IGR 204 - Mastère Big Data 2018

# M5 - REPORT

# Nouveaux-nés : Les prénoms les plus populaires en France

https://github.com/ThibaultChristophe/projetdatavisIGR204

Marine Ferrary - Amaël Chaigneau - François Lecerf - Jérome Divac et Christophe Thibault

## 1. Introduction

Nous allons présenter dans ce rapport le projet de Data Visualization que nous avons effectué dans le cadre du module IGR204. Nous articulerons cette présentation en abordant différents points successifs.

D'abord nous présenterons l'idée globale de notre visualisation, avec les différentes interactions possibles, ainsi que les données plus en détails, et expliquerons ce que nous avons visualisé et souhaité extraire comme information de ce jeu de données, ainsi que le type d'utilisateurs ciblés. Nous expliquerons ensuite les tâches représentatives possibles, ainsi que le fonctionnement du Bubbles chart, et l'architecture de la carte de France que nous avons retenue. Enfin, nous aborderons la partie relative à l'optimisation des vues. Enfin nous élargirons les perspectives en évoquant quelques pistes d'améliorations.

L'idée globale est de regarder à travers un Bubbles chart les prénoms donnés, pour une année sélectionnée, et de voir quelle est la répartition de ce prénom pour cette même année, à travers les différents départements français. Nous avons fait en sorte que plus un prénom est donné et plus la taille de la bulle correspondante est grande. Nous souhaitions également offrir à l'utilisateur la possibilité de ne choisir que les prénoms masculins, que les prénoms féminins, ou les deux à la fois. Nous avons fait en sorte que

les couleurs soient intuitives à l'œil, en choisissant des couleurs de bulles qui indiquent si le prénom est en hausse ou en diminution par rapport à l'année précédente. De plus, les départements français sont affectées d'un gradient de couleurs selon la proportion du prénom selectionné. Ces différents éléments sont résumés dans le sketch ci-dessous (voir Fig.1).

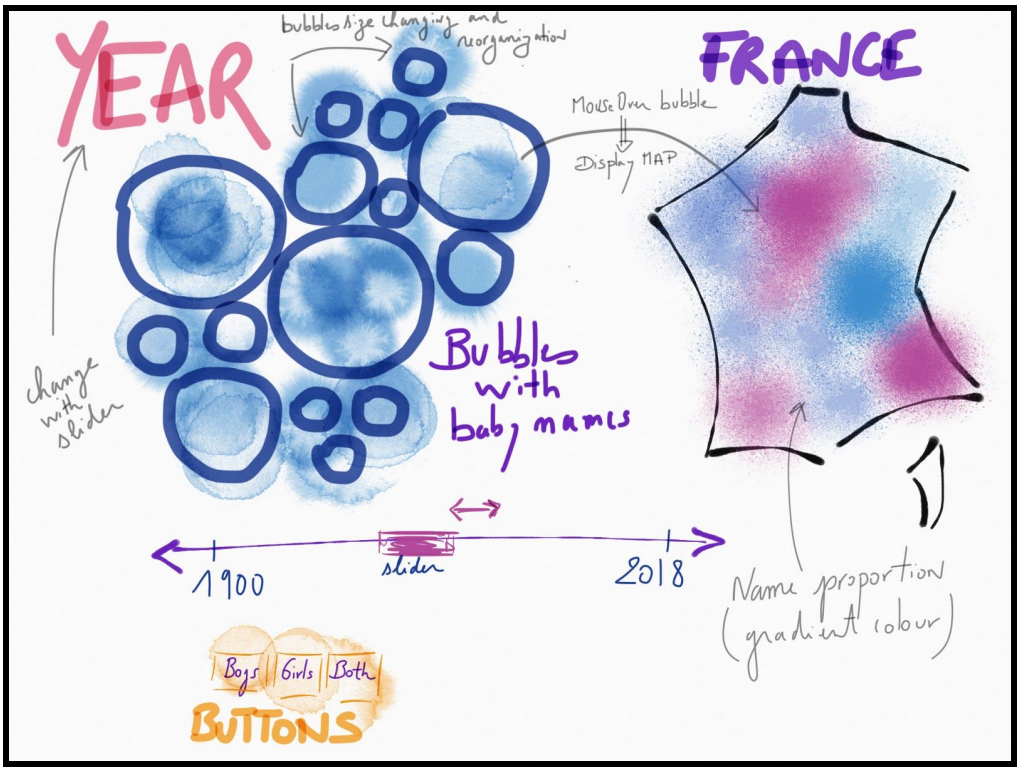


Fig.1 : sketch du projet

## 2. Les données

Les données que nous avons utilisées proviennent du site de l'INSEE (https://www.insee.fr/fr/statistiques/2540004), et se répartissent sur un fichier départemental. Le fichier en question contient les prénoms donnés aux nouveaux-nés, en France entre 1900 et 2016, ainsi que le nombre associé pour chaque sexe regroupés au niveau du département de naissance.

