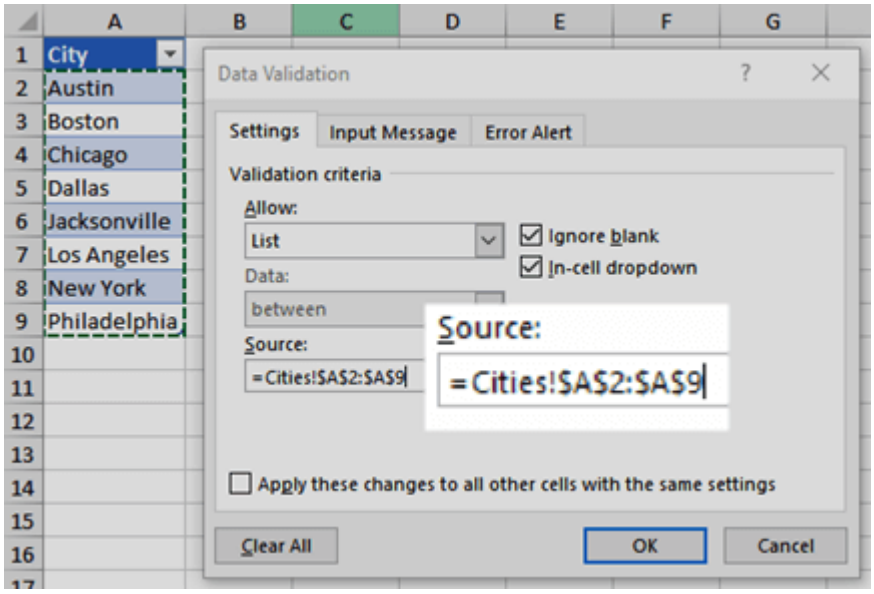
Maintenant - seulement pour les nouvelles données - si un utilisateur essaie d'entrer une valeur qui n'est pas valide, une fenêtre pop-up apparaît avec le message "Cette valeur ne correspond pas aux restrictions de validation des données pour cette cellule".

A noter que les listes sont un type un peu particulier. Avant de pouvoir valider un type « liste », vous devez spécifier des termes valables pour cette liste. Dans Excel, cela nécessite une série d'étapes supplémentaires.

1. Créez une nouvelle feuille de calcul dans Excel, et mettez-y la liste des termes que vous voulez définir comme valeurs valides pour la liste. Par exemple :

	A
1	City
2	Austin
3	Boston
4	Chicago
5	Dallas
6	Jacksonville
7	Los Angeles
8	New York
9	Philadelphia

2. Sélectionnez les données de votre liste et formatez-les sous forme de table (**Accueil > Styles > Mettre sous forme de tableau**).
3. Vous pouvez nommer votre table à partir de l'onglet **outils** de la table - celle-ci pourrait s'appeler "CityTable". Cela vous aidera à garder une trace de plusieurs tables.
4. Dans le processus de validation, sélectionnez le type « Liste », puis ajoutez une plage nommée ou un nom de table pour votre liste.
5. Indiquez une source pour vos termes via l'onglet **Données > Validation de Données > Autoriser > Liste > Source**. Ensuite, spécifiez votre liste de termes suivant une des méthodes suivantes :

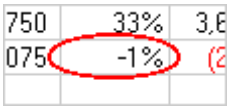


(Cities dans l'exemple de l'image est le nom de feuille qui contient la liste)

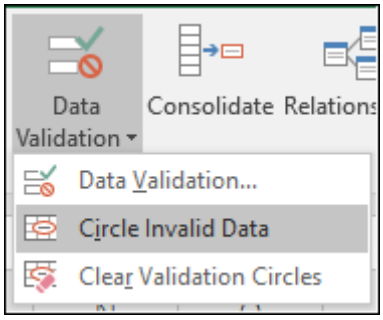
- Vous pouvez sélectionner directement la feuille de liste et la plage (par exemple =Sheet1 ! A4:A10)
- Convertissez votre liste en une table avec **Ctrl+L**, puis à partir de l'onglet **Création de Tableau** donnez un nom à votre table, ce qui vous permet de faire référence au nom de la table et à la colonne (par exemple =CityTable[City])
- Dans l'onglet **Formules**, sélectionnez **Gestionnaire de Nom**, créez un **Nouveau** élément avec un nom approprié (par exemple CityList), et référencez les cellules (par exemple =Sheet1!A4:A10), ce qui vous permet ensuite de référencer votre liste n'importe où (par exemple =CityList)

3.2.2. Application d'un schéma de données et détection des problèmes :

Une fois que l'ensemble des contraintes sur les valeurs est fixé, Excel n'applique pas les vérifications sur les cellules existantes, mais uniquement sur les nouvelles données. Il y a cependant un moyen rapide de mettre en évidence les cellules non valides existantes en encerclant les valeurs comme dans l'exemple ci-dessous :



1. Pour appliquer les cercles, sélectionnez les cellules que vous voulez évaluer et allez à **Données > Outils de données > Validation de Données > Entourer les Données non valides**.



2. Si vous corrigez une entrée non valable, le cercle disparaît automatiquement.
3. Pour supprimer la validation des données pour une cellule, sélectionnez-la, puis allez **à Données > Outils de données > Validation de Données > Options > Effacer Tout**.

3.3. Anonymisation des données :

L'objectif de cette section est de présenter les techniques existantes d'anonymisation de données qui doivent être appliquées avant publication lorsque le jeu de données contient des informations qui rentre dans le cadre de 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. Il existe un large éventail de techniques pour anonymiser des données. Elles se répartissent en deux grandes catégories :

- **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.
- **Les techniques d'agrégation** : Les données sont agrégées délibérément pour garantir l'anonymat des données aberrantes.

Il est important de noter que, d'une part, les outils comme Excel ne permettent pas d'appliquer les techniques d'agrégation, et que, d'autre part, le choix d'une technique spécifique d'anonymisation et son application est un processus extrêmement complexe qui nécessite des compétences spécifiques en statistiques. Il est hors du cadre de ce manuel et de la fonction de GdD de maitriser ces techniques et les appliquer. L'objectif de cette section est de présenter les techniques les plus simples qui peuvent être suffisantes pour traiter des jeux de données simples.

