



Royaume du Maroc
Chef du Gouvernement

Ministère de la Transition
Numérique et de la Réforme
de l'Administration

#ADD

وكالة التنمية الرقمية

ⵜⴰⵎⴰⵔⵜ ⵜⴰⵏⵓⵎⴰⵔⵜ ⵜⴰⵔⵉⵎⴰⵏⵜ ⵜⴰⵔⵉⵎⴰⵏⵜ

Agence de Développement du Digital

Atelier de formation sur l'Open Data au profit des organismes et institutions publics relevant de la Région Fès-Meknès

Data.gov.ma

Région Fès-Meknès

23/01/2025

Session de formation sur l'Open Data

Séance 3: Standards et plateformes techniques

11h40-12h30

Séance 2:

Plateformes et standards techniques

Section 3.1:

Plateformes techniques Open Data



Session 3: Standards et Plateformes techniques

Recommandations

- ❑ Disposer des moyens techniques et humains pour mettre en place et maintenir une plateforme de publication de données
- ❑ Utiliser une plateforme à l'échelle nationale
- ❑ Potentiellement interfacier la plateforme nationale aux différentes plateformes régionales
- ❑ Utiliser une solution open source
- ❑ Possibilité de création de catalogues, ensembles de données, et métadonnées associées
- ❑ La plateforme doit permettre la participation et l'interaction avec les utilisateurs et/ou visiteurs :
 - Demandes d'ajout de nouvelles données
 - Pouvoir signaler un jeu de données
 - Les commentaires, les discussions
 - Les cas de réutilisation de données
 - Les suggestions d'amélioration

Session 3: Standards et Plateformes techniques

Les accès

les APIs sont un levier essentiel pour maximiser l'utilisation et l'impact des données ouvertes

Accès simple et automatisé

Permet aux utilisateurs d'accéder facilement et automatiquement aux données, sans nécessiter de téléchargement manuel

Mise à jour en temps réel

Les utilisateurs peuvent accéder aux données les plus récentes, sans attendre des publications manuelles

Interopérabilité




Elles facilitent l'intégration des données ouvertes avec d'autres systèmes, applications ou plateformes

Encouragement à l'innovation

Permet aux développeurs de créer des applications, des visualisations ou des services innovants en exploitant les données ouvertes




Session 3: Standards et Plateformes techniques

Comparatif des plateformes de publication

Plateforme	 Description	 Caractéristiques clés	 Cas d'usage
CKAN	<ul style="list-style-type: none">Le portail de données open source le plus célèbre est CKAN. Développé à l'origine par l'Open Knowledge Foundation et est utilisé par les gouvernements américain, britannique et espagnol, entre autres	<ul style="list-style-type: none">Solution matureNombreux modules et extensionsAPI puissanteGrande communauté de développeurs	Catalogues nationaux et projets à grande échelle
DKAN	<ul style="list-style-type: none">Basée sur Drupal (un système de gestion de contenu écrit en PHP)Axée sur la personnalisation et la visualisation	<ul style="list-style-type: none">User friendlyVisualisations intéressantes disponiblesIntégration native avec Drupal	Petites et moyennes administrations locales

Session 3: Standards et Plateformes techniques

Comparatif des plateformes de publication (suite)

Plateforme	 Description	 Caractéristiques clés	 Cas d'usage
uData	<ul style="list-style-type: none">• Développée par Etalab, spécialisée pour les besoins des gouvernements• Développée en python• Peut être installée sur un serveur linux ou MacOS	<ul style="list-style-type: none">• Conçue pour les données publiques• Interface moderne	Portails gouvernementaux nationaux ou locaux
GeoNode	<ul style="list-style-type: none">• Spécialisée pour la gestion de données géospatiales	<ul style="list-style-type: none">• Gestion avancée de données SIG• Cartographie et analyse géographique intégrées• Peut être intégrée dans une autre plateforme existante	Projets nécessitant une gestion de données géospatiales (urbanisme, environnement)

Session 3: Standards et Plateformes techniques

Différences clés



- **CKAN** : Meilleur choix pour un catalogue généraliste robuste avec une grande communauté de maintenance



- **DKAN** : Idéal pour les projets nécessitant une interface conviviale et la personnalisation avec Drupal. Cela peut également avoir ses avantages lorsque Drupal est également utilisé pour le site Web du gouvernement local ou régional



- **uData** : Axé sur les gouvernements, bon choix pour les initiatives de données publiques nationales



- **GeoNode** : Incontournable pour les projets SIG nécessitant des outils cartographiques avancés.

Session 3: Standards et Plateformes techniques

Architecture CKAN

Trois types d'installation

- Installation de CKAN à partir du package
- Installation de CKAN à partir des sources
- Installation de CKAN avec Docker Compose

Pour un portail au niveau d'une ville avec peu de trafic :

