

Guide méthodologique pour l'inventaire des données publiques

Introduction

Ce document a pour objectif de présenter un cadre méthodologique permettant de réaliser un inventaire des données publiques. Il introduit d'abord les différents concepts et cadre juridique, il présente ensuite l'organisation à adopter, et décrit enfin le déroulement de l'opération d'inventaire.

En annexe, des exemples illustratifs d'inventaires et de jeux de données sont fournis.

SOMMAIRE

Introduction.....	1
1. Contexte et objectifs.....	4
2. Définition des concepts et cadre juridique.....	5
2.1 Définition des concepts.....	5
2.2 Cadre juridique.....	8
3. Cadre organisationnel.....	9
4. Inventorier les données.....	10
4.1 Identification des jeux de données.....	10
4.2 Renseignement des métadonnées.....	11
5. Maintenir l’inventaire des données.....	14
6. Annexes.....	15
6.1 Format de publication des données.....	15
6.2 Le modèle d’inventaire de données.....	16
6.3 Exemples de jeux de données.....	16

1. Contexte et objectifs

La Tunisie a adhéré à l'Open Government Partnership (OGP) en 2014, regroupant différents pays qui se sont inscrits dans une démarche œuvrant pour une gouvernance ouverte, favorisant la transparence et encourageant la participation citoyenne. Chaque Etat membre doit réaliser un plan d'actions tous les deux ans.

L'ouverture des données publiques (Open Data) est considérée par le gouvernement tunisien comme un objectif majeur dans sa stratégie de modernisation de l'action publique et de développement de l'économie numérique. La Tunisie dispose pour cela d'un cadre juridique propice à l'ouverture des données publiques et a mis en place un portail national Open Data <http://www.data.gov.tn> ainsi que des portails sectoriels Open Data.

L'objectif à présent est de disposer d'un inventaire des données publiques qui soit le plus exhaustif possible et ce pour tous les organismes publics. Pour ce faire un portail dédié a été mis en place leur permettant d'y référencer les données qu'ils produisent et collectent. L'UAE est chargée d'assister l'ensemble des structures publiques pour cela et d'assurer la coordination entre elles, avec l'assistance technique de la Banque Mondiale.

Pourquoi inventorier ses données ?

En plus de l'engagement qu'a pris la Tunisie dans le cadre de l'OGP, il est primordial pour chaque structure de l'administration tunisienne de :

- ▶ Maîtriser son patrimoine de données existant en ayant un référencement exhaustif des jeux de données disponibles ;
- ▶ Organiser et prioriser la publication des données (les données publiées seront un sous ensemble du patrimoine qui aura été référencé dans l'étape précédente)
- ▶ Constituer un catalogue de jeux de données pour les utilisateurs/réutilisateurs publics, privés ou citoyens de l'open data en Tunisie ou à l'étranger ;
- ▶ Identifier les redondances dans l'administration et aider à la constitution de référentiels nationaux ouverts.

2. Définition des concepts et cadre juridique

2.1 Définition des concepts

L'Open Data consiste en la publication des données publiques, en vue notamment de leur réutilisation. Ci-dessous une définition des principaux concepts qui y sont liés.

Donnée publique	C'est une donnée produite ou collectée par un organisme public, dans le cadre de ses activités de service public.
Donnée ouverte	C'est une donnée publique librement accessible et réutilisable par le public, disponible sous un format ouvert, lisible par machine et non soumise à des restrictions liées aux droits d'autrui dont la protection de sa vie privée, ses données personnelles, sa propriété intellectuelle et d'autres droits couverts par une protection légale. (cf. les 10 concepts ci-dessous qui caractérisent une donnée ouverte)
Jeux de données	C'est une collection d'éléments connexes de données associées entre elles et accessibles individuellement ou de façon combinée, ou gérées comme une entité.
Licence d'ouverture des données	Il s'agit d'un document lié aux jeux de données publiques ouvertes publiées qui spécifie les règles de leur disponibilité et de leur utilisation.
Producteur	Il s'agit de la structure publique qui produit une donnée publique et l'ouvre à la réutilisation (conformément aux conditions prévues par la licence)
Réutilisateur	Il s'agit de toute personne physique ou morale qui réutilise une donnée publique (conformément aux conditions prévues par la licence)
Données brutes	Ce sont les données non interprétées émanant d'une source primaire, ayant des caractéristiques liées à celle-ci et qui n'ont été soumises à aucun traitement ou toute autre manipulation
Inventaire de données	La liste des jeux de données produits ou collectés par un organisme dans le cadre de l'exercice de ses activités et l'ensemble des informations y afférentes comme la classification, la périodicité de collecte, le propriétaire, et diverses métadonnées.

