Schémas

Etalab

etalab gouv.fr

Table des matières

Introduction

> **Version bêta**
>
> Ce guide a été publié initialement fin mars 2020. Il est amené à être ajusté suite aux retours de nos partenaires et lecteurs. Vous êtes invités à entrer en contact avec nous.

> **Lexique : Schémas de données**
>
> Les schémas de données permettent de décrire la structure d'un jeu de données. Ils indiquent clairement quels sont les différents champs, comment sont représentées les données, quelles sont les valeurs possibles etc.
>
> Synonymes : *modèle de données*, *modèle logique de données*, *schéma*.

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 schéma de données 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 schéma de données facilite le travail d'agrégation des données - ce schéma devient donc un standard pour votre communauté ;
- La formalisation d'un schéma de données assure une pérennité des fichiers dans le temps ;
- La documentation d'un schéma de données existant est déjà rédigée et accessible.

Il est également possible de vérifier la conformité d'un jeu de données vis-à-vis d'un schéma de données, ce qui permet de valider un premier niveau de qualité de votre jeu de données. Par ailleurs, il est aussi possible de générer des jeux de données d'exemple ou de proposer des formulaires de saisie standardisés.

> **schema.data.gouv.fr**
>
> Le site schema.data.gouv.fr est l'initiative de la plateforme data.gouv.fr. L'objectif de ce site est de référencer les schémas de données publiques existants en France.

À qui s'adresse ce guide ?

Ce guide s'adresse à **des personnes susceptibles de créer des schémas de données**. Vous pouvez vous trouver dans cette situation si vous envisagez de partager des données avec des partenaires ou à tout le monde en open data.

À quoi sert-il ?

Ce guide propose de vous accompagner lors des phases nécessaires à la création d'un schéma de données et à son référencement sur schema.data.gouv.fr le cas échéant.

1. **Phase d'investigation** : envisager de créer un schéma de données ;
2. **Phase de concertation** : rassembler plusieurs parties prenantes pour créer un schéma de données ;
3. **Phase de construction** : implémenter le schéma de données obtenu après la phase de concertation.

Il propose un processus à suivre, des bonnes pratiques et des outils.

> **Conseil de lecture**
>
> Nous vous recommandons de lire une première fois ce guide **en intégralité** afin de prendre connaissance des différentes phases. Vous pourrez ensuite vous référer aux pages pertinentes au fur et à mesure de votre avancée.

Sources

Ce guide s'inspire du contenu rédigé par de nombreux partenaires, listés par ordre alphabétique :

— Charles Nepote
— Datactivist
— La FING
— OpenDataFrance

Merci à eux !

Phase d'investigation

> **Lexique : Phase d'investigation**
>
> La phase d'investigation est la première phase de la création d'un schéma de données. Cette phase a pour finalité de s'asssurer que la création d'un schéma est pertinente et vise à aboutir à la décision de continuer ou de choisir une autre alternative.

Étapes à suivre

Afin de déterminer s'il est nécessaire de créer ou non un schéma de données, nous vous recommandons de suivre les étapes suivantes :

1. Lire attentivement les différentes sections de ce guide ;
2. Organiser une réunion réunissant des acteurs métiers, techniques et de potentiels réutilisateurs. Lors de cette réunion, vous débatterez de la pertinence de la création de votre schéma de données;
3. Entrer en contact avec les équipes d'Etalab et leurs partenaires en référençant votre schéma, afin de bénéficier de conseils lors de la création de votre schéma de données, d'une visibilité accrue pour celui-ci et d'une assistance d'experts.

Exemples

Situations favorables à la création d'un schéma de données

Ces situations sont des exemples où il est pertinent de créer un schéma de données :

> — Le ministère chargé des transports souhaite consolider une base nationale des lieux pouvant servir de points de covoiturage. Les collectivités territoriales sont en charge de la création, recensement et aménagement de ces lieux.

Il est pertinent de créer un schéma de données car un grand nombre de producteurs de données doivent produire le même jeu de données. Un schéma facilitera la diffusion des prérequis, permettra la validation des données et facilitera l'agrégation nationale.