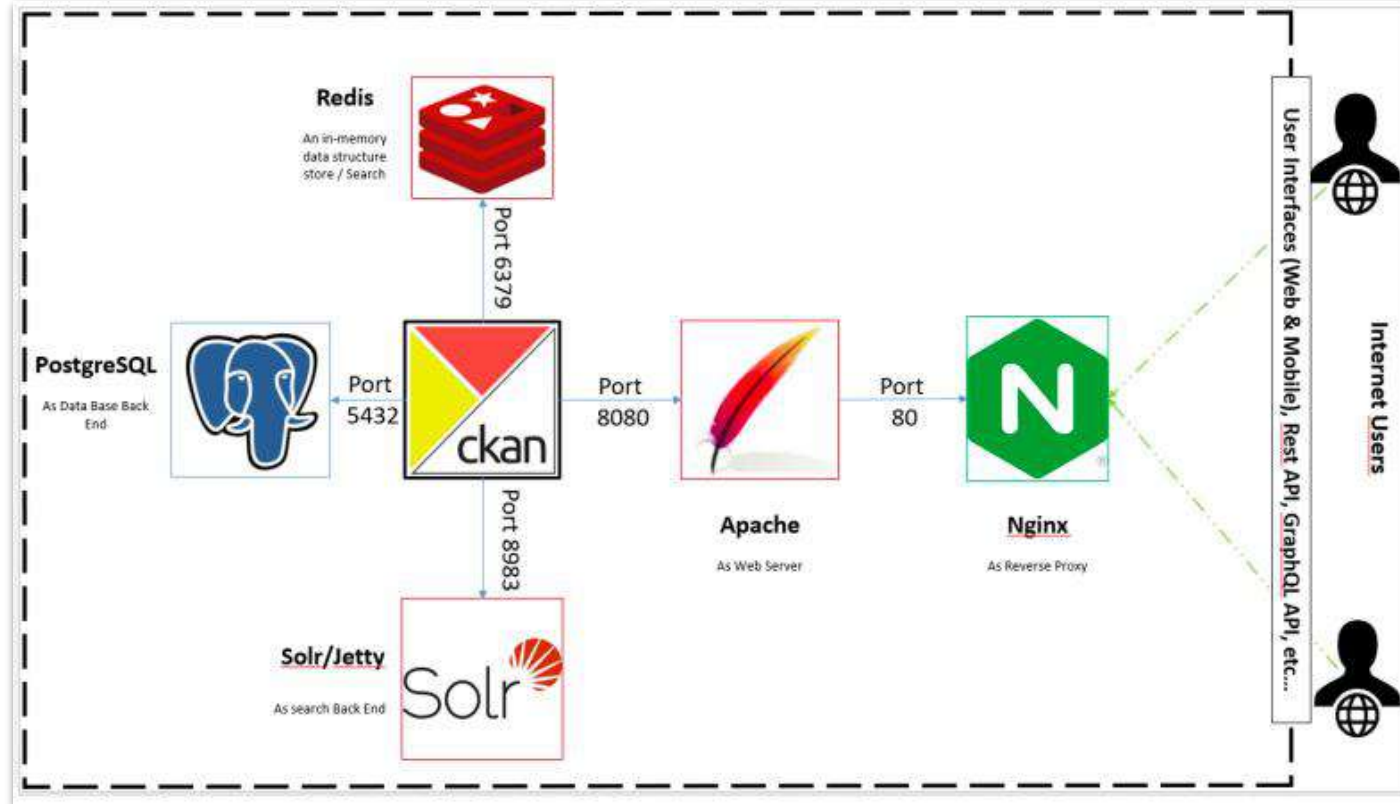
- 2 * Serveurs avec 2 Go de RAM (Web et DB/solr)
- Disque dur de 80 Go sur les deux.
- Processeurs double cœur

Pour un grand portail national à fort trafic :

- 2 * serveurs avec 8 Go de RAM (un pour le Web et un pour la base de données/solr)
- Disque dur de 160 Go sur les deux.
- Processeurs quad core (plus rapide)

Session 3: Standards et Plateformes techniques

Architecture CKAN



Session 3: Standards et Plateformes techniques

Dépendances CKAN (from package)

Service	Port	Used for
NGINX	80	Proxy
uWSGI	8080	Web Server
uWSGI	8800	DataPusher
Solr/Jetty	8983	Search
PostgreSQL	5432	Database
Redis	6379	Search

Session 3: Standards et Plateformes techniques

Dépendances CKAN (from package)

Package	Description
Python	The Python programming language, v3.6 or newer (or v2.7)
PostgreSQL	The PostgreSQL database system, v9.5 or newer
libpq	The C programmer's interface to PostgreSQL
pip	A tool for installing and managing Python packages
python3-venv	The Python3 virtual environment builder (or for Python 2 use 'virtualenv' instead)
Git	A distributed version control system
Apache Solr	A search platform
Jetty	An HTTP server (used for Solr).
OpenJDK JDK	The Java Development Kit (used by Jetty)
Redis	An in-memory data structure store

Session 3: Standards et Plateformes techniques

Dépendances CKAN (from package)

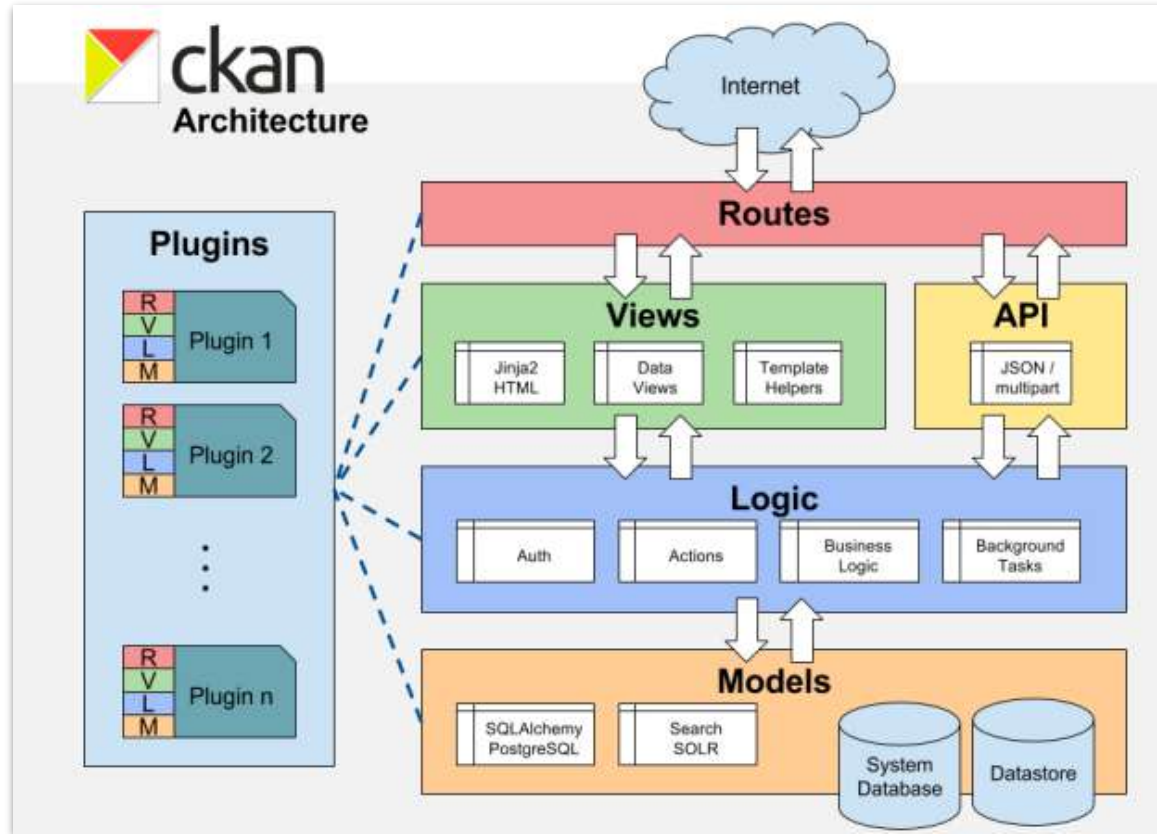
Test system	Prod (minimal config)	Prod (recommended config)	
CPU	1 CPU / core	4 CPU / core	4-8 CPU / core
RAM	4 GB	4 GB	8 GB
Hard disk	20 GB	20 GB	20-40 GB

