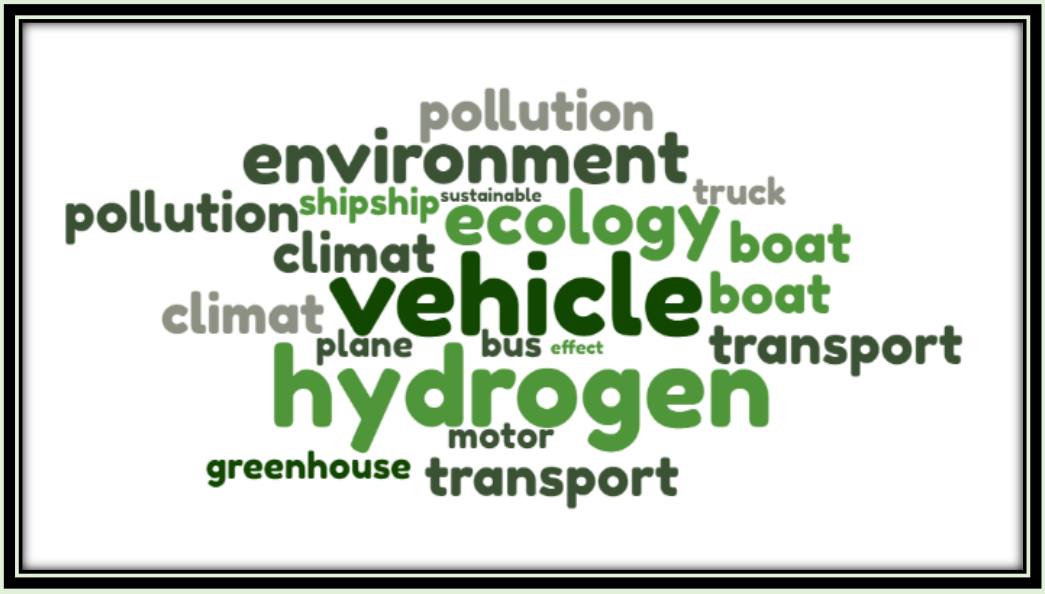
Université **T**oulouse **III** – **P**aul **S**abatier
Département Mathématiques
Formation **S**tatistique et **I**nformatique **D**écisionnelle

Rapport Projet M1 SID - Tableau de bord
Les véhicules à hydrogène



Vincent BARUDIO, Damien CHALIGNE,
Théo SACCAREAU et **Stévan STRICOT**

Clients :
Wahiba BAHSOUN et **Riad MOKKADEM**

Année Universitaire 2020 - 2021

Historique des modifications

Le tableau ci-dessous récapitule l'évolution du rapport en indiquant pour chaque version la cause de la modification.

Référence	Projet M1 SID - Tableau de Bord			
Date	**Version**	**Chapitre(s) concerné(s)**	**Objet de la modification**	**Auteur(s)**
15/02/21	Rapport_V0	Tous	Création du plan type	L'ensemble de l'équipe de projet
17/02/21	Rapport_V1	Chapitre 1	Rédaction	L'ensemble de l'équipe de projet
03/03/21	Rapport_V2	Chapitres 3 à 6	Rédaction	L'ensemble de l'équipe de projet
10/03/21	Rapport_V3	Chapitre 4	Construction des schémas SADT	L'ensemble de l'équipe de projet
18/03/21	Rapport_V4	Chapitres 3 à 6	Mise à jour	L'ensemble de l'équipe de projet
19/03/21	Rapport_V5	Chapitre 2	Ajout des documents de références	L'ensemble de l'équipe de projet
20/03/21	Rapport_V6	Chapitre 7	Rédaction de la partie « Bilan de projet »	L'ensemble de l'équipe de projet
21/03/21	Rapport_V7	Tous	Finalisation du rapport et ajout des annexes	L'ensemble de l'équipe de projet
22/03/21	Rapport_V8	Tous	Relecture	L'ensemble de l'équipe de projet