> — L'INSEE souhaite diffuser le Code Officiel Géographique. Il rassemble des données sur des communes, des cantons, des arrondissements, des départements, des régions et des pays. Ce fichier est actualisé tous les ans.

Il est pertinent de créer un schéma car ces données sont des données de référence. Un grand nombre de réutilisateurs est susceptible d'utiliser ces données. Il est primordial que ces réutilisateurs aient accès à une documentation de qualité, que la structure des fichiers des données reste stable dans le temps et que les données mises à disposition soient de bonne qualité.

Situations où le référencement d'un schéma sur schema.data.gouv.fr ne semble pas nécessaire

Ces situations sont des exemples où il ne semble pas pertinent de créer ou diffuser un schéma :

> — Une administration centrale diffuse des statistiques d'activité d'un bureau, en open data, de manière annuelle.

Avec ces seules informations, la création d'un schéma ne semble pas nécessaire. En effet, il n'y a qu'un seul producteur et le potentiel de réutilisation semble limité.

Bénéfices des schémas de données en interne

Bien qu'il ne paraisse pas nécessaire dans certaines situations de créer et diffuser un schéma, vous pouvez choisir de le faire. En effet, les schémas de données comportent de nombreux avantages (documentation, montée en qualité, réutilisations, etc.) qui sont bénéfiques, même lorsque les données sont utilisées uniquement en interne.

Points de sortie

À l'issue de cette phase, vous devriez :

- Connaître les schémas de données ;
- Être en mesure de décider si votre projet requiert la création d'un schéma de données ;
- Savoir si votre schéma de données devra être référencé à terme sur schema.data.gouv.fr.

Phase de concertation

Lexique : Phase de concertation

La phase de concertation est la phase centrale de la création d'un schéma de données. C'est l'étape où plusieurs parties prenantes (producteurs, réutilisateurs, experts métiers et techniques) se rassemblent pour définir et spécifier les éléments essentiels à la constitution de ce schéma.

Pour spécifier un schéma de données, il est nécessaire de définir :

- les champs ;
- les types associés de ces champs (une date, un nombre, une chaîne de caractère etc.) ;
- les contraintes de chaque champ (entier positif, texte dans une liste fermée etc.) ;
- la description de chaque champ ;
- une documentation associée au schéma de données décrivant le contexte, les acteurs, les cas d'usages.

Procédure de collaboration

Nous conseillons de mener cette phase de concertation en travaillant sur un document partagé, accessible en ligne, tel qu'un Framapad ou Google Doc. L'important est que plusieurs contributeurs puissent contribuer (modifier ou mettre des commentaires) sans avoir besoin d'être présents physiquement ou de recevoir des versions intermédiaires par e-mails.

En complément de ce document partagé, nous vous conseillons d'organiser plusieurs réunions afin de débattre du schéma de données à produire. L'implication d'une multitude d'acteurs est clé : vous devez rassembler des producteurs, experts métiers, experts techniques et réutilisateurs. La richesse des profils et des enjeux permettra d'aboutir à une solution la plus adaptée.

> **Référencer votre schéma**
>
> Référencer votre schéma sur schema.data.gouv.fr vous permettra de bénéficier de conseils de la part d'Etalab et de ses partenaires institutionnels et associatifs. Découvrez comment référencer votre schéma en cours de concertation.

Grands principes

Nous avons listé ci-dessous plusieurs conseils qui vous permettront de construire un schéma de données de qualité.

- **Profiter de l'existant.** De nombreux standards existent déjà, qu'ils concernent des formats de données ou des formats de champs. Certains standards sont devenus incontournables aujourd'hui, comme ISO-8601 pour les dates ou WGS 84 pour les coordonnées géographiques.
- **Identifier et associer l'écosystème.** Les personnes/organisations que vous associez sont la meilleure garantie d'un schéma de données efficace et largement adopté, permettant d'aboutir à un véritable standard.
  - Les producteurs d'une part qui connaissent la réalité de leurs données, de la collecte, etc. et qui ont leurs propres usages.
  - Les usagers d'autre part, leurs besoins et leurs difficultés d'autres part, qu'ils soient déjà connus, « sous le radar » ou en devenir.
- **Prendre le temps.** Un schéma de données est susceptible de concerner beaucoup de producteurs et d'usagers. Sa modification peut avoir un impact important. Il est donc crucial de prendre le temps d'obtenir tous les retours avant de publier un schéma utilisable par le plus grand nombre. Un schéma de données devrait être publié quand il est prêt, non pas en fonction d'un impératif de délai.
- **Lever les implicites et les ambiguïtés.** Le diable est dans les détails… Toutes les spécifications d'un schéma de données doivent être les plus claires possibles, y compris pour des cas/données