Métadonnées	Les informations associées à un jeu de données pour le décrire et le définir en précisant la source de ces données, la date à laquelle elles ont été préparées ou collectées et publiées, l'emplacement de son hébergement, et d'autres informations d'identification s'y rapportant.
Format ouvert	C'est un format de données structuré, lisible par machine et interopérable qui peut être utilisé sans restriction d'accès ou de réutilisation.
Nomenclature	C'est un ensemble exhaustif et structuré de catégories mutuellement exclusives et bien décrites, souvent présentées selon une hiérarchie qui est reflétée au moyen de codes numériques ou alphabétiques qui leur sont assignés et qui est utilisée en vue de normaliser des concepts et de compiler des données.
Anonymisation des données	C'est une modification du contenu ou de la structure de ces données afin de rendre très difficile ou impossible l'identification des personnes ou des entités concernées.
API	C'est une interface logicielle de programmation permettant un accès informatique automatisé à des contenus.
Uniform Resource Identifier	C'est une adresse unique utilisée pour identifier le contenu sur le réseau, et sa structure répond aux normes Internet définies par le World Wide Web. Cet identifiant permet une identification permanente de la ressource, même si cette ressource est déplacée ou supprimée.

Une donnée est réputée ouverte si elle respecte ces dix principes établis par l'ONG « Sunlight Foundation » qui font office de référence internationalement :

#	Principes	Descriptions
1	Complètes	Le jeu de données publié ne doit pas être tronqué
2	Primaires	Le jeu de données publié doit provenir de la source qui l'a produite, sans aucun traitement de la donnée
3	Fraîches	Il doit y avoir une mise à jour régulière des données publiées
4	Accessibles	Il ne doit y avoir aucune barrière compliquant l'accès aux données tel que formulaire à remplir, etc... L'accès doit être direct avec une URL
5	Electroniquement lisibles par une machine	Utiliser le format requis (cf. annexe)
6	Accessibles sans discriminations	Tout le monde doit pouvoir y accéder
7	Disponibles sous des formats ouverts	Utiliser le format requis (cf. annexe)
8	Disponibles sous licences ouvertes	Utiliser la licence nationale (cf. data.gov.tn)
9	Accessibles de façon pérenne en ligne	Les données doivent rester accessibles
10	Sans coût d'utilisation	La donnée doit être disponible gratuitement sans redevance à payer.

Le producteur d'une donnée publique doit garantir autant que possible le respect de ces principes d'ouverture. Par ailleurs, il ne garantit pas l'absence de défauts, d'erreurs ou d'irrégularités viciant la donnée. De ce point de vue, l'expérience d'autres pays a montré que l'ouverture de la donnée a été une opportunité pour fiabiliser la donnée par les différentes interactions avec la communauté des réutilisateurs des données publiques.

2.2 Cadre juridique

La Tunisie a adopté le 26 Janvier 2014 une nouvelle constitution qui fournit un cadre juridique propice aux principes de la gouvernance ouverte et participative, notamment avec l'article 32 qui garantit le droit d'accès à l'information.

Ceci a été appuyé par une nouvelle loi d'accès à l'information n° 2016-22 adoptée le 26 Mars 2016 en remplacement du décret-loi n°2011-41, qui garantit aux citoyens un cadre légal pour toute demande d'accès aux données et informations produites par l'administration dans le cadre de leur mission de service public.

Enfin, le décret n°2021-3 du 6 Janvier 2021, relatif aux données publiques ouvertes, est venu compléter le cadre juridique en apportant toutes les précisions utiles permettant d'organiser la publication des données ouvertes. Il précise notamment que chaque organisme public doit préparer et mettre à jour un inventaire des données publiques (art. 6) ainsi que les modalités de publication de ces données.

