
Préparer les données à l'ouverture et la circulation

Etalab

etalab^{gouv.fr}

26/09/2021

Table des matières

1	Introduction	4
1.1	Le partage des données	4
1.2	Le critère de qualité des données	4
2	Préparer le jeu de données	5
2.1	Extraire le jeu de données du système d'information de l'organisation	5
2.2	La structure du jeu de données	5
2.2.1	Cas 1 - La structure du jeu de données correspond à un schéma de données existant	6
2.2.2	Cas 2 - La structure du jeu de données ne correspond à aucun schéma de données existant	10
2.3	Le choix du format du jeu de données	11
2.4	Le contenu du jeu	13
2.4.1	Le titre du jeu de données	13
2.4.2	L'encodage du fichier	13
2.4.3	L'entête des colonnes (pour le format tabulaire)	14
2.4.4	Le séparateur (pour le format tabulaire)	14
2.4.5	Gestion des champs non attribués	14
3	Lier les données à un référentiel	14
3.1	Avantages	15
3.2	Exemples de référentiels	15
3.2.1	Le service public de la donnée	15
3.2.2	Les autres référentiels	17
3.2.3	Partager ses propres référentiels	20
3.3	Le cas spécifique des adresses	21
3.3.1	Le géocodage	22
4	Documenter les données	22
4.1	Description générale du jeu de données	23
4.2	Description du mode de production des données	24
4.3	Description du modèle de données	25
4.4	Description des métadonnées	27
4.5	Description des changements majeurs	28
4.6	Points de contact	29

5	Éléments organisationnels	29
5.1	Les prochaines étapes	30
5.2	Informations complémentaires	31

1 Introduction

Comment faciliter la réutilisation d'un jeu de données ?

Ce guide s'adresse aux acteurs publics ou privés qui souhaitent préparer leurs jeux de données à leur circulation — que ce soit en interne d'une organisation ou en open data. Des lignes directrices sont proposées afin de faciliter la prise en main, le croisement et l'exploitation des jeux de données par le plus grand nombre d'acteurs.

1.1 Le partage des données

Le partage de données entre acteurs, que ce soit à l'intérieur ou l'extérieur d'une organisation, est devenu un enjeu économique, politique et culturel. La circulation des données démultiplie leur potentiel d'usage et rend possible leur réutilisation pour des finalités qui n'étaient pas envisagées lors de leur production.

En France, le mouvement de l'ouverture des données publiques se fonde sur ces principes depuis 2011. En juin 2019, la plateforme data.gouv.fr comptait plus de 30 000 jeux de données pour 2 350 organisations. En interne, les organisations ont également pris conscience de l'intérêt que représente la circulation et l'exploitation croisées des données pour leurs activités.

1.2 Le critère de qualité des données

Pour autant, la circulation des jeux de données n'entraîne pas directement leur réutilisation. Par exemple, il a été constaté que seuls certains jeux de données publiés sur la plateforme data.gouv.fr étaient régulièrement réutilisés. De la même manière, des organisations constatent que la création d'un *data lake* n'entraîne pas forcément l'exploitation des données par des équipes tierces.

Ce constat s'explique notamment par les difficultés que rencontrent les réutilisateurs lorsqu'ils souhaitent s'approprier les données partagées. De manière générale, les jeux de données publiés sont produits dans un contexte propre à un processus métier et pour un usage particulier. Par exemple, la *base de données des demandes de valeur foncière* est historiquement produite par la Direction Générale des Finances publiques dans l'objectif de tenir un fichier immobilier et collecter l'impôt. Cet environnement métier, connu et compris par les agents de l'équipe productrice, n'est pas forcément familier aux individus tiers, qu'ils soient internes ou externes à l'organisation. Ces réutilisateurs peuvent rencontrer de nombreuses difficultés dans la compréhension de la structure du jeu de données et des données elles-mêmes.

Il est indispensable de prendre en compte les pratiques des réutilisateurs en amont de la production des jeux de données. Pour ce faire, une réflexion sur la structure des jeux de données, sur le format des

fichiers ou encore sur la documentation doit être menée systématiquement. Ce travail de réflexion facilitera l'appropriation des données par des acteurs tiers et fera gagner du temps à l'organisation productrice, qui n'aura plus à répondre à de nombreuses questions.

2 Préparer le jeu de données

2.1 Extraire le jeu de données du système d'information de l'organisation

Il est possible que les données que vous souhaitez faire circuler ne soient pas structurées sous la forme d'un jeu de données. Dans cette situation, il est nécessaire de réaliser une extraction des données depuis le système d'information où elles sont stockées. Cette extraction permet d'obtenir un jeu de données structuré, qui ordonne les données en fonction de différentes caractéristiques.

Lorsque vous cherchez à extraire des données d'un système d'information, plusieurs situations peuvent se présenter :

- Le système d'information propose un outil qui permet d'exporter l'ensemble des données depuis le système d'information. Il est nécessaire de sélectionner les données éligibles à la circulation en aval de l'export;
- Le système d'information propose un outil qui permet d'exporter l'ensemble des données ou de sélectionner un sous ensemble des données à exporter depuis le système d'information;
- Le système d'information ne prévoit pas d'outil d'exportation des données. Dans ce cas, il est nécessaire de réaliser une opération technique qui permette de réaliser l'export des données. Cette opération est directement liée aux spécificités du système d'information utilisé.

Quel que soit le mode d'export des données, il est recommandé d'automatiser l'opération réalisée afin de faciliter la mise à jour des données publiées. Cette automatisation instaure un processus sur le long terme et fait gagner du temps à l'organisation.

2.2 La structure du jeu de données

Les jeux de données qui ont vocation à circuler seront réutilisés par des acteurs tiers qui ne connaissent pas l'environnement de votre administration. Il est nécessaire de proposer une structure de jeu de données compréhensible et appropriable par tous.

Deux approches sont envisageables :

- La structure de votre jeu de données correspond à un schéma de données existant;
- La structure de votre jeu de données ne correspond pas à aucun schéma de données existant. Un travail de modélisation est nécessaire en amont de la création du jeu de données.

