

# Open Data : Inventaire des données publiques & Référentiels de données

20 Octobre 2022 – Séance n°4



UNITED NATIONS

الاسواق  
ESCWA

Shared Prosperity **Dignified Life**





Shared Prosperity **Dignified Life**



# L'inventaire des données publiques

# L'inventaire des données

**Un inventaire des données publiques != Un inventaire des données ouvertes**

**Un inventaire des données publiques != Un inventaire des documents administratifs**

**C'est quoi une donnée publique ?**

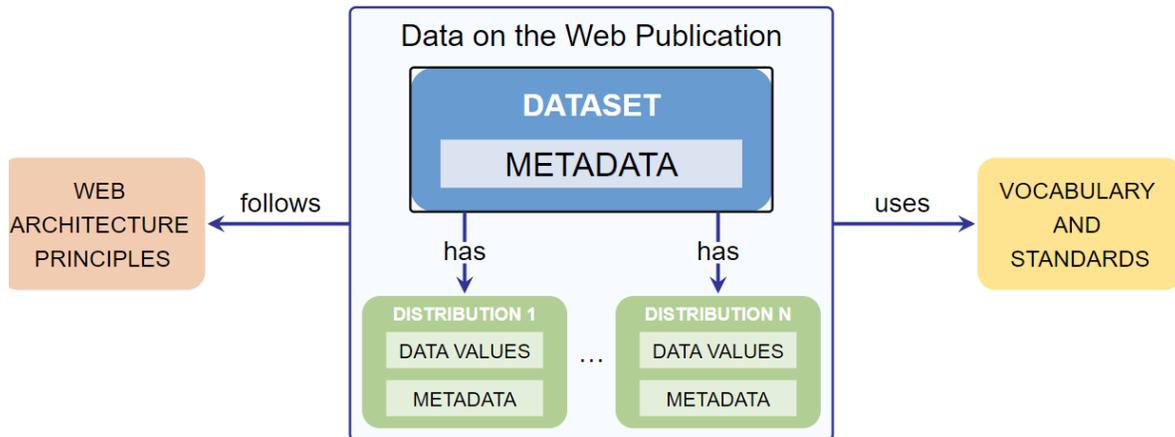
Toutes les données produites ou bien collectées par les organisations publiques dans le cadre de l'exercice de leurs activités.

Un inventaire des données publiques est un registre/catalogue contenant des données sur les données publiques →  
« **Uniquement des métadonnées** et sans données réelles » (Contrairement à l'Open Data !)

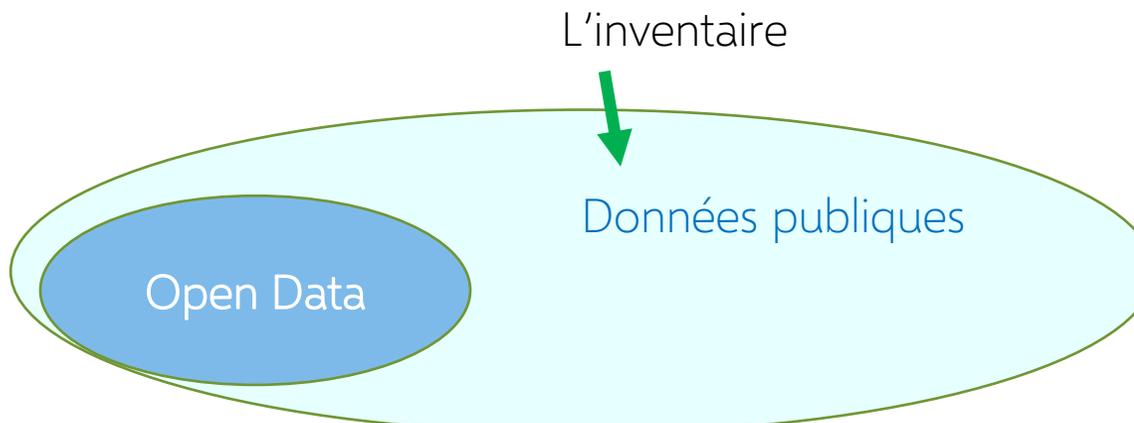
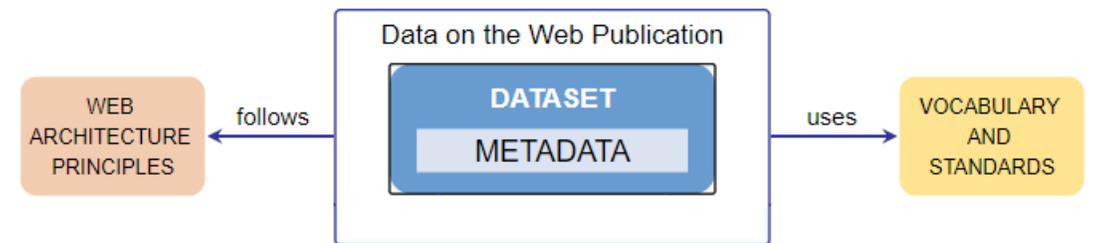
**Attention :** Il ne faut pas exclure aucun jeu de données pour des raisons de confidentialité !

