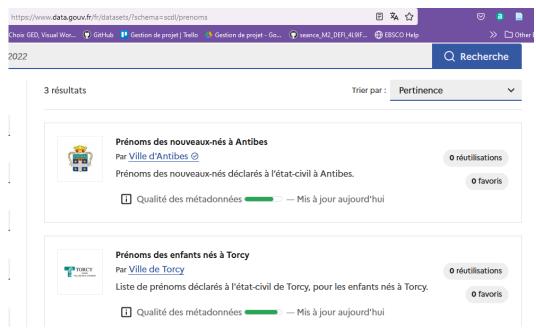
I-Choix du Schema

Tout d'abord, nous avons choisi un schema qui n'est ni en construction, ni uniquement publié mais qui a été adopté afin de garantir un minimum de cohérence et de qualité dans les données choisies.

Ensuite, pour être à peu près sûre qu'on n'aura pas beaucoup de problème de nettoyage : nous avons regardé dans la partie « données » du schéma proposé, les jeux de données qui ont adopté ce schéma. Le schéma choisi est « Prénoms des nouveaux-nés » (accessible via le lien ci-après https://schema.data.gouv.fr/scdl/prenoms/ vérifié le 04-12-2023).

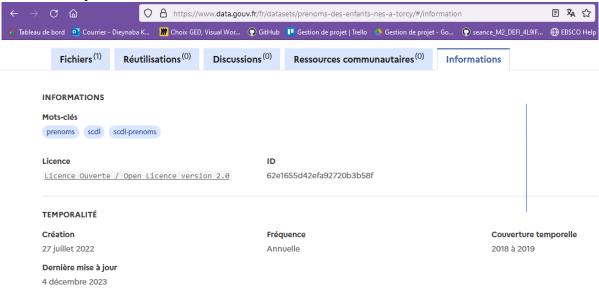


Nous avons repéré quelques noms de villes qui ont adopté le schéma « Prénoms des nouveaux nés ». On a repéré sur Open data la ville de Torcy et d'Antibes par exemple.



Nous avons ensuite vérifié les informations sur les jeux de données et on a choisi Antibes car la période couverte était plus large (2011-2022) que celle de Torcy (2018-2019)





Ville d'Antibes

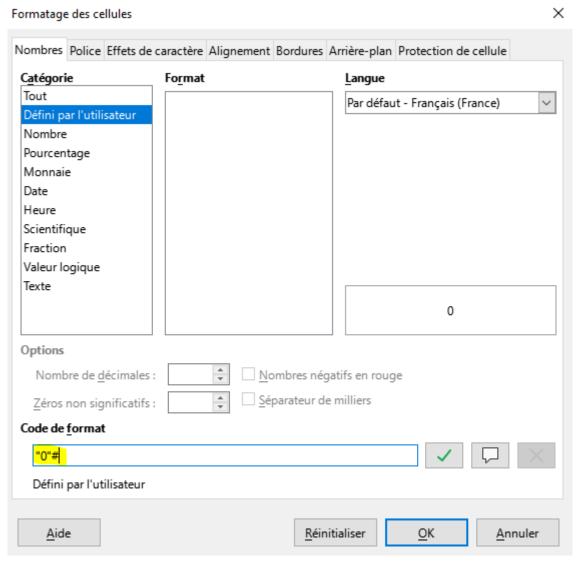


Et enfin on s'est rendu sur le site de la ville pour retrouver le jeu de données. Prénoms des nouveaux nés de la ville d'Antibes accessible via le lien ci-après : https://trouver.datasud.fr/dataset/prenoms-des-nouveaux-nes-de-la-ville-antibes-juan-les-pins/ resource/c49b3327-31d1-4d34-b90b-2d1585a20327

II-Analyse du jeu de données

Nous avons procédé au Sprint Qualité du jeu de données puis nous l'avons vérifié sur https://publier.etalab.studio/fr/. Nous avons nommé le fichier corrigé « antibes_prenoms_naissance_corrige.csv ».