2.2.1 Cas 1 - La structure du jeu de données correspond à un schéma de données existant

Lexique : Schéma de données

Un schéma de données est un modèle qui permet de décrire de manière précise et univoque les différents champs et valeurs possibles qui composent un jeu de données.

Il permet notamment de valider qu'un jeu de données se conforme à un schéma, de générer de la documentation automatiquement, de générer des jeux de données d'exemple ou de proposer des formulaires de saisie standardisés. Ces schémas facilitent la montée en qualité et le croisement des données proposées en open data, surtout lorsque plusieurs producteurs de données sont amenés à produire un même jeu de données.

 Consultez [notre guide à destination des producteurs de schémas](#)

Les schémas existants peuvent avoir été définis par voie :

- **Réglementaire** : un modèle de données a été défini de manière réglementaire, par décret ou arrêté. Un schéma est un moyen de faciliter l'adoption de ces modèles par les producteurs de données. Par exemple, le schéma de données relatif à la publication des données essentielles dans la commande publique est fixé par [arrêté depuis le 14 avril 2017](#).
- **D'usage** : la réutilisation des données décrites par le schéma bénéficie à un grand nombre de réutilisateurs ou de nombreux producteurs sont amenés à utiliser ce schéma.

2.2.1.1 Pourquoi utiliser un schéma de données ? La création d'un jeu de données en conformité avec un schéma de données existant apporte plusieurs bénéfices :

- Le jeu de données créé peut être facilement croisé avec d'autres jeux de données conformes au standard utilisé ;
- L'interopérabilité des données et leur croisement est simplifié ;
- Si le jeu de données que vous créez est une agrégation de plusieurs fichiers produits par différents acteurs, la formalisation et le partage d'un standard de données facilite le travail d'agrégation des données ;
- La formalisation d'un standard de données assure une pérennité des fichiers dans le temps ;
- La documentation d'un standard de données existant est déjà rédigée et accessible.

2.2.1.2 Comment identifier un schéma de données déjà existant ? Le site schema.data.gouv.fr référence une liste de schémas de données existants. Il offre également la possibilité à tout utilisateur de soumettre de nouveaux schémas de données. Lorsque les données que vous souhaitez faire circuler correspondent à un schéma existant, nous vous conseillons de l'appliquer au plus près. Le site

schema.data.gouv.fr permet d'intégrer les schémas de données et documentations associées dans d'autres systèmes informatiques.

2.2.1.3 Comment produire un jeu de données conforme à un schéma de données ? Si le jeu de données n'est pas extrait d'un système d'information mais saisi manuellement, l'outil **CSV-GG** peut assister le processus de production. À partir d'un schéma de données sélectionné, il est possible de saisir les valeurs de chaque information et ainsi produire un fichier exhaustif et conforme.

CSV-GG mode d'emploi

Cet outil vous permet de créer un fichier CSV en vous assurant qu'il est conforme à un schéma, c'est-à-dire que ses données sont complètes, valides et structurées.

1. Sélectionnez le schéma qui vous intéresse dans la liste déroulante, les schémas disponibles ici étant ceux référencés sur schema.data.gouv.fr.
2. Remplissez le formulaire à l'aide des descriptions des différents champs et des valeurs d'exemples. Les champs indiqués par un astérisque rouge doivent obligatoirement être renseignés au moment de la saisie.
3. L'outil vous prévient d'éventuelles erreurs de validation, le cas échéant vous pouvez les corriger.
4. Une fois votre formulaire valide, les valeurs apparaissent sous la forme d'une ligne dans un tableau récapitulatif.
5. Vous pouvez alors choisir d'ajouter une ou plusieurs lignes (répétez les étapes 2 à 4) ou télécharger le fichier CSV correspondant au tableau récapitulatif.

2.2.1.4 Comment valider la conformité d'un jeu de données avec un schéma de données ? Il est possible de valider la conformité d'un jeu de données à un schéma de données existant grâce à différents outils.

Tout d'abord, il est possible d'indiquer que votre jeu de données correspond à un schéma depuis l'interface d'administration de data.gouv.fr. Lorsque vous déposez ou éditez une ressource, vous pouvez sélectionner le schéma correspondant à vos données dans une liste déroulante.

Dernière version consolidée

Titre : Dernière version consolidée

Type : Fichier principal

Description : [Rich text editor with toolbar: Bold, Italic, Highlight, Link, Image, Table, List, Code, Quote, Preview, Undo]

Date de publication : [Empty field]

Schema :

- Base adresse locale
- Budget des collectivités et établissements publics locaux
- Catalogue simplifié
- Equipements
- Lieux de covoiturage
- Lieux de stationnement
- ✓ Schéma IRVE
- Schéma SCDL Délibérations
- Schéma SCDL Subventions
- Spécification du fichier de déclaration de profil d'acheteur

URL : [Empty field]

Taille : Taille

Format : csv

Type Mime : text/csv

Somme de contrôle : sha1 Somme de contrôle

Annuler **Enregistrer**

FIG. 1 : Capture d'écran de la sélection d'un schéma depuis l'interface d'administration de data.gouv.fr

Le fait d'indiquer que votre ressource est censée respecter un schéma permet de bénéficier de vérifications de la qualité des données et d'indiquer aux réutilisateurs que vos données respectent un référentiel.




Schéma associé

Cette ressource est associée au schéma **etalab/schema-irve** ce qui signifie qu'elle doit en respecter la spécification.

Documentation


Vous pouvez consulter la documentation de ce schéma, découvrir le contexte et les recommandations destinées aux producteurs de données.

 **LIRE LA DOCUMENTATION**

Validation

La validation d'une ressource par rapport à un schéma consiste à vérifier que la ressource est bien conforme au schéma qu'elle est censée respecter. En cas d'erreurs, un rapport de validation indique les erreurs trouvées : colonnes mal nommées, valeurs non conformes etc.

data.gouv.fr met à votre disposition un outil pour valider vos fichiers par rapport à un schéma et corriger les éventuelles erreurs.