En termes de volumétrie, les données représentent environ 3,5 millions de lignes, avec environ 15,000 prénoms masculins et 18,000 prénoms féminins. Concernant la dimension, il y a plusieurs variables fournies.

Le fichier contient les variables suivantes :

- SEXE : 1 pour masculin, 2 pour féminin
- PREUSUEL : premier prénom
- ANNAIS : année de naissance
- DPT : département de naissance
- NOMBRE : nombre de naissance

Ce fichier de données permet donc d'extraire un certain nombre d'information. Par exemple, pour une année donnée, il est alors possible de voir quel est le nombre total de personnes qui se sont vues attribuer ce prénom. Il est également possible de voir quelle est la répartition d'un prénom en France, et ce à l'échelle de chaque département et de voir par exemple dans quel département un prénom est, en proportion, le plus fortement représenté. Une autre information que l'on peut extraire est de pouvoir visualiser dans quelle mesure le nombre de prénoms données va en augmentant au fur et à mesure du temps. Cela se voit de façon très rapide par exemple en observant le nombre de bulles présente au début 1900, alors que le nombre de bulles est largement plus important dans les années 2000. Cela met donc en évidence la diversification des prénoms donnés.

Par contre, le design que nous avons retenu ne permet pas forcément de suivre l'évolution d'un prénom donné durant le temps, du moins facilement. Mais cela est plus dû au fait que ce n'est pas l'information la plus importante que nous voulions modéliser à travers cette interaction.

## 3. Les utilisateurs ciblés

Notre site s'adresse essentiellement à des professionnels, plus spécifiquement aux équipes commerciales et chefs de produits qui commercialisent des produits porteurs d'un prénom afin de les aider à optimiser leurs ventes à travers la France. Ces utilisateurs recherchent, étant donnée une année ou une époque, les prénoms les plus couramment attribués et leur répartition à travers la France. Notre site leur permet de rapidement percevoir les prénoms qui les intéressent ( les bulles les plus grosses se détachent nettement de la visualisation et sont donc immédiatement repérées par l'utilisateur). Ce dernier peut aisément se déplacer dans le temps grâce au slider qui modifie instantanément la représentation des données et peut facilement appliquer un filtre sur les données grâce aux boutons prévus à cet effet.

Notre site peut également s'avérer intéressant pour des jeunes parents en recherche de prénom pour leur bébé ou des sociologues/historien s'intéressant à l'évolution des prénoms en France dans le temps et à travers le territoire.

## 4. Tâches représentatives

### a. Layette portant le nom d'un enfant

Le premier cas d'usage que nous avons imaginé concerne une marque de vêtements pour enfants qui voudrait commercialiser des gigoteuses, des bavoirs, des peluches ou divers autres objets porteurs d'un prénom. Par exemple, la chaîne de magasin de vêtements pour bébés et enfant "Okaidi" possède 452 boutiques en France, dispersées dans tout le pays, l'enseigne "Petit bateau" en possède 395, et d'autres marques de vêtements et accessoires pour enfants telles que Jacadi, Du Pareil Au Même… sont ainsi présentes dans de nombreux départements de France et sont ciblées par notre application.

Toutes ces enseignes possèdent des listes de naissance. Il devient donc possible pour les proches d'offrir aux nouveaux parents de la layette, des bavoirs ou jouets portant le prénom de l'enfant, en optimisant les chances de trouver le prénom en magasin.

Pour l'équipe commerciale/ le chef de produit, la démarche est la suivante : utiliser notre site pour connaître les tendances actuelles en termes de prénoms donnés. En utilisant les couleurs des bulles il peut évaluer si la tendance pour un prénom donné est à la hausse ou à la baisse. Par exemple, il évitera de commercialiser un bavoir brodé d'un prénom qui semble être en forte perte de vitesse. Ainsi, Le prénom "Louis" a été beaucoup attribué en 2015 ( la dernière année pour laquelle nous disposons de données), 4749 fois exactement mais a diminué de 40% par rapport à l'année précédente (Fig.2). C'est une information à prendre en compte par une équipe commerciale qui voudrait mettre en place une stratégie à moyen terme.