Sommaire

Historique des modifications 3

I. Objectif, but et domaine d'application du document 6

I.1 Objectif 6

I.2 But 6

I.3 Domaine d'application du document 7

II. Documents de référence et terminologie 8

II.1 Documents de référence 8

II.2 Documents applicables 9

II.3 Terminologie 9

III. Organisation de développement 11

III.1 Ressources humaines 11

III.2 Gestion de projet 11

III.3 Ressources matérielles 13

IV. Démarche de développement 14

IV.1 Processus A-0 – Réaliser un tableau de bord 15

IV.2 Processus A0 – Réaliser un tableau de bord 16

IV.3 Processus A1 – Collecter les données 17

IV.4 Processus A2 – Extraire les données 21

IV.5 Processus A3 – Administrer 26

IV.6 Processus A4 – Exploiter les données 27

IV.7 Processus A5 – Valorisation des données 29

IV.7.1 Partie 1 – Analyses générales sur les articles 30

IV.7.2 Partie 2 – Analyses sur les véhicules 38

IV.7.3 Partie 3 - Analyses temporelles 41

IV.7.4 Partie 4 – Analyses sur les pays 45

V. Gestion de configuration 50

VI. Assurance et contrôle qualité 53

VII. Bilan du projet 56

Annexes 58

Table des figures

Figure 1 - Diagramme de Gantt 12
Figure 2 - Légende du diagramme de Gantt 12
Figure 3 - Règles sémantiques permettant de représenter une activité avec la méthode SADT. 14
Figure 4 - Diagramme A-0 - Réaliser d'un tableau de bord 15
Figure 5 - Diagramme A0 - Réaliser d'un tableau de bord 16
Figure 6- Diagramme A1 - Collecter les données 17
Figure 7 - Diagramme A2 - Extraire les données 21
Figure 8 - Modèle Conceptuel de Données_V1.2 22
Figure 9 - Modèle Logique de Données_V0.2 25
Figure 10 - Diagramme A4 - Exploiter les données 27
Figure 11 - Diagramme A5 - Valorisation des données 29
Figure 12 – Proportion de citations par type de véhicule 30
Figure 13 – Proportion de citations par marque de véhicules. 31
Figure 14 - Proportion de citations par type de production 32
Figure 15 – Proportion de citations par source de production 33
Figure 16 – Proportion de citations par type de stockage 34
Figure 17 – Proportion de citations par type de moteur 35
Figure 18 – Proportion de citations par domaine. 36
Figure 19 – Nuage de mots représentant la fréquence d'apparition des mots-clés 37
Figure 20 - Répartition des types de stockage en fonction des types de véhicule. 38
Figure 21 - Répartition des sources de production en fonction des types de véhicule. 39
Figure 22 - Répartition des types de moteur par types de véhicule. 40
Figure 23 - Evolution du nombre d'articles sur les véhicules hydrogènes depuis 1945. 41
Figure 24 - Pays le plus mentionné chaque année depuis 1990 43
Figure 25 - Type de véhicule le plus mentionné chaque année depuis 1990 43
Figure 26 – Source d'hydrogène la plus évoquée chaque année depuis 1990 44
Figure 27 - Type de stockage le plus cité chaque année depuis 1990 44
Figure 28 - Proportion de citations par pays 45
Figure 29 - Nombre de fois qu'un pays est évoqué dans un article sur les véhicules hydrogènes. 46
Figure 30 - Nombre de fois qu'un pays est évoqué dans un article scientifique sur les véhicules hydrogènes (USA exclus). 47
Figure 31 - Top 3 des véhicules les plus mentionnées pour les 10 pays les plus évoqués 47
Figure 32 - Top 3 des pays les plus intéressés par types de véhicule. 48
Figure 33 - Charte qualité 65
Figure 34 - Dictionnaire de données 66

I. Objectif, but et domaine d'application du document

I.1 Objectif

Ce document est un rapport de projet réalisé au cours de l'année universitaire 2020-2021 par Vincent BARUDIO, Damien CHALIGNE, Théo SACCAREAU et Stévan STRICOT, étudiants en 1ère année de master en Statistique et Informatique Décisionnelle à l'université Toulouse III – Paul Sabatier.

Réalisé dans le cadre de l'UE « EMMAB2EM - Projet Tableau de Bord », l'objectif de ce projet est de développer un système d'aide à la décision sur le thème des véhicules à hydrogène. Le système s'appuie sur des analyses détaillées à partir de la visualisation de données textuelles via un tableau de bord. Les données textuelles proviennent d'articles scientifiques disponibles sur le site Core et sont stockées sur une base de données créée sur l'environnement de développement Oracle SQL Developer.

Ce projet nous a permis d'appliquer nos connaissances acquises au cours de notre formation, notamment à travers les enseignements en gestion de bases de données, en collecte et valorisation de données ou encore en génie logiciel et gestion de projet. Par ailleurs, il nous a permis d'acquérir une nouvelle expérience de collaboration en équipe. Enfin, c'est une occasion d'appliquer en situation concrète la relation Client-Fournisseur, très utile avant de rentrer dans le monde actif dans quelques mois.

Ce rapport a pour objectif de retracer l'évolution chronologique du projet en détaillant les choix effectués, la démarche et l'organisation de développement adoptées, les méthodes employées et les résultats obtenus. Nous expliquerons également les difficultés rencontrées et la manière dont nous les avons surmontées.

I.2 But

Ce projet est réalisé à l'intention de nos professeurs encadrant dans ce module qui représentent donc les clients : Madame Wahiba BAHSOUN et Monsieur Riad MOKADEM.

Ce projet a débuté le 18 Janvier 2021. Sa livraison est prévue pour le 22 Mars 2021 avec une présentation le 25 Mars 2021.

I.3 Domaine d'application du document

Nous avons choisi de construire notre projet autour des **véhicules hydrogènes** car, étant dans l'ère des nouvelles technologies, les véhicules hydrogènes sont un transport d'avenir. De ce fait, ils sont depuis quelques années en plein développement. Il nous semblait intéressant de découvrir de nouveaux aspects sur ce sujet.