 **VALIDER LA RESSOURCE**

À propos des schémas

Les schémas de données permettent de décrire des modèles de données : quels sont les différents champs, comment sont représentées les données, quelles sont les valeurs possibles. Découvrez comment les schémas améliorent la qualité des données et quels sont les cas d'usages possibles sur schema.data.gouv.fr.

Fig. 2 : Capture d'écran de data.gouv.fr des informations disponibles sur la page d'un jeu de données lorsqu'un schéma est spécifié sur une ressource

D'autres solutions en dehors de data.gouv.fr existent. Des solutions disponibles en anglais comme goodtables.io ou [CSV Lint](https://csvlint.com) proposent des validateurs de jeux de données. Enfin, il est possible

d'intégrer une fonction de validation d'un jeu directement dans la procédure de publication. C'est le cas pour les données d'adresses locales qui font l'objet d'une validation directement sur le site adresse.data.gouv.fr.

2.2.2 Cas 2 - La structure du jeu de données ne correspond à aucun schéma de données existant

Si les données que vous souhaitez faire circuler ne correspondent à aucun schéma de données existant, il est nécessaire de réfléchir en amont à la meilleure structure pour vos données.

Tant que les données de votre administration sont dans un environnement logiciel, leur usage reste adapté à des problématiques métiers spécifiques. L'ouverture de ces données en dehors de leur environnement logiciel les émancipent de ce contexte métier. La structure du jeu de données doit alors être pensée en fonction des attentes des réutilisateurs et non plus en fonction des besoins propres à l'organisation.

Les bonnes pratiques à suivre sont les suivantes :

- Occulter l'ensemble des colonnes dont les champs contiennent des données couvertes par un secret légal (se référer au guide juridique pour plus d'information);
- Occulter l'ensemble des colonnes dont les champs contiennent des données à caractère personnel dont la publication n'est pas nécessaire à l'information du public (se référer au guide juridique pour plus d'information);
- Privilégier la présence de variables pivots. Ces variables proposent des identifiants communs qui permettent de lier plusieurs jeux de données entre eux (ex. Le numéro d'identification SIRET de la base Sirene). Pour plus de détails, consultez la page [Lier les données à un référentiel](#).

Il est également nécessaire de mener une réflexion sur la granularité de votre jeu de données :

- Faut-il proposer des données fines ou agrégées?
- Faut-il proposer un export quotidien, mensuel, trimestriel ou annuel?

Ces questions doivent être posées en amont de l'automatisation des exports. Un dialogue avec les réutilisateurs est conseillé afin de comprendre leurs besoins. Certains utilisateurs peuvent souhaiter manipuler des données granulaires tandis que d'autres préfèrent disposer d'agrégats qui permettent une réutilisation simple et rapide. A minima, il est conseillé de proposer un fichier complet unique qui contient l'ensemble des données historiques.

2.3 Le choix du format du jeu de données

Afin que le maximum d'utilisateurs, internes ou externes à votre organisation, puisse s'approprier les données, il est conseillé de les faire circuler dans un format :

- **Ouvert** : un format ouvert n'impose pas de spécifications techniques qui entraveraient l'exploitation des données (par exemple l'utilisation d'un logiciel payant);
- **Aisément réutilisable** : un format aisément réutilisable sous-entend que toute personne ou machine peut réutiliser facilement le jeu de données;
- **Exploitable par un système de traitement automatisé** : un système de traitement automatisé permet de réaliser des opérations par des moyens automatiques, relatifs à l'exploitation des données. Par exemple, un fichier CSV est aisément exploitable par un système de traitement automatisé contrairement à un fichier PDF.

Rappel juridique

Toute organisation de plus de 50 agents chargée d'une mission de service public (les administrations, les collectivités de plus de 3500 habitants et les délégations de service public) est tenue de publier ses jeux de données dans un format ouvert, utilisable et exploitable par un système de traitement automatisé.

Les formats ouverts et communément acceptés sont les suivants :

Type de données	Formats conseillés	Documentation	Description
Données tabulaires	CSV	Ici	Un fichier CSV est constitué de lignes de données, où chaque champ est séparé par une virgule. Ce format est le standard le plus réutilisable, car ouvert et facilement exploitable par une machine.

Type de données	Formats conseillés	Documentation	Description
Données statiques de transport	GTFS/NeTEx	Ici	Le format GTFS est le format le plus utilisé en France par les services de mobilité d'information voyageur. Le format NeTEx est le format de référence européen qui vise l'interopérabilité des données entre États membres.
Données géographiques	GeoJSON, Shapefile, MapInfo MIF/MID, MapInfo TAB et GML, pour les vecteurs / ECW, JPEG2000 et GeoTIFF, pour les données pixelisées (raster)	Ici	Les données géographiques sont organisées sous forme d'ensemble de données hiérarchisées. Les formats proposés sont conçus spécifiquement pour être largement exploitables et être intégrés facilement dans des outils de cartographie.

Type de données	Formats conseillés	Documentation	Description
Données hiérarchiques	JSON / XML / YAML	indisponible	Les données hiérarchiques décrivent des relations hiérarchiques entre différentes données. Le format JSON est préconisé lorsque les données sont liées entre elles sous forme d'arbres verticaux.

2.4 Le contenu du jeu

2.4.1 Le titre du jeu de données

Le titre de votre jeu de données doit pouvoir renseigner n'importe quel réutilisateur sur le contenu du fichier. Pour cela, il est nécessaire de :

- Ne pas donner un titre trop générique qui obligerait le réutilisateur à ouvrir le jeu de données pour comprendre son contenu (Par exemple "liste.csv" ou encore "balance comptable" sans indiquer l'organisation concernée);
- Ne pas donner un titre trop long qui rendrait la manipulation du fichier difficile. Par exemple le titre du jeu de données "Fichier consolidés des données essentielles de la commande publique" est suffisamment générique pour ne pas revenir sur toutes les sources de données utilisées pour agréger le jeu de données;
- Ne pas donner un titre contenant des accents ou caractères spéciaux qui poseraient des problèmes d'interopérabilité des fichiers;
- Ne pas donner de titre trop technique issu de nomenclatures métier.