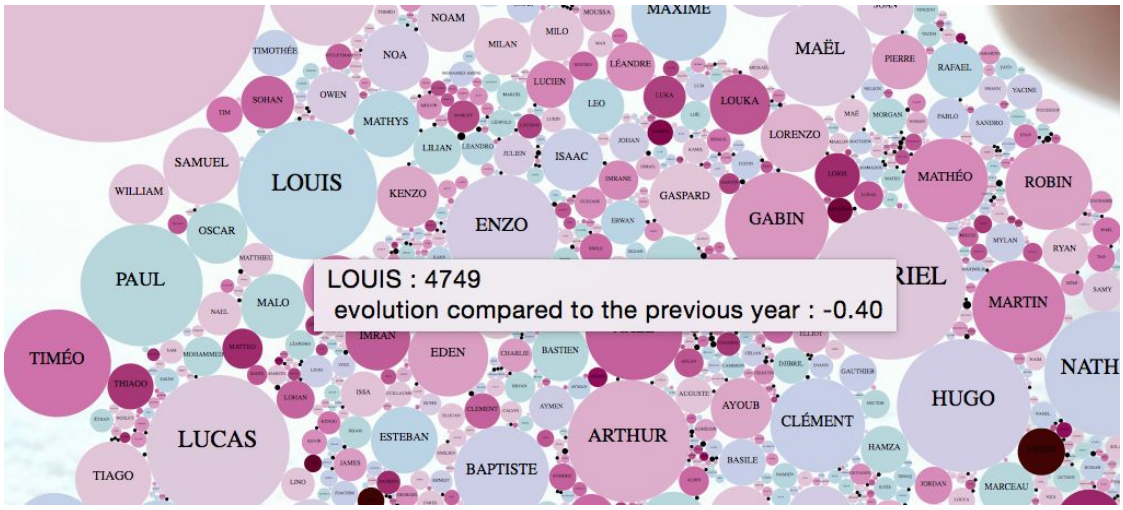


Fig.2: Choix du prénom "Louis" sur le bubbles chart

A l'inverse, un prénom moyennement attribué, mais en pleine ascension de popularité serait intéressante à commercialiser (Fig.3).

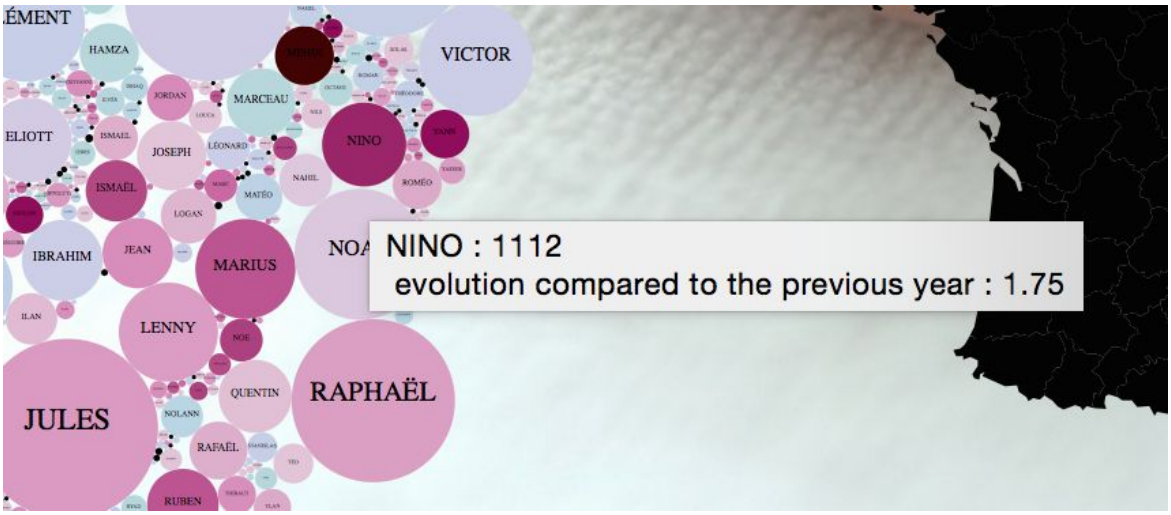


Fig.3 : Choix du prénom “Nino” sur le bubbles chart

Une fois, les prénoms intéressants identifiés, l’équipe peut affiner son ciblage commercial en adaptant l’offre dans chaque département dans lesquels elle possède des enseignes, en utilisant pour chacun de ces prénoms, leur nombre et leur proportion dans les naissances de chaque département (voir Fig.4).

