Recommandations pour favoriser l'interopérabilité des données open data

Normalisation des éléments de base

Version : OpenDataFrance, janvier 2018, v1.0, Licence : CC-BY-SA-NC

Table des matières

1 - Cadrage

Quelque soit le jeu de données publiés par les collectivités, certains champs doivent avoir un format unifié au niveau national pour éviter les disparités dans leur présentation.
Dans quelques cas, il est aussi indispensable que des champs “pivots” soient respectés pour faciliter les liens entre les bases et la récupération de nombreuses données contextuelles.
Enfin, de bonnes pratiques émergent pour produire des jeux de données identifiables, exploitables et de qualité.

Bien qu’un peu éloigné de l’open data, on pourra aussi consulter le Référentiel Général d’Intéropérabilité, ou RGI, produit par l’Etat (DINSIC) qui décrit des standards d’interopérabilité : information sur les formats d’échange (synthèse des standards retenu pour l’interopérabilité syntaxique), interopérabilité sémantique (définition à retenir par tous pour les principaux concepts liés à l’interopérabilité, etc.).

Nous rassemblons ici quelques règles générales et bonnes pratiques, sans prétendre à l'exhaustivité.

2 - Format des données

Des nombreuses données sont soumises à des normes (plus ou moins connues ou appliquées) nationales ou internationales. Les lister ici serait fastidieux, probablement incomplet et peu opérationnel.
Nous indiquons ci-après des champs courants dans les jeux de données produits par les collectivités qui gagneraient à respecter des standards reconnus. Leur normalisation favorise l'interopérabilité et la compréhension des données, donc leur qualité et réutilisabilité.

- Date : format AAAA-MM-JJ (2016-07-23), conforme à la norme internationale ISO8601
  (voir aussi https://fr.wikipedia.org/wiki/Heure)

- Heure (pour les données temps-réel notamment), heure/minutes/seconde et période de temps : format HH:MM:SS (éventuellement suivi d'une virgule ',' puis de décimales de seconde), conforme à la norme internationale ISO8601.
  (voir aussi https://fr.wikipedia.org/wiki/Heure)

- Nombre : dans le cas des données décimales, on privilégiera le séparateur décimal point "." (la virgule étant réservée à la séparation des champs dans un fichier CSV)

Il existe aussi des référentiels pour codifier les langues (ISO 639-1, cas des données régionales), pour les pays (ISO3166), pour les monnaies (ISO 4217), etc. que nous ne détaillons pas ici.

- Adresse : la norme officielle pour coder les adresses est assez lourde. Elle est décrite dans le document :
  En pratique, nous recommandons de décrire une adresse avec les champs suivants :

Champ	Objet	Obligatoire	Type / Format / Exemple	Commentaires et Références
NUMERO + SUFFIXE	Numéro de l'adresse dans la voie	OUI	Numéro d'adresse dans la voie et, dans le cas des voies sans adresse, la valeur "99999" est attendue ex : 32 Bis	Voir le détail de cette spécification dans le modèle de données établi par l'AITF

TYPE + VOIE_NOM	Type et nom de la voie	OUI	Avenue Saint-Jérome	Voir le détail de cette spécification dans le modèle de données établi par l'AITF
COMMUNE_CP	Code Postal de la Commune	OUI	Texte / codifié CodePostaux / Ex : 56190, 2B100	
COMMUNE_NOM	Nom officiel de la commune	OUI	ex : Arzal	
Coordonnées de géolocalisation	X	NON	Selon le référentiel choisi (voir ci-dessous)	
Coordonnées de géolocalisation	Y	NON	Selon le référentiel choisi (voir ci-dessous)	
CLE_INTEROP	Clé nationale d'interopérabilité	NON	Voir travaux de l'AITF https://cms.geobretagne.fr/sites/default/files/documents/aitf-sig-topo-adresse-fichier-echange-simplifie-v_1.1_0.pdf	L'utilisation de ce champ est encore peu généralisé mais les acteurs de l'information géographique préconisent sa généralisation à des fins d'interopérabiité.