2.4.2 L'encodage du fichier

L'encodage d'un fichier est la norme utilisée pour coder chaque caractère par une suite de 0 et de 1 compréhensible par une machine. Lorsque l'encodage est mal choisi, le réutilisateur des données est souvent contraint de convertir le fichier, notamment afin de faire apparaître les accents et caractères spéciaux.

Il est conseillé d'utiliser l'encodage UTF-8. Cet encodage permet d'encoder l'ensemble des caractères du répertoire universel de caractères codés (notamment les caractères contenant des accents ou des caractères spéciaux).

2.4.3 L'entête des colonnes (pour le format tabulaire)

Dans un fichier tabulaire, la première ligne du fichier peut être utilisée pour nommer chaque colonne et donner des informations sur les données associées. Plutôt que d'indiquer "Colonne n°X", il est conseillé de donner un nom de colonne explicite. Le nom des colonnes doit être sans majuscule, abréviation, accents, ni espaces (préférez le caractère `_`) afin de faciliter la manipulation des fichiers.

2.4.4 Le séparateur (pour le format tabulaire)

Dans un fichier tabulaire, le séparateur permet de structurer les données sous forme de cellules. Il est conseillé d'utiliser la virgule comme séparateur.

:::warning Séparateurs décimaux Dans un fichier CSV, la virgule n'est pas considérée comme un séparateur décimal. Si votre fichier contient des valeurs décimales, il est nécessaire d'encapsuler chaque champ entre des guillemets. La plupart des tableurs (Excel, OpenOffice Calc, etc) proposent l'encapsulation des champs entre guillemets. Une seconde solution consiste à convertir l'ensemble des virgules utilisées pour des valeurs décimales par un point.

2.4.5 Gestion des champs non attribués

Il est possible qu'un champ de votre jeu de données ne soit pas attribué. Il convient de laisser ce champ vide plutôt que d'attribuer la valeur 0. Le zéro correspond à une valeur, qui peut dénaturer le sens de votre fichier.

3 Lier les données à un référentiel

Comme explicité dans la page [Préparer le jeu de données - Cas 2 - La structure du jeu de données ne correspond à aucun schéma de données existant](#), il est important d'intégrer dans vos jeux de données des données pivots relevant d'un référentiel.

Exemple

Mon jeu de données est une liste d'actions culturelles menées par ma région. Certaines de ces actions sont gérées par des associations. Il peut être intéressant de publier un jeu de données re-

censant ces actions avec un champ correspondant à l'identification des associations. Cet identifiant existe et est standardisé, il s'agit du numéro RNA, identifiant national des associations dont le répertoire est opéré par le ministère de l'Intérieur.

3.1 Avantages

L'intégration dans votre jeu de données de données pivots qui correspondent à un référentiel offre plusieurs avantages :

- **Une meilleure formalisation** : en se basant sur un référentiel, le producteur de données a l'assurance d'utiliser un format de données standard et partagé par un grand nombre de jeux de données;
- **Une meilleure synthèse** : en se basant sur un référentiel, le producteur évite l'abondance de détails et va à l'essentiel. L'obtention d'informations complémentaires se fera par le biais de la consultation du référentiel en lui-même;
- **Une meilleure compréhension** : en intégrant dans son jeu de données des données correspondant à un référentiel, le producteur facilite la compréhension de celui-ci par des éventuels utilisateurs car il se réfère à un standard largement adopté;
- **Une meilleure réutilisation** : intégrer des données liées à un référentiel facilitera la réutilisation du jeu de données et permettra son enrichissement avec d'autres données partageant la même donnée pivot;
- **Une meilleure interopérabilité** : intégrer des données pivots facilite le lien avec des données de référence fiables et à jour.

3.2 Exemples de référentiels

3.2.1 Le service public de la donnée

Parmi les nombreux référentiels existants, nous pouvons citer le service public de la donnée (SPD) qui référence 9 jeux de données maintenus de manière pérenne par des administrations.

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.

 Base Adresse Nationale (BAN) 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	 Base Sirene des entreprises et de leurs établissements (SIREN, SIRET) Pour vous abonner à notre lettre d'information Sirene open data actualités, suivez ce lien : 15 7 27 France	 Code Officiel Géographique (COG) 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
 Plan Cadastral Informatisé (PCI) 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	 Registre parcellaire graphique (RPG) : contours des parcelles et îlots cultureux et leur groupe de cultures majoritaire 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	 Référentiel de l'organisation administrative de l'Etat 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
 Référentiel à grande échelle (RGE) 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	 Répertoire National des Associations (RNA) 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	 Répertoire Opérationnel des Métiers et des Emplois (ROME) Dans un contexte marqué par de fortes mutations de l'environnement économique et social, le ROME (Répertoire Opérationnel 5 2 9

Vous pouvez consulter sur le site data.gouv.fr la liste [des données du service public de la donnée](#).

Parmi ces jeux de données, nous pouvons citer entre autres :

- La [base Sirene](#) qui liste l'ensemble des entreprises françaises. Les identifiants principaux liés à cette base sont le N° SIREN (identifiant d'une unité légale) et le numéro SIRET (identifiant d'un établissement) ;
- La [base du répertoire national des associations \(RNA\)](#), qui répertorie l'ensemble des associations relevant de [la loi du 1er juillet 1901 relative au contrat d'association](#). L'identifiant des associations est le numéro RNA ;
- Le [code officiel géographique \(COG\)](#) qui décrit les codes et libellés des communes, des cantons, des arrondissements, des départements, des régions et des pays et territoires étrangers.

Exemple

Afin de lister l'ensemble des actions culturelles de ma région, nous avons vu que le numéro RNA pouvait être utile pour identifier les associations. Grâce à celui-ci, il est également possible de récupérer le numéro SIRET de l'association si celle-ci en possède un. Il est également possible de détailler dans le jeu de données le code commune et le code département de chaque action. Pour cela, il convient de se référer au Code officiel géographique. **Attention à bien respecter celui-ci. Par exemple, le code département de l'Ariège est le "09" et pas le "9". Ce type d'erreur pourrait entraîner des difficultés lors de la réutilisation des données.**

3.2.2 Les autres référentiels

Au-delà du service public de la donnée, il existe un grand nombre de jeux de données standardisées et communément partagées avec le plus grand nombre. Il est conseillé d'intégrer les données pivots de ces référentiels dans vos jeux de données, si vous pensez que la démarche est pertinente.