Configuration par VM

- 1 CPU with 4 cores
- RAM: 4GB
- Disk space: 20GB

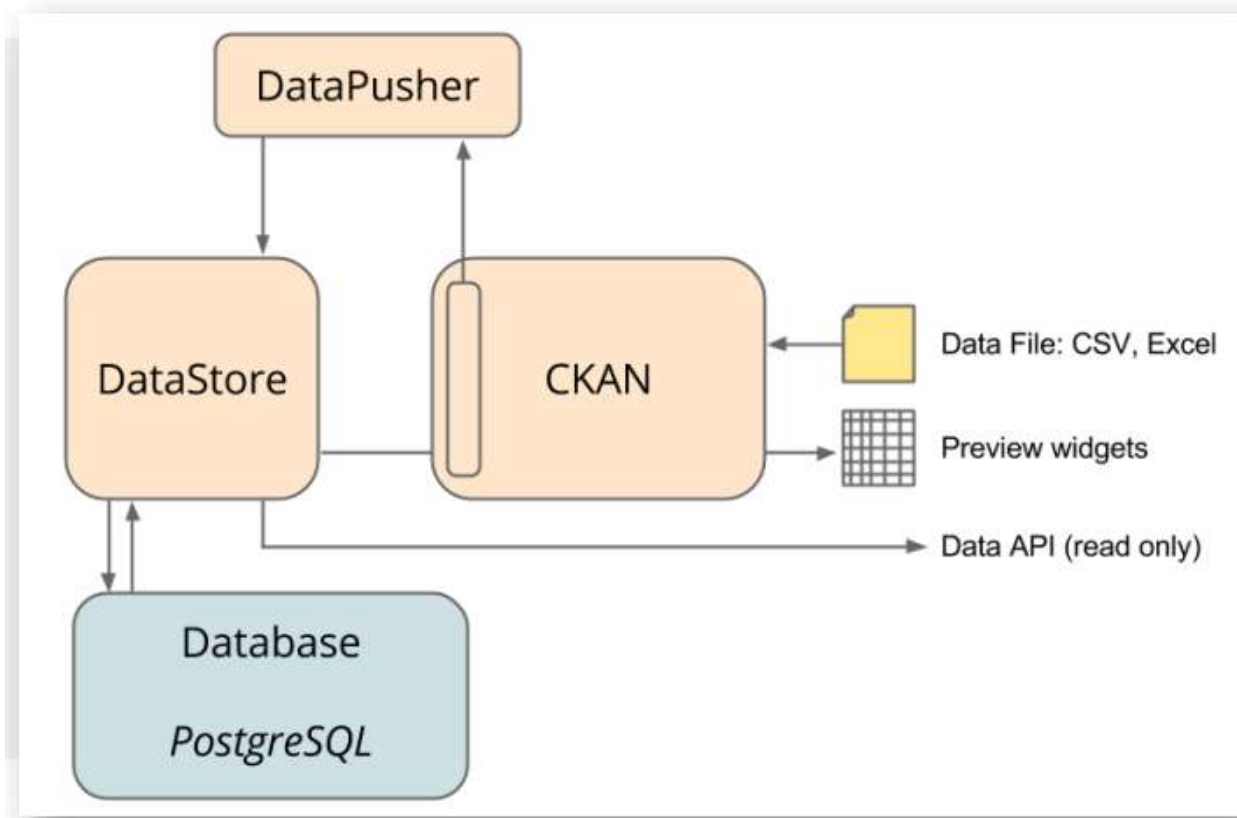
Session 3: Standards et Plateformes techniques

Architecture CKAN



Session 3: Standards et Plateformes techniques

Architecture CKAN



Session 3: Standards et Plateformes techniques

Extensions CKAN

DataStore + Data Pusher / xloader

L'extension CKAN DataStore fournit une base de données ad hoc pour le stockage de données structurées à partir de distributions d'un jeu de données CKAN.

Les données peuvent être extraites des fichiers et stockées dans le DataStore.

Lorsqu'une distribution est ajoutée au DataStore, on obtient :

- Aperçu automatique des données sur la page de la distribution, à l'aide de l'extension Data Explorer
- API Data : permettant de chercher, filtrer et mettre à jour les données, sans télécharger et uploader l'intégralité du fichier de données

Le DataStore est intégré à l'API CKAN et au système d'autorisation.

Le DataStore est généralement utilisé avec le DataPusher, qui télécharge automatiquement les données vers le DataStore à partir de fichiers appropriés, qu'ils soient téléchargés via le FileStore de CKAN ou via des liens

Session 3: Standards et Plateformes techniques

Architecture CKAN

<https://github.com/open-data/ckanext-scheming>

Scheming est une extension permettant de personnaliser des schémas des jeux de données, des groupes et d'organisation dans CKAN.

L'extension des schémas se fait à l'aide de fichiers JSON ou YAML qui incluent des règles de validation, la modification d'extraits de modèle de formulaire et des extraits d'affichage.

Session 3: Standards et Plateformes techniques

Extensions CKAN

<https://github.com/open-data/ckanext-fluent>

Fluent est une extension permettant d'ajouter du texte et des balises multilingues aux jeux de données, aux groupes et organisations de CKAN.

Fluent fonctionne avec Scheming en fournissant des validateurs et des extraits de code pouvant être utilisés dans les schémas personnalisés.

Fluent affiche des zones d'édition pour toutes les langues pour chaque champ dans le formulaire d'édition, mais par défaut n'affiche que la valeur de la langue de l'utilisateur lors de l'affichage d'un jeu de données.

Toutes les valeurs multilingues sont renvoyées et créées/mises à jour via l'API à l'aide d'un objet JSON.

Par exemple : un champ Fluent "label" peut avoir la valeur : ... "label": { "ar": "كتب", "fr": "Livres"}, ...

Session 3: Standards et Plateformes techniques

Extensions CKAN

Les autres extensions

webpage_view / pdf_view / officedocs_view / basic-charts

Spatial_metadata / spatial_query navigablemap / geojson_view

Harvest / dcat

Pages / googleanalytics / showcase / datarequests / disqus / rating / contact

Session 3: Standards et Plateformes techniques

Publication sur la plateforme

Trois scénarios possibles pour la publication d'un jeu de données

SCÉNARIO 1

Généralement, dans le cas de :

- Faible nombre de jeux de données avec une faible fréquence de mise à jour,
- Si le producteur des données ne dispose pas de systèmes d'information
- Si l'environnement ne favorise pas l'extraction et la mise à jour automatique

Les envoyer à l'ADD par mail pour que l'ADD les publie sur l'espace du producteur concerné

Session 3: Standards et Plateformes techniques

Publication sur la plateforme

Trois scénarios possibles pour la publication d'un jeu de données

SCÉNARIO 2

- Plus de 10 jeux de données avec une fréquence de mise à jour régulière

Le producteur peut obtenir son propre compte public directement ses jeux de données sur son espace en utilisant le login et le mot de passe que l'ADD lui communique.