# L'inventaire des données publiques

## Jeu de données en Open Data



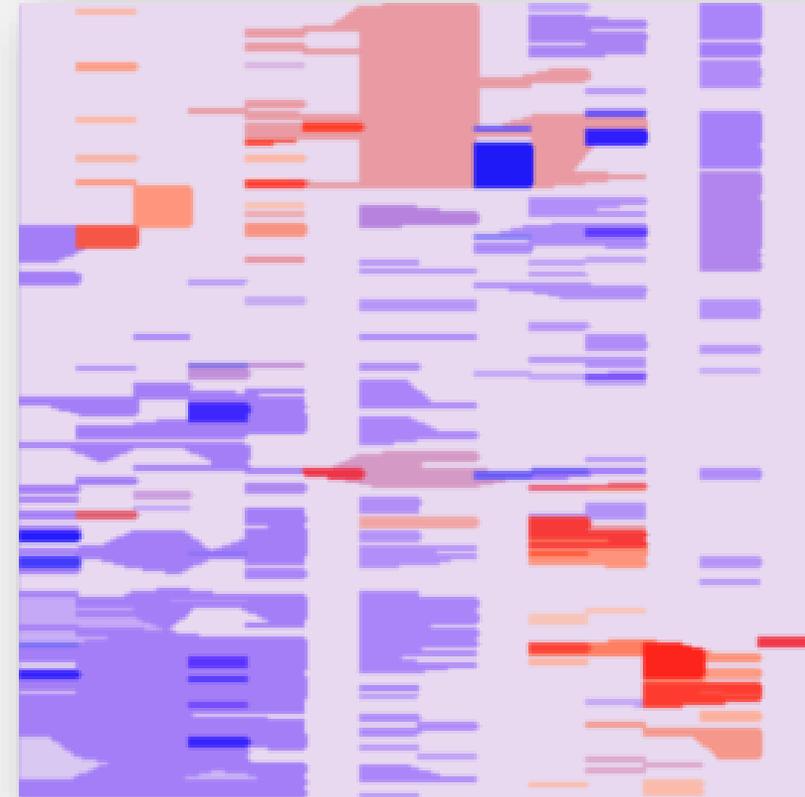
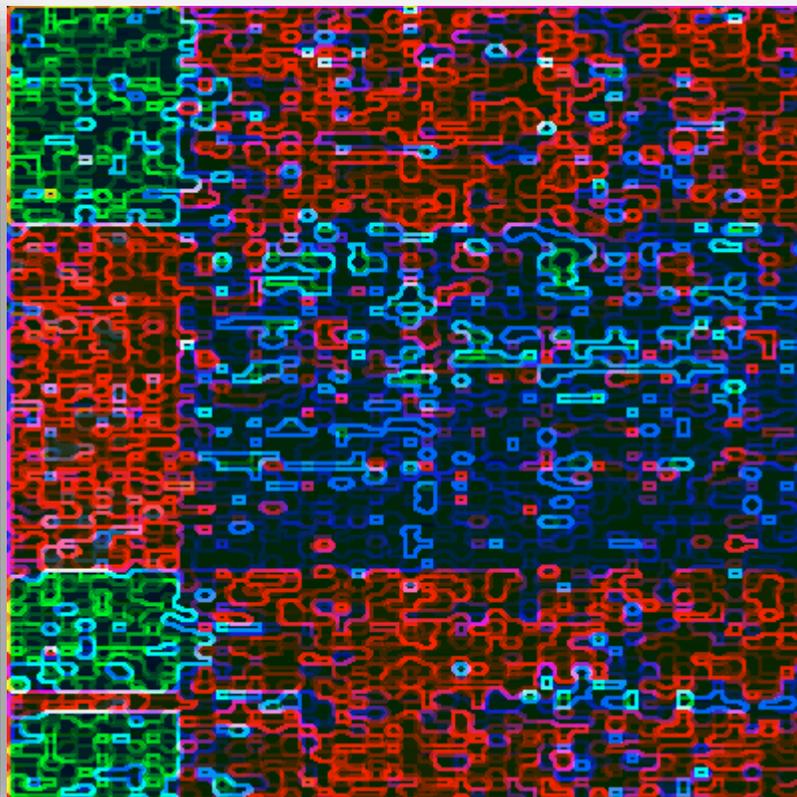
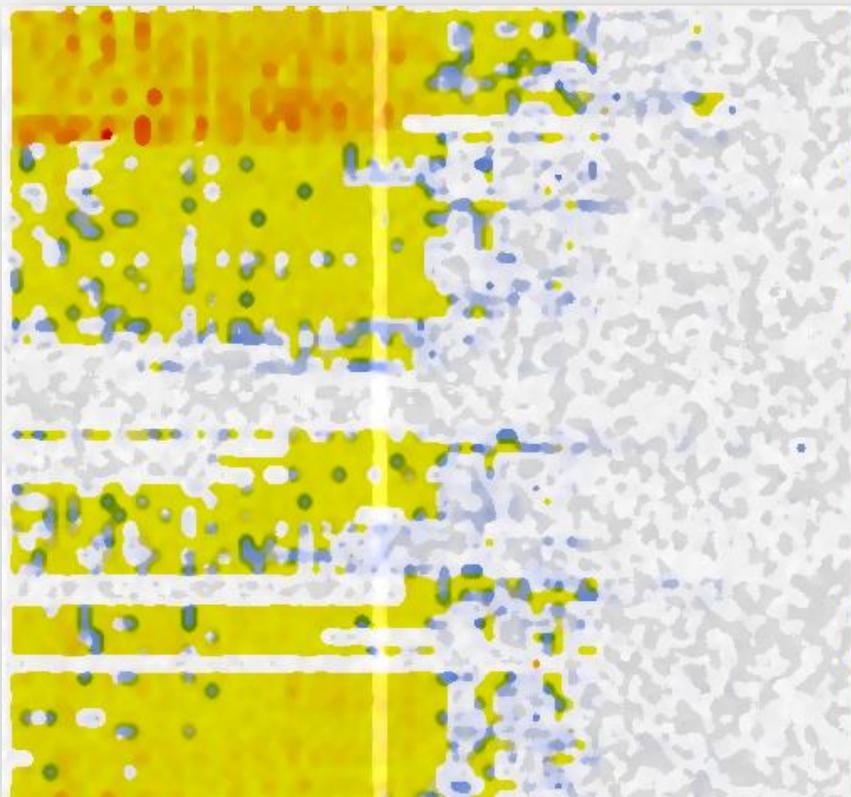
## Jeu de données dans l'inventaire des données publiques