Exemple

L'identifiant unique d'une certification professionnelle est le [numéro RNCP](#). Ce jeu de données ne fait pas partie du service public de la donnée mais est largement partagé par les acteurs du domaine de la formation professionnelle.

Vous trouverez ci-dessous un recensement, non exhaustif, de référentiels sur lesquels vous pouvez vous appuyer pour l'intégration de variables pivots :

3.2.2.1 Référentiels métiers

Nom du jeu de données	Variable(s) pivot(s)	Description	Producteur
Base SIRENE	SIRET, SIREN	Liste des établissements (SIRET) et unités légales (SIREN) françaises	INSEE
Base Adresse Nationale (BAN)	BAN	Référencement de l'intégralité des adresses du territoire français	BAN
Code Officiel Géographique (COG)	Codes et libellés	Liste des communes, cantons, arrondissements, départements, régions, pays et territoires étrangers	INSEE
Plan Cadastral Informatisé (PCI)	Identifiant	Représentation de chacune des sections du cadastre français	Ministère de l'Économie et des Finances

Nom du jeu de données	Variable(s) pivot(s)	Description	Producteur
Registre parcellaire graphique (RPG)	Identifiant	Base de données géographique de référence pour l'instruction des aides de la politique agricole commune (PAC)	IGN
Référentiel de l'organisation administrative de l'Etat	Identifiant	Liste des institutions régies par la Constitution de la Ve république ainsi que les administrations qui en dépendent	DILA
Référentiel à grande échelle (RGE)	Identifiant	Composantes ortho-photographique, topographique et adresse, parcellaire et altimétrique des territoires de l'Etat français	IGN
Répertoire National des Associations (RNA)	N° RNA / N° Waldec	Ensemble des associations relevant de la loi du 1er juillet 1901 relative au contrat d'association, dont le siège est en France	Ministère de l'Intérieur

Nom du jeu de données	Variable(s) pivot(s)	Description	Producteur
Répertoire Opérationnel des Métiers et des Emplois (ROME)	Code ROME	Inventaire des dénominations d'emplois/métiers les plus courantes, analyse des activités et compétences, regroupement des emplois selon un principe d'équivalence ou de proximité	Pôle Emploi
Nomenclature d'activités française (NAF)	Code NAF	Nomenclature des activités économiques productives, principalement élaborée pour faciliter l'organisation de l'information économique et sociale	INSEE
Répertoire National des Certifications Professionnelles (RNCP) et Répertoire Spécifique (RS)	N°RNCP / N°RS	Répertoire des certifications officielles inscrites au RNCP et au RS	France Compétences
Fichier FANTOIR des voies et lieux-dits	N° FANTOIR	Nom des lieux-dits et des voies pour chaque commune, y compris celles situées dans les lotissements et les copropriétés	Ministère de l'Économie et des Finances

Nom du jeu de données	Variable(s) pivot(s)	Description	Producteur
Etats et capitales du monde	Code Pays	Liste des états indépendants reconnus par la France	Ministère de l'Europe et des Affaires Etrangères
Nomenclatures des professions et catégories socioprofessionnelles	Code PCS / Code PCS-ESE	Nomenclatures des professions et catégories socioprofessionnelles	INSEE
Liste des établissements d'enseignements supérieurs Liste des établissements d'enseignements secondaires	N°UAI	Liste des unités administratives immatriculées	ONISEP

3.2.2.2 Référentiels techniques Les référentiels techniques n'ont pas de significations métiers. Cependant, ils permettent de décrire une donnée de manière standardisée. Ces standards permettent aux utilisateurs et aux algorithmes de pouvoir interpréter automatiquement la donnée de manière correcte. C'est le cas principalement des dates et des coordonnées géographiques pour lesquels sont listés ci-dessous deux exemples de standards largement adoptés.

Nom du référentiel	Description	Information
WGS84	Coordonnées géodésiques d'un lieu	Wikipedia
ISO8601	Représentation numérique d'une date et d'une heure	Wikipedia

3.2.3 Partager ses propres référentiels

::: info Cadre Commun d'Architecture des référentiels de données de l'État Le [Cadre Commun d'Architecture des référentiels de données de l'État](#) fait spécifiquement mention de l'importance des variables pivots dans le partage et la publication de données. Il stipule notamment que :

- Les données sont un bien, un actif de l'État, elles doivent être gérées et valorisées en conséquence;
- Les données doivent être standardisées, définies sur la base d'un vocabulaire commun, contextualisées, et combinables les unes aux autres;
- Les données doivent être facilement réutilisables, partageables et accessibles à travers les frontières des administrations;
- Les données publiques doivent être mises à disposition librement et ouvertement sur internet;
- La sécurité et l'archivage des données doit être assuré.

Pour favoriser au mieux le partage et l'interopérabilité des données, les acteurs sont encouragés à mettre en place leurs propres référentiels internes ou à les partager s'ils existent déjà.

Ainsi, il est pertinent de diffuser, en même temps que votre jeu de données, la liste des valeurs possibles correspondant à votre propre référentiel métier. Celui-ci sera connu et potentiellement réutilisé par d'autres acteurs.

La mise en place de référentiels fait également parti d'une stratégie de montée en qualité de la donnée. Néanmoins ce n'est souvent pas suffisant : il est ensuite nécessaire de diffuser, former et vérifier que les données produites intègrent ces référentiels et n'en dérivent pas (à partir d'un contrôle humain ou de tests automatiques).

Exemple

J'utilise en interne un numéro unique permettant d'identifier chaque type d'action culturelle (arts du spectacle, cirque, arts plastiques...). Il peut être pertinent de diffuser en parallèle à la diffusion de mon jeu de données la liste de mon référentiel. Des communes de ma région pourraient potentiellement le réutiliser pour décrire leurs actions culturelles à une maille plus fine.

3.3 Le cas spécifique des adresses

Il est fréquent que les producteurs de données souhaitent ajouter des adresses dans leurs jeux de données.