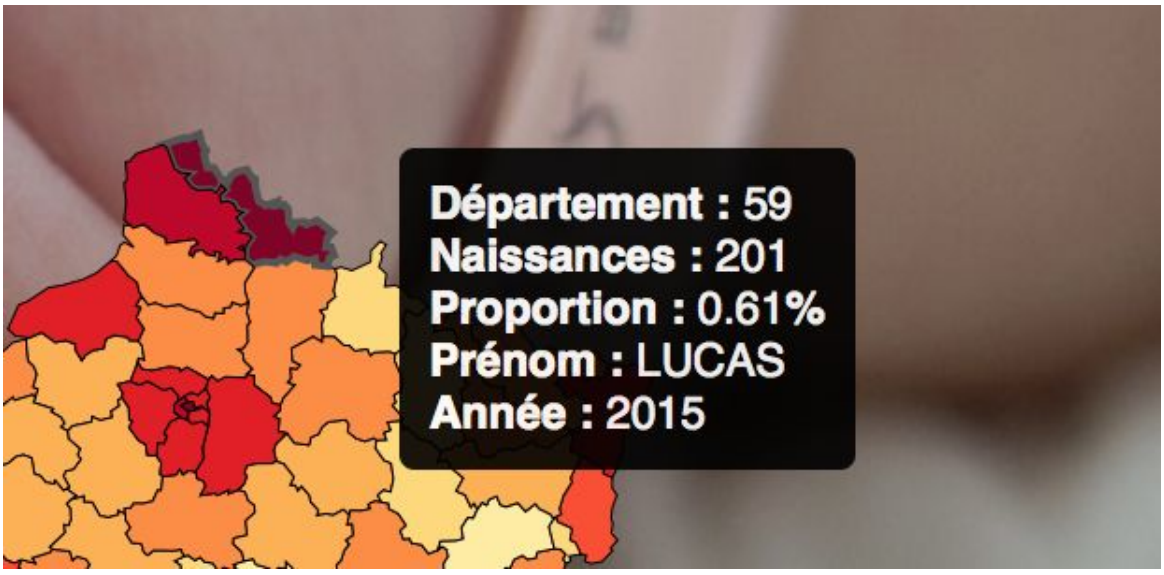


Fig.4: Choix du prénom “Lucas” dans le département 59 “Nord”

## b.Gourmettes pour adultes

Le deuxième cas d'usage auquel nous avons pensé concerne la vente de gourmettes pour adultes. Ces bijoux en acier argent ou en or, gravés du prénom de la personne sont essentiellement portés par des hommes, entre 30 et 60 ans.


Elles sont commercialisées en grande surface, et sur les présentoirs des petites boutiques.

Les commerçants souhaitant commander des lots de gourmettes peuvent utiliser notre site pour identifier les prénoms masculins les plus donnés entre 1960 et 1990. Ensuite, en fonction du département où ils sont implantés, ils peuvent utiliser la carte de France pour affiner leur commande et ainsi minimiser le risque d'invendus.

Au delà de ces deux cas d'usage, d'autres cas peuvent être imaginés, comme par exemple pour des bijoutiers, la possibilité de pré-graver des médailles de baptême, mais de manière générale notre site s'avère utile pour la commercialisation de tout objet portant un prénom (mugs, tee-shirts) avec ciblage d'une tranche d'âge donnée.