Un véhicule à hydrogène est un moyen de transport qui utilise une transformation chimique du dihydrogène comme énergie de propulsion. L'hydrogène est l'élément le plus abondant de l'Univers (92 % en nombre d'atomes) mais il ne représente que 0,22 % des atomes sur Terre.

Toutefois, le véhicule à hydrogène présente un intérêt grandissant dans divers secteurs puisqu'il est souvent présenté comme « décarboné » car sa motorisation n'émet pas directement de gaz à effet de serre. Cependant, cet avantage peut être nuancé par le fait que la production d'hydrogène peut provenir d'hydrocarbures, fortement polluants.

Ainsi, l'objectif de notre système d'aide à la décision est de répondre à la problématique suivante : *Dans un contexte où l'écologie devient de plus en plus importante, comment évolue la conception et les technologies des véhicules hydrogènes à travers le monde ?*

II. Documents de référence et terminologie

II.1 Documents de référence

Pour mener à bien ce projet, nous avons dû nous documenter. Voici les principales ressources qui nous ont aidées tout au long du projet.

Tout d'abord, les cours reçus grâce à notre formation en Statistique et Informatique Décisionnelle (SID) à l'université Paul Sabatier. Plus précisément :

- Les cours de Madame Wahiba BAHSOUN à travers les UEs de « Développement Logiciel » en Licence 3 et de « Génie Logiciel et Gestion de projet » en Master 1.
- Les cours de Monsieur Riad MOKADEM, Monsieur Franck MORVAN et Madame Shaoyi YIN pour l'UE « Concepts Fondamentaux des Bases de Données » en Licence 3.
- Les cours de Madame Lynda LECHAMINE-TAMINE pour l'UE « Langages de Requêtes » en Licence 3.

En complément de ces cours, il a fallu nous documenter sur des notions non abordées au cours de notre formation. Voici nos principales sources :

- Pour passer du sujet à une équation de recherche, nous avons suivi les conseils et les étapes du site CoopIST.
- Pour la réalisation du diagramme de GANNT, nous nous sommes inspirés de la ressource offerte par OfficeToolTips.
- Pour réaliser le scraping, les quatre vidéos Youtube de la page Ninja Scripter nous ont été d'une aide précieuse.
- Pour la documentation sur les véhicules hydrogènes (différents types de stockages, types de production, etc.), nous avons utilisé les informations de Wikipédia et de Air Liquide.
- Pour le langage Python, les documentations des librairies Numpy, Pandas, et ReGex étaient indispensables. Tout comme les sites d'aide : Delftstack ou Python Simple.
- Enfin, pour la visualisation de données, la page de TowardsDataScience nous a permis de nous guider sur la réalisation d'un graphique empilé à 100% et la page de Xavier Dupré nous a été utile pour la réalisation de carte.

II.2 Documents applicables

Durant la réalisation du projet, nous avons dû respecter les exigences du client par rapport au produit final que nous lui fournirons. Pour cela, nous avons suivi le cahier des charges que les clients, Madame Wahiba BAHSOUN et Monsieur Riad MOKADEM, nous ont délivré lors du lancement du projet. Ce document est disponible en annexe (Annexe 1 – Cahier des charges).

Par ailleurs, comme dans tout projet informatique, il était important de respecter certaines normes de codage à la fois pour le langage Python que pour le langage SQL. Pour que notre code soit lisible, compréhensible et homogène, nous avons défini une charte de codage définissant des règles à respecter par tous les membres de l'équipe. Ce document est a retrouvé en annexe (Annexe 2 – Charte de codage).

II.3 Terminologie

Nous présentons dans cette partie les différents termes techniques et abréviations utilisés dans ce rapport permettant une meilleure compréhension de celui-ci.

<u>- Partie base de données :</u>

* <u>*MCD*</u> : *Modèle Conceptuel de Données*. Correspond à la première étape lors de la conception d'une base de données. Il permet d'avoir une notation unifiée pour représenter les informations gérées par les systèmes de gestion de bases de données. Pour cela, il fournit une description graphique pour représenter des modèles de données sous la forme de diagrammes contenant des entités et des associations.

On parle également parfois de schéma Entité-Association ou de modèle Entité-Association.

* <u>*MLD*</u> : *Modèle Logique de Données*. Le modèle précédent (MCD) ne peut être implanté dans une base de données sans modification, il est obligatoire de le transformer. On dit qu'on effectue un passage du modèle conceptuel de données vers le modèle logique de données. Après ce passage, le MLD pourra être implanté dans une base de données relationnelle. Pour réaliser ce changement, il faut suivre dans l'ordre trois règles d'or que nous développerons par la suite.

Le modèle logique de données est également appelé modèle (ou schéma) relationnel (uniquement lorsqu'on travaille avec une base de données relationnelle).

- Partie programmation :

* *Scraping* : Aussi appelé *Web Scraping*, c'est technique d'extraction du contenu de sites Web, via un script ou un programme, dans le but de le transformer pour permettre son utilisation dans un autre contexte (ici la recherche d'informations).

- Partie sujet :

* *Hydrogène* : C'est l'élément le plus léger dans l'univers et le premier élément dans le tableau périodique, l'hydrogène est aussi le plus abondant, constituant les étoiles comme les nuages moléculaires. Il compose aussi l'une des molécules vitales pour la vie : l'eau. Son symbole est H.