3.3.1. Méthodes de rédaction :

Les techniques de rédaction sont des techniques de destruction de données pour les rendre anonymes. Il n'est pas possible après l'application de ces techniques de pouvoir retrouver les données initiales à partir des données transformées. Il existe plusieurs techniques présentées ci-dessous :

- **Suppression d'attributs** : Cette méthode consiste à la suppression d'un champ (colonne) entier. C'est l'une des premières mesures que nous pouvons prendre, et la plus facile. Il convient donc de premier abord de :
    - Supprimer les données inutiles par rapport au jeu de données.

    - Supprimer les éléments d'identification directe qui ne sont pas utiles à l'exploitation des données.

  Il s'agit d'une mesure destructrice puisque la suppression supprime les données originales.

- **Pseudonymisation** : La pseudonymisation est le remplacement des données d'identification par des valeurs aléatoires. Elle peut être réversible, si vous créez une clé entre les données et les valeurs générées, ou irréversible si vous ne conservez pas les clés. Les pseudonymes persistants permettent de relier un même individu (ou un même objet comme un bâtiment, une route, etc.) à travers différents ensembles de données.
- **Généralisation** : La généralisation est une réduction délibérée de la précision des données, telle que la conversion de l'âge d'une personne en une fourchette, ou d'un lieu précis en un lieu moins précis (par exemple passer d'une rue à un quartier, ou une ville).
- **Brassage** : Le brassage consiste à réorganiser les données de manière à ce que les valeurs des attributs individuels soient toujours représentées dans l'ensemble de données, mais ne correspondent pas aux enregistrements originaux. Cette méthode ne convient pas à toutes les données. Par exemple, l'échange de maladies entre différents patients rendra certainement les données anonymes, mais confondra également toute analyse épidémiologique.

Les trois premières techniques sont facilement implémentables avec Excel, la dernière est plus difficile.

3.3.2. Méthodes d'agrégation :

L'agrégation est bien plus destructrice que la rédaction. 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.

Là où la rédaction est guidée presque exclusivement par les données, l'agrégation est guidée par les objectifs de recherche sur les données. Toute forme d'agrégation limitera ce qui peut être fait et il est essentiel de prendre conscience de ces limites. Les données de recensement sont généralement agrégées de cette manière, les micro-données individuelles (réponses de chaque ménage) étant uniquement mises à la disposition des chercheurs accrédités, tandis que les versions agrégées sont mises à la disposition du public.

L'objectif sera de créer des groupes de données et de procéder ensuite à des agrégations sur chaque groupe. L'éventail des agrégats possibles comprend notamment :

- **Le comptage** : comptage des membres individuels du groupe.
- **Les totaux** : sommes des valeurs et sommes des sous-groupes à l'intérieur des valeurs (par exemple, durée totale de la maladie et durée de chaque type de maladie).

- **Les moyennes** : Moyennes au sein du groupe qui peut également être enrichi des valeurs médianes, ou des écart-types.
- **Les distributions** : plusieurs types de distributions peuvent être fournis comme les quantiles.

À bien des égards, un cours complet de statistiques est nécessaire pour bien effectuer les agrégations. De plus, il est important de noter que les agrégations nécessitent une certaine familiarité et expérience avec les données à agréger. Il est très difficile de se contenter d'un ensemble de données aléatoires et de savoir comment les agréger d'une manière qui favorise l'analyse et en extrait le sens. Il est peu probable qu'un GdD soit responsable de l'agrégation de données dont il n'a pas l'expérience, et lorsqu'il a l'expérience, savoir comment l'agréger devient plus clair.

Pour ceux qui veulent aller plus loin sur le sujet de l'anonymisation, un document en ligne et en français[16] détaille toutes les techniques d'anonymisation, leurs points forts et leurs limitations avec des exemples concrets.

3.4. Génération d'un fichier CSV et publication sur le portail Open Data :

Une fois que la préparation des données est terminée, que la validation et les corrections d'erreurs sont effectuées, et le cas échéant, que la ou les techniques d'anonymisation sont appliquées, il n'y a plus qu'à sauvegarder le jeu de données sous un format ouvert et le publier sur le portail.

Le format de préférence pour les données ouvertes tabulaires numériques est le format CSV. Les fichiers de valeurs séparées par des virgules (.csv) sont des fichiers texte dans lesquels le caractère de virgule « , » sépare chaque champ de texte. Lorsqu'une virgule apparaît dans la valeur - qu'il s'agisse d'une chaine de caractère ou d'un nombre - la valeur est alors entourée de guillemets, par exemple 100, 200, "20,000" indique trois valeurs dans trois champs séparés.

Vous pouvez modifier le caractère de séparation utilisé dans les fichiers texte délimités et .csv, et il existe un large éventail de formats (par exemple « ; » ou «*»).

Dans Excel, vous pouvez exporter une feuille de calcul au format CSV en utilisant « **Save As/Sauvegarder sous »** comme suit :

1. Allez à **Fichier > Enregistrer une copie**.
2. Cliquez sur **Parcourir**.
3. Dans la boîte de dialogue **Enregistrer sous**, dans la case **Type**, choisissez le format de fichier texte pour la feuille de travail ; par exemple, cliquez **sur CSV UTF-8 (délimité par des virgules)** ou **CSV (séparateur : point-virgule)** (les options peuvent variées en fonction de la version d'Excel).

[16] http://www.benjamin-nguyen.fr/papers/ss.pdf

4. Naviguez jusqu'à l'emplacement où vous souhaitez enregistrer le nouveau fichier texte, puis cliquez sur Enregistrer.

A noter que vous ne pouvez exporter que la feuille de calcul actuelle (c'est-à-dire celle qui est en vue lorsque vous exécutez le processus de sauvegarde) vers le nouveau fichier CSV. Vous pouvez enregistrer d'autres feuilles de calcul en tant que fichiers texte séparés en répétant cette procédure pour chaque feuille de calcul.

Toutes les caractéristiques spécifiques de la feuille de calcul seront perdues. Le formatage (gras, couleurs, etc.), les formules et les critères de validation seront supprimés, ne laissant que les données dans un fichier texte.