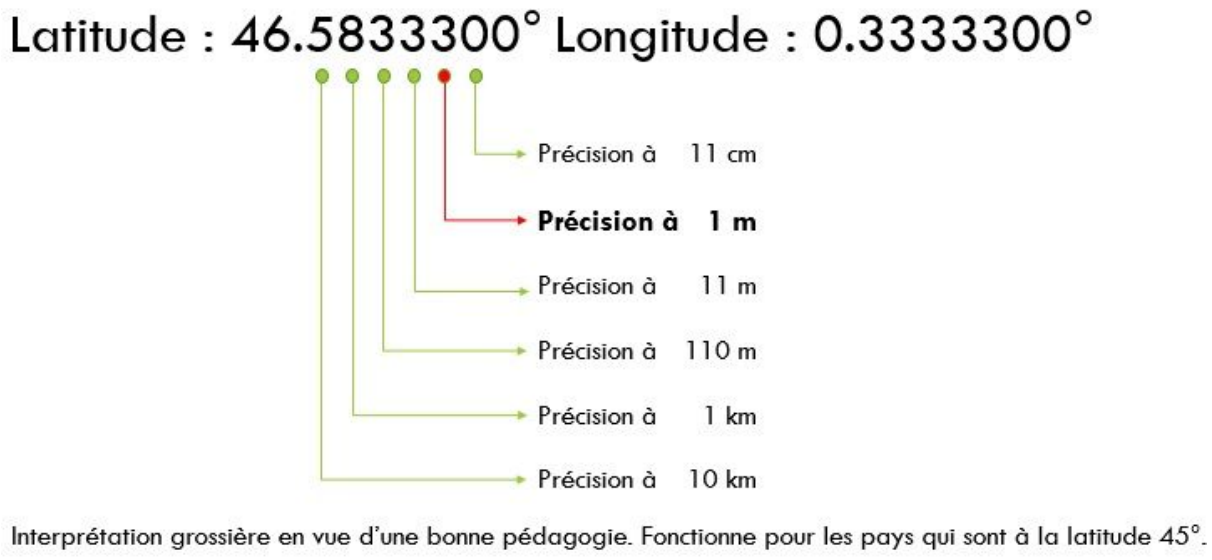
- Géolocalisation :

  Les adresses peuvent être exprimées dans un des deux référentiels de référence :

  ➔ WGS84
  Ce référentiel est proposé car cela correspond à un usage très courant, notamment un contexte international et la plupart des équipements et services de géolocalisation (GPS). Il s'applique à certaines zones du territoire français (DOM-TOM).

  ➔ RGF93 (ou Lambert93)
  Ce référentiel est préconisé par la réglementation pour le territoire métropolitain français. Il est cependant peu utilisé dans un contexte international.
  Réf : https://frama.link/referentiel_geographique

- Cas des Longitudes et Latitudes exprimées dans le format WGS84
  - Exprimé LAT (Latitude) et LONG (Longitude)
  - Nombre de décimale :
    Le nombre de décimale dépend de la précision de la localisation.
    5 décimales correspondent à une précision au mètre et permettent de couvrir la majorité des usages.
  - Exemple : Latitude : 48.87079, Longitude : 2.31689



Interprétation grossière en vue d'une bonne pédagogie. Fonctionne pour les pays qui sont à la latitude 45°.

- Cas des Longitudes et Latitudes exprimées dans le format RGF93

  Généralement appelé X (pour la longitude) et Y (pour la latitude)

  Exemple : Centre-ville de la ville de Tours :
  X=1525240,99 (ou EST) et Y =6246398,33 (ou NORD) dans RGF93-CC47
  Ce qui correspond dans le référentiel WGS84 à :
  LAT=47,39414(59) LONG=0,68467(36)

- Les outils géomatiques savent générer ou convertir les coordonnées d'un point dans un référentiel ou un autre.

3 - Format des données "pivots"

Certains champs peuvent servir de "pivots" au niveau national. Ce sont des champs qui permettent de retrouver de façon univoque, claire ou complète des données dans des bases de référence nationale (en particulier dans les bases du Service Public de la Données).

On rappelle les principaux ci-après :

- Identification d'une collectivité
  - Code COG (connu aussi sous le nom de code INSEE) :
    - Le Code officiel géographique (COG) est la référence légale éditée par l'Insee, qui rassemble les codes et libellés des communes, des cantons, des arrondissements, des départements, des régions, des collectivités d'outre-mer, des pays et territoires étrangers. On se réfère souvent à l'un ou l'autre de ces codes géographiques en parlant de code Insee :
      - Le *code région* contient deux chiffres.
      - Le *code département* contient deux à trois chiffres ou lettres. Il se retrouve sur les plaques d'immatriculation.
      - Le *code arrondissement* contient un chiffre.
      - Le *code canton* contient deux chiffres.
      - Le *code commune* contient cinq chiffres ou lettres (concaténation du code département et de la codification de la commune de deux à trois chiffres). Il est le plus employé.

      Attention : un code issu du COG n'est pas unique quelque soit les niveaux hiérarchiques des collectivités. Par exemple le code de la Région Champagne-Ardenne est 21, comme celui du département de la Côte-d'Or. De ce fait, il ne faut pas mélanger des régions et des départements dans la même colonne; une pratique constatée est de donner le COG dans un champ et le type de collectivité dans un autre (comme par exemple dans les données de marchés publics).