* *Dihydrogène* : Le dihydrogène est la forme moléculaire de l'élément hydrogène qui existe à l'état gazeux aux conditions normales de température et de pression. Les molécules comportent deux atomes d'hydrogène, sa formule chimique est H2. Il est également appelé « hydrogène moléculaire ».

* *Véhicule à hydrogène* : C'est est un moyen de transport qui utilise une transformation chimique du dihydrogène comme énergie de propulsion.

* *Energies renouvelables* : Energies que la nature constitue ou reconstitue plus rapidement que l'Homme ne les utilise. Elles peuvent ainsi être considérées comme inépuisables à l'échelle du temps humain. Elles ne produisent pas (ou peu) de CO2.

* *Energies fossiles* : Energies dont les réserves se sont constituées il y a très longtemps (des millions d'années) par décomposition sédimentaire de matériaux organiques. Des réserves, en d'autres mots, composées principalement de carbone. Et des réserves non renouvelables, car elles ne peuvent être reconstituées qu'à des échelles de temps géologiques.

* *Gaz à effet de serre* : Gaz qui participent à l'effet de serre. L'augmentation de leur concentration dans l'atmosphère terrestre est à l'origine du réchauffement climatique. Les principaux gaz à effet de serre émis par l'activité humaine sont le dioxyde de carbone, méthane, protoxyde d'azote, la vapeur d'eau, l'ozone et trois substituts des chlorofluorocarbones

* *Effet de serre* : Phénomène thermique bien connu sur les planètes comme la Terre et Vénus, où l'atmosphère laisse passer une partie du rayonnement solaire qui vient frapper le sol. Réchauffé, celui-ci émet un rayonnement infrarouge en partie ou totalement piégé par l'atmosphère rendue « imperméable » par la présence de gaz.

* *Hydrocarbures* : Dans le langage courant, on parle d'hydrocarbure pour faire référence au pétrole, au charbon ou encore au gaz naturel.

III. Organisation de développement

III.1 Ressources humaines

Notre équipe pour ce projet était composée de 4 étudiants de Master 1 SID, à savoir Vincent BARUDIO, Damien CHALIGNE, Théo SACCAREAU, Stévan STRICOT.

Pour mener à bien le développement du produit demandé, des rôles ont été attribués :

- Chefs de projet : Théo SACCAREAU et Stévan STRICOT
- Responsable de gestion de configuration : Damien CHALIGNE
- Responsable assurance et contrôle qualité : Vincent BARUDIO

Les chefs de projet ont pour rôle de gérer les différentes itérations effectuées ainsi que de coordonner le projet dans sa globalité en gérant notamment le temps passé sur chaque tâche.

Damien CHALIGNE, en tant que responsable de gestion de configuration, a pour mission de vérifier le passage d'une version à l'autre du système ainsi que de gérer l'ensemble des modifications apportées au cours de l'évolution du système. En d'autres termes, il est responsable de l'ensemble des composants permettant d'assurer la conformité d'un produit aux exigences tout au long du projet.

Vincent BARUDIO étant le responsable assurance et contrôle qualité, a pour responsabilités de concevoir et de mettre en œuvre la solution et de réaliser les tests qui doivent répondre aux besoins du projet et de la problématique. Il doit également vérifier la lisibilité des différents codes et programmes (Python/SQL) ainsi que la qualité globale du projet.

III.2 Gestion de projet

La gestion de projet, c'est l'art de coordonner le temps, les événements (la panne de matériel par exemple), les ressources humaines, et le matériel (le budget n'étant pas une variable intervenant dans notre projet).

Il est alors essentiel de créer un planning, de définir des tâches et d'affecter des ressources. Parfois des tâches devaient être réalisées en parallèle avec une ou plusieurs ressources humaines, nécessitant de construire un planning prévisionnel et d'essayer de le respecter au mieux.

Comme nous l'avons vu dans le cours de gestion de projet, il faut suivre quatre étapes pour construire un planning. Tout d'abord, il faut identifier les tâches, puis estimer leur durée et leur enchainement et enfin affecter les ressources.

Une fois que nous avions listé les différentes tâches à effectuer lors du projet ainsi qu'estimé leur durée et leur enchainement, nous avons procédé à la répartition en se basant sur nos goûts mais aussi en fonction des rôles évoqués précédemment.

Pour représenter ce planning, les exigences du client nous imposaient de choisir la méthode de Gantt. Les diagrammes de Gantt sont des outils utilisés en ordonnancement et en gestion de projet permettant de visualiser dans le temps les diverses tâches composant un projet.