Pourquoi publier les données publiques ?

- ▶ **Favoriser l'innovation économique:** L'ouverture des données contribue à stimuler l'innovation par le développement de nouveaux services par les entreprises et la société civile
- ▶ **Améliorer l'efficacité de l'action publique :** L'ouverture des données publiques permet à l'administration d'envisager de nouvelles méthodes de travail et de rompre les silos entre les différentes structures
- ▶ **Transparence et redevabilité de l'action publique :** L'ouverture des données publiques permet de renforcer le lien de confiance entre élus et citoyens en partageant une information claire sur les décisions prises et l'utilisation des moyens financiers
- ▶ **Participation à la vie publique :** L'ouverture des données publiques permet d'encourager la participation du citoyen à l'élaboration des politiques publiques ainsi qu'au suivi de leur mise en œuvre et leur évaluation.

3. Cadre organisationnel

S'engager dans une démarche d'ouverture des données publiques est une action qui doit s'inscrire dans la durée, de manière continue et itérative. Pour ce faire, nous recommandons la mise en place d'une organisation propice à la réussite du projet.

Tout d'abord, nous préconisons la mise en place d'un comité multidisciplinaire au sein de chaque ministère, chargé du suivi d'avancement du projet. Ce comité serait constitué des profils suivants :

- Un(e) chargé(e) de l'open data
- Un(e) responsable juridique
- Un(e) représentant(e) de la direction informatique
- Un(e) représentant(e) de la direction des archives
- Un(e) représentant(e) de la direction des études et des statistiques

Pour chaque structure sous tutelle, il faut également définir a minima un référent métier et un référent technique qui assurent le suivi et la coordination avec le chargé open data du ministère. Les organismes sous tutelle du ministère peuvent également mettre en place un groupe de travail équivalent à celui énoncé ci-dessus pour travailler sur leur inventaire.

Le point focal chargé de l'Open Data dans le ministère, qui est le vis-à-vis de l'UAE, doit être chargé de collecter les informations depuis les différents organismes sous tutelle et de maintenir un inventaire de données à jour au niveau ministériel.

Facteurs clés de succès

- ▶ **La sensibilisation de toutes les directions aux enjeux et intérêts de l'Open Data :** L'adhésion de tous les participants conditionne fortement le succès du projet
- ▶ **Implication « métier » sur toutes les phases du projet:** Leur connaissance de leur administration et leurs compétences sont essentielles pour mener à bien le projet
- ▶ **Mener le projet avec une démarche progressive et évolutive :** L'obtention de résultats concrets et visible rapidement est de nature à faciliter la poursuite du projet

4. Inventorier les données

La réalisation de l'inventaire se base sur deux étapes clés : identification des jeux de données et renseignement de leurs métadonnées.

4.1 Identification des jeux de données

Pour identifier tous les jeux de données produits ou collectés, nous préconisons au groupe de travail de procéder par des ateliers brainstorming en vous laissant guider par ces étapes de priorisation :

1. Toutes les données qui sont déjà publiées dans le site interne de votre structure ou dans le portail national 'open data'
2. Les données non encore publiées et qui doivent faire l'objet d'une publication proactive selon l'article 6 de la loi relative au droit d'accès à l'information n°2016-22
3. Les données non encore publiées et qui font l'objet de demandes récurrentes d'accès à l'information par les citoyens (y compris journalistes, chercheurs, etc...)
4. Les données non encore publiées qui peuvent présenter un intérêt économique en vue de leur réutilisation
5. Les données non encore publiées et qui peuvent faire l'objet de demandes par d'autres entités publiques
6. Les données non encore publiées et qui sont utilisées dans les rapports et études réalisées par votre structure
7. Analyse comparative avec les données publiées par des structures analogues à la votre à l'international.

Il est important, que l'inventaire soit le plus exhaustif possible, pour chacune des structures sous tutelle du ministère. Vous pouvez également partir des textes définissant la mission de service public de votre structure, cela contribue à mieux identifier les données qui sont produites dans le cadre de ce service public. Les bases de données des différents systèmes (ou encore les rapports d'activités ou autres rapports) peuvent constituer aussi une source d'identification des données à inventorier.