Session 3: Standards et Plateformes techniques

Publication sur la plateforme

Trois scénarios possibles pour la publication d'un jeu de données

SCÉNARIO 3

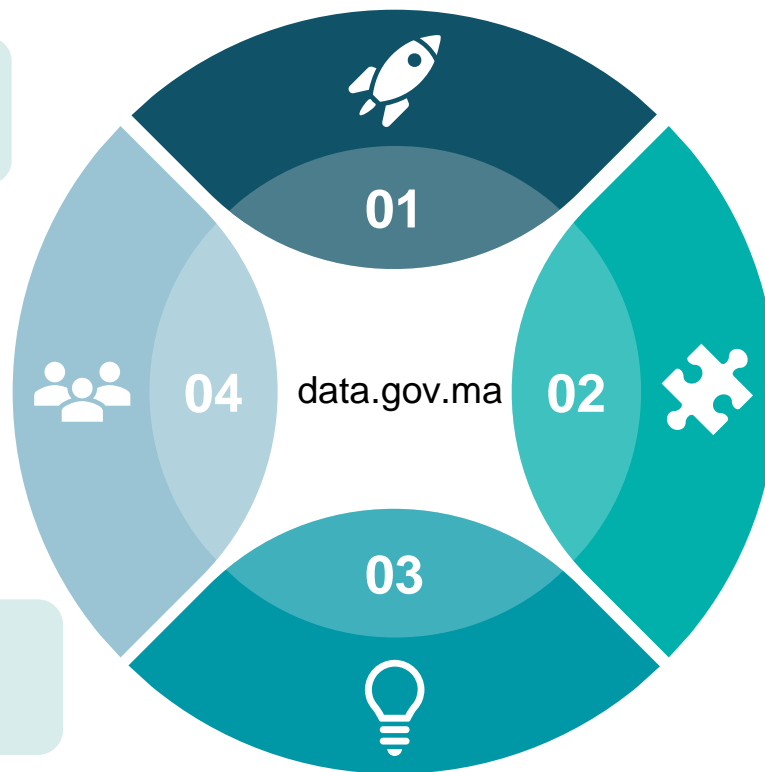
Dans le cas où le nombre de jeux de données est important avec une actualisation fréquente, et une maturité du système d'information et des mécanismes d'automatisation

Activer les APIs entre le SI du producteur et le portail Open Data pour permettre la mise à jour automatique des jeux de données

Session 3: Standards et Plateformes techniques

Cas du Royaume du Maroc

Plateforme technique :
CKAN 2.9.3 + Drupal 9 en
évolution continue par l'ADD



Mises à jour des données
en automatique qui se base
sur l'exploitation des APIs
(d'écriture)

Mises à jour des données en
manuel avec insertion des
données directement

Fonctionnalité de demande
d'ajout de nouveaux jeux de
données

Session 3: Standards et Plateformes techniques

Cas du Royaume du Maroc

Se connecter - Portail Open Data

data.gov.ma/data/fr/user/login

Etape 1 : Cliquer sur "Espace producteur" pour se connecter

Espace Producteurs العربية Donner votre avis

DONNÉES THÈMES PRODUCTEURS DOCUMENTATION ACTUALITÉS STATISTIQUES À PROPOS

Se connecter

Rénitialiser votre mot de passe.

Utilisez notre formulaire de régénération du mot de passe pour le réinitialiser.

Mot de passe oublié ?

Se connecter

Nom de l'utilisateur:

Mot de passe:

Se souvenir de moi

Se connecter

Session 3: Standards et Plateformes techniques

Cas du Royaume du Maroc

The screenshot shows a web browser window with the URL `data.gov.ma/data/fr/organization/agence-de-developpement-du-digital`. The page header includes the Data.gov.ma logo and navigation links: `Tableau de bord`, `العربية`, `Donnez votre avis`, `DONNÉES`, `THÈMES`, `PRODUCTEURS`, `DOCUMENTATION`, `ACTUALITÉS`, `STATISTIQUES`, and `À PROPOS`.

The main content area is titled `Organisations / ADD`. On the left, there is a sidebar for the ADD organization, featuring the logo `#ADD` and the text `Agence de Développement du Digital`. Below the logo, it says `ADD` and `Données ouvertes de l'Agence de Développement du Digital`. There is a `lire plus` link and a `Jeux de données` section showing `7` datasets and a `S'abonner` button.

The main content area shows a search bar with the text `Rechercher les jeux des données...` and a `Gérer` button. Below the search bar, there is a red annotation: `Etape 2 : Une fois connecté on a le bouton "Ajouter un jeu de données" qui s'affiche`. Below the search bar, it says `7 jeux de données trouvés` and `Par Ordre: Pertinence`.

Session 3: Standards et Plateformes techniques

Cas du Royaume du Maroc

Jeu de données / Créer le jeu de données

Etape 3 : On doit renseigner les métadonnées : Titre en FR et AR, et la description en FR et AR

Que sont les jeux de données ?

A CKAN Dataset is a collection of data resources (such as files), together with a description and other information, at a fixed URL. Datasets are what users see when searching for data.