Voici le diagramme de Gantt réalisé pour notre projet :

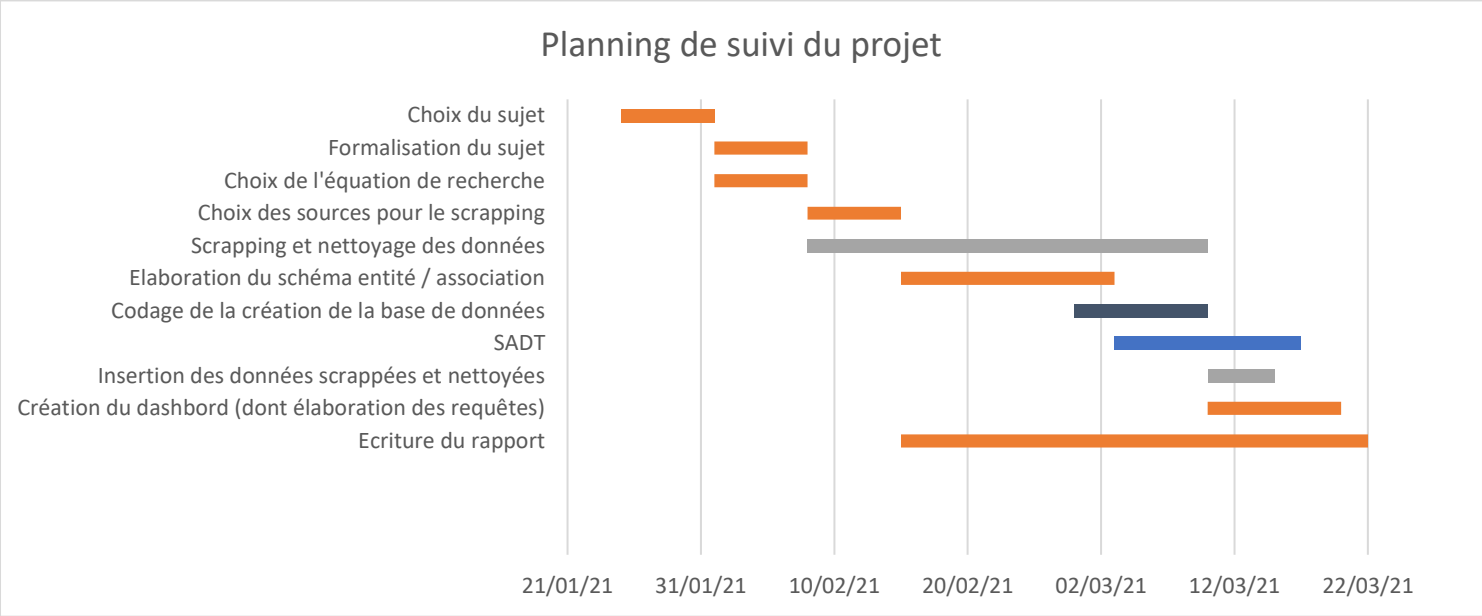


Figure 1 - Diagramme de Gantt

Couleur	Ressources mobilisées
	Tout le monde
	Barudio Vincent / Chaligné Damien
	Stricot Stévan
	Saccareau Théo

Figure 2 - Légende du diagramme de Gantt

Pour plus de détail, il est possible de consulter le tableau qui nous a permis de faire ce diagramme en annexe (Annexe 4 – Détail diagramme de Gantt).

A noter que la prévision initiale que nous avons faite était bonne, nous avons réussi à bien gérer notre temps et à respecter les délais que nous nous étions imposés. Ainsi, il n'y a pas vraiment eu de différences entre le planning prévisionnel et le planning réel.

III.3 Ressources matérielles

Comme évoqué précédemment, la gestion de projet nécessite aussi la gestion du matériel. Dans le contexte d'un projet essentiellement en distanciel, la possession d'un ordinateur pour chaque membre était indispensable.

Par ailleurs, pour le côté informatique, nous nous sommes appuyés sur plusieurs logiciels.

Pour commencer, comme nous venons de le voir, le logiciel Excel pour la réalisation du diagramme de Gantt.

Ensuite, le langage de programmation Python, avec ses librairies « BeautifulSoup » et « Request », nous a permis d'effectuer le scraping des informations à partir du site « Core ». Il nous a également servi à insérer les données dans la base de données grâce à l'interface « Cx_Oracle ». Enfin, c'est toujours à partir du langage Python que nous avons mis en valeur les données grâce à des graphiques produits grâce aux librairies « MatPlotLib » et « Cartopy ».

Pour le stockage des informations sur une base de données, nous avons utilisé l'environnement de développement Oracle SQL Developer.

Pour finir, la plateforme Google Drive nous a permis de stocker, d'échanger et de se partager les différentes versions de fichiers tout au long du projet. En ce qui concerne la communication, nous avons utilisé les plateforme Zomm et Discord.

IV. Démarche de développement

La démarche de développement, aussi appelée processus de développement, d'un système informatique correspond à l'ensemble des étapes partiellement ordonnées qui concourent à la création de ce système (ou son évolution s'il est déjà existant). L'objectif étant de produire des logiciels de qualité (répondant aux besoins utilisateurs) et dans des temps prévisibles.

Pour illustrer notre démarche de développement pour ce projet, nous avons utilisé la méthode SADT, choix imposé par le client. La méthode SADT (en anglais Structured Analysis and Design Technics), aussi connue sous le label IDEF0, est une méthode systémique de modélisation d'un système. En d'autres termes, cette méthode nous permettra de modéliser nos différentes étapes du projet.