## 5. Bubbles chart

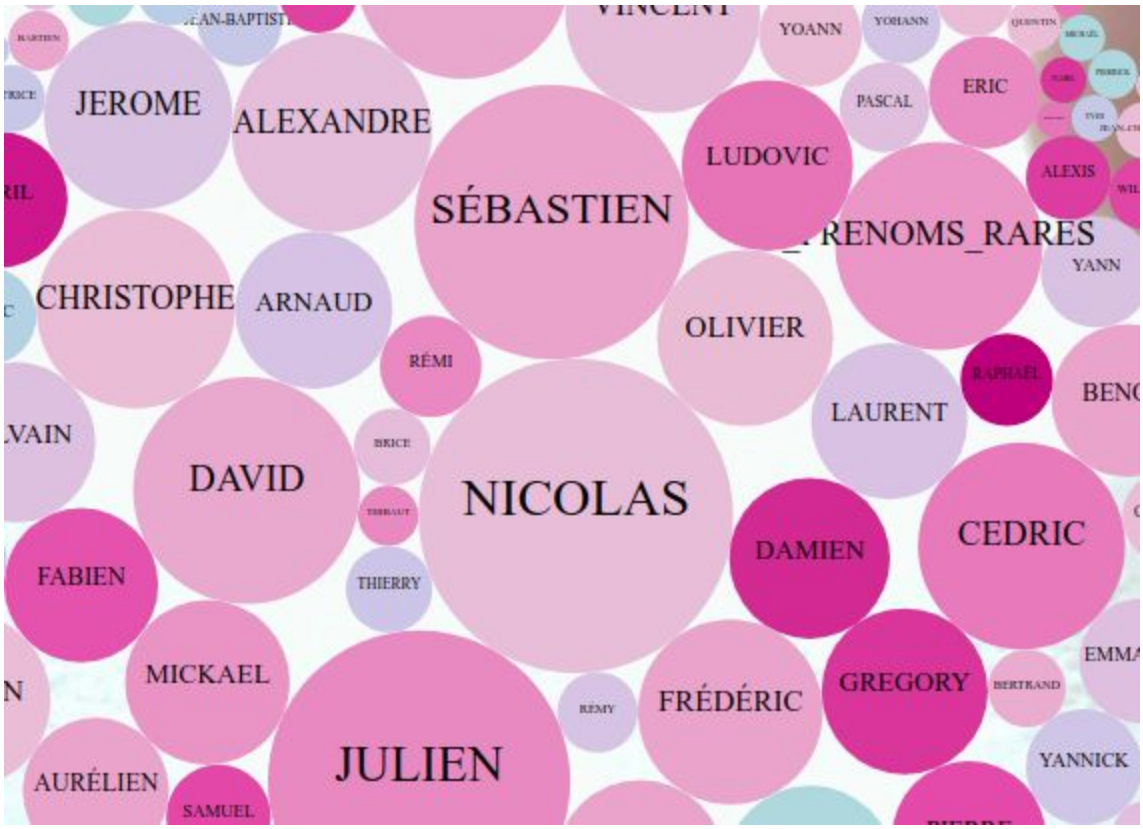


Fig.5: Bubbles chart

Le bubbles chart (Fig.5) est la première vue que l'on remarque lorsque l'on arrive sur la visualisation. Celle ci a pour but de donner rapidement beaucoup d'informations à l'utilisateur sur les prénoms d'une année, à savoir:

- L'ensemble des prénoms les plus utilisés pour une année donnée
- Les prénoms les plus importants (encodage sur l'aire des bulles)
- La dynamique de chaque prénom (encodage sur la couleur des bulles)

En plus de cela, nous souhaitions avoir une présentation qui attire l'oeil de l'utilisateur, c'est pour cela que nous avons également travaillé les transitions d'une année sur l'autre.

## Utilisation de d3.js

### Préparation du dataset:

Les données nationales ont été agrégées dans un dictionnaire permettant d'avoir, pour chaque année, et pour chaque filtre sélectionné par l'utilisateur (Boys / Girls / Mixed) les prénoms principaux associés ainsi que le rayon du cercle associé


Le paramètre seuil nous donne la proportion de prénoms à exclure sur chaque (année, filtre) (en volumétrie de la population globale).

Ensuite nous calculons le rayon de chaque prénom/filtre de l'année concernée : les contraintes sont les suivantes:

- Les aires des cercles sont proportionnelles au nombre de naissances
- La superficie totale des cercles doit correspondre, à un facteur de perte près, à la surface du container SVG

$$r_i = \sqrt{\frac{W * h * pop_i}{\Pi * pop_{year,filtre}}}$$

Ces traitements sont préparés en amont pour éviter de ralentir la visualisation lors de la manipulation des filtres et des années.

### Positionnement des bulles:

A chaque rafraichissement de la visualisation, on utilise la fonction packSiblings de D3 qui calcule le positionnement des cercles en fonction de leurs rayons, de telle façon à ce que ceux ci ne se chevauchent pas.

On gère les créations, modifications et suppressions de bulles sur le SVG via le pattern enter/update/exit de D3 (https://bost.ocks.org/mike/join/)

### Gestion des transitions

- L'arrivée de nouvelles bulles (enter) est fortement inspirée d'une visualisation de Mike Bostock:
  https://bl.ocks.org/mbostock/b07f8ae91c5e9e45719c
  La position d'arrivée des cercles a déjà été calculée via packSiblings. On calcule l'angle associé et on repousse le point de départ de la transition à l'extérieur du SVG.
- La mise à jour de bulles existantes (update) se fait en déplaçant/resizant la bulle de sa position actuelle à sa position calculée par packSiblings
- La suppression de bulle (exit) se fait en mettant à jour le rayon de la bulle à 0 sans la déplacer

## Encodage: choix des couleurs

Nous voulions utiliser des couleurs :

- Associées à la symbolique chaud / froid: la couleur est d'autant plus rouge que le gradient temporel du prénom est élevé, d'autant plus bleu dans le cas inverse.
- Différentes de celles utilisées par la carte de France (on encode une information différente)

## Interactivité avec la souris (tooltip)

Lorsque l'utilisateur survole un nom, un pop-up lui affiche le nombre de naissances associées ainsi que l'évolution par rapport à l'année précédente.

Lorsque l'utilisateur clique sur le prénom en question, la carte de France se met à jour sur le prénom / l'année associée.

### Remarques

Si nous avions eu plus de temps à notre disposition, nous aurions :

- Utilisé des champs de force pour contraindre les disques à rester à l'intérieur du container svg, ainsi que pour faire en sorte que les bulles ne se déplacent pas trop d'une année sur l'autre, ce qui permettrait de suivre mieux l'évolution globale des prénoms d'une année sur l'autre.
- Clusterisé l'espace en faisant en sorte que les prénoms commençant par une lettre particulière soient toujours au même endroit, ce qui permettrait à l'utilisateur de se savoir rapidement où chercher un prénom
- Mis en évidence la bulle / le prénom sélectionné, de telle façon à ce que l'évolution de celle ci soit visible immédiatement par l'utilisateur
- Fluidifié l'expérience utilisateur: gestion du changement d'année via la roulette de la souris

## 6. Carte de France (carte choroplèthe)

Une des vues les plus intéressantes à représenter sur ce projet a été de construire une carte de France pour répondre à certaines questions : - dans quels départements les prénoms sont les plus donnés et voir leur répartition géographique, - avoir la proportion des prénoms pour un département donné. De plus, cette carte de France des prénoms est intéractive (Linked Views) avec le bubbles chart, c'est à dire qu'elle se met à jour en fonction des choix faits sur le bubbles chart (choix du prénom). Enfin, un curseur permet de filtrer les données (Dynamic Query Interface) afin de choisir une année en particulier.