1 Créer le jeu de données

2 Ajouter des données

Titre:

par ex. un titre explicite

* URL: <data.gov.ma/data/fr/dataset/<dataset>> Modifier

Title (ar):

par ex. un titre explicite

Description:

par ex. un commentaire utile au sujet de cette donnée

You can use Markdown formatting here

Description (ar):

par ex. un commentaire utile au sujet de cette donnée

You can use Markdown formatting here

Mots-clés:

Session 3: Standards et Plateformes techniques

Cas du Royaume du Maroc

**Etape 3 : Renseigner les mots clés (tags).
La licence et le nom de l'organisme (producteur) sont sélectionnés par défaut.
La visibilité permet de publier le jeu de donnée sur le portail ou de le garder en mode privé**

Mots-clés:

par.ox. économie, santé mentale, gouvernement

Licence:

Open Data Commons Open Database License (ODbL)

License definitions and additional information can be found at <https://opendatacommons.org>

Producteurs:

ADD

Visibilité:

Privé

Source:

http://example.com/dataset.json

Version:

1.0

Producteurs:

Jean Dupont

Courriel de l'auteur:

joe@example.com

Mainteneur:

Jean Dupont

Courriel du mainteneur:

joe@example.com

Session 3: Standards et Plateformes techniques

Cas du Royaume du Maroc

Créer le jeu de données - Porta

data.gov.ma/data/fr/dataset/new?group=9703ee9c-cc16-4a03-97bb-0ab1118e9a56

Mainteneur:

Courriel du mainteneur:

Champ personnalisé:

Clé:	<input type="text"/>	Valeur:	<input type="text"/>
------	----------------------	---------	----------------------

Champ personnalisé:

Clé:	<input type="text"/>	Valeur:	<input type="text"/>
------	----------------------	---------	----------------------

Champ personnalisé:

Clé:	<input type="text"/>	Valeur:	<input type="text"/>
------	----------------------	---------	----------------------

The data license you select above only applies to the contents of any resource files that you add to this dataset. By submitting this form, you agree to release the metadata values that you enter into the form under the Open Database License.

* Required field

Etape 4 : Cliquer sur suivant pour ajouter le fichier

Suivant : Ajouter des données

Session 3: Standards et Plateformes techniques

Cas du Royaume du Maroc

Organisations / ADD / testtt / Modifier / Ajouter une Nouvelle Ressource

Qu'est-ce qu'une ressource ?

Une ressource peut-être un fichier ou un lien vers un fichier contenant des données utiles.

1 Créer le jeu de données

2 Ajouter des données

Etape 5 : Sélectionner le fichier à partager et cliquer sur "Terminer"

Data:

Envoi Link

Nom:

par exemple : Prix de l'or en janvier 2011

Description:

Quelques notes utiles à propos des données

You can use Markdown formatting here

Format:

par exemple : CSV, XML ou JSON

This will be guessed automatically. Leave blank if you wish

Précédent Enregistrer & ajouter un autre **Terminer**

Session 3: Standards et Plateformes techniques

Cas du Royaume du Maroc

The screenshot shows a web browser window with the URL `data.gov.ma/data/fr/dataset/testttt`. The page title is "testttt" and the breadcrumb is "Organisations / ADD / testtt".

On the left sidebar, there is a "testttt" header, a "S'abonner" button, a "Producteur" section with the ADD logo (Agence de Développement du Digital), and social media links for Twitter and Facebook.

The main content area has a navigation bar with "Jeu de données", "Groupes", "Flux d'activité", "Reutilisations", and "Gérer". A red warning message reads: "Le jeu de données s'affiche sur l'espace producteur, il reste à l'affecter au thème correspondant".

Below the warning, there is a "Données et ressources" section with a file "BDD STARTUPS - ADD_2024.xlsx" and an "Explorer" button.

The "Info additionnelle" section contains a table with the following data:

Champ	Valeur
État	actif
Last Updated	22 janvier 2025, 21:47 (UTC+01:00)
Créé le	22 janvier 2025, 21:46 (UTC+01:00)

At the bottom, there is a "Laissez un commentaire" section with a "Contenu:" label and an empty text input field.

Session 3: Standards et Plateformes techniques

Cas du Royaume du Maroc

testtt - Jeu de données - Portail x +

data.gov.ma/data/fr/dataset/groups/testtt

Logo: **Data.gov.ma** (البيانات الحرة MOROCCO) | DONNÉES | THÈMES | PRODUCTEURS | DOCUMENTATION | ACTUALITÉS | STATISTIQUES | À PROPOS

Organisations ADD testtt

Sous l'onglet "Groupe" on affecte le thème au jeu de donnée

testtt

S'abonner

Producteur

#ADD
Agence de Développement du Digital

ADD
Données ouvertes de l'Agence de Développement du Digital.
lire plus

Social

Twitter

Jeu de données | **Groupe** | Flux d'activité | Reutilisations | Gérer


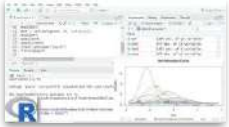
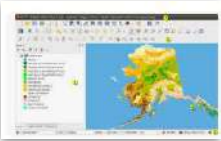
Agriculture

Add to group

There are no groups associated with this dataset

Session 3: Standards et Plateformes techniques

Outils d'analyse de données

Outil	Description	Caractéristiques clés
<p>Pandas (Python)</p> 	<ul style="list-style-type: none">• Nettoyage des données• Analyse des données tabulaires• Création de modèles et simulations	<ul style="list-style-type: none">• Très populaire dans l'écosystème machine learning et big data• Supporte une vaste gamme de graphiques, y compris les graphiques interactifs
<p>R</p> 	<ul style="list-style-type: none">• Statistiques et visualisation avancées• Création de modèles et simulations	<ul style="list-style-type: none">• Facilité d'utilisation• Peut gérer de grands volumes de données• Fortement adopté par les statisticiens et chercheurs en analyse de données
<p>QGIS</p> 	<ul style="list-style-type: none">• Analyse géospatiale pour les données SIG	<ul style="list-style-type: none">• Supporte de nombreux formats de fichiers et bases de données spatiales (Shapefile, GeoJSON, PostGIS, etc.)

Séance 2:

Plateformes et standards techniques

Section 3.2:

Standards techniques et méthodes d'anonymisation



Session 3: Standards et Plateformes techniques

Standards techniques: un pilier essentiel

Les standards techniques jouent un rôle clé pour garantir que les données ouvertes soient exploitables, fiables et réutilisables par tous les utilisateurs. Voici les raisons majeures qui justifient leur importance :

Interopérabilité



Permet à différents systèmes, plateformes et applications d'échanger et de traiter des données sans obstacles techniques.

Réutilisabilité



Les données conformes aux standards peuvent être manipulées et analysées à l'aide d'outils standardisés.

Qualité et Transparence



Les standards garantissent une présentation claire et une structuration fiable des données, augmentant leur crédibilité.

Efficacité



Évite de réinventer des processus pour chaque publication, ce qui diminue les coûts et efforts techniques.

Innovation et Développement de Nouveaux Services



Les standards techniques stimulent la créativité en facilitant la conception de nouvelles applications et services innovants.

Prévisibilité et Planification



Les données conformes aux standards permettent une analyse fiable des tendances pour la prise de décision stratégique.