Pas de distributions  
Pas de données,  
Uniquement  
des métadonnées

# L'inventaire des données publiques

Pourquoi réaliser un inventaire des données publiques ? « Big Picture » / « Vue Globale »



# L'inventaire des données

Gouvernance des Données  
Données ouvertes

## Pourquoi réaliser un inventaire des données publiques ?

- 
- La mise en place d'un système de gouvernance des données au sein des organismes publics,
  - Savoir qui détient quelles données publiques,
  - Identifier les redondances dans l'administration,
  - Identifier et créer des référentiels nationaux,
  - Mettre en place un registre national des données produites ou bien collectées par les acteurs publics,
  - Stimuler les demandes de priorisation d'ouverture,
  - Prioriser l'ouverture des données,
  - Établir un calendrier annuel pour l'ouverture des données (nécessaire à l'élaboration des plans de d'action annuels),

# L'inventaire des données

- ✓ L'utopie de l'inventaire **exhaustif**
- ✓ Une démarche **progressive** et **exploratoire**

## Comment conduire un inventaire des données publiques ?

- Définir un modèle d'inventaire de données (les métadonnées à renseigner),
- Définir une méthodologie claire,
- Identifier le périmètre (Les organisations à impliquer),
- Fixer un calendrier précis pour la première phase,
- Constitution des équipes et définition des responsabilités (les responsables Open Data, les collaborateurs, les collaborateurs IT, etc... ).

# L'inventaire des données

Numéro ou code du jeu de données
Titre du jeu de données en arabe
Titre du jeu de données en français
Description du jeu de données en arabe
Description du jeu de données en français
Mot clés en arabe
Mot clés en français
Périodicité de la production ou de la collecte des données
Année de la première collecte / production de la donnée
Année de fin de la collecte / production de la donnée
La structure publique propriétaire du jeu de données
Précision de la source des données
Contact (nom, courriel) du responsable Métier
Contact (nom, courriel) du responsable technique
Niveau de classification des données (public, protégé, confidentiel)
Énumération des colonnes contenant les données sensibles (données personnelles, propriété intellectuelle, commerciale, sécurité nationale, etc.)
Langues disponibles (concernant les données)

# L'inventaire des données

## Formulaire en ligne

- (+) C'est le moyen le plus rapide,
- (-) La qualité n'est pas garantie.

## Les entretiens one-to-one

- (-) période de temps plus longue,
- (+) Une meilleure qualité,

## Approche hybride

- (+) Concentrer les efforts sur les secteurs ou les domaines prioritaires
- (+) Envoyer le questionnaire en ligne à ceux qui maîtrisent le processus d'inventaire

# L'inventaire des données

- ❑ Mettre en place une plateforme nationale dédiée à l'inventaire des données publiques
  - ✓ Le cas de l'USA,
    - ➔ CKAN pour l'inventaire national des données publiques
  - ✓ Le cas de la Jordanie,
    - ➔ Une plateforme spécifique pour la classification et l'inventaire des données publiques
  - ✓ Le cas de la Tunisie,
    - ➔ Une plateforme spécifique pour l'inventaire,

# L'inventaire des données

## Cas de l'USA

Au niveau national : <https://inventory.data.gov/dataset>

- L'inventaire de données publiques est une obligation selon la « OPEN Government Data Act »
- General Services Administration (GSA) a mis en place une plateforme d'inventaire des données à base de CKAN pour aider les agences à se conformer à la loi « OPEN Government Data Act ».

La plateforme est utilisée par les agences ne disposant pas de système d'inventaire de données!

Exp :

- La ville de Philadelphia a développé son propre système d'inventaire de données et il est accessible à tous : <https://metadata.phila.gov/>

- Ville de San Francisco a développé une méthodologie pour la conduite des données ouvertes accessibles [https://docs.google.com/document/d/1W5C5oO2TrVnmOgLe8I\\_KYgmbghj6hDs9-4SC-ygMDV4/edit](https://docs.google.com/document/d/1W5C5oO2TrVnmOgLe8I_KYgmbghj6hDs9-4SC-ygMDV4/edit)