A l'issue de ce processus, vous avez les éléments suivants

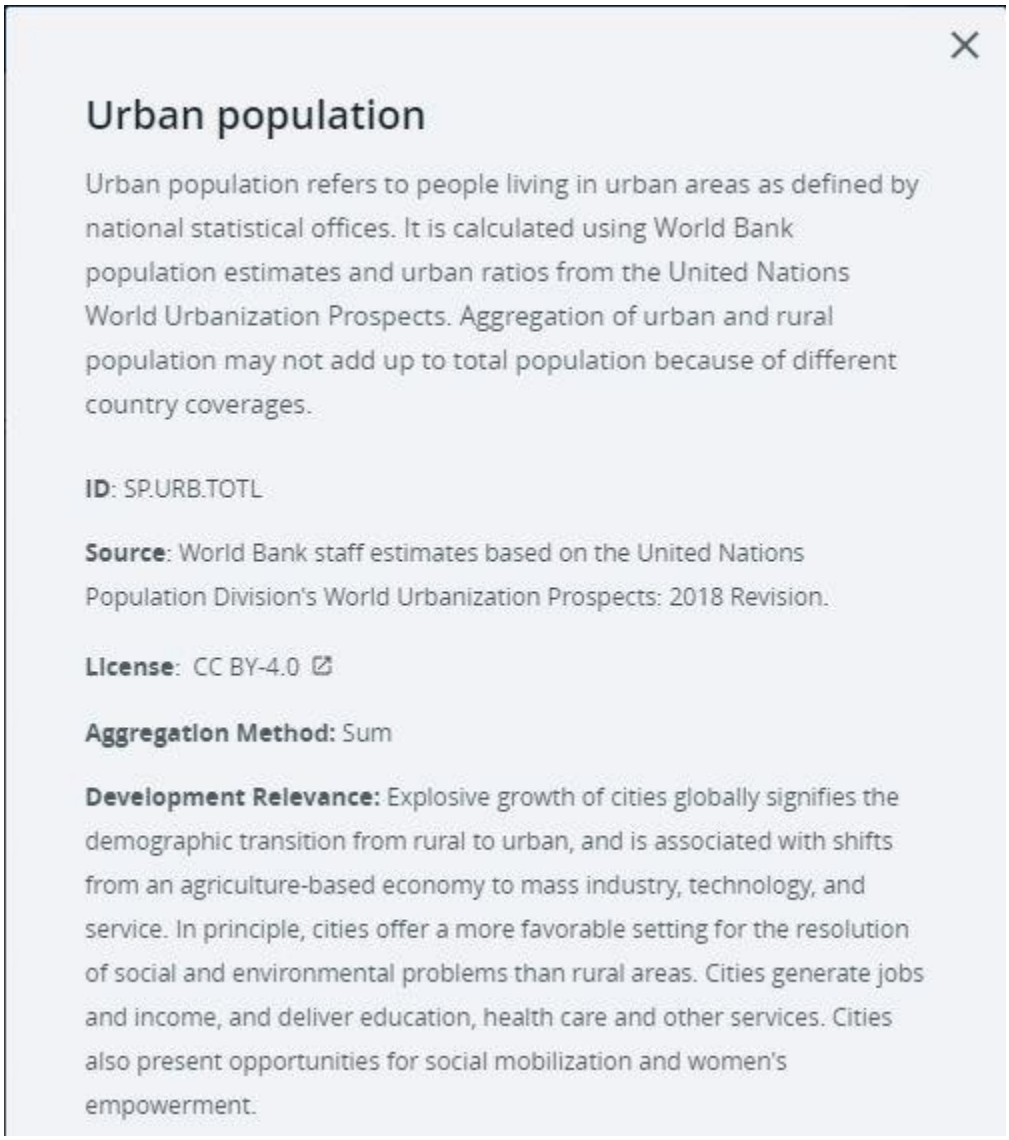
- Le fichier csv qui constitue le fichier à publier.
- Le fichier de schéma des données qui devra être renseigné dans les métadonnées lors de la publication sur le portail Open Data.
- La liste des informations de source que vous avez noté dans votre feuille « Source ».

Il ne reste plus qu'à définir le contenu des métadonnées structurelles (Cf. section 3.1.1). Une fois cette tâche terminée, vous êtes prêt à publier sur le portail national Open Data http://www.data.gov.ma ou sur le portail de la structure publique concernée. Rapprochez-vous du Responsable Open Data de votre structure pour obtenir vos informations de connexion sur le portail et la procédure à suivre pour publier l'ensemble des informations (données et métadonnées).

Annexe A - Exemple complet de préparation de données :

L'exemple suivant a été écrit par Lisa Charlotte Rost @ Datawrapper.de[17] et est utilisé avec son autorisation. Il a été légèrement modifié pour être conforme à l'approche adoptée dans ce manuel. Il s'agit d'un exemple relativement simple, mais qui vous guidera à travers les principaux défis que vous rencontrerez.

Le tutoriel travaillé utilise des données de la Banque Mondiale (https://data.worldbank.org/indicator/SP.URB.TOTL?view=chart). Un lien vers les métadonnées ouvre la fenêtre suivante :



Urban population

Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects. Aggregation of urban and rural population may not add up to total population because of different country coverages.

ID: SP.URB.TOTL

Source: World Bank staff estimates based on the United Nations Population Division's World Urbanization Prospects: 2018 Revision.

