

Le Dictionnaire de Données : Explication

Le **dictionnaire de données** est un élément clé dans tout projet de traitement et d'analyse de données. Il sert à **définir, structurer et documenter** les données utilisées afin d'assurer une compréhension claire et uniforme.

- De préférence 1 dictionnaire de données pour toutes les variables du projet.
- Classer les variables de la plus importante à la moins importante (ID - Localisation - Date - Complémentaires...)

Nom	Libellé	Type	Contrainte	Mode	Fichier téléchargé	Site d'origine	Commentaires
id_station	Identifiant unique de la station Vélib'	Entier	Unique, non nul	Identification	velib-disponibilite-en-temps-reel.csv (téléchargé le 12/03/2025)	data.gouv.fr	Données mises à jour en temps réel
fonctionnement	Station en service ou non	Booléen	Oui/Non	État/Fonctionnement	velib-disponibilite-en-temps-reel.csv (téléchargé le 12/03/2025)	data.gouv.fr	1 = en service, 0 = hors service

Résumé : Structure Complète du Dictionnaire de Données

Un bon dictionnaire de données doit contenir :

- **Nom de la variable** : Comment elle est nommée dans la base.
- **Libellé** : Ce qu'elle représente.
- **Type** : Son format (entier, texte, booléen...).
- **Contraintes** : Règles à respecter (unique, non nul...).
- **Mode** : Son rôle dans l'analyse.
- **Source (Fichier téléchargé - Site d'origine (lien))** : D'où vient la donnée.
- **Commentaires** : Informations supplémentaires.

1) Les types de données possibles

Chaque variable a un **type de données** qui détermine la manière dont elle est stockée et utilisée. Voici les principaux types :

Type	Définition	Exemples
Entier (integer)	Nombre entier	123, 42, 6789
Décimal (float/double)	Nombre avec décimales	12.45, 0.89, 3.14159
Texte (string)	Chaîne de caractères	"Paris", "Station A"
Booléen (boolean)	Valeur vraie ou fausse	Oui/Non, True/False, 1/0
Date (date/time)	Date ou heure	"2024-03-10", "14:30:00"
Catégorie (enum)	Liste de valeurs possibles	["Homme", "Femme", "Autre"]
Clé étrangère (foreign key)	Fait référence à une autre table	id_station lié à id_commune

2) Les contraintes possibles sur les données

Pour garantir la qualité des données, on peut leur appliquer certaines **contraintes** :

Contrainte	Définition	Exemple
Unique	Une valeur ne peut pas se répéter	id_station doit être unique
Non nul (not null)	Une valeur doit toujours être renseignée	ville ne peut pas être vide
Valeurs limitées (enum)	La variable ne peut prendre que certaines valeurs	fonctionnement = Oui ou Non
Relation (foreign key)	La valeur doit exister dans une autre table	codegeo doit correspondre à un arrondissement existant
Plage de valeurs	Une valeur doit être comprise entre un minimum et un maximum	p21_pop ≥ 0 (une population ne peut pas être négative)

3) Le Mode

Le **mode** permet de classifier les variables selon leur rôle dans l'analyse. Voici les principales catégories :

Mode	Explication	Exemples
Identification	Données permettant d'identifier une entité unique	id_station, codegeo
Localisation	Données permettant de situer un élément dans l'espace	ville, coordonnees_geographiques
État/Fonctionnement	Données indiquant si un élément est actif ou non	fonctionnement (Oui/Non)
Démographie	Données sur la population et la répartition des habitants	p21_pop, c21_pop15p
Statistique/Calcul	Données utilisées pour des traitements et analyses	densite_population, ratio_stations_par_habitant



4) Les différentes façons d'indiquer la source

Type de Source	Explication	Exemple
Nom du fichier	Si les données proviennent d'un fichier spécifique dans le projet	velib-disponibilite-en-temps-reel.csv
Nom du site ou plateforme	Si les données sont téléchargées depuis un site ou une API	data.gouv.fr, Paris Open Data
Nom de l'organisme ou institution	Si les données sont issues d'une base officielle	INSEE, Mairie de Paris, IGN
Nom de la base de données	Si les données proviennent d'une base interne	BDD_Transport_2024
API utilisée	Si les données sont récupérées en temps réel depuis une API	API Vélib', Google Maps API

Quelle source choisir selon le cas ?

- **Si les données viennent d'un fichier CSV, Excel, JSON, etc., mentionne le nom du fichier** (velib-disponibilite-en-temps-reel.csv).
- **Si elles viennent d'un site open data ou d'un organisme public, indique le site et l'organisme** (data.gouv.fr, INSEE).
- **Si elles sont récupérées en direct via une API, précise le nom de l'API** (API Vélib').
- **Si elles proviennent d'une base de données interne, indique le nom de la base** (BDD_Transport_2024).

Recommandation :

 Fichier téléchargé	 Site d'origine
Monfichier.csv (téléchargé le 12/03/2025)	data.gouv.fr