# L'inventaire des données

## Cas de l'USA

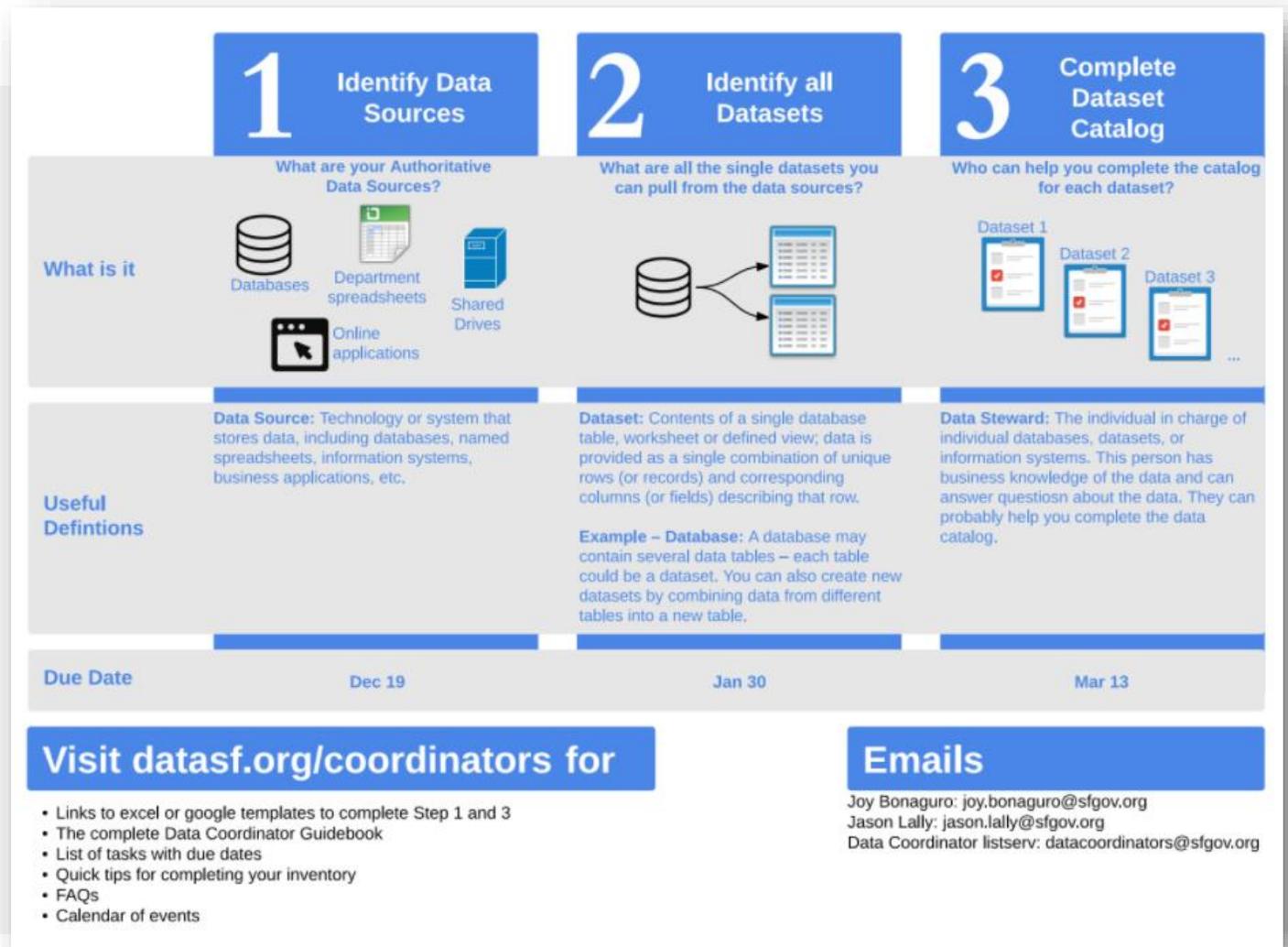
### DataSF Guidebook: Detailed Inventory Guide

Le guide prend en considération le niveau de complexité organisationnel :

- Faible complexité d'inventaire,
- Complexité moyenne d'inventaire,
- Complexité élevée d'inventaire,

Le guide propose l'organisation des réunions de brainstorming pour l'identification des données !

De point technologique → des feuilles Excel



# L'inventaire des données

## Cas de la Jordanie

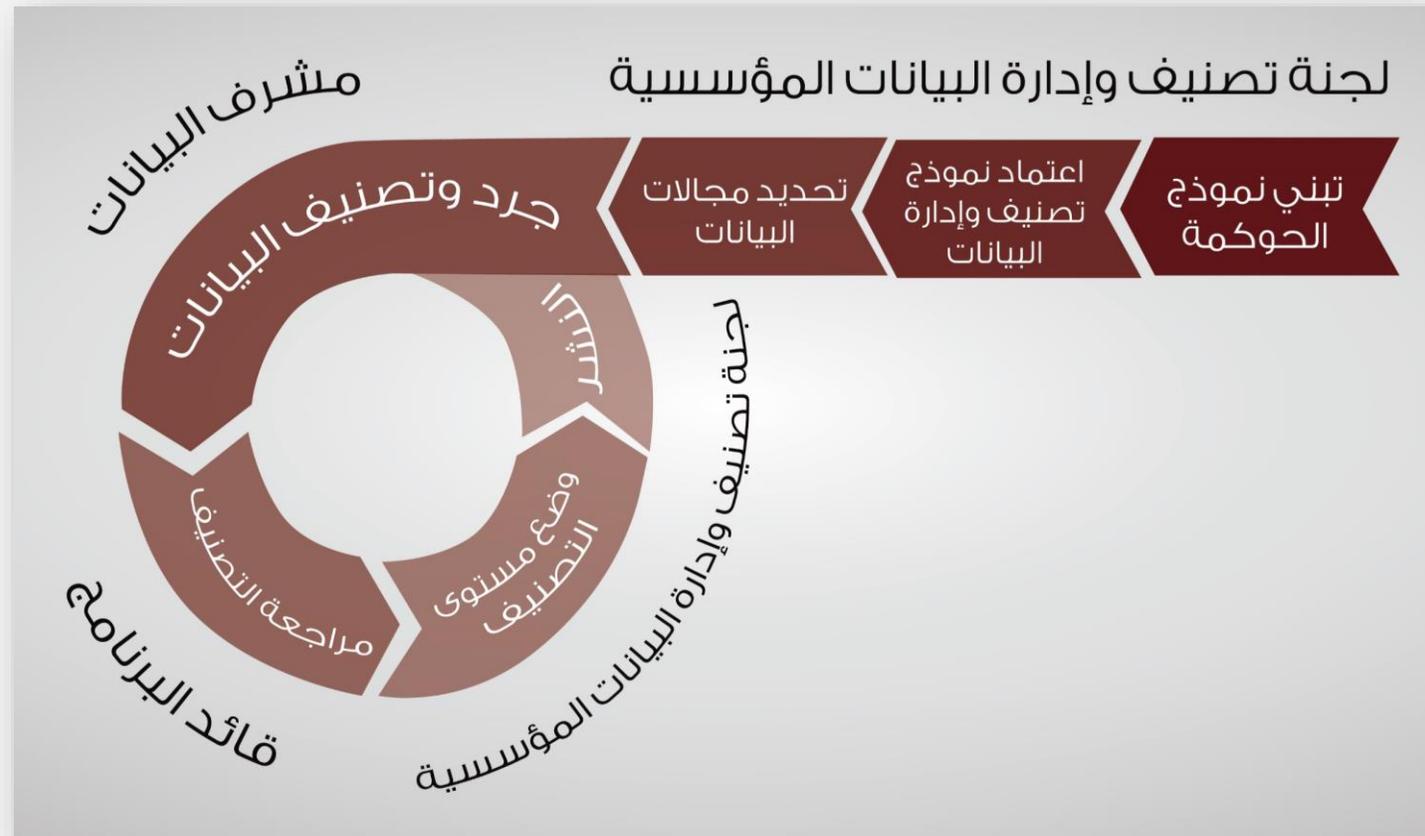
Politique de classification des données (2020):