      Le Code officiel géographique est révisé chaque année, en fonction notamment des fusions et associations de communes ou de territoires, et des changements de dénomination.
  - Code SIRENE :
    - Cet identifiant est plus précis qu'un code INSEE lorsqu'un désigne un acteur public, car il n'existe pas de code INSEE pour un EPCI (Etablissement Public de Coopération Intercommunale)
    - Identifiant du Système d'Identification du Répertoire des Entreprises (et dans notre cas aussi des Collectivités)
    - Chaîne numérique 9 (correspondant à l'entité juridique)
    - Ex : 797681236
  - Code SIRET :
    - Cet identifiant est à choisir lorsque l'on veut désigner un établissement au sein d'une organisation. Cela est souvent pertinent lorsque l'on s'intéresse à un budget ou une décision locale.
    - Identifiant du Système d'Identification du Répertoire des Établissements

        - Chaîne numérique 9+5 (ces 5 derniers chiffres correspondant à l'établissement de l'entité juridique indiquée par la première chaîne de 9 chiffres)
        - Ex : 79768123600015

- Identification d'une entité juridique (entreprises, organismes -dont les collectivités- et associations)

    - Code SIRENE ou SIRET selon si l'on parle de l'entité juridique ou de l'établissement (une personne, un budget) :
        - Voir ci-dessus

    Se voient attribuer un numéro SIREN par l'INSEE :

    - des personnes physiques :
        - toutes celles exerçant une profession non salariée de façon indépendante (professions libérales, commerçants, etc.) ;
        - des loueurs de biens immobiliers non inscrits au RCS mais signalées à l'Insee à la demande des centres des impôts ;
        - et, à des fins de gestion donc hors accès public, les associés-gérants signalés à l'Insee par les URSSAF
    - les groupements de droit privé non dotés de la personnalité morale (indivisions, sociétés de fait, sociétés en participation, etc.)
    - et toutes les personnes morales, c'est-à-dire :
        - les personnes morales de droit privé (SA, SARL, GIE, sociétés civiles, associations, syndicats, etc.) ;
        - les personnes morales de droit public soumises au droit commercial (entreprises publiques) ;
        - les personnes morales de droit étranger ayant un établissement ou un bureau de liaison en France ;
        - les personnes morales (ou organismes assimilés comme telles) soumises au droit administratif (comme les institutions et services de l'État, les collectivités territoriales, etc.) ;

    Il existe cependant une exception pour les personnes morales : les associations loi de 1901 n'ont l'obligation d'avoir un SIREN que dans trois cas seulement :

    - si elles sont employeuses ;
    - si elles sont soumises à la TVA ;
    - si elles souhaitent l'obtention de subventions auprès d'administrations publiques.

    - Code RNA (RépertoireNational des Associations) :
        - L'inscription d'une association au RNA donne lieu à une immatriculation sous la forme d'un "numéro RNA", composé de la lettre W suivie de 9 chiffres. (ex. W123456789)

- Cet identifiant est unique mais les données associées dans la base RNA ne sont pas toujours mises à jour (il s'agit plutôt de l'acte de déclaration à la création de l'association). La base SIRENE fournit généralement des informations plus fiables.

4 - Produire des données de qualité

- Produire un fichier CSV de qualité :
  - Un travail approfondi a été mené par plusieurs acteurs sous la conduite de la Fing pour énoncer des règles de base de production et de contrôle d'un fichier tabulaire au format CSV :
    - https://docs.google.com/document/d/1KK515FwKJ4UEg27MkDHrIlzWAp_7skcSYU1JmepqXQw/edit#heading=h.uv9wks7mugeg

- Documenter en annexe du jeu de donnée (fichier séparé ou autre) le sens des données contenus dans le jeu de donnée :
  - nom explicite de chaque champ (pas uniquement des codes plus ou moins compréhensibles),
  - portée de chaque donnée et condition particulière de production,
  - sens des codifications lorsque le cas se présente

  On trouvera un exemple de documentation dans les jeux de données du Socle Commun des Données Locales (spécifications et onglet "métamodèle" des exemples de jeux de données).

- Produire des données selon plusieurs formats.
  - Données tabulaires
    - Le principe de l'opendata est de publier des données à un format ouvert non prioritaire, on privilégiera donc le format "CSV" et le format "ODT" (lisible par un logiciel libre tel que Libre Office)
    - On acceptera le format "XLS" qui, bien que de propriétaire (Microsoft), peut être lisible par des logiciels libre comme LibreOffice.
    - On pourra aussi produire les données au format XMLl" qui est un standard ouvert et reconnu pour les jeux de données de structure un peu plus complexe (xml)

  - Données textuelles
    - Format "TXT" :
    - Format "JSON" : JSON est un format de données textuelles dérivé de la notation des objets du langage JavaScript. Le principal avantage de JSON est qu'il est simple à mettre en œuvre par un développeur tout

en étant complet. Au rang des avantages, on peut également citer : peu verbeux, ce qui le rend lisible aussi bien par un humain que par une machine.
    - ODF (ou ODT) : OpenDocument est un format ouvert de données pour les applications bureautiques : traitements de texte, tableurs, présentations, diagrammes, dessins et base de données bureautique. OpenDocument est la désignation d'usage d'une norme dont l'appellation officielle est OASIS (Open Document Format for Office Applications, également abrégée par le sigle ODF)
    - RDF

- Données géographiques
    - GéoJSON
    - Shapefile
    - OWS
    - KML
    - ...

- Données temps-réel
    - (rédaction ultérieure)