Nous avons essayé de partir d'un fichier xls qu'on a enregistré sous le format CSV pour palier au problème d'affichage des caractères accentués (voir notre rendu sprint qualité). Mais cela n'était pas la bonne approche car ce fichier ne pouvait s'ouvrir qu'avec Microsoft Excel. Nous avons donc désinstaller LibrOffice et l'avons réinstallé avec un JRE 64 bit car c'était un warning du logiciel. Ce qui explique pourquoi il y avait des problèmes d'affichage. Une fois ces mises à jour faites, il n'y avait plus de problème avec l'accentuation. Nous avons changé le format de la colonne « COLL_INSEE » afin que cela puisse correspondre au format du code postal d'Antibes.



Ensuite, nous nous sommes rendus sur https://publier.etalab.studio/fr/upload?schema=scdl %2Fprenoms . D'après l'analyse faite, il y avait 2 *erreurs dans la colonne « ENFANT_PRENOM »* et on les trouvées grâce au rapport d'erreurs proposé :

- 2 syntaxes de prénoms non valides aux lignes 7310, 7355.

▲ Votre fichier contient des erreurs.

Nous avons trouvé 2 erreur(s). Vous pouvez voir la liste des erreurs et corriger le fichier.

Corriger le fichier 🧳

Voir le rapport d'erreur ≡

Erreurs de contenu (2)

• Format incorrect: colonne ENFANT_PRENOM, ligne 7310

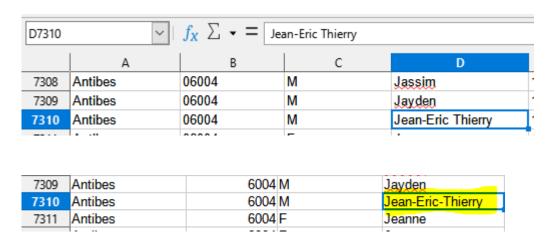
Prénom de nouveau(x)-né(s) constaté comme premier prénom dans les actes d'état-civil de l'année correspondante. Un acte de naissance peut désigner un nouveau-né avec plusieurs prénoms et le législateur a prévu que "tout prénom inscrit dans l'acte de naissance peut être choisi comme prénom usuel". La plupart du temps le premier prénom est le prénom d'usage initialement choisi par le(s) parent(s). Cette spécification ne retient donc que le premier prénom : si un nouveau-né est appelé 'Armelle Julia Blanche', seul 'Armelle' sera retenu pour constituer ce jeu de données. Un prénom composé comme 'Marie-Jeanne' compte pour un prénom complet. Attention, "l'alphabet utilisé doit être celui qui sert à l'écriture du français. Les caractères alphabétiques étrangers ne sont donc pas autorisés (par exemple le « l' »)". Outre les caractères alphabétiques, un prénom peut posséder un trait d'union, voire deux, comme dans 'Lou-Anne' ou 'Mohamed-El-Amine'. Des prénoms peuvent posséder une apostrophe comme dans Gwenc'Hlan ou N'Deye, voire peut-être deux. Nous considérons aussi qu'un prénom pourrait se terminer, voire débuter, par une apostrophe - cette dernière étant parfois utilisée en français pour marquer la suppression de la finale ou le début d'un mot, comme dans Boul' Mich'.

• Format incorrect : colonne ENFANT_PRENOM, ligne 7355

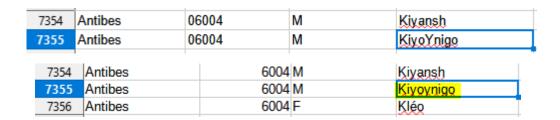
Correction des syntaxes incorrectes:

Sachant que c'était le même problème de contenus (syntaxe des prénoms) pour les 2 lignes citées plus haut, on a rectifié selon le problème .

- A la ligne 7310 : un prénom composé séparé par un espace et pas par deux tirets comme spécifié dans la documentation du schema. Donc, on a remplacé « Jean-Eric Thierry » par « Jean-Eric-Thierry ».



- A la ligne 7355 : on a gardé uniquement la première lettre du prénom en majuscule.



Conslusion:

Après la correction des 2 erreurs : on a eu un jeu de données qui valide le schéma « Prénoms des nouveaux-nés » et qui est donc conforme, prêt à être utilisé.