qui n'existent pas encore mais pourraient apparaître à l'avenir.
- **Éviter la redondance mais sans l'exclure absolument.** Trois champs pour définir une latitude et une longitude (`latitude`, `longitude`, `lat-lon`) est inutilement redondant. Toutefois, préciser le nom d'une commune en plus de son code INSEE rend les données plus faciles à lire et à exploiter.
- **Utiliser des données pivot relevant d'un référentiel ouvert** pour relier les données à d'autres données, par exemple l'utilisation du numéro SIREN pour identifier des organisations. Ce principe permet aussi d'éviter l'abondance de détails et d'aller à l'essentiel : l'obtention d'informations complémentaires se fera par le biais d'un autre référentiel.

> **Exemples à votre disposition**
>
> Vous pouvez parcourir des fichiers de schémas sur schema.data.gouv.fr pour faciliter votre travail. Consultez par exemple le schéma des lieux de stationnement.
>
> En complément, nous avons rédigé un guide dédié à la préparation de jeux de données qui pourrait vous être utile pour définir votre schéma.

Points de sortie

À l'issue de cette phase, vous devriez :

- Avoir réuni divers partenaires afin de collaborer sur votre schéma de données;
- Avoir décidé des différents champs de votre schéma de données, leurs types et définitions et produit une documentation associée.

Phase de construction

> **Lexique : Phase de construction**
>
> La phase de construction consiste à implémenter techniquement le schéma de données obtenu après la phase de concertation. Pour cela, il est nécessaire de choisir un standard technique, créer les fichiers requis, les tester et les diffuser.

Durant cette phase, vous devez mobiliser des personnes possédant des compétences techniques. Cette phase consiste à transcrire les décisions prises lors de la phase de concertation pour un schéma de données.

Choisir un standard technique pour la description de votre schéma de données

> **Lexique : Standard**
>
> On utilise les termes « normes » et « standards » pour décrire un référentiel commun et documenté destiné à harmoniser l'activité d'un secteur.

Il existe plusieurs standards techniques pour les schémas de données. Le standard est à choisir en fonction de la nature des données concernées et des habitudes de l'écosystème produisant ou réutilisant les données liées au schéma.

Les principaux standards techniques sont les suivants :

- Table Schema : adapté pour la description de données tabulaires (sous forme de tableurs ou de CSV). Ce standard technique utilise le format JSON
- JSON Schema : adapté pour la description de données avec une notion de hiérarchie. Ce standard utilise le format JSON
- XML Schema Definition (XSD) : adapté pour la description de données avec une notion de hiérarchie. Ce standard utilise le format XML

Notez que tous ces standards techniques sont supportés par schema.data.gouv.fr.

> **Aller au-delà de la documentation texte**
>
> Un schéma de données décrit uniquement par du texte ou par un tableau se prive de nombreux avantages, notamment celui de l'interopérabilité entre différents systèmes informatiques.
>
> Les schémas de données décrits par des standards techniques permettent, en plus d'une documentation textuelle ou sous forme d'un tableau, de valider que des données correspondent à un modèle de données, d'agréger des données similaires, de générer automatiquement des données respectant un schéma.

Créer votre schéma de données

Une fois un standard technique choisi, il faudra créer les fichiers requis pour modéliser vos données. La documentation de chaque standard technique décrit le contenu des fichiers à renseigner. Reportez-vous aux documentations respectives pour tirer parti des fonctionnalités avancées offertes : types de données et contraintes sur les valeurs en particulier.