License: CC BY-4.0

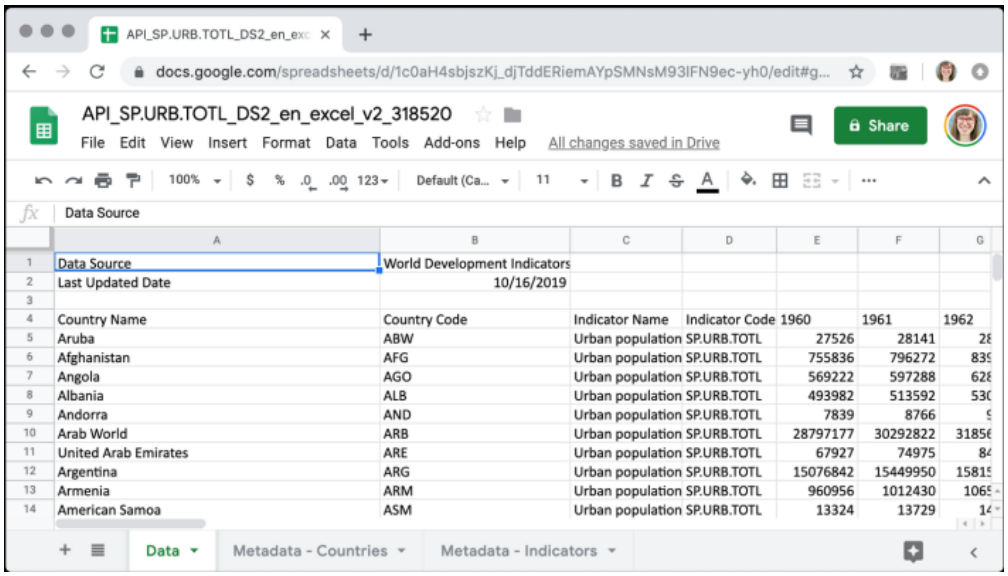
Aggregation Method: Sum

Development Relevance: Explosive growth of cities globally signifies the demographic transition from rural to urban, and is associated with shifts from an agriculture-based economy to mass industry, technology, and service. In principle, cities offer a more favorable setting for the resolution of social and environmental problems than rural areas. Cities generate jobs and income, and deliver education, health care and other services. Cities also present opportunities for social mobilization and women's empowerment.

[17] https://blog.datawrapper.de/prepare-and-clean-up-data-for-data-visualization/

Etape 1 - Examiner les données :

Téléchargez l'ensemble de données à http://api.worldbank.org/v2/en/indicator/SP.URB.TOTL?downloadformat=excel et consultez la feuille principale.



	A	B	C	D	E	F	G
1	Data Source	World Development Indicators					
2	Last Updated Date	10/16/2019					
3							
4	Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962
5	Aruba	ABW	Urban population	SP.URB.TOTL	27526	28141	28
6	Afghanistan	AFG	Urban population	SP.URB.TOTL	755836	796272	839
7	Angola	AGO	Urban population	SP.URB.TOTL	569222	597288	628
8	Albania	ALB	Urban population	SP.URB.TOTL	493982	513592	530
9	Andorra	AND	Urban population	SP.URB.TOTL	7839	8766	9
10	Arab World	ARB	Urban population	SP.URB.TOTL	28797177	30292822	31856
11	United Arab Emirates	ARE	Urban population	SP.URB.TOTL	67927	74975	84
12	Argentina	ARG	Urban population	SP.URB.TOTL	15076842	15449950	15815
13	Armenia	ARM	Urban population	SP.URB.TOTL	960956	1012430	1065
14	American Samoa	ASM	Urban population	SP.URB.TOTL	13324	13729	14

Lorsque vous téléchargez un fichier Excel, il comporte souvent plusieurs feuilles. Notre ensemble de données en comporte trois, comme on peut le voir en bas : "Données", "Métadonnées - Pays" et "Métadonnées - Indicateurs". Regardez toutes vos feuilles et assurez-vous que vous comprenez ce que vous y voyez. Les en-têtes, le nom du fichier et/ou les données elles-mêmes indiquent-ils que vous avez téléchargé le bon fichier ? Y a-t-il des notes de bas de page ? Que vous disent-elles ? Avez-vous peut-être affaire à de nombreuses estimations ? (Cela signifie-t-il peut-être que vous devez chercher d'autres données ?) Si vous ne trouvez pas de notes dans les données, assurez-vous de les chercher sur le site web de votre source.

Notre exemple de données semble correct. Il n'y a pas d'estimations dont nous devrions nous préoccuper. Et nous obtenons une explication claire de la "population urbaine" dans les "Métadonnées - Indicateurs", qui commence par "La population urbaine se réfère aux personnes vivant dans des zones urbaines telles que définies par les bureaux nationaux de statistiques...".

Etape 2 - Renommer votre fichier :

Maintenant que nous savons à quoi nous avons affaire, assurons-nous de pouvoir le faire encore dans six mois. Le nom original « API_SP.URB.TOTL_DS2_fr_excel_v2_318520 » est parfaitement incompréhensible à part peut-être pour les quelques employés de la Banque Mondiale qui gèrent ce fichier. Il convient donc de renommer ce fichier en quelque chose de mémorisable et de précis : Banque mondiale_population_urbaine_par-pays, par exemple.

Etape 3 - Dupliquer la ou les fiches de données et ne plus jamais y toucher :

C'est l'une des parties les plus importantes de l'ensemble du processus : Avant de modifier quoi que ce soit dans les données, dupliquez votre feuille de données dans le même fichier. Pensez à renommer vos deux feuilles, par exemple en "Données brutes" et "Données" ou en "Données originales" et "Données modifiées". Si vous avez un énorme fichier Excel avec beaucoup de feuilles, vous pouvez également créer un fichier de sauvegarde du document original et le stocker dans un dossier "source".

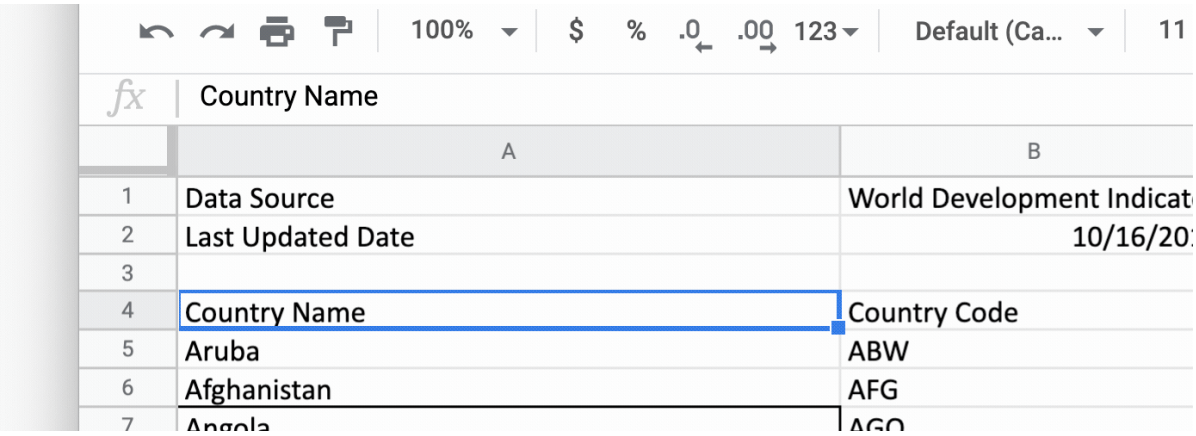
Cette étape est importante parce que vous allez modifier considérablement les données et il est toujours utile de pouvoir se référer aux données de départ. Il est très fortement recommandé de ne jamais travailler sur les données originales.