Il faudra prendre soin d'identifier les jeux de données qui contiennent des données personnelles, ou qui présentent un risque pour la sûreté de l'Etat.

Une fois tous les jeux de données produits ou collectés, ont été identifiés, il convient de les renseigner dans l'inventaire. Pour ce faire, nous vous préconisons de vous appuyer sur cette méthodologie reposant sur une série de questions à se poser afin de pouvoir décrire toutes les caractéristiques des jeux de données identifiés.

Qui ?

- Qui a produit les données ?
- Qui possède les données ?
- Qui peut publier les données ?
- Qui est le public ciblé ?

Quoi ?

- Quel est le titre ?
- Quel type de données ?
- Quel volume de données ?

Quand ?

- Quand les données ont été produites ?
- Quand ont-elles été publiées ?
- Quand ont-elles été mises à jour ?
- Quand devraient-elles être mises à jour la prochaine fois ?

Comment ?

- Sous quel format ?
- Avec quelle licence ?
- Quelles contraintes à prendre en compte ? (vie privée, sécurité ...)
- Y a-t-il une redevance à payer ?

Où ?

- Où les données ont-elles été produites ?
- Où sont-elles décrites ?
- Où sont-elles physiquement localisées ?

Ceci dans le but de renseigner toutes les métadonnées utiles à l'inventaire décrites ci-dessous.

4.2 Renseignement des métadonnées

Les métadonnées peuvent être obligatoires ou facultatives et ont été regroupées en 3 catégories :

- Les métadonnées clés
 - Il s'agit des métadonnées essentielles qui permettent de caractériser le jeu de données, décrire la source de la donnée, l'organisme propriétaire, sa fréquence de production, et sur quoi le jeu de données porte.
- Les métadonnées pour l'évaluation de la priorisation interne
 - Il s'agit de métadonnées qui permettent d'évaluer la qualité du jeu de données, son potentiel de réutilisation et son niveau de préparation technique quant à l'éventualité de sa publication. L'idée étant de prioriser les données de bonne qualité, avec un fort potentiel de réutilisation et techniquement prêts. Ces jeux de données peuvent être renseignés avec l'UAE.
- Les métadonnées avancées
 - Il s'agit de métadonnées qui permettent d'évaluer l'effort requis pour préparer le jeu de données à son ouverture et sa maintenance et préciser le format et la date de publication prévue.

Ainsi pour chaque jeu de données identifié, qui est produit par un des établissements dépendants de votre ministère, il faudra renseigner les attributs suivants (en priorité vous pouvez vous concentrer sur les métadonnées clés) :

Métadonnées clés

- **Titre** (*en arabe et en français*): Permet de renseigner de manière courte sur quoi porte le jeu de données ;
- **Description** (*en arabe et en français*): Il s'agit d'un complément au titre qui permet de décrire plus longuement le contenu du jeu de données ;
- **Mots clés** (*en arabe et en français*) : De 1 à 5 mots qui permettent de décrire le jeu de données.
- **Périodicité de le production/collecte des données** : Définit la fréquence à laquelle le jeu de donnée est mis à jour (ex : quotidiennement, mensuellement, annuellement,...) ;
- **Date début de production** : Il s'agit de l'année où la donnée a été produite ou collectée la première fois
- **Date fin de production** : Il s'agit de l'année où la donnée a cessé d'être produite ou collectée. (Non obligatoire).
- **Propriétaire** : Préciser la structure ou l'établissement sous tutelle qui a collecté ou produit le jeu de donnée. Il s'agit de la source primaire de la donnée ;
- **Source** : Pour les données informatisées, il est demandé de préciser le format dans lequel existe le jeu de donnée (ex : Excel, CSV, PDF, ...). Pour les données non informatisées, il s'agit de préciser le support (manuscrit, ...) ;
- **Contact responsable métier** : Nom et éventuellement adresse mail de la personne qui est à considérer comme le point de contact pour ce jeu de donnée ;
- **Contact responsable technique** : Nom et éventuellement adresse mail du responsable technique qui est à considérer comme le point de contact pour ce jeu de donnée ;
- **Niveau de classification des données** : Définit le niveau d'accès de ce jeu de donnée. Public, c'est à dire qu'il peut être consulté par tout le monde. Protégé : C'est à dire qu'il serait concerné par des restrictions d'accès. Confidentiel : Ne peut-être publié.
- **Langue** : La langue dans laquelle est publié le jeu de données (arabe, français, anglais);
- **Thème au niveau de l'organisation**: Cet attribut caractérise le domaine ou le thème concerné par le jeu de donnée. (la liste déroulante sera propre à chaque structure);
- **Thème au niveau du portail national Open Data**: Cet attribut caractérise le domaine ou le thème concerné par le jeu de donnée. (la liste déroulante sera définie par l'UAE);
- **Lien vers le jeu de données publié** : Il s'agit de préciser un lien URL si la donnée est déjà publiée
- **Couverture géographique** : Précise la région géographique concernée par le jeu de donnée (ex : Gouvernorat Tunis, Tunisie, Ville de Bizerte, etc....) ;