De la même manière, il existe des référentiels pour décrire une adresse de manière unique. Le référentiel officiel d'adresse s'intitule la [base d'adresse nationale \(ou BAN précédemment listé\)](#). Si vous partez de zéro pour constituer votre jeu de données, il est pertinent de partir de cette base de données pour décrire vos adresses.

Cependant, il arrive souvent qu'un producteur travaille sur un jeu de données qui contient déjà des adresses saisies. Il peut s'avérer fastidieux de corriger manuellement l'ensemble des adresses erronées. Heureusement, il existe un certain nombre d'outils pour obtenir une base d'adresse normalisée!

3.3.1 Le géocodage

Le géocodage consiste à affecter des coordonnées géographiques à une adresse postale. Cette opération peut être en partie automatisée grâce à des outils proposés par Etalab.

Le site <https://adresse.data.gouv.fr/> permet de géocoder une liste d'adresse via un appel à une API ou par le dépôt de fichier csv.

En utilisant l'outil avec un jeu de données contenant des adresses déjà saisies, l'appliquatif nous retourne un jeu de données enrichi :

- De coordonnées géographiques (longitude/latitude) ;
- Des adresses « corrigées » récupérées de la BAN ;

Géocodage massif

Le site adresse.data.gouv.fr est limité à des utilisations ponctuelles et des volumétries de données considérées faibles (moins d'un million de lignes). Si vous souhaitez géocoder un très gros jeu de données (plusieurs millions de lignes), il est recommandé d'installer votre propre environnement de géocodage, en utilisant par exemple le géocodeur [Addok](#). Vous pouvez consulter des ressources sur [GitHub](#) pour vous aider dans l'installation de votre environnement.

Quelle que soit la méthode utilisée, le processus de géocodage retournera une liste d'adresses standardisées avec leurs coordonnées géographiques associées. De plus, vous aurez également accès à une information `geo_score` correspondant au score de confiance que le géocodeur accorde à l'adresse retournée. Cet indicateur peut être utile à garder dans votre jeu de données final, puisqu'il donnera une indication aux utilisateurs sur la performance du géocodage pour chaque adresse.

4 Documenter les données

Les données issues de votre organisation ont été produites dans un contexte métier particulier. Un individu externe à l'organisation n'est pas forcément familier avec cet environnement métier, ce qui peut le freiner dans l'exploitation des données diffusées.

La documentation de vos jeux de données a une visée pédagogique et facilite la réutilisation des données. Elle décrit les données et la structure des fichiers publiés.

Il est conseillé de proposer votre documentation en ligne et non sous format PDF. Une documentation en ligne permet de s'assurer que les réutilisateurs des données disposent toujours de la version la plus à jour. Des portails de données, tels que www.data.gouv.fr, proposent des espaces dédiés à la documentation du jeu de données. Vous pouvez également héberger votre documentation sur des sites web statiques. Si le jeu de données a pour vocation de circuler en interne de votre organisation, nous

vous conseillons à minima de proposer une documentation dans un fichier séparé aux données. Le fichier contenant les données doit être réservé à la manipulation de ces dernières. Le fichier contenant la documentation a lui pour vocation d'informer sur la nature des données et sur la structure des fichiers.

Dans le cadre de la publication des données de sauvetage en mer (opérations coordonnées par les CROSS), un [site statique](#) a été créé afin de présenter la documentation du jeu de données.

SECMAR [Rechercher] [Données brutes] [Cartographie]

Accueil
Préambule et avertissement
Changements sur le jeu de données

Concepts métier ▶

Modèle de données
Schéma de données
operations
resultats_humain
flotteurs
operations_stats
Tables de codes
CROSS
Événements déclenchants
Bilan humain
Autorités des moyens
Types précis de flotteurs
Ports de référence pour la marée

» Cacher le menu

Jeu de données SECMAR

Ce jeu de données contient toutes les données statistiques disponibles informatiquement sur les interventions d'assistance et de sauvetage coordonnées par les CROSS (Centres régionaux opérationnels de surveillance et de sauvetage). Il renseigne, pour chaque opération d'assistance ou de secours coordonnée en eaux françaises :

- quel était le motif d'intervention ;
- quand, comment et par qui l'alerte a été donnée ;
- le contexte météo et géographique de l'opération ;
- quels flotteurs étaient impliqués ;
- quels moyens aériens, nautiques ou terrestres ont été engagés ;
- quel a été le bilan humain de l'opération.

Ce jeu de données contient des données propres à un métier complexe et chaque fichier comporte plusieurs centaines de milliers de lignes. Nous vous recommandons de lire attentivement [la documentation](#) mise à votre disposition avant toute analyse.

La publication des moyens engagés pour chaque opération sera réalisée ultérieurement.

Ce jeu de données est produit par la [Direction des Affaires Maritimes](#).

4.1 Description générale du jeu de données

Afin de donner un aperçu rapide des informations mises à disposition, il est conseillé de commencer la documentation par une description synthétique du jeu de données.

La description peut couvrir les points suivants :

- Une description générale des données ;
- La liste des fichiers mis à disposition ;
- La description du format des fichiers ;
- La fréquence de mise à jour.

Description générale du jeu de données du Répertoire national des élus
Description générale du jeu de données du Répertoire national des élus

Données du Répertoire national des élus

Ce jeu de données provient d'un service public certifié

Le Répertoire National des Elus (RNE) a pour finalité le suivi des titulaires d'un mandat préfectoral et par les services du ministère de l'intérieur, notamment sur la base des données d'enregistrement des candidatures.

Les données du RNE sont structurées par mandat. Neuf fichiers sont publiés ici :

- 1- les conseillers municipaux ;
- 2- les conseillers communautaires ;
- 3- les conseillers départementaux ;
- 4- les conseillers régionaux ;
- 5- les conseillers de l'Assemblée de Corse ;
- 6- les représentants au Parlement européen ;
- 7- les sénateurs ;
- 8- les députés ;
- 9- les maires.

Un élu qui dispose de plusieurs mandats figurera dans plusieurs fichiers.

Les fichiers mis en ligne sont actualisés trimestriellement.

Les données relatives à la profession sont déclaratives.