Concrètement, c'est une démarche descendante qui se présente sous la forme de pyramide avec plusieurs niveaux. Le niveau le plus haut (en haut de la pyramide) permet de décrire l'activité globale. Ensuite, à partir du niveau initial, on raffine jusqu'à atteindre un niveau où l'on comprend la notion d'activité à elle seule, c'est le niveau le plus fin (le plus détaillé) qui se trouve à la base de la pyramide. Les niveaux s'emboitent les uns aux autres.

Une activité est représentée par des rectangles et des flèches avec des règles de sémantiques à respecter. Les voici :

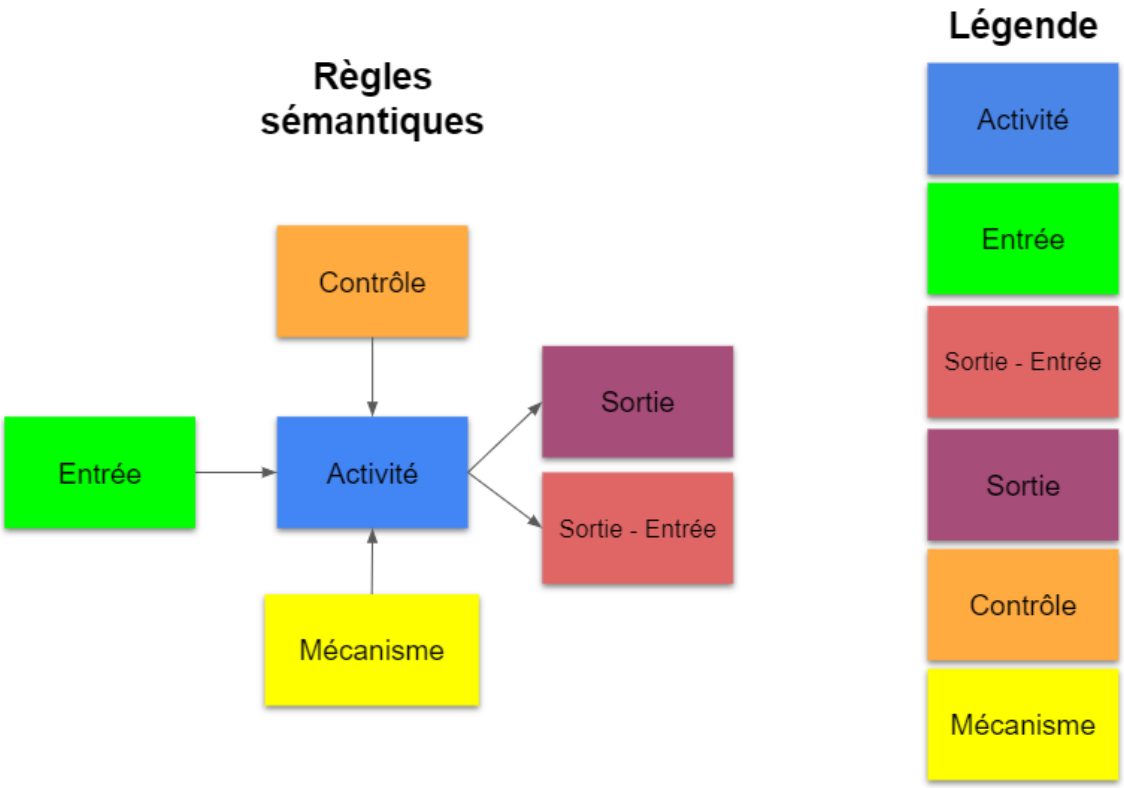


Figure 3 - Règles sémantiques permettant de représenter une activité avec la méthode SADT.

Une activité est donc représentée par un rectangle bleu au centre. Elle reçoit à sa gauche des composants en entrée et les résultats produits se trouve à sa droite (la sortie). Une sortie d'une activité peut constituer une entrée pour une activité suivante.

Les flèches du dessous correspondent aux mécanismes, ce sont les moyens qui permettent à l'activité de produire un résultat. Ils peuvent aussi bien être humains (l'équipe de projet par exemple) que matériels.

Enfin, au-dessus ce sont les contrôles. Un résultat ne peut être produit en sortie s'il ne respecte pas les contrôles (normes, exigences du client, etc.).

Nous pouvons désormais voir la démarche de développement que nous avons adoptée.

IV.1 Processus A-0 – Réaliser un tableau de bord

Le premier diagramme, correspondant au niveau le plus général, reprend le nom du projet.