[https://gdc.gov.jo/GDC\\_Folders/InformationCenter?AttachmentType=1](https://gdc.gov.jo/GDC_Folders/InformationCenter?AttachmentType=1)

Une plateforme a été développée au niveau national : <https://gdc.gov.jo>

L'inventaire de données publiques se fait dans le cadre de la classification des données publiques.

Par défaut, toutes les données sont classées au niveau « Public » (il existe 4 niveaux) !



# L'inventaire des données

## Cas de la Jordanie

تستعرض هذه القائمة كافة البيانات المتوفرة لدى الجهات الحكومية والمصنفة بمستوى عادي.

اسم الجهة

اسم المؤسسة

#	اسم المجموعة باللغة العربية	اسم المجموعة باللغة الإنجليزية	اسم الجهة	نوع ملف البيانات
1	منهجيات شؤون الموظفين	methodologies for employees	وزارة الخارجية وشؤون المغتربين	وثائق وملفات
2	منهجيات تتعلق بالعمليات والخدمات	Methodologies for operations and services	وزارة الخارجية وشؤون المغتربين	وثائق وملفات
3	منهجيات تتعلق بالأمور المالية	methodologies for finance	وزارة الخارجية وشؤون المغتربين	وثائق وملفات
4	مجموعة القوانين والانظمة	set of laws and regulations	صندوق تسليف النفقة	وثائق وملفات
5	متابعة الاستراتيجية الوطنية للاعوام(2021-2023)	The strategy 2021-2023	هيئة تنظيم النقل /هيئة تنظيم الطيران المدني	وثائق وملفات
6	خطة الشراء لسنة 2021	Purchase plan for 2021	صندوق تسليف النفقة	وثائق وملفات
7	نموذج اقتراح تعديل وثيقة	250F-0203	هيئة تنظيم النقل /هيئة تنظيم الطيران المدني	وثائق وملفات
8	تقرير الصندوق الخاص بتطبيق معايير حقوق الانسان	Report of the Fund on the implementation of human rights standards	صندوق تسليف النفقة	وثائق وملفات
9	اعداد المستلفين من الصندوق	Number of borrowers from the fund	صندوق تسليف النفقة	وثائق وملفات
10	اعداد الطلبات المقدمة	number submitted requests	صندوق تسليف النفقة	وثائق وملفات

... 10987654321

Les données classées dans le niveau « public » une fois validées, leurs métadonnées seront publiées et accessibles à tous.

Les rôles dans chaque organisation publique :

- Un Comité de pilotage pour la classification et la gestion des données,
- Un responsable du programme OGD,
- Des coordinateurs de données,

# L'inventaire des données

## Cas de la Jordanie

### أداة تحديد مستوى التصنيف المساندة



تهدف أداة تحديد مستوى التصنيف المساندة الى مساعدة الجهات الحكومية على تحديد مستوى التصنيف المناسب حسب المستويات الأربعة المعتمدة في نظام التصنيف الوارد في سياسة تصنيف وإدارة البيانات الحكومية 2020.

لذلك قامت الوزارة بإعداد أداة مساندة تعتمد على طرح مجموعة من الأسئلة ليتمكن مشرفي البيانات وقادة البرامج في الجهات الحكومية من تحديد مستوى التصنيف الصحيح بكل يسر وسهولة.

تتضمن الأداة مجموعة من الأسئلة التي يجب أخذها بعين الاعتبار عند تحديد مستوى التصنيف بناء على الأثر / للضرر المترتب عن الإفصاح أو الاختراق الأمني أو الوصول الغير المصرح به للبيانات / للمعلومات، حيث يتم تصنيف المعلومات أو البيانات وفقاً لأعلى مستوى تصنيف وارد بعد الإجابة على الأسئلة

هل تتضمن بيانات / معلومات يؤدي الإفصاح أو الاختراق الأمني أو الوصول الغير المصرح به إلى إلحاق الضرر بالدفاع الوطني أو أمن الدولة، وسياستها الخارجية؟

نعم

لا

Un outil en ligne accessible à tous pour la détermination du niveau de classification vise à aider les agences gouvernementales à déterminer le niveau de classification le plus approprié pour un jeu de données.

L'outil comprend un ensemble de questions qui doivent être prises en considération lors de la détermination du niveau de classification en fonction de l'impact / des dommages résultant de la divulgation / de l'atteinte à la sécurité / d'un accès non autorisé aux données, etc...