Un guide des standards Open Data élaboré par l'ADD est fourni ici: [Lien vers le guide](#)

Session 3: Standards et Plateformes techniques

Critère 1: Formats de fichiers

Pour permettre au plus grand nombre d'utiliser un fichier de données, il est nécessaire que ce fichier soit dans un format documenté et public, aussi appelé « format ouvert », manipulable par des outils standard et ne nécessitant pas l'acquisition de logiciels spécifiques.

Pour les données tabulaires

CSV ✓

XLS ✓ ⓘ

Pour les images

PNG ✓

PSD ✗

JPEG ✓

TIFF ✗

Pour les images vectorielles

SVG ✓

AI ✗

Pour les API

OAS ✓

REST ✓

Pour les données géospatiales

Geojson ✓

DXF ✗

Geotiff ✓

GML ✓

KML ✓

WFS ✓

Voir également les
normes NM ISO 1913126
et ISO 19119

Pour les données structurées

XML ✓

AZW ✗

JSON ✓

RDF ✓

ODF ✓

EPUB ✓

Session 3: Standards et Plateformes techniques

Formats de fichiers courants



Simple, lisible par des humains et des machines, adapté aux grandes bases de données.

Exemple: Données démographiques sur la population de Fès (source : portail national des données ouvertes).

CSV (Comma-Separated Values)

Parfait pour les APIs et les applications web grâce à sa légèreté et son format hiérarchique.

Exemple : Infrastructures publiques à Guelmim publiées en JSON pour des applications mobiles.



JSON (JavaScript Object Notation)



Format spécifique pour les données géographiques, idéal pour les cartes interactives.

Exemple : Cartographie des ressources hydriques dans la région de Guelmim.

GeoJSON

Gère des structures complexes avec une bonne compatibilité inter-applications.

Exemple: Déclarations administratives numériques à Fès disponibles en XML.



XML (Extensible Markup Language)

Session 3: Standards et Plateformes techniques

Critère 2: Données tabulaires

Les données tabulaires doivent être exploitables informatiquement et une structure de données compatible avec les besoins de calcul et d'archivage (données longues vs. données larges)

Exploitable informatiquement:

- Permettre le traitement directe de la donnée, en opérant des filtres de lignes/colonnes, agrégats, ... Il est essentiel que les cellules ne soient pas fusionnées (cas de l'excel)
- Uniformité des formats de données

Tableau 2 : Données non exploitables informatiquement

Structure tabulaire adaptée

- Privilégier les structures par insertion de lignes (vertical) plutôt que l'ajout de colonnes (horizontal)

Country Name	Country Code	Indicator Name	Indicator Code	1990	1991	1992	1993	1994	1995
Aruba	ABW	Urban population SP.URB.TOTL		27526	28141	28532	28761	28924	29682
Afghanistan	AFG	Urban population SP.URB.TOTL		750836	796272	888385	885228	834335	886074
Angola	AGO	Urban population SP.URB.TOTL		509222	597288	628381	660180	691552	721552
Albania	ALB	Urban population SP.URB.TOTL		493982	513592	530766	547928	565248	582374
Andorra	AND	Urban population SP.URB.TOTL		7839	8768	9754	10811	11915	13067
Arab World	ARB	Urban population SP.URB.TOTL		28707177	30292822	31856717	33513046	35275337	36923
United Arab Emirates	ARE	Urban population SP.URB.TOTL		67927	71975	84367	95215	10611	6471
Argentina	ARG	Urban population SP.URB.TOTL		15676942	15449950	15815562	16163085	16552517	1723103
Armenia	ARM	Urban population SP.URB.TOTL		960956	1012430	1065431	1119586	1174560	1229980
American Samoa	ASM	Urban population SP.URB.TOTL		18324	13729	14254	14871	15522	16176
Antigua and Barbuda	ATG	Urban population SP.URB.TOTL		21466	21472	21458	21441	21449	21489

Department	Entity	Date	Expense T	Expense A	Supplier	Transactio	Amount
Departme	Departme	#####	CASH FUN	FINANCE	{ NOTTINGH	HAFS-796	45000000
Departme	Departme	#####	CASH FUN	FINANCE	{ NOTTINGH	HAFS-796	85000000
Departme	Departme	#####	CASH FUN	FINANCE	{ OLDHAM	HAFS-796	28000000
Departme	Departme	#####	CASH FUN	FINANCE	{ OXFORDS	HAFS-797	75000000
Departme	Departme	#####	CASH FUN	FINANCE	{ PETERBOF	HAFS-797	30000000
Departme	Departme	#####	CASH FUN	FINANCE	{ PLYMOUTH	HAFS-797	30000000
Departme	Departme	#####	CASH FUN	FINANCE	{ PORTSMO	HAFS-797	27000000
Departme	Departme	#####	CASH FUN	FINANCE	{ REDBRIDG	HAFS-797	31000000

Session 3: Standards et Plateformes techniques

Critère 3: Encodage des données

Parmi une multitude d'encodages possibles (ASCII, UTF), l'encodage défini au niveau international est UTF-8

Encodage ASCII / Latin-1 (ISO8895)

- 128 caractères dédiés à l'anglais ou spécifique aux langues latines
- Source de problèmes d'incompatibilité des langues

Œ sont les caractères accentués ?

Encodage UTF-8

- 2 millions d'entrées

Où sont les caractères accentués ?

Pour éviter ces problèmes, UTF-8 est l'encodage défini au niveau international qui permet de couvrir l'ensemble des caractères de tous les langages.

C'est également l'encodage recommandé dans le CGI.

Session 3: Standards et Plateformes techniques

Critère 4 : Types de données (valeurs numériques)

Certains types de données présentent un défi particulier car peuvent être présentées de façons différentes. Il est donc essentiel de standardiser les structures dont sont présentées certaines informations

Cas des valeurs numériques,

par exemple le nombre 1.5

- Dans les pays anglosaxons: **“1.5”**
- Dans les pays francophones: **“1,5”**

par exemple le nombre 2 500 000

- Dans les pays anglosaxons: **“2,500,00”**
- Dans les pays francophones: **“2 500 000”**

l'organisme de standardisation IEEE a défini une norme internationale (IEEE 754-201938) qui:

- impose le « . » comme séparateur de décimal
- interdit l'utilisation de séparateur de milliers

(par exemple « 1001.5 » respecte cette norme).