Etape 4 - Enregistrez votre source dans une feuille supplémentaire :

Il est important de garder la trace de la source des données que vous modifiez. Créez une nouvelle feuille, nommez-la "Source" et ajoutez des liens vers toutes les sources de données que vous utilisez dans votre document. Ajoutez la date de téléchargement du fichier.

Etape 5 - Supprimer tout ce qui se trouve au-dessus de l'en-tête :

Les fichiers Excel contiennent souvent des informations dans des lignes supplémentaires au-dessus des données réelles. Dans notre cas, le GdD de la Banque mondiale veut nous faire savoir que la source des données est "Indicateurs du développement dans le monde", et que la dernière mise à jour des données remonte à Octobre 2019. Ce sont des informations importantes qu'il convient de conserver mais ces lignes supplémentaires nous empêchent de trier ou de filtrer les données. Il convient donc supprimer toutes les lignes vides et toutes les informations au-dessus de l'en-tête. Supprimez-les (vous pouvez toujours les retrouver dans votre feuille de "données originales") ou copiez-collez les dans votre feuille "Source" qui contient des informations sur les données.



Etape 6 - Séparer les cellules fusionnées et enlever les entêtes sur deux ou plusieurs lignes :

Parfois, vous rencontrerez des en-têtes qui sont sur deux lignes, et non une seule. Surtout lorsque le tableau est créé pour communiquer, et non pour analyser, les en-têtes à deux lignes peuvent aider à donner un sens à l'information. Mais ils vous gêneront lorsque vous voudrez éventuellement supprimer des lignes ou des colonnes. Les données exploitables par une machine doivent comporter une ligne d'en-tête, et une seule ligne d'en-tête.

Par exemple, il est important de transformer un tableau qui contiendrait les cellules suivantes :

Afghanistan		Angola	
Population	Urban population	Population	Urban population

En :

Afghanistan Population	Afgahnistan Urban population	Angola Population	Angola Urban population

De la même manière, il convient d'enlever les lignes fusionnées comme dans l'exemple suivant :

Afghanistan	1960	755836
	1961	796272
	1962	839385
	1963	885228
	1964	934135
	1965	986074

Il faut dédoubler la case 'Afghanistan' en 6 lignes et mettre le nom du pays dans chaque.

Etape 7 - Mettre des mesures dans l'en-tête et supprimer les notes de bas de page :

Pour que les outils de données tels que Excel reconnaissent les chiffres, assurez-vous que vous disposez de chiffres « propres » c'est-à-dire ne contenant pas de caractère qui empêcherait Excel de reconnaitre un chiffre à cause par exemple de la présence de l'unité. Enlevez de vos cellules de données de tous les €, $, kg, %, km/h, etc. Mettez-les plutôt dans les en-têtes.

Cela nécessitera souvent une série d'étapes de recherche - remplacement, mais assurez-vous toujours d'avoir créé une colonne supplémentaire capturant ces unités afin de ne pas les perdre. Cette opération est similaire à la division des valeurs qui contiennent plusieurs termes.

Par exemple, il conviendra de transformer le tableau de gauche ci-dessous pour arriver au tableau de droite :

	fake GDP	fake share
Afghanistan	293 Euro	46.48%
Angola	383 Euro	29.40%
Albania	919 Euro	19.01%
Andorra	294 Euro	31.07%

	fake GDP in Euro	fake share in %
Afghanistan	293	46.48
Angola	383	29.40
Albania	919	19.01
Andorra	294	31.07

Une chose à noter dans l'exemple ci-dessus est que le % serait normalement une proportion de 1 (par exemple, 48% est 0,48), et donc les données dans cette colonne sont susceptibles d'être ambiguës. Il convient de corriger cette situation pour garantir la validité de vos données.

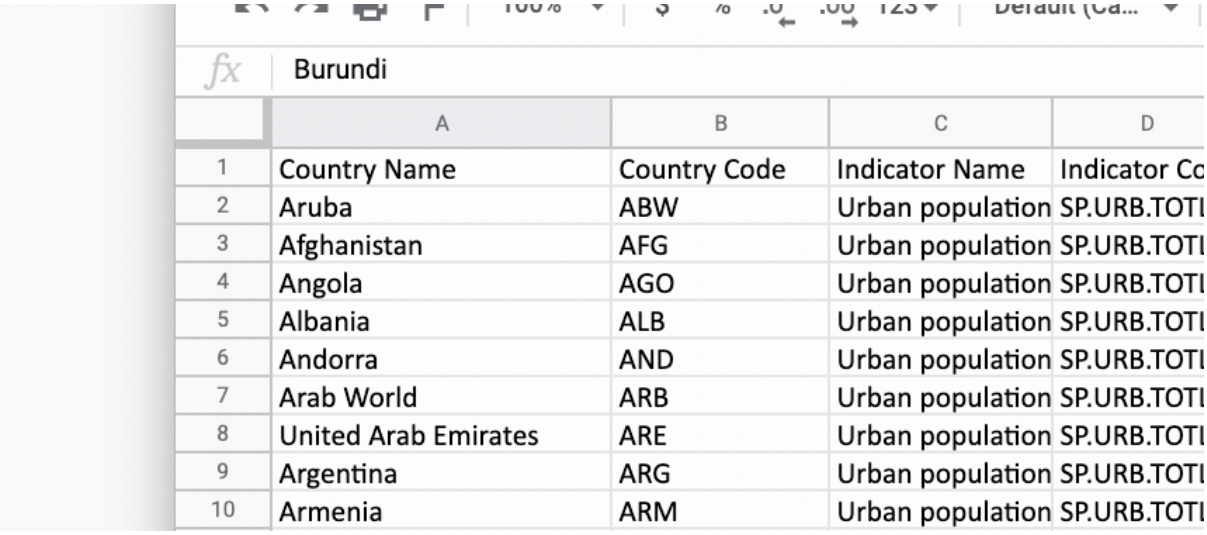
Vous devrez également extraire les notes de bas de page. Des valeurs comme 28.394† ou 1.39[^1] ne seront pas reconnues comme des chiffres. Avant de les supprimer, assurez-vous que vous comprenez le schéma des données : Les points de données de 2019 sont-ils tous des estimations ? (Devriez-vous peut-être exclure cette année-là ?) Ou bien les données d'un certain pays sont-elles mesurées différemment ? Dans tous ces cas, veillez à le faire savoir au lecteur de votre tableau final. Les notes de bas de page dans les données que vous utilisez doivent toujours se traduire par des définitions spécifiques dans le fichier de métadonnées. Vos utilisateurs en auront besoin.