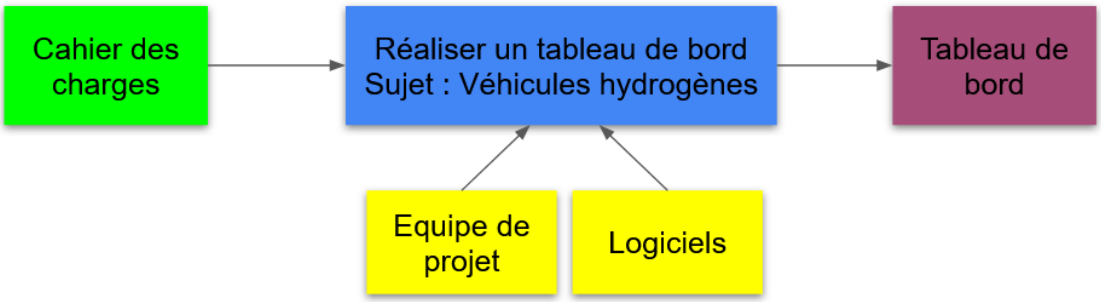


Diagramme A-0 : Réaliser d'un tableau de bord

Figure 4 - Diagramme A-0 - Réaliser d'un tableau de bord

En entrée, nous retrouvons le cahier des charges. Point du départ du projet, il récapitule les exigences du client et les objectifs à atteindre. En sortie, le résultat produit à l'issue du projet sera un tableau de bord dont le sujet porte sur les véhicules hydrogènes. Tout au long du projet, c'est l'équipe de projet et différents logiciels qui permettront à l'activité de produire le résultat attendu.

Commençons le « raffinement », en descendant d'un niveau avec le diagramme A0 qui va présenter les cinq principales activités au cours du projet.

IV.2 Processus A0 – Réaliser un tableau de bord

Dans ce graphique, nous retrouvons les éléments essentiels du graphique précédent : en entrée le cahier des charges, en sortie le tableau de bord et tout au long du projet l'équipe de projet et les logiciels qui jouent le rôle de mécanisme.

La différence c'est que l'activité « Réaliser un tableau de bord » a été divisée en 5 sous-activités : collecter, extraire, administrer, exploiter et valoriser. Nous avons donc un diagramme plus détaillé (plus fin) que le précédent.

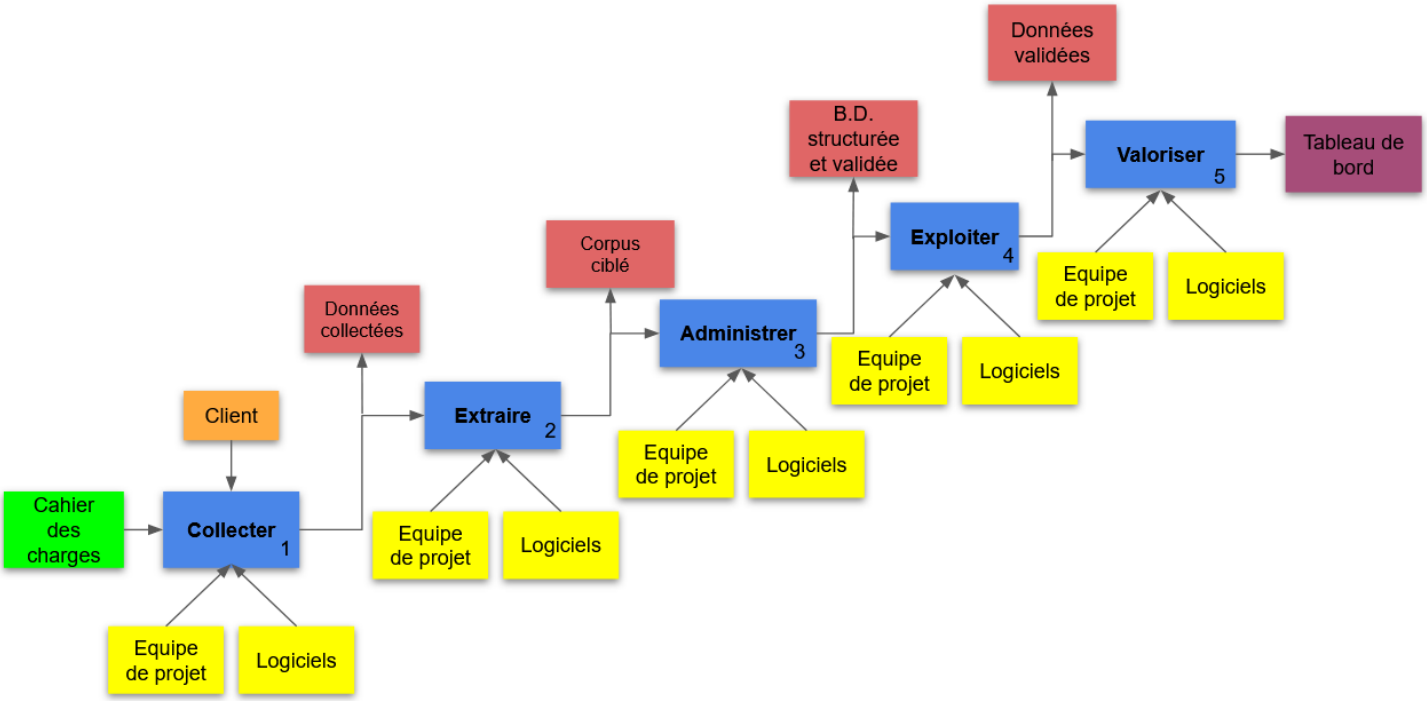


Diagramme A0 : Réaliser d'un tableau de bord

Figure 5 - Diagramme A0 - Réaliser d'un tableau de bord

L'activité « collecter » permet de récupérer les informations des articles scientifiques sur le site web choisi, c'est le scraping. En sortie, on retrouve naturellement les données collectées, elles ne sont pas encore nettoyées. Le client joue le rôle de contrôle. En effet, il veille à ce que les données collectées soient suffisamment nombreuses, pertinentes et en adéquation avec le sujet choisi.