Il est souvent possible de vérifier qu'un fichier correspond à un standard à l'aide d'outils en ligne ou en ligne de commande. Utilisez ces outils pour vérifier que vos productions correspondent au standard.

> **Exemples à votre disposition**
>
> Pour un schéma au format Table Schema, nous mettons à votre disposition un modèle de départ

> pour créer un dépôt Git contenant un schéma au format Table Schema.
>
> Pour les autres formats de schémas, nous vous recommandons de consulter les schémas et dépôts Git listés sur schema.data.gouv.fr.

Documenter votre schéma de données

En complément du fichier du schéma de données, nous vous conseillons de rédiger a minima deux documents complémentaires :

- **une documentation générale** : vous indiquerez le contexte, les modalités de production des données, le cadre juridique, la finalité, les cas d'usage etc. Ce fichier est traditionnellement rédigé en Markdown et nommé `README.md` ;
- **un fichier répertoriant les changements** : permettant de suivre les modifications, d'une version à une autre. Ce fichier est traditionnellement rédigé en Markdown et nommé `CHANGELOG.md`.

La présence de ces fichiers représente un package complet (documentation, liste des changements et schéma de données décrit dans un standard technique), apprécié des réutilisateurs. schema.data.gouv.fr se repose sur ces éléments pour intégrer votre documentation et votre liste de changements sur une page web.

> **Exemples à votre disposition**
>
> Vous pouvez consulter la documentation et la liste des changements du schéma des lieux de stationnement.

Publier et diffuser votre schéma de données

Une fois votre schéma de données créé, il est nécessaire de le publier et de le diffuser pour que d'autres personnes puissent en bénéficier. Nous vous recommandons de publier vos schémas de données en tant que logiciels libres, sur votre forge de développement ou par le biais de GitLab ou GitHub.

Vous bénéficierez alors des avantages habituels des dépôts de code Git en ligne : historique des modifications, fonctionnalités de tickets ou de demandes de modifications. Utilisez un compte d'organisation (dédié à votre entreprise, direction, service, ministère) et non votre compte personnel afin d'assurer une URL stable dans le temps.

> **Exemples à votre disposition**

> Vous trouverez plusieurs dépôts Git de schémas sur schema.data.gouv.fr. Consultez par exemple le dépôt Git décrivant les lieux de stationnement à l'aide d'un schéma TableSchema sur GitHub.

Référencer votre schéma de données sur schema.data.gouv.fr

Pour faciliter la découverte de votre schéma de données et des données sous-jacentes, nous vous recommandons de le faire référencer sur schema.data.gouv.fr. Nous avons rédigé une page dédiée à ce sujet décrivant les plus-values, prérequis et démarches à suivre.

Faire évoluer votre schéma de données

Une fois votre schéma de données défini et implémenté, le travail ne s'arrête pas là. Au-delà du besoin de diffusion et de promotion, il est probable que vous deviez faire des modifications : clarifications de la documentation, corrections d'erreurs, évolutions du cadre réglementaire, etc. Autant de raisons où il est nécessaire de mettre en œuvre une nouvelle version.

Posséder un dépôt Git pour votre schéma de données vous permettra d'avoir plusieurs versions et tags. Notez que schema.data.gouv.fr supporte plusieurs versions pour un même schéma de données et affiche les modifications effectuées au fur et à mesure, dès lors que ces modifications sont renseignées dans un fichier dédié.

Points de sortie

À l'issue de cette phase, vous devriez :

- Avoir implémenté votre schéma de données dans un des standards reconnus ;
- Avoir publié votre travail en ligne, dans un répetoire Git dédié ;
- Avoir pris contact avec les équipes de schema.data.gouv.fr dans le but de référencer votre schéma de données si nécessaire.

Intégration avec schema.data.gouv.fr

schema.data.gouv.fr est l'initiative de data.gouv.fr de référencement des schémas de données publiques pour la France. Cette plateforme de référencement nationale permet un accès aux schémas produits par différents acteurs et facilite l'intégration avec des systèmes informatiques par le biais de standards, d'URLs stables, de processus de validation et d'API.