Etape 8 - Vérifier si le contenu de l'en-tête a un sens :

Après avoir effectué ces tâches techniques, il est important de vérifier que les noms des en-têtes de colonnes que vous n'avez pas encore touchés ont vraiment un sens. Peut-être ne sont-ils que des codes comme SP.URB.TOTL ? Si c'est le cas, retournez à votre source et découvrez la signification de ces codes. Ou peut-être sont-ils trop longs ? Par exemple, « Nom du pays » peut facilement être réduit à « Pays ». Renommez les en-têtes pour qu'il soit facile pour les réutilisateurs de leur donner un sens : courts, mais précis et uniques. (Si vous voyez des colonnes que vous avez l'intention de supprimer, ne vous embêtez pas à les renommer). Rappelez-vous les conventions de dénomination dont nous avons parlé précédemment. Vous pouvez aussi choisir de convertir Nom du pays en nom_du_pays, mais soyez cohérent dans la façon dont vous nommez tous vos champs.

Etape 9 – Rendre fixe la première ligne (et la première colonne)

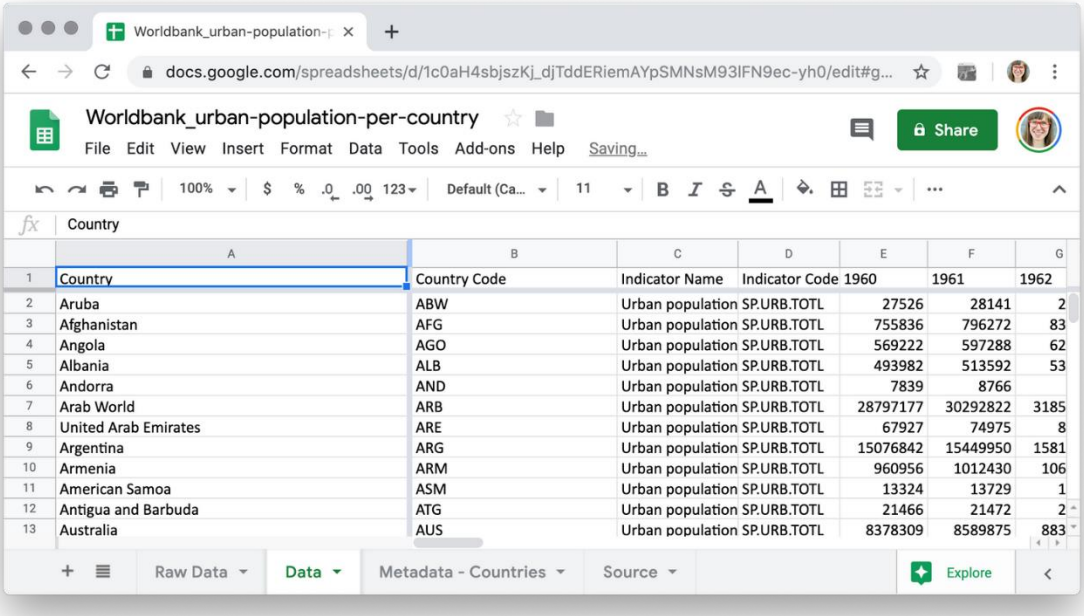
Vous devriez maintenant être à un point où vos en-têtes sont corrigés. Il est important pour les transformations suivantes de toujours les avoir en vue. Il faut pour ça les « figer » pour arriver à une situation comme présentée dans l’animation ci-dessous :



Burundi

	A	B	C	D
1	Country Name	Country Code	Indicator Name	Indicator Co
2	Aruba	ABW	Urban population	SP.URB.TOTI
3	Afghanistan	AFG	Urban population	SP.URB.TOTI
4	Angola	AGO	Urban population	SP.URB.TOTI
5	Albania	ALB	Urban population	SP.URB.TOTI
6	Andorra	AND	Urban population	SP.URB.TOTI
7	Arab World	ARB	Urban population	SP.URB.TOTI
8	United Arab Emirates	ARE	Urban population	SP.URB.TOTI
9	Argentina	ARG	Urban population	SP.URB.TOTI
10	Armenia	ARM	Urban population	SP.URB.TOTI

Voilà à quoi devrait ressembler les données de départ maintenant :