Métadonnées pour l'évaluation de la priorisation interne

- **Potentiel de réutilisation des données** : Un score de 1 à 10 pour évaluer le potentiel de réutilisation.
- **Niveau de demande interne** : Score de 1 à 10. Ajouter +1 pour chaque demande identifiée au sein de l'administration en général.
- **Niveau de la qualité de données** : Score de 1 à 10 permettant d'évaluer la qualité du jeu de données
- **Niveau de préparation technique** : Score de 1 à 10, permettant d'évaluer la capacité de lire automatiquement les données.

Métadonnées avancées *(non obligatoires)*

- **Estimation des activités nécessaires pour l'ouverture et la maintenance du jeu de données** : Un petit descriptif de ce qu'il faut faire pour être en situation de pouvoir publier les données
- **Estimation du coût de l'ouverture et de la maintenance du jeu de données** : Montant en dinars du coût estimé
- **Liens de publications existantes de la donnée (non open data)** : Lister les éventuels liens où la donnée est déjà publiée mais en format non ouvert
- **Système d'extraction ETL existant O/N** : Oui ou Non s'il existe un système d'extraction de données (ETL).
- **Description de l'ETL existant** : Un petit descriptif du système s'il existe.
- **Date d'ouverture planifiée** : Date prévisionnelle macro en trimestre. Ex : T2 2022
- **Formats de publication** : Préciser le(s) format(s) de publication, parmi CSV, JSON, API, etc...
- **Date de dernière mise à jour**: date de dernière mise à jour des métadonnées pour le jeu de données.

Le renseignement de l'inventaire se fera directement dans le portail dédié à cela. Dans les phases intermédiaires il est possible de remplir le template excel, puis de le charger sur le portail.

5. Maintenir l'inventaire des données

L'inventaire doit être maintenu à jour avec les informations du ministère et de ses établissements sous tutelles. Ce travail de collecte et de synchronisation est de la responsabilité du point focal chargé de l'Open Data dans le ministère. Ce dernier pourra faire le suivi de la saisie et la mise à jour du fichier d'inventaire lors du comité pluridisciplinaire si nécessaire ou tout autre moyen qui le juge utile.

Il est essentiel que l'inventaire des données soit maintenu et mis à jour régulièrement afin que son utilisation demeure pertinente. La fréquence de mise à jour reste à l'appréciation de chaque ministère en fonction de ses spécificités et la nature des données produites et manipulées. Mais il est d'usage que la fréquence de mise à jour d'un tel inventaire soit a minima une fréquence annuelle.