# L'inventaire des données

## Cas de la Tunisie

Au niveau national : une plateforme [PDIMS](#)

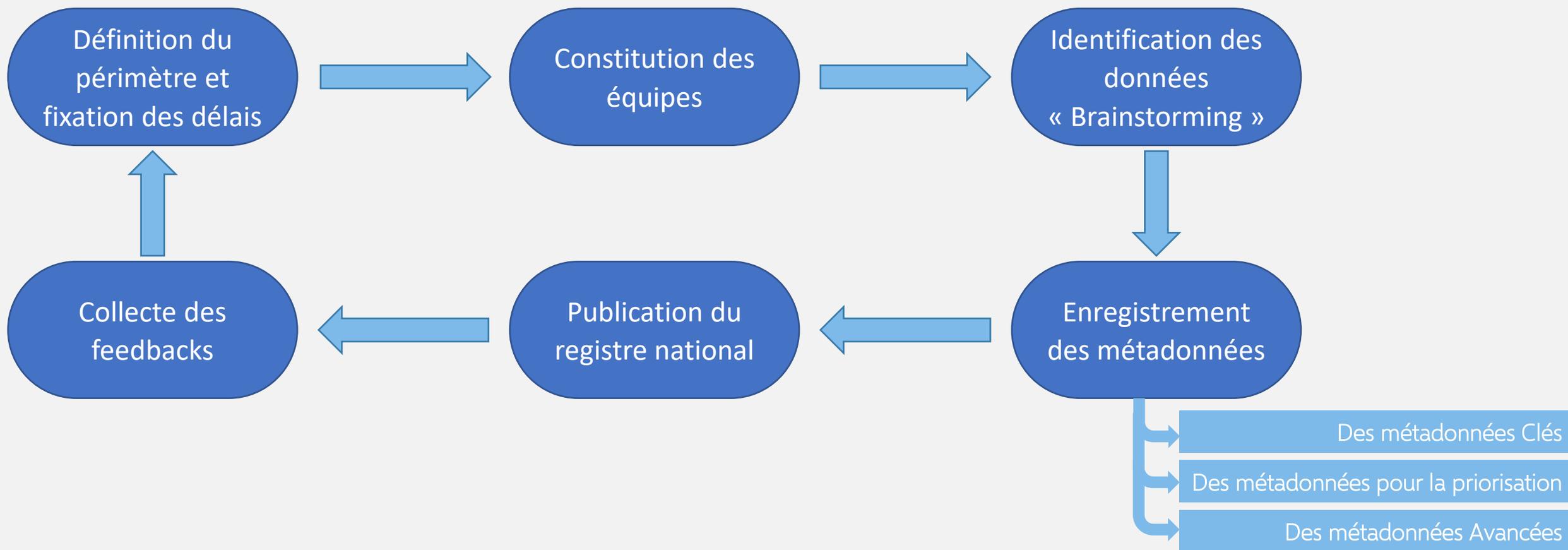
L'inventaire de données publiques est une obligation selon la « Décret Gouvernemental n° 2021-3 du 6 janvier 2021, relatif aux données publiques ouvertes »

Chaque organisation est tenue de faire son inventaire de données,

L'inventaire de données doit être publié sous format ouvert sur la plateforme nationale Open Data.

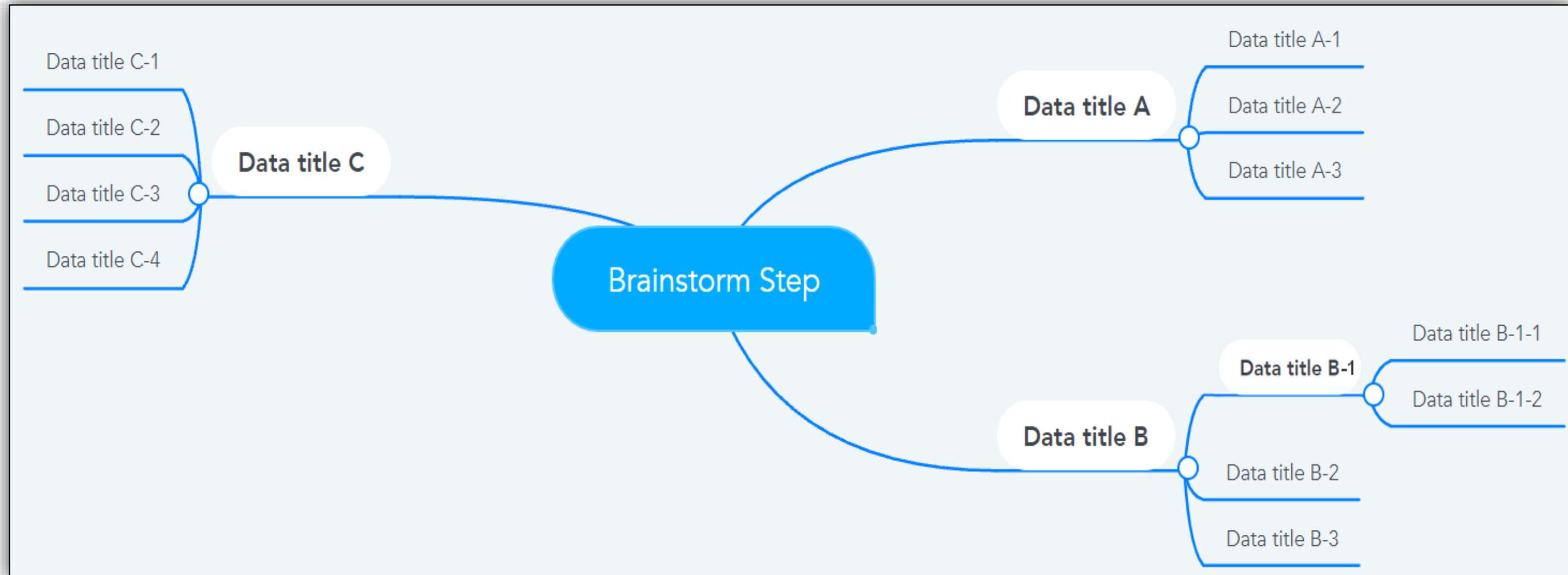
# L'inventaire des données

## Cas de la Tunisie



# L'inventaire des données

## Cas de la Tunisie





Shared Prosperity **Dignified Life**



# La classification des données

# La classification des données

## Loi N°05-20 du 25 Juillet 2020 :

Une loi qui définit la classification des données en fonction de l'impact potentiel des incidents de cybersécurité.