Les demandes de rectification doivent être adressées directement par courriel à la préfecture du département (contact@nom-du-departement.gouv.fr). Aucune demande de rectification adressée sur un autre canal ne sera prise en compte. Les rectifications apportées dans le RNE ne seront pas reportées immédiatement sur les données publiées. Les rectifications seront prises en compte pour la publication suivante.

4.2 Description du mode de production des données

La structure de votre jeu de données ainsi que son contenu sont intrinsèquement liés au contexte de production des données. La description de cet environnement métier au réutilisateur des données est indispensable :

- Comment ont été produites les données (saisie manuelle, collecte automatique, etc.) ?
- Quels sont les acteurs producteurs des données ? Si les données sont produites par plusieurs acteurs, quel modèle de gouvernance est mis en place pour centraliser les données ?
- Les données sont-elles exhaustives ? Présentent-elles des limites dans leur qualité ?

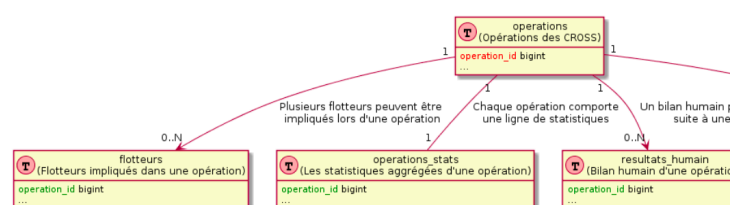
Il est également conseillé de préciser les points d'attention et précautions d'usage relatifs à la manipulation des données. Certains jeux de données ne peuvent pas être utilisés dans certains contextes, ou possèdent des limitations qui rendent impossible certaines analyses. Par exemple, conformément à l'article R112 A-3 du Livre des procédures fiscales, la réutilisation du jeu de données « Demandes de valeurs foncières » ne peut pas avoir ni pour objet ni pour effet de permettre la ré-identification des personnes liés à des transactions immobilières.

À partir de la description du mode de production du jeu de données, le réutilisateur comprend la structure du jeu, la nature des données mais aussi les possibles manques ou incohérences figurant dans le fichier.

4.3 Description du modèle de données

Le jeu de données que vous mettez à disposition peut se composer de plusieurs éléments liés entre eux. Le modèle de données décrit la structure logique du jeu de données. Il est conseillé de faire apparaître ce modèle de données à l'aide de schémas et de tableaux. Si votre jeu de données se compose de plusieurs fichiers, il est souhaitable de faire apparaître les relations entre eux.

La [documentation](#) du jeu de données des opérations de sauvetage en mer décrit le modèle de données utilisé. Ce modèle de données permet de comprendre rapidement les relations qui



Le schéma ci-dessus fait apparaître les différentes tables représentant le jeu de données SECMAR. Des tableaux détaillant tous les fichiers et toutes les relations sont disponibles ci-dessous.

unissent les différents fichiers du jeu de données.

Si vous publiez des données tabulaires, vous pouvez produire un tableau récapitulatif indiquant, pour chaque colonne :

- le nom de la colonne
- son type de données (entier, chaîne de caractères, nombre décimal etc.)
- la description de la donnée contenue dans cette colonne
- une ou plusieurs valeurs d'exemple.

La documentation du jeu de données de sauvetage en mer présente un tableau récapitulatif des différentes colonnes. La description des champs permet de faire le lien avec le fichier de données,

ce qui facilite la lecture des données.

Colonne	Type	Description	E
operation_id	integer(\$int64)	Le numéro unique de l'opération.	11199203
type_operation	string	Le type d'opération coordonné par le CROSS. SAR (search and rescue) : vie humaine en danger ; MAS (maritime assistance service) : assistance aux navires ; SUR : sûreté des navires ; DIV : autres cas.	SAR
pourquoi_alerte	string	Pourquoi l'alerte a-t-elle été donnée	Événement

Les termes employés dans un jeu de données sont propres à un environnement métier. S'il existe des termes complexes ou des énumérations, il est conseillé de faire apparaître un lexique de ces valeurs. Cet effort de définition fait gagner un temps considérable au réutilisateur lorsqu'il s'approprie le jeu de données. De plus, le partage de définitions formalisées et uniques permet de prévenir des contresens dans l'exploitation des données.

La base de données de demande de valeur foncière recense l'ensemble des transactions immobilières intervenues au cours des cinq dernières années. Le vocabulaire utilisé dans ce jeu de données est issu d'un environnement administratif, parfois difficile à appréhender. La Direction générale des Finances publiques met à disposition une [documentation](#) qui comprend notamment un lexique de définition des termes rencontrés. Ce lexique facilite l'appropriation et la réutilisation des données par des acteurs tiers.

Lexique immobilier DGFIP

Nature et valeur de la mutation

- **Disposition** : une disposition constitue une unité d'analyse juridique. Un document peut en comporter plusieurs, mais seules celles concernant les mutations à titre onéreux sont restituées. Ainsi, une vente simple est représentée par une seule disposition rattachée à un prix (cf : valeur foncière) pour laquelle sont identifiés les rôles et les droits détenus pour chacune des parties sur chaque immeuble. Un document comportant une vente ainsi qu'une division de parcelle contribue à la création de deux dispositions, l'une concernant la division et l'autre la vente car il s'agit de 2 unités d'analyse juridique distinctes.
- **Document ou acte** : entité retraçant un ou plusieurs événements juridiques et/ou fiscaux qui portent sur une ou plusieurs personnes et/ou un ou plusieurs biens. Un document est composé de une ou plusieurs dispositions.
- **Mutation à titre onéreux** : transfert de propriété moyennant une contrepartie.
- **Mutation immobilière** : transfert de la propriété d'un immeuble pouvant intervenir à titre onéreux (vente) ou gratuit (donation).
- **VEFA** : vente en état futur d'achèvement. Cette vente dite souvent 'sur plan' rend l'acquéreur propriétaire des sols, des constructions existantes et à venir jusqu'à achèvement de son bien.
- **Valeur foncière déclarée** : il s'agit du prix du ou des biens immobiliers déclarés dans le cadre d'une mutation à titre onéreux. À chaque disposition correspond un prix.