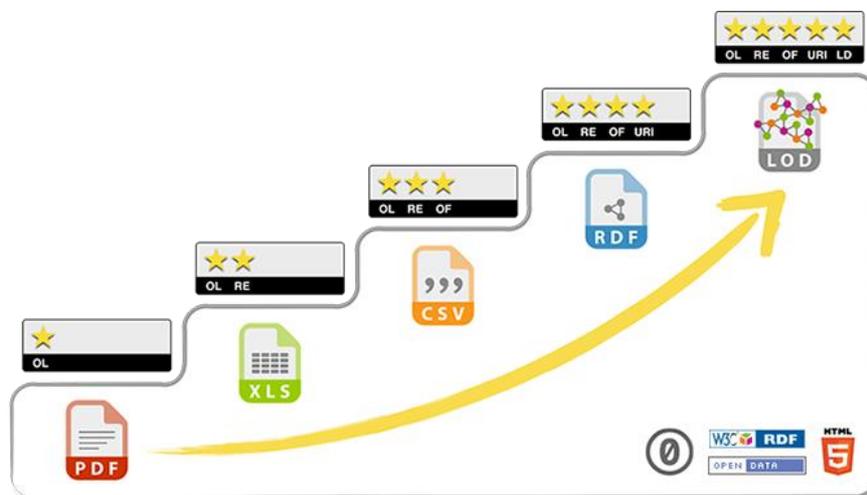
6. Annexes

6.1 Format de publication des données

Le format dans lequel doit être ouverte une donnée publique doit être structuré, libre d'accès et pouvoir être lu par une machine.

Quand la donnée est publiée dans un format le plus ouvert, elle aura plus de chance d'être réutilisée et exploitée par l'écosystème.

Ci-dessous un schéma montrant la notation des portails Open Data basée sur le format des données :



- Le format PDF étant le moins bien noté car il n'est pas exploitable facilement.
- Le format XLS (excel) ou DOC/DOCX (word) n'est pas très bien noté car il nécessite les outils bureautiques payants de Microsoft Office. Et donc cela limite la réutilisation de ces données à une certaine tranche de la population qui n'a pas ces licences.
- Le format CSV commence à être bien noté car gratuit, réutilisable par des machines.
- Le format RDF est encore mieux noté car il est plus structuré et standardisé pour les données sémantiques du web et donc plus facilement retraité/consommé par des machines.
- Le format LOD est le mieux noté car en plus de se baser sur le format RDF il offre d'autres avantages pour la réutilisation des données en les référençant par des URI, les interconnectant et en les rendant accessible par le protocole http.

Autres formats possibles

D'autres formats sont préconisés notamment pour les données géospatiales tel que : KML, GeoJSON.

6.2 Le modèle d'inventaire de données

L'inventaire dispose d'un portail dédié qui en permet la saisie directement en ligne, et pourra ainsi être public et mis à disposition directement.

Les chargés de l'open data de chaque ministère auront des accès au portail pour gérer et coordonner la saisie de l'inventaire. Il peut aussi dans le portail déléguer le renseignement de l'inventaire à des tierces personnes, notamment des référents pour les différents organismes sous tutelle.

Par ailleurs nous mettons à disposition un modèle de l'inventaire sous le format d'une feuille de calcul EXCEL. Le format du modèle fourni doit être respecté et ne doit pas être modifié.

(Il sera fourni conjointement à ce guide méthodologique dans un document séparé)

Autre exemple d'un inventaire de données

L'inventaire des données de Los Angeles :

<https://docs.google.com/spreadsheets/d/1uNtA4GbBwky8PPdNUvmXXZCI1GLtH5cGF-Q0FqD90w0/edit?usp=sharing>

6.3 Exemples de jeux de données

Pour illustrer davantage, voici quelques exemples de jeux de données publiques à l'état brut, produits ou collectés par une structure publique :

Exemple n°1 : Liste des médicaments et substances dopantes, publiée par le Ministère de la ville, de la jeunesse et des sports (**France**)

<https://www.data.gouv.fr/fr/datasets/liste-des-medicaments-et-des-substances-dopantes/>

Exemple n°2 : Liste des unités médicales ayant autorisation d'exercer, publiée par le ministère de la santé (**Ouganda**)

<http://catalog.data.ug/dataset/licensed-health-units-as-at-19th-june-2019>

Fin du document

Pour toute question ou demande de précision :

Mme Rim GARNAOUI, Directeur à l'Unité de l'Administration Electronique :
E-mail : rim.garnaoui@pm.gov.tn

Mme Sonia Gharbi, S-Directeur à l'Unité de l'Administration Electronique :
E-mail : Gharbi.sonia@pm.gov.tn