ART. 13. – Les actifs informationnels de type « données » classifiés sur la base de l'échelle de l'analyse des impacts prévue à l'article 11 ci-dessus sont classés suivant leur degré de sensibilité en termes de confidentialité selon les niveaux ci-après :

- « TRES SECRET », si un incident de cybersécurité portant sur la confidentialité a un impact très grave ;
- « SECRET », si un incident de cybersécurité portant sur la confidentialité a un impact grave ;
- « CONFIDENTIEL », si un incident de cybersécurité portant sur la confidentialité a un impact modéré ;
- « DIFFUSION RESTREINTE », si un incident de cybersécurité portant sur la confidentialité a un impact limité.

# La classification des données

## **La loi n°31-13 du 22 Février 2018, relative au droit d'accès à l'information :**

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 :**

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.



Shared Prosperity **Dignified Life**



# Les référentiels des données

# Lier les données ouvertes à un référentiel

## Pourquoi un référentiel de données ?

En tant que ROD d'une commune, je souhaite publier un jeu de données contenant la liste des entreprises ayant participé à un appel d'offres pour la réalisation d'un projet spécifique dans ma région.

→ Il est intéressant d'ajouter un champ correspondant à l'identification des sociétés.

Cet identifiant existe, il s'agit de l'ICE « **Identifiant Commun de l'entreprise** ».

Mais, il y a d'autres identifiants légaux : le numéro de registre de commerce, l'identifiant fiscal et le numéro de CNSS !

**→ Ajouter un identifiant en lui-même n'est pas suffisant → il est important de spécifier le référentiel utilisé !**

**→ Il est important d'intégrer dans un jeu de données des données pivots relevant d'un référentiel (si elles existent).**

# Lier les données ouvertes à un référentiel

## Avantages :

- ☐ Une meilleure **formalisation**;
- ☐ Une meilleure **synthèse**;
- ☐ Une meilleure **compréhension**;
- ☐ Une meilleure **réutilisation**;
- ☐ Une meilleure **interopérabilité**.

<https://guides.etalab.gouv.fr/qualite/lier-les-donnees-a-un-referentiel/#avantages>

# Pour quoi un référentiel de données

## Pourquoi des référentiels de données ?

- ❑ Les systèmes d'information des acteurs publics sont construits le plus souvent de manière indépendante les uns des autres (sans concertation/ en silo),
- ❑ Les acteurs publics utilisent souvent leur propre système de codification pour décrire leurs données,
- ❑ Les structures techniques des données sont souvent différentes,
  
- ➔ Perte de ressources, de temps et d'énergie pour harmoniser et combiner les jeux de données les uns aux autres.
- ➔ Freine considérablement la réutilisation des jeux de données (surtout l'automatisation de la réutilisation).
- ➔ Nécessite la mise en place des tables de correspondances (difficile à synchroniser)

# Pour quoi un référentiel de données

## Principaux enjeux !

- Identification des données de référence :
  - ✓ Identifier de façon unique, pérenne et partagée les objets (c.-a.-d : les données de référence),
  - ✓ Identifier l'unique point de vérité pour toutes les données de référence (l'organisme responsable de la donnée),
  
- Faciliter le partage et l'interopérabilité :
  - ✓ S'appuyer sur les standards,
  - ✓ Modéliser et formaliser,
  - ✓ Mettre en place un système d'accès aux données de références,
  
- Faciliter la consolidation et le suivi dans le temps :
  - Mettre en place un système de mise à jour,
  - Prévoir les mécanismes de gestion des erreurs ou de mauvaise qualité,

# Exemple de référentiels

- ***Le référentiel national des arrêts***

Identifier de façon unique, pérenne et partagée les arrêts de transport en commun

- ***Le répertoire national des sociétés, associations, etc...***

Lister toutes les informations relatives aux sociétés et associations et les codes d'identification

- ***Les référentiels des codes géographiques***

- Garantir une bonne superposition des couches géographiques,
- Echanger des données géographiques avec des acteurs internes et externes :

# Le Cas de La France

## LOI no 2016-1321 du 7 octobre 2016 pour une République numérique

### Article 14

I. – Le chapitre I<sup>er</sup> du titre II du livre III du même code est complété par un article L. 321-4 ainsi rédigé :

« Art. L. 321-4. – I. – La mise à disposition des données de référence en vue de faciliter leur réutilisation constitue une mission de service public relevant de l'Etat. Toutes les administrations mentionnées au premier alinéa de l'article L. 300-2 concourent à cette mission.

« II. – Sont des données de référence les informations publiques mentionnées à l'article L. 321-1 qui satisfont aux conditions suivantes :

« 1<sup>o</sup> Elles constituent une référence commune pour nommer ou identifier des produits, des services, des territoires ou des personnes ;

« 2<sup>o</sup> Elles sont réutilisées fréquemment par des personnes publiques ou privées autres que l'administration qui les détient ;

« 3<sup>o</sup> Leur réutilisation nécessite qu'elles soient mises à disposition avec un niveau élevé de qualité.

« III. – Un décret en Conseil d'Etat précise les modalités de participation et de coordination des différentes administrations. Il fixe les critères de qualité que doit respecter la mise à disposition des données de référence. Il dresse la liste des données de référence et désigne les administrations responsables de leur production et de leur mise à disposition. »

II. – Le I du présent article entre en vigueur à la date de publication du décret mentionné au III de l'article L. 321-4 du code des relations entre le public et l'administration et, au plus tard, six mois après la promulgation de la présente loi.