Worldbank_urban-population-per-country

	A	B	C	D	E	F	G
1	Country	Country Code	Indicator Name	Indicator Code	1960	1961	1962
2	Aruba	ABW	Urban population	SP.URB.TOTL	27526	28141	2
3	Afghanistan	AFG	Urban population	SP.URB.TOTL	755836	796272	83
4	Angola	AGO	Urban population	SP.URB.TOTL	569222	597288	62
5	Albania	ALB	Urban population	SP.URB.TOTL	493982	513592	53
6	Andorra	AND	Urban population	SP.URB.TOTL	7839	8766	
7	Arab World	ARB	Urban population	SP.URB.TOTL	28797177	30292822	3185
8	United Arab Emirates	ARE	Urban population	SP.URB.TOTL	67927	74975	8
9	Argentina	ARG	Urban population	SP.URB.TOTL	15076842	15449950	1581
10	Armenia	ARM	Urban population	SP.URB.TOTL	960956	1012430	106
11	American Samoa	ASM	Urban population	SP.URB.TOTL	13324	13729	1
12	Antigua and Barbuda	ATG	Urban population	SP.URB.TOTL	21466	21472	2
13	Australia	AUS	Urban population	SP.URB.TOTL	8378309	8589875	883

Etape 10 - Supprimer les colonnes et lignes inutiles :

Vos données sont destinées à être des données sources, il devrait donc être rare que vous ayez besoin de supprimer quoi que ce soit, mais parfois les données sont redondantes. Soit parce que les données sont répétitives, soit parce qu'elles fournissent des résumés qui ne sont pas nécessaires parce qu'ils peuvent être calculés. Par exemple, si vous disposez de données sur les dépenses par poste, avez-vous également besoin d'une ligne pour le total ? C'est une décision que vous devrez prendre, mais n'oubliez pas de la consigner dans votre fichier de métadonnées descriptives.

Etape 11 - Supprimer les milliers de séparateurs :

Les milliers de séparateurs sont des caractères (« , » en anglais, « . » en allemand, parfois c'est juste un espace) qui permettent de reconnaître facilement la magnitude d'un nombre. Par

exemple, 38.394.105 s'arrondit plus vite à 38 millions dans notre esprit que 38394105. Mais si ces marques sont utiles pour les humains, ils sont terribles pour les ordinateurs et conduisent à des interprétations ambiguës. Il convient donc d'enlever tous les séparateurs qui pourraient être présents.

Par exemple le tableau de gauche ci-dessous devrait être transformé pour arriver au tableau de droite :

755,836	796 272
569,222	597 288
493,982	513 592
7,839	8 766
28,797,177	30 292 822
67,927	74 975

755836	796272
569222	597288
493982	513592
7839	8766
28797177	30292822
67927	74975

Etape 12 - Standardiser le texte :

Examinez les noms et les textes (tels que les pays, les questions de sondage, etc.) et voyez s'ils doivent être corrigés ou s'ils doivent être conformes à une norme. Il peut y avoir plusieurs façons d'épeler le nom d'un lieu particulier, mais il est probable qu'il n'y ait qu'une seule orthographe officielle (voir le référentiel de noms officiels). Si vous savez que les utilisateurs peuvent trouver cette orthographe peu claire, vous devrez peut-être créer une nouvelle colonne avec un code d'identification standard. Par exemple, les Nations unies ont des codes de pays officiels pour chaque pays, ce qui peut être très utile pour le croisement de données.

Etape 13 – Correction des Dates :

Il est difficile de travailler avec des dates. Il existe de nombreux formats de date différents ("1er novembre 2019", "1/11/19", etc.). Excel, Google Sheets, etc. ne les enregistrent pas dans l'un d'eux, mais comme "numéro de série". Le 1er novembre 2019 devient 43770 si vous changez le format de la cellule en un numéro. Si jamais vous rencontrez un nombre aussi étrange que celui-ci, il vous suffit de changer le format de la cellule en date. Pour ce faire, cliquez sur Format > Nombre > Date.

Il est essentiel de toujours utiliser le format ISO pour les dates (c'est-à-dire AAAA-MM-JJ). Cette méthode est meilleure pour l'analyse des données, car elle permet un tri automatique, et elle élimine tout doute. Certains formats de dates impliquent une abréviation du nom des mois. Dans différents pays, sur plusieurs sites et dans plusieurs langues, cela peut entraîner une confusion et une incohérence inutiles. Il est toujours préférable d'éliminer tout doute, et les formats de date sont une cause perpétuelle d'erreurs dans les grandes séries de données.

Etape 14 - Séparer les variables en colonnes séparées :

Parfois, vous verrez plusieurs variables dans une colonne. Comme une colonne avec les États américains au format US-TX. Ou une colonne avec les entreprises et le produit qu'elles vendent. Lorsque vous voulez analyser ou visualiser les données, il sera impossible de faire des visualisations fines. Il est donc nécessaire de séparer toutes les colonnes contenant plus

d'une information. Pour ce faire, vous pouvez utiliser la formule =SPLIT(B1, "-") ou les formules =LEFT(), =RIGHT() et =MID()[18].

Etape 15 - Ajouter des colonnes avec des informations supplémentaires (par exemple, des ID géographiques) :

Parfois, vous avez des données dans deux documents Excel différents. Ou dans deux feuilles différentes du même fichier Excel. Parfois, ces informations sont nécessaires pour visualiser les données - par exemple lorsque vous avez besoin des identifiants géographiques pour créer une carte choroplèthe, mais qu'ils ne figurent pas dans votre source de données d'origine. Ces champs sont connus sous le nom de "clés étrangères" dans une base de données et ils permettent de joindre des données diverses sur cette colonne commune en raison de ces identités uniques.

Dans notre échantillon de données, nous avons aussi quelques informations supplémentaires. Dans la fiche "Métadonnées - Pays", les données de la Banque mondiale expliquent dans quelle région et dans quel groupe de revenu se trouvent tous les pays listés. Pour intégrer la région et/ou le groupe de revenu dans la feuille des données, il est possible de trier la colonne des pays par ordre alphabétique dans les deux feuilles, puis copier et coller la colonne des régions et des groupes de revenus de leur feuille dans la feuille des données modifiées. Cette approche ne fonctionne que lorsque nous avons des valeurs uniques (sinon le tri devient peu fiable), et si nous sommes sûrs à 100 % que les valeurs sont les mêmes. Dès que les deux colonnes diffèrent en longueur, il est nécessaire d'utiliser une autre méthode présentée ci-dessous.

La deuxième méthode consiste à utiliser la formule =VLOOKUP(B1,A1:A100,2,FALSE)[19] . Cette formule n'est pas très facile à utiliser, mais Microsoft a fait un assez bon travail en l'expliquant en détail en ligne[20] .