Le choix technique pour représenter la carte de France s'est porté sur la librairie d3.js, qui permet d'obtenir assez facilement des cartes géographiques à l'aide d'un fichier JSON ou GeoJSON. Le fichier JSON qui nous avons utilisé contient les contours de tous les départements sur le territoire français. Grâce à cela, nous pouvons représenter sur la carte de France les relations spatiales entre les différents éléments du jeu de données

(spatial relationships). Les raisons du choix d'une carte Choroplèthe (choropleth map) sont relativement simples car elles permettent de lier des données (densité, nombre de naissances, proportions, etc.) à des données géographiques en les encodant suivant différentes couleurs (colour map) - généralement un dégradé de couleurs.

## Utilisation de d3.js

Comme nous l'avons dit précédemment, la carte de France représente les différents départements français. Elle est construite à partir d'un fichier JSON. Ce fichier, en plus de contenir les polygones de chaque département, fournit également le nom du département, et son numéro (pour simplifier, nous n'avons pas utilisé les Dom-Tom). Sans rentrer dans les détails, nous sommes partis sur un choix de projection assez classique pour la cartographie. Nous centrons cette projection sur la France (latitude & longitude) et puis agrandissons la projection pour finalement la centrer (Fig.6).

Fig.6: Carte de la France

## Encodage: choix des couleurs

Une fois que le contours de la carte est tracée, il est important de lier les données que nous voulons représenter avec les départements tracés. Nous allons donc faire appel à un encodage formé par un jeu de couleurs progressives. Ici, nous avons décidé de choisir neuf couleurs différentes allant du beige clair, en passant par orange pour finir sur du rouge bordeaux.  Il faut trouver un bon compromis entre un nombre de couleurs pas trop conséquent et aussi pouvoir représenter clairement la distribution des données à afficher. En effet, choisir trop de couleurs donnerait l'impression que tous les départements ont des couleurs différentes et choisir trop peu de couleurs montrerait que peu de différences entre les départements. La méthode utilisée a été de prendre le min et le max des valeurs à représenter et de diviser cette différence par neuf, ce qui permet de voir rapidement où les valeurs les plus élevées sont. Grâce à cela, on voit visuellement la répartition du nombre des naissances dans les différents départements français en fonction de la palette de couleurs (Fig.7).

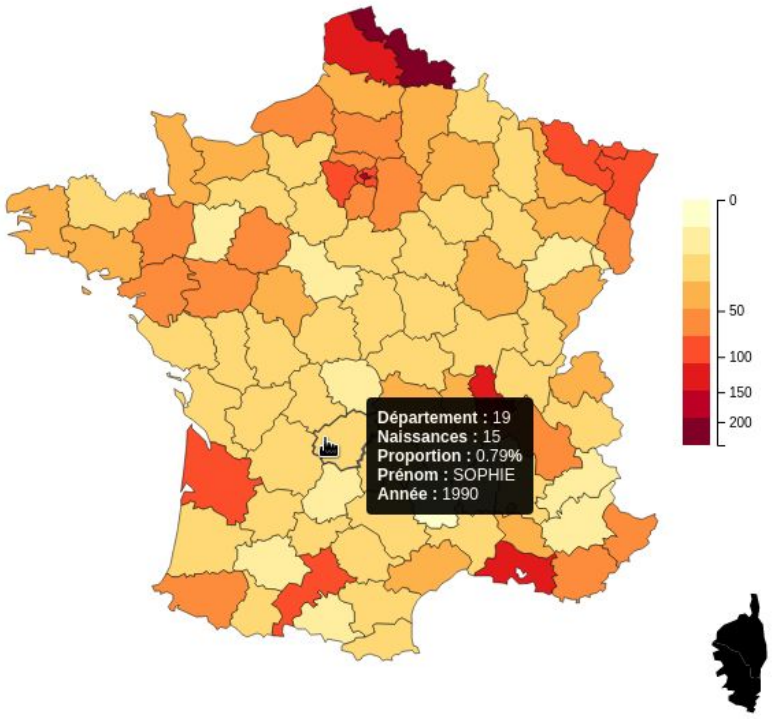


Fig. 7: Choroplèthe de la France

## Interactivité avec la souris (tooltip)

Nous avons voulu notre visualisation interactive lors du passage de la souris sur le svg, et nous avons choisi d'utiliser un tooltip qui apparaît au dessus de la carte (Fig.8) et qui devra faire apparaître certaines informations que l'on veut afficher à l'écran :