L'étape « extraire » a pour but de récupérer les informations du scraping stockées dans un fichier, de les nettoyer avant de les insérer dans la base.

Dans la partie administration de la base de données, il faut entre autres mettre en place la base de données, anticiper les volumes de données à stocker, créer le dictionnaire de données, etc.

Vient ensuite l'exploitation des données où l'on sélectionne et on interroge les données de la base via des requêtes SQL. On en profite pour analyser leur pertinence avant de mettre en place les graphiques.

La création des graphiques intervient effectivement dans la dernière partie dédiée à la valorisation des données. Une fois les graphiques créés nous sommes en mesure d'en tirer des informations pertinentes et de faire des analyses permettant de mettre en place notre système d'aide à la décision et de conclure notre projet.

La modélisation et la description des cinq activités que nous venons de faire n'est pas suffisante pour pouvoir réaliser le projet, le niveau n'est pas assez fin. On continue donc le « raffinement » activité par activité.

IV.3 Processus A1 – Collecter les données

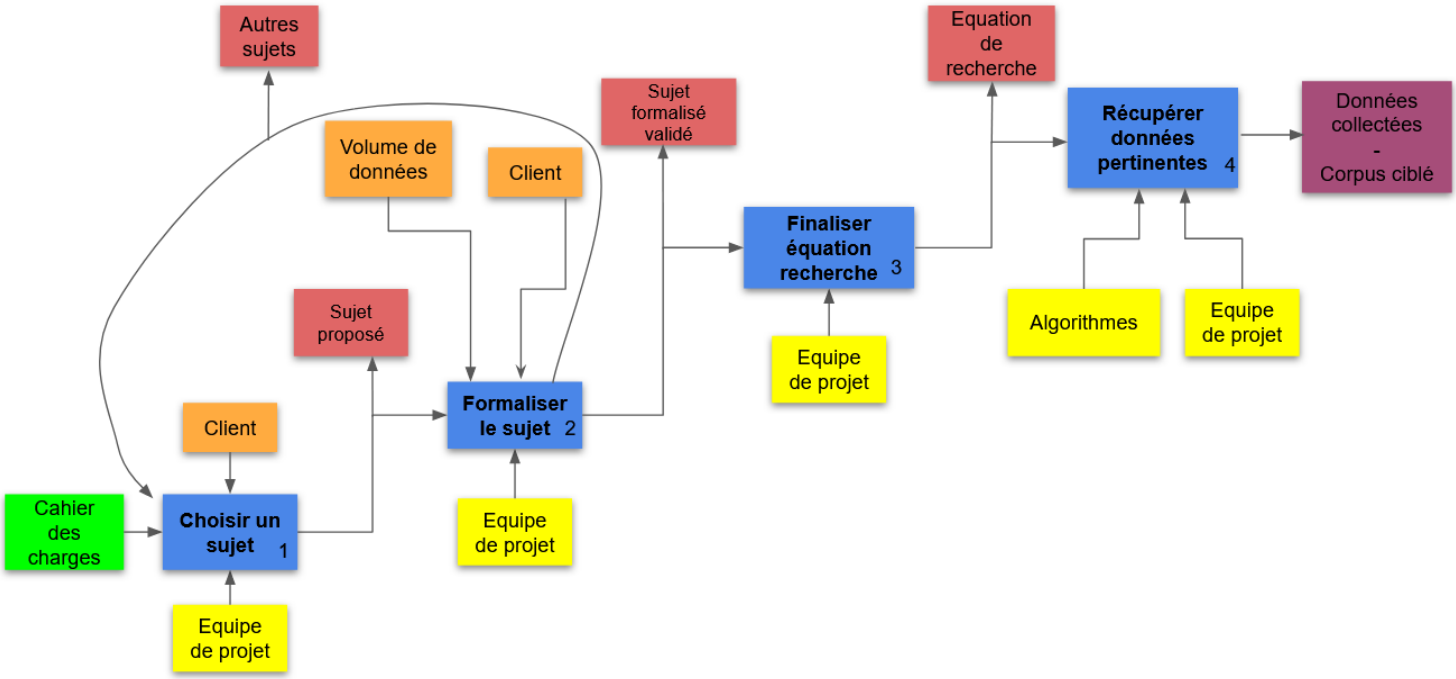


Figure 6- Diagramme A1 - Collecter les données

La première principale activité du projet est la collecte de données. Le diagramme A1 permet de la décomposer est en quatre nouvelles sous-étapes, permettant une meilleure compréhension et un gain de temps dans sa réalisation.

Comme on peut