Session 3: Standards et Plateformes techniques

Critère 4 : Types de données (Dates)

Certains types de données présentent un défi particulier car peuvent être présentées de façons différentes. Il est donc essentiel de standardiser les structures dont sont présentées certaines informations

Cas des valeurs dates,

par exemple le 12 janvier 2005

- Dans les pays anglosaxons: **“01/12/2005”** ou **“01-12-2005”**
- Dans les pays francophones: **“12/01/2005”** ou **“12-01-2005”**

Plusieurs formats de dates longues existent

- Par exemple 12/05/2022 2:57pm , 2002-05-12 12:27:58 ou 2022-05-12T14:57:00Z
- Nécessité de tenir compte des fuseaux horaires

L'organisme de standardisation ISO a défini une norme (ISO 860139), norme également adoptée par l'IMANOR (NM ISO 860140) qui fixe un format de date universel

AAAA-MM-JJTHH:MM:SS(.sss)Z

exemple : 2022-05-12T14:57:00Z pour le 12 mai 2022 à 14:57:00 UTC).

Ce format est référencé dans le CGI.

Session 3: Standards et Plateformes techniques

Critère 4 : Types de données (n° téléphone)

Certains types de données présentent un défi particulier car peuvent être présentées de façons différentes. Il est donc essentiel de standardiser les structures dont sont présentées certaines informations

Cas des valeurs numéros de téléphone,

Groupages différents selon les pays

- Au Maroc: 05 37 11 11 11
- Aux Etats Unis: 555-333-1234

Indicatifs internationaux

- +212 661 111 123 ou +212 (6) 61 111 123 ou 00212661111123, etc...

L'Union Internationale des Télécommunications (UIT) a adopté un standard (ITU E.16441) qui

impose une représentation unique des numéros de téléphone avec les caractéristiques suivantes

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

Les exemples précédents s'écriraient donc :
212661111123

Session 3: Standards et Plateformes techniques

Critère 4 : Types de données (Coordonnées géographiques)

Certains types de données présentent un défi particulier car peuvent être présentées de façons différentes. Il est donc essentiel de standardiser les structures dont sont présentées certaines informations

Cas des valeurs coordonnées géographiques,

Il existe 2 façons de transcrire des coordonnées

- Format décimal : 34.0209° N, -6.8416° W
- Format degrés, minutes, secondes (DMS) : 34° 1' 15" N, 6° 50' 30" W

Il existe plusieurs systèmes géodésiques

- WGS 84, NAD 83, PZ-90, GCI-02, BD-09

Données atomiques: il est recommandé de fournir ce type de données séparées sous plusieurs colonnes

Le plus utilisé est WGS 84:

Standard global : Utilisé pour les systèmes de navigation par satellite, notamment **GPS**.

Universalité : Compatible avec la majorité des cartes et systèmes de localisation modernes.

Précision : Offre une représentation précise de la forme de la Terre (ellipsoïde) et des coordonnées géographiques.

Adoption internationale : Reconnu comme référence par de nombreuses organisations, y compris l'ONU, l'aviation civile (OACI), et les services maritimes.

Session 3: Standards et Plateformes techniques

Récapitulatif des critères techniques

Dimension	Critère	Standard recommandé
Format de fichiers	Données tabulaires	CSV - RFC 4180
	Image bitmap	PNG(*), JPEG (*), TIFF (*)
	Image vectorielle	SVG (*)
	Texte	TXT(*), RTF (*), HTML(*)
	Données géospatiales	Geojson, geotiff ou shapefile
	Données structurées/hiérarchiques	XML(*), JSON (*), EPUB
Encodage	Encodage des fichiers	UTF-8 (*)
Type de données	Valeurs numériques	IEEE 754-2019
	Date	NM ISO 8601 (*)
	Numéro de téléphone	ITU E.164
	Coordonnées géographiques	Système géodésiques WGS 84 (GPS)
	Données textuelles	Utilisation de référentiel
	Données atomiques	Les données ayant plusieurs composantes comme les données géographiques sont séparées (une colonne par composante)
	Valeur manquante	Cellule/champs vide
Unités	Unité de mesure	Système international de données - ISO 80000-1

Session 3: Standards et Plateformes techniques

Transformer la donnée pour la rendre homogène

Il est nécessaire de garder en tête que ces données sont destinées à être lues par des machines, et non pas à être ergonomiques pour un humain. Il faut transformer/formater les données selon des standards définis

Normaliser

Les espaces, les majuscules, les accents, la ponctuation

Coordonnées géographiques

Système géodésiques WGS 84 (GPS)
Séparer les valeurs en colonnes

Valeurs numériques

Appliquer la norme IEE 754-2019
Décimales avec le '.', pas de séparateurs de milliers

Données textuelles

Privilégier l'utilisation de référentiels (ex. registre de commerce, codes postaux, ..)

Dates

Harmoniser les structures de date
Utiliser la norme 8601
Préciser le fuseau horaire
Préférer le fuseau UTC

Données atomiques

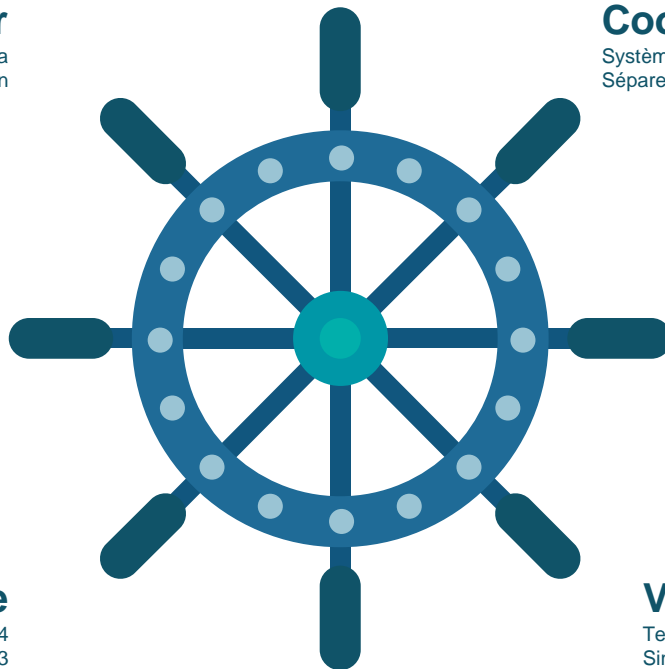
Séparer les données composées de sous données (ex. GPS) en plusieurs colonnes

Numéros de téléphone

Appliquer la norme ITU E.164
Ex. 212661123123

Valeurs manquantes

Tenter de récupérer la donnée
Sinon laisser vide



Session 3: Standards et Plateformes techniques

Rendre lisible par machine

Il est nécessaire de garder en tête que ces données sont destinées à être lues par des machines, et non pas à être ergonomiques pour un humain