- département (un nombre)
- naissances (nombre, nombre de naissances de dans le département)
- proportion (nombre pourcentage, proportion des naissances sur la totalité des naissances dans le département)
- prénom (caractères)
- année (nombre compris entre 1900 et 2016)

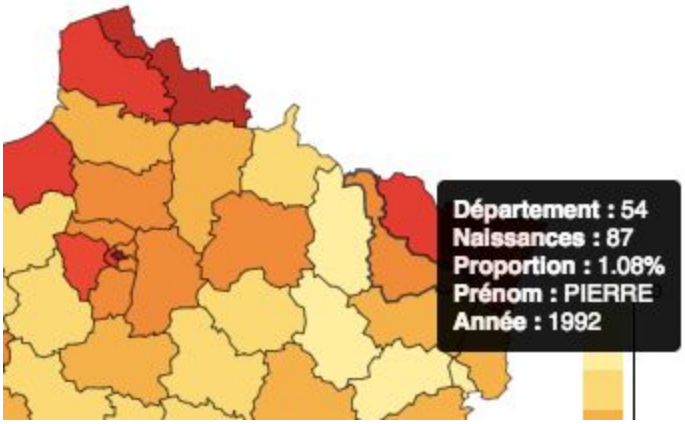


Fig.8 : effet du passage de la souris sur un département

La tooltip apparaît lorsque la souris se déplace au dessus d'un département et disparaît lorsque la souris est au dessus d'un autre département (ou lorsqu'elle est située hors de la carte). Ce tooltip résume les données pour un département donné et un prénom, et donne comme information supplémentaire aux couleurs des naissances la proportion des

naissances (nombre de naissances par rapport aux naissances totales dans le département) dans ce département.

## Linked Views

Nous avons vu précédemment que la carte de France intéractive nous indique le nombre de naissances dans chacun des départements, et ceci pour un prénom et une année données. Dans notre outil de visualisation, le prénom est fixé par une autre vue, celle des bubbles chart. En effet, le choix du prénom se fixe lorsque l'on clique sur une bulle qui contient chacun des prénoms. On peut choisir une prénom qui nous intéresse, et dynamiquement et automatiquement, la répartition des naissances sur la carte de France change en fonction du choix du prénom.

## Dynamic Query Interface

Pour ajouter plus d'interactivité à nos vues, nous avons couplé à la fois le bubbles chart et la carte de France à un curseur (slider) qui permet à l'utilisateur de choisir une année précise (comprise entre 1900 et 2016). Les bulles se réorganisent suivant l'année, ainsi que la réparation des naissances pour chacun des départements. Cet outil ajoute une fonctionnalité supplémentaire quant au côté exploratoire de l'outil, car il permet de voir les changements des naissances au fil des années - diminution ou augmentation.

## Remarques générales

Lorsque l'utilisateur explore les données grâce à la carte de France, il remarque que certains départements apparaissent en "noir". Pour la Corse, les naissances dans les deux départements sont étiquetées avec la valeur "20" alors que dans le fichier JSON, la Corse du Sud est étiquetée "2A" et la haute Corse "2B". Autres départements sans résultats, certains départements d'Ile de France qui ont été créés lors de la réorganisation de la région parisienne en 1964-1968 (disparition et création de départements). Ces départements apparaissent en "noir" avant 1964, alors qu'ils apparaissent colorés après la réforme.

L'avantage de cette représentation est qu'on voit rapidement la répartition des données que l'on veut afficher. L'identification est rapide et il suffit de connaître la géographie pour se repérer. Le principal inconvénient de ce type de représentation vient des couleurs choisies. En effet, il est souvent difficile de distinguer différentes nuances d'une même couleur. Dans la légende, on les reconnaît généralement bien parce qu'elles sont classées de la plus sombre vers la plus claire ou inversement. Par contre, sur la carte, une même couleur peut paraître plus sombre ou plus claire suivant les couleurs qui l'entourent. Il convient de ne pas prendre une légende continue sous peine de se faire "piéger".

## 7. Traitement des datasets

Deux fichiers étaient mis à disposition sur le site de l'INSEE:

- un fichier national
- un fichier par département

Pour optimiser le traitement de l'application, nous ne chargeons que le fichier par département: le fichier national est simplement une agrégation du fichier par département.

Pour encoder les différentes informations que nous souhaitons présenter, nous devons générer différents datasets à partir du fichier initial (Fig.9).