4.4 Description des métadonnées

Qu'est-ce qu'une métadonnée ?

Une métadonnée est une donnée qui décrit ou définit une autre donnée. Dans la vie courante, l'étiquette d'un produit fournit des informations/métadonnées sur le produit (origine, composition, date de péremption, etc.). Appliqué aux jeux de données, les métadonnées sont des descriptions normalisées du contenu du jeu

Des formats standards de métadonnées existent afin de faciliter leur collecte, leur recherche et leur traitement automatique. La plateforme data.gouv.fr propose un module qui renseigne directement les métadonnées d'un jeu de données. Les métadonnées retenues sont les suivantes :

- Titre
- Sigle
- Description
- Licence
- Fréquence de mise à jour

- Mots clés
- Couverture temporelle
- Couverture spatiale
- Granularité spatiale
- Mode privé

Il est possible de consulter le guide de publication d'un jeu de données sur data.gouv.fr pour avoir des informations complémentaires sur ces métadonnées. La description des métadonnées apportera à votre jeu de données une meilleure visibilité sur les catalogues.

4.5 Description des changements majeurs

En pratique, il est souhaitable que le modèle de données et la nature de vos données n'évoluent pas au fil du temps. Toutefois, des changements dans la structure des données, dans le mode de collecte ou dans les dispositions réglementaires peuvent affecter le jeu de données. Dans cette situation, il est conseillé de maintenir une liste de ces changements. Cette liste peut faire figurer la date, la version des données (si vous versionnez vos données) et la nature du changement. Si nécessaire, il est possible d'indiquer des liens, comme par exemple lorsque des changements sont introduits par une modification du code de transformation des données.

La [documentation](#) du jeu de données des sauvetages en mer comporte une section "Changement sur le jeu de données". Cette section référence les changements du jeu de données en renseignant les informations suivantes :

- La date du changement
- La nature du changement

Changements sur le jeu de données

Les changements de schéma du jeu de données SECMAR sont répertoriés ci-dessous.

2018-11-13

Ajout de la colonne `immatriculation_omi` dans `flotteurs` et `nombre_navires_mas_omi` dans `operations_stats`.

Commits: [predisauvetage@a0858f](#), [predisauvetage@6a6e55](#)

2018-11-01

Modification de la colonne `concerne_snosan` dans `operations_stats` pour prendre en compte les opérations qui n'ont pas de flotteurs impliqués mais qui sont pour des événements qui s'appellent "jeu de données" à du loisir.

Issue: [Issue #123](#)

- Les liens associés au changement

4.6 Points de contact

Les réutilisateurs des données peuvent avoir des questions à propos des fichiers mis à disposition. Il est conseillé de proposer un espace d'échange entre les producteurs et réutilisateurs des données. Il est préférable que cet espace d'échange soit public afin qu'il puisse bénéficier aux personnes qui auraient des questions similaires. De plus, la collecte des retours d'usage permettra d'améliorer votre documentation de manière incrémentale.

5 Éléments organisationnels

La démarche de partage des données implique un engagement des équipes métiers. Afin de faciliter le processus et d'en tirer le maximum de bénéfices, il est indispensable de mettre en place une stratégie organisationnelle :

- **Identifier le coordinateur de la démarche** : il a pour mission de publier les jeux de données, de s'assurer que leurs mises à jour sont effectuées et d'animer la vie des jeux de données sur la plateforme (répondre aux commentaires, mettre en valeur les réutilisations, etc.). Le coordinateur travaille en lien direct avec les équipes métiers afin de comprendre les problématiques techniques.

- **Élaborer un processus de rétroaction** : lors de l'exploitation des jeux de données, les réutilisateurs peuvent identifier des anomalies ou des problèmes de qualité ou encore proposer des améliorations aux jeux de données publiés. Il est nécessaire d'instaurer un canal de rétroaction afin d'intégrer ces remarques dans les processus métiers et ainsi améliorer la qualité des jeux de données. Lorsque le jeu de données est publié sur un catalogue de données en ligne, le coordinateur a intérêt de communiquer aux équipes métiers les anomalies ou les suggestions proposées dans les commentaires du jeu de données. Si le jeu de données est partagé entre équipes ou organisations, le coordinateur a intérêt de proposer un point de contact (adresse e-mail, outil partagé de communication, etc.) afin de collecter les différents retours et faire remonter les informations auprès des équipes métiers.
- **Valoriser la publication de nouveaux jeux de données à fort impact ou la mise en place d'une démarche de circulation des données en interne** : la rédaction d'articles de blog, le relais sur les réseaux sociaux et l'organisation de *data session* et de hackathons sont des moyens de mettre en avant la démarche de circulation, de valoriser le travail des équipes métiers et de favoriser les réutilisations.
- **Animer la communauté de réutilisateurs** : il est nécessaire d'engager un dialogue constant avec la communauté de réutilisateurs des données afin de faciliter la réutilisation des données et évaluer les impacts. Il est notamment conseillé de proposer un espace d'échange en ligne (sous forme de modules de commentaire ou de forum) et d'organiser régulièrement des ateliers de discussion entre producteurs et réutilisateurs.

5.1 Les prochaines étapes

La préparation de votre jeu de données à la circulation est le début de votre démarche d'ouverture et de circulation des données. Afin de poursuivre cette démarche, plusieurs questions peuvent se poser :

- Les jeux de données doivent-ils circuler en open data ou entre organisations ?
- Comment publier les jeux de données sur data.gouv.fr ?
- Comment faire circuler les jeux de données entre organisations ?

Etalab propose différents guides pour vous accompagner sur ces sujets :

- Cadre légal de l'open data
- Publication des données sur data.gouv.fr
- Circulation des données entre acteurs

5.2 Informations complémentaires

Ce guide est à l'attention des acteurs qui souhaitent préparer leurs jeux de données à la circulation. Si vous avez des remarques sur son contenu, si vous souhaitez proposer des améliorations ou si vous ne trouvez pas de réponse à vos questions, vous pouvez nous contacter à l'adresse suivante : ouvert@data.gouv.fr