# Le Cas de La France

Le service public de la donnée (SPD) en France a publié 9 jeux de données de référence maintenus de manière pérenne par des administrations sur la plateforme open data nationale.

Sur la plateforme Open Data « data.gouv.fr » il existe plus de 2000 jeux de données proposant des référentiels de données.

Une grande partie de ces référentiels est proposée par des acteurs non gouvernementaux.

## Les données de référence

À ce jour, neuf jeux de données, qui couvrent un large champ thématique ont été identifiés comme des données de référence.

 <b>Base Adresse Nationale (BAN)</b> La Base Adresse Nationale est une base de données qui a pour but de référencer l'intégralité des adresses du territoire 1 33 40 France	 <b>Base Sirene des entreprises et de leurs établissements (SIREN, SIRET)</b> Pour vous abonner à notre lettre d'information Sirene open data actualités, suivez ce lien : 15 7 27 France	 <b>Code Officiel Géographique (COG)</b> Chaque année, l'Insee met à disposition sur son site (insee.fr) le code officiel géographique qui rassemble les codes et 35 13 22 France Commune française
 <b>Plan Cadastral Informatisé (PCI)</b> Plan Cadastral Informatisé PCI Ce jeu de données provient d'un service public certifié Données de référence Le plan cadastral est 3 9 57 France	 <b>Registre parcellaire graphique (RPG) : contours des parcelles et îlots cultureux et leur groupe de cultures majoritaire</b> Le registre parcellaire graphique est une base de données géographiques servant de référence à l'instruction des aides de la 20 2 12 2013-2017 France	 <b>Référentiel de l'organisation administrative de l'Etat</b> Le décret du 14 mars 2017 a institué le Service Public de la Donnée. Celui-ci met à la disposition du public 9 jeux de données 5 4 16 2016-2026 France
 <b>Référentiel à grande échelle (RGE)</b> L'État a confié à l'IGN le développement du référentiel à grande échelle (RGE). Pour ce faire, il fait appel à ses moyens propres 12 2 8 France	 <b>Répertoire National des Associations (RNA)</b> Que contient le RNA ? Le RNA répertorie l'ensemble des associations relevant de la loi du 1er juillet 1901 relative au contrat 19 8 28 1901-2017 France Point	 <b>Répertoire Opérationnel des Métiers et des Emplois (ROME)</b> Dans un contexte marqué par de fortes mutations de l'environnement économique et social, le ROME (Répertoire Opérationnel) 5 2 9

# Référentiel de données : Méthodologie

Les étapes ci-dessous représentent une démarche à suivre pour la mise en place d'un référentiel de données :

**Prérequis** : Identifier la donnée de référence ? Une donnée de référence est principalement caractérisée par la stabilité de sa définition et par un fort accès consultatif.

- Mettre en place le mode de gouvernance (pilotage, services métiers à impliquer, etc...);
- Définir le périmètre & la sémantique ;
- Définir le mode d'alimentation, d'actualisation et d'amélioration de la qualité;
- Publier et ouvrir;

# Référentiel de données : Méthodologie

## I- Mettre en place le mode de gouvernance :

Préciser le mode de pilotage;

Identifier le groupe des experts (INSEA, IMANOR, Equipe de l'open data, etc...);

Identifier la communauté (acteurs externes et internes) producteurs et consommateurs de la donnée;

Assurer la planification et le suivi des travaux ;

Communiquer et diffuser;

→ Une approche itérative de mise en place est fortement recommandée

# Référentiel de données : Méthodologie

## 2- Définir le périmètre & Formaliser la sémantique :

**Définir le périmètre** : On s'intéresse ici au contenant – la sémantique de la donnée de référence – Définir pour comprendre et se faire comprendre – Le dictionnaire de données (schéma de données) – Préciser l'identifiant à utiliser

**Ex** : Les adresses des administrations publiques ?

**Définir la couverture** : Il est question ici du contenu – On s'intéresse aux données , aux instances de l'objet métier (le concept) du référentiel

**Définir la portée** : pour quel usage le référentiel est il prévu ? On cherche à identifier les acteurs pour lesquels le référentiel peut être utilisé.

# Référentiel de données : Méthodologie

## 2- Définir le périmètre & Formaliser la sémantique :

Pour le choix d'un identifiant → les URLs

<https://{domain}/{type}/{concept}/{reference}>

Ex : <https://reference.data.gov.ma/id/company/000015465321>

**{domain}** est une combinaison du nom de l'hôte et du secteur concerné.

**{type}** devrait être l'un d'une petite série de valeurs possibles qui déclarent le type de ressource identifié. Les exemples typiques incluent: - 'id' ou 'item' pour les objets du monde réel; - 'doc' pour les documents qui décrivent ces objets; - 'def' pour des concepts; - 'set' pour les jeux de données; ect...

**{concept}** pourrait être une collection, le type d'objet du monde réel ou le nom du schéma conceptuel;

**{reference}** est un élément ou bien concept spécifique.

# Référentiel de données : Méthodologie

## 3- Définir le mode d'alimentation, d'actualisation et d'amélioration de la qualité :

Mettre en place les mécanismes de création et de mise à jour des données (sous forme de services SAO, API Rest, IHM),

Des systèmes d'alimentation collaboratifs pour le cas des référentiels transverses,

Définir les indicateurs de qualité et leur mode de calcul,

L'initialisation et le premier chargement,

La documentation et la formalisation,

Les mécanismes d'ajustement et de collecte des feedbacks

# Référentiel de données : Méthodologie

## 4- Publier & Ouvrir le référentiel de données sous forme d'un jeu de données

Définir les métadonnées descriptives et structurelles,

Joindre la documentation nécessaire,

Associer une licence,

Insérer le jeu de données dans la catégorie de « référentiels de donnée »



Shared Prosperity **Dignified Life**



# Discussion