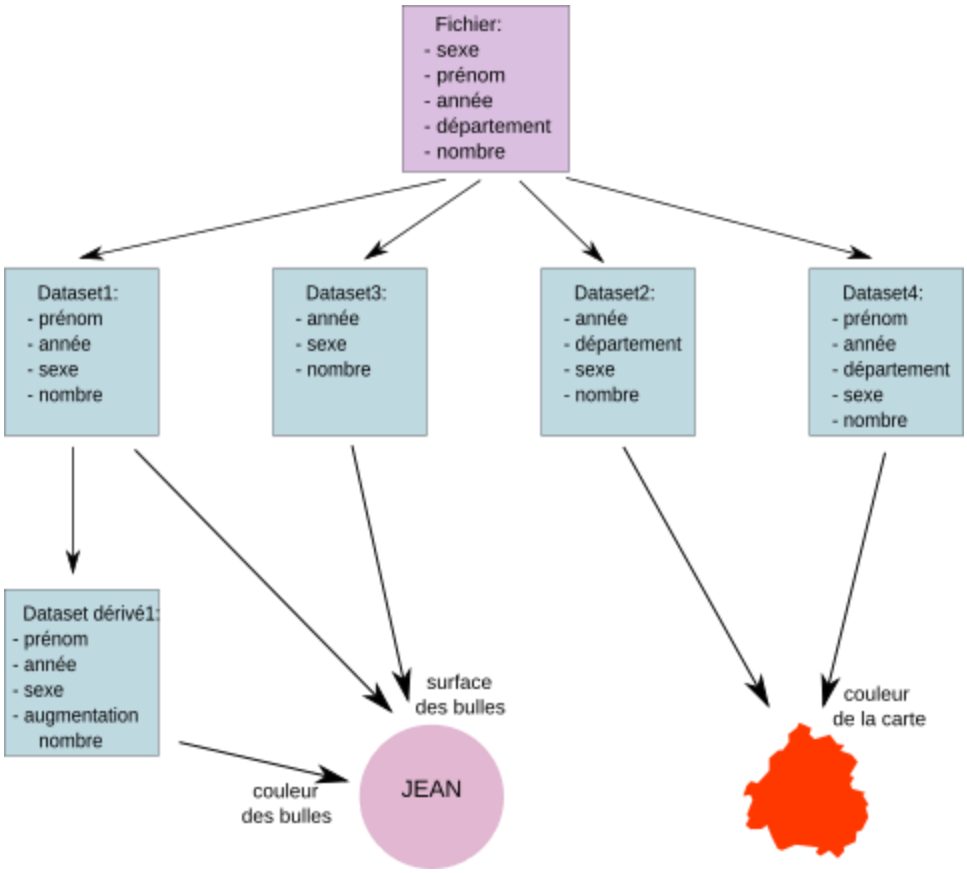


Fig.9: Traitement des datasets

## a. Proportion des prénoms à une date donnée

Cette information encodée par la surface des bulles requiert deux datasets:

- le nombre de naissances par année (permet de dimensionner la surface totale des bulles)
- le nombre de naissance par année et par prénom

A partir des ces données, on peut calculer le rayon de chaque bulles.

### b. Evolution du nombre de prénoms donnés

Pour rappel, on encode, pour un prénom à une année donnée, l'évolution du nombre d'attribution de ce prénom avec la couleur de la bulle. On a calculé les dérivés du dataset du nombre des prénoms par année. La valeur de cette dérivée est encodée par la couleur de la bulle.

### c. Proportions par département

Pour un prénom à une date donnée, on encode la proportion du nombre d'attributions de ce prénom en regard de tous les autres. Cette information est encodée avec la couleur du département. On  a besoin pour cela de la population totale par département.

Le schéma suivant présente l'ensemble des datasets requis pour encoder l'ensemble des informations d'intérêt.

## 8. Améliorations

Au niveau des améliorations que l'on pourrait apporter à notre outil, une textbox paraît la plus évidente (ou un menu textuel avec tous les prénoms présents dans la base de données). En effet, pouvoir choisir un prénom en le tapant directement est une manière beaucoup plus efficace d'explorer la répartition de ce prénom sur la carte de France.  La recherche d'un prénom grâce à la vue bubbles chart n'est pas des plus efficace. Enfin, la carte de France pourrait être agrandie, voire de la même taille que le bubbles chart pour plus de lisibilité et permettre une meilleure exploration des données.

## 9. Conclusion

Dans ce projet, nous avons implémenté  en d3.js une vue originale couplant deux vues distinctes : le bubbles chart et une carte choroplèthe de la France. L'intérêt des bubbles chart est de pouvoir explorer rapidement quels sont les prénoms qui se distinguent par rapport aux autres (prénoms les plus répandus) mais aussi de pouvoir voir grâce aux

couleurs choisies s'il y a une augmentation ou une diminution de ces prénoms au fil des ans. L'aspect géographique apparaît quant à lui grâce à la carte choroplèthe , qui permet d'observer la répartition d'un prénom pour une année donnée sur l'ensemble des départements français. Les différents cas de figure - déterminer les départements où certains prénoms sont les plus nombreux, ainsi que leurs âges - peuvent être étudiés grâce à l'interaction entre ces deux vues. Enfin ce projet nous a permis pour la plupart d'entre nous à appréhender, à coder en JavaScript et plus précisément la librairie d3.js.