Etape 16 - Mettre les données au format large :

On peut disposer les mêmes données en lignes et en colonnes dans des formats différents. Notre échantillon de données de la Banque mondiale est dans un format large : Les valeurs de chaque année se trouvent dans une nouvelle colonne, de sorte qu'une ligne comporte plusieurs valeurs. Comme indiqué précédemment, il est généralement plus facile de manipuler des données au format long, qui évite de devoir changer le schéma de métadonnées lorsque des nouvelles données sont ajoutées. De la même manière, le format long permet de conserver une structure stable tout en rajoutant des données, ce qui facilite les réutilisations et l'automatisation des d'analyses.

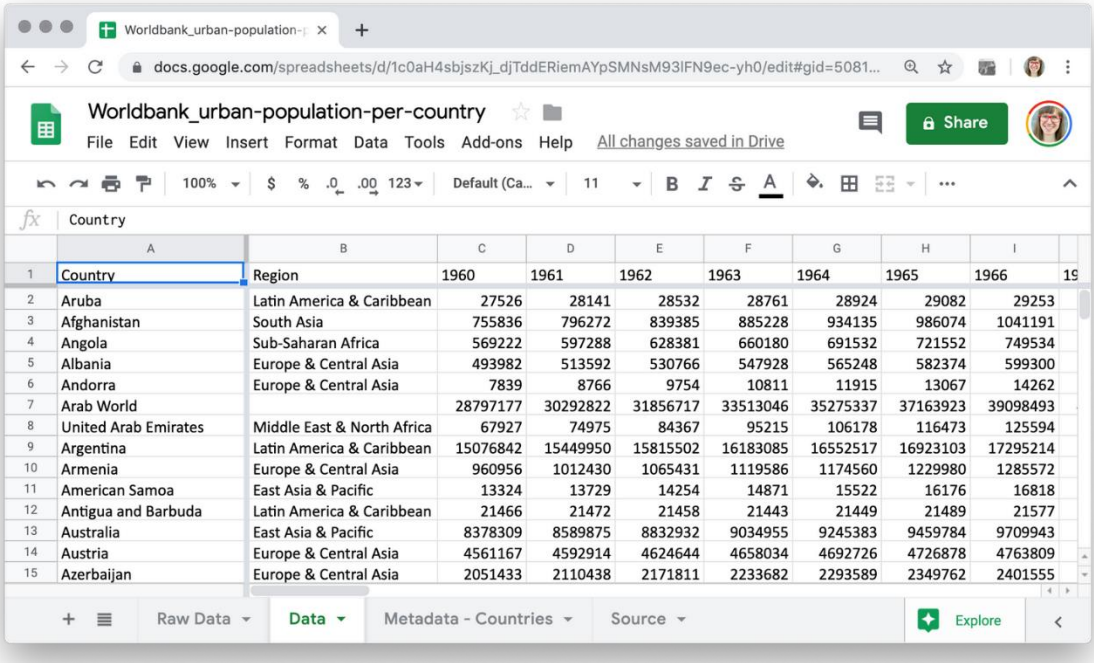
---

[18] Attention suivant les langues d'installation d'Excel, les noms de fonctions changent. Voir la correspondance des fonctions anglaises et françaises à http://fr.excelfunctions.eu/.

[19] Excel utilise des noms de fonctions différents suivant la langue dans laquelle est installé le logiciel. Voir la correspondance des fonctions anglaises et françaises à http://fr.excelfunctions.eu/

[20] https://support.microsoft.com/fr-fr/office/recherchev-recherchev-fonction-0bbc8083-26fe-4963-8ab8-93a18ad188a1

Pour transformer des données au format long, il faut d'abord réfléchir à la structure du fichier et aux informations à transformer des colonnes en lignes. Par exemple à partir du fichier de données sous la forme suivante



	A	B	C	D	E	F	G	H	I
1	Country	Region	1960	1961	1962	1963	1964	1965	1966
2	Aruba	Latin America & Caribbean	27526	28141	28532	28761	28924	29082	29253
3	Afghanistan	South Asia	755836	796272	839385	885228	934135	986074	1041191
4	Angola	Sub-Saharan Africa	569222	597288	628381	660180	691532	721552	749534
5	Albania	Europe & Central Asia	493982	513592	530766	547928	565248	582374	599300
6	Andorra	Europe & Central Asia	7839	8766	9754	10811	11915	13067	14262
7	Arab World		28797177	30292822	31856717	33513046	35275337	37163923	39098493
8	United Arab Emirates	Middle East & North Africa	67927	74975	84367	95215	106178	116473	125594
9	Argentina	Latin America & Caribbean	15076842	15449950	15815502	16183085	16552517	16923103	17295214
10	Armenia	Europe & Central Asia	960956	1012430	1065431	1119586	1174560	1229980	1285572
11	American Samoa	East Asia & Pacific	13324	13729	14254	14871	15522	16176	16818
12	Antigua and Barbuda	Latin America & Caribbean	21466	21472	21458	21443	21449	21489	21577
13	Australia	East Asia & Pacific	8378309	8589875	8832932	9034955	9245383	9459784	9709943
14	Austria	Europe & Central Asia	4561167	4592914	4624644	4658034	4692726	4726878	4763809
15	Azerbaijan	Europe & Central Asia	2051433	2110438	2171811	2233682	2293589	2349762	2401555

La transformation en format long nécessite que Les années deviennent des éléments d'une ligne, donc la transformation au format long nécessite les actions suivantes :

- Créer 2 colonnes « année » et « valeur »
- Transformer chaque ligne en autant de lignes que d'années. Les deux premières colonnes (pays et région) ne changeront pas et chaque couple « en-tête d'années – valeur » deviendra une ligne

<u>Conclusion</u>

A l'issue de toutes ces étapes, vous obtenez un fichier avec la première ligne contenant les en-têtes de colonnes et toutes les autres lignes sont les lignes de données au format long. Ce fichier est un fichier correspondant au meilleur format possible pour la publication en données ouvertes. Avant de le publier, il convient maintenant de vérifier son contenu. Ce processus est présenté dans la section 3.2.