Vous trouverez ci-dessous une capture d'écran de l'interface de schema.data.gouv.fr pour le schéma dédié aux lieux de covoiturage.

FIGURE 1 – Capture d'écran de l'interface de schema.data.gouv.fr

Qui peut référencer des schémas ?

Tout acteur est libre de proposer le référencement de schémas.

Concrètement, vous pouvez être une administration, une entreprise privée, une association, un citoyen etc.

Quels schémas de données sont acceptés ?

schema.data.gouv.fr accepte des schémas de données décrivant des données publiques.

Les schémas de données sont acceptés dès lors que leur l'existence est justifiée par voie :

— **réglementaire** : c'est une disposition réglementaire qui est à l'origine de la définition du schéma de données ;
— **d'usage** : la réutilisation des données décrites par le schéma bénéficie à un grand nombre ou de nombreux producteurs sont amenés à utiliser ce schéma de données.

Etalab se réserve le droit de refuser l'ajout de schémas en motivant son refus. Nous vous encourageons à initier une discussion préalablement à l'ouverture d'une *pull request*.

Quand référencer son schéma de données ?

Nous vous invitions à référencer votre schéma de données le plus tôt possible, **dès la phase d'investigation**. En référençant celui-ci en amont, vous bénéficierez de l'accompagnement d'Etalab et de partenaires tout au long de la création de votre schéma de données : de l'investigation à la publication sur schema.data.gouv.fr.

Vous pouvez référencer votre schéma de données en ouvrant un ticket sur GitHub ou en entrant en contact avec notre équipe par e-mail Nous avons créé une page dédiée pour détailler la procédure. Nous tenons à jour une liste de schémas actuellement en phase d'investigation ou de construction sur cette même page.

Quels schémas de données sont acceptés ?

schema.data.gouv.fr accepte des schémas de données décrit par un standard technique (voir la page "Phase de construction" de ce présent guide). Les schémas de données décrits uniquement par de la documentation textuelle ou des tableaux ne sont pas acceptés.

Standards techniques supportés

Les standards techniques de schémas de données actuellement supportés sont les suivants :

— Table Schema : adapté pour la description de données tabulaires (sous forme de tableurs ou de CSV). Ce standard technique utilise le format JSON
— JSON Schema : adapté pour la description de données avec une notion de hiérarchie. Ce standard utilise le format JSON

— XML Schema Definition (XSD) : adapté pour la description de données avec une notion de hiérarchie. Ce standard utilise le format XML

Prérequis de validation des schémas de données

> **Lexique : Validation d'un schéma de données**
>
> La validation d'un schéma de données est l'étape qui permet de vérifier si celui-ci est conforme au standard technique sélectionné et aux prérequis de schema.data.gouv.fr. Cette étape s'intéresse uniquement au schéma de données et à la façon dont il est publié.
>
> Il ne faut pas confondre la validation d'un schéma avec le fait de vérifier que des données correspondent à un schéma.

Pour tous les types de schéma de données, il faut que :

— votre schéma de données soit sur un dépôt Git, à raison d'un dépôt par schéma. Ce dépôt doit pouvoir être cloné depuis Internet sans authentification préalable ;
— votre dépôt Git doit comporter des tags indiquant les versions de votre schéma de données. Ces versions doivent respecter la gestion sémantique de version semver, sous la forme `1.3.2` par exemple ;
— votre dépôt doit comporter un fichier `README.md` à la racine contenant une documentation du schéma de données indiquant par exemple le contexte de production, la gouvernance ;
— passer avec succès les tests spécifiques au type de schéma de données que votre dépôt contient.

> **Critères complets de validation**
>
> Cette page présente les grands principes de validation des schémas de données. Pour connaître en détail les prérequis propres à chaque type de schéma de données et accéder à des exemples, consultez la page dédiée à la validation des schémas de données.

Points de sortie

À l'issue de cette phase, vous devriez :

— Avoir pris connaissance des procédures de validation en place sur schema.data.gouv.fr ;
— Avoir un dépôt Git conforme aux prérequis de schema.data.gouv.fr ;
— Avoir effectué votre demande de référencement.