Table 4. Gender Development Index									
HDI rank	Country	Gender Development Index		Human Development Index (HDI)		Life expectancy at birth (years)		Expected years of school (years)	
		Value	Score	Female	Male	Female	Male	Female	Male
		2007	2011	2011	2011	2011	2011	2011	2011

2011 Rank (2011)	Country	Score	Value	Score	Value	Score	Value	Score	Value	Score	Value	Score	Value	Score	Value
148	Algeria	0.645	0.629	0.810	6.813	0.619	0.813	0.841	0.841	0.841	0.841	0.841	0.841	0.841	0.841
149	Algeria	0.577	0.561	0.587	6.591	0.595	0.608	0.617	0.627	0.636	0.644	0.652	0.660	0.668	0.676

Department	Entity	Date	Expense T	Expense #	Supplier	Transactio	Amount
Departme	Departme	#####	CASH FUN	FINANCE	(NOTTING	HAFS-796	45000000
Departme	Departme	#####	CASH FUN	FINANCE	(NOTTING	HAFS-796	83000000
Departme	Departme	#####	CASH FUN	FINANCE	(OLDHAM	HAFS-796	28000000
Departme	Departme	#####	CASH FUN	FINANCE	(OXFORDS	HAFS-797	75000000
Departme	Departme	#####	CASH FUN	FINANCE	(PETERBOR	HAFS-797	22000000
Departme	Departme	#####	CASH FUN	FINANCE	(PLYMOUTH	HAFS-797	35000000
Departme	Departme	#####	CASH FUN	FINANCE	(PORTSMO	HAFS-797	27000000
Departme	Departme	#####	CASH FUN	FINANCE	(REDBRIDG	HAFS-797	31000000
Departme	Departme	#####	CASH FUN	FINANCE	(REDCAR A	HAFS-797	18000000
Departme	Departme	#####	CASH FUN	FINANCE	(RICHMON	HAFS-797	21000000
Departme	Departme	#####	CASH FUN	FINANCE	(ROTHERH	HAFS-797	29500000
Departme	Departme	#####	CASH FUN	FINANCE	(SALFORD	HAFS-797	31000000
Departme	Departme	#####	CASH FUN	FINANCE	(SANDWEL	HAFS-797	43000000
Departme	Departme	#####	CASH FUN	FINANCE	(SEFTON P	HAFS-798	38000000
Departme	Departme	#####	CASH FUN	FINANCE	(SHEFFIELD	HAFS-798	70500000
Departme	Departme	#####	CASH FUN	FINANCE	(SHROPSH	HAFS-798	32400000
Departme	Departme	#####	CASH FUN	FINANCE	(SOLIHULL	HAFS-798	26200000
Departme	Departme	#####	CASH FUN	FINANCE	(SOMERSE	HAFS-798	65000000
Departme	Departme	#####	CASH FUN	FINANCE	(SOUTH B	HAFS-798	49500000

- Pas de cellules fusionnées
- Un seul en tête

- Données longues plutôt que larges
- Plus propice pour archivage et manipulations BDD

Session 3: Standards et Plateformes techniques

Anonymiser les données

Les données permettant d'identifier une personne, directement ou indirectement doivent être anonymisées, en vertu de la loi 09-08

01

Pseudonymisation

Transformation des données

Champs : Noms, identifiants uniques (ID), adresses e-mail

Exemple (avant -> après) :

John Doe
-> User123



02

Masquage

Suppression partielle

Champs: Numéros de téléphone, cartes de crédit

Exemple (avant -> après) :

1234-5678-9012-3456 ->
XXXX-XXXX-XXXX-3456



03

Généralisation

Réduction de précision

Champs : Adresses, dates de naissance

Exemple (avant -> après) :

12 Avenue des FAR ->
Avenue des FAR



04

Perturbation

Ajout de bruit

Champs : Salaires, données financières

Exemple (avant -> après) :

45,000
-> 44,987



04

Suppression

Élimination complète

Champs : Champs sensibles inutiles

Exemple (avant -> après) :

123 Avenue des FAR -
> (vide)



04

K-Anonymisation

Regroupement

Champs : Âge, code postal

Exemple (avant -> après) :

29
-> 20-30



05

Randomisation

Modification aléatoire

Champs : Adresses e-mail, noms

Exemple (avant -> après) :

jane.doe@example.com ->
zane.qwe@example.com



06

Tokenisation

Remplacement par un jeton

Champs : Identifiants uniques, numéros de compte

Exemple (avant -> après) :

AB123456
-> TK982345



Session 3: Standards et Plateformes techniques

Choix de la méthode d'anonymisation

La bonne méthode d'anonymisation selon le contexte

Selon le type de data, le niveau de réidentification, et l'usage cible prévu, il est possible d'appliquer la bonne méthode:



Type de données

Les types de données définissent la sensibilité et le besoin de protection :

Si ce sont par exemple des données permettant directement l'identification (Nom/Prénom, Adresse, Numéro de téléphone, ...)

Techniques associées :

- **Suppression** : Retirer les informations directement identifiables (ex. : noms dans un fichier médical).
- **Pseudonymisation** : Remplacer les identifiants directs par des codes (ex. : un numéro aléatoire au lieu d'un nom).
- **Généralisation** : Réduire la granularité des données (ex. : transformer une date de naissance en année de naissance).



Niveau de risque de réidentification

Le risque de réidentification dépend de la sensibilité des données et des techniques de croisement:

Si par exemple on a un risque de réidentification à partir des éléments croisés (ex. profession + localisation permettent une identification).

Techniques associées :

- **K-anonymité** : Assure qu'au moins K individus partagent les mêmes attributs pour éviter la réidentification (ex. : au moins 10 personnes avec la même combinaison de sexe et âge).
- **Perturbation** : Ajouter du bruit statistique (ex. : altérer légèrement les valeurs numériques pour masquer les données exactes).



Usage prévu des données

L'usage final des données détermine la précision nécessaire et les contraintes éthiques :

Si par exemple l'usage prévu est à des fins statistiques :

Techniques associées :

- **Agrégation** : Transformer des données individuelles en statistiques globales (ex. : taux moyen d'une maladie par région).

Session 3: Standards et Plateformes techniques

Choix de méthode d'anonymisation: Exemples

Infrastructures techniques



Contenu : Nom des équipements techniques installés (antennes, transformateurs, ..)

Technique d'anonymisation : Perturbation des coordonnées GPS (précision réduite au quartier).

Données de santé par région



Contenu : Nombre de patients par maladie et région.

Technique d'anonymisation : Agrégation par région et tranche d'âge.

Données économiques



Contenu : Revenus moyens par profession.

Technique d'anonymisation : K-anonymité pour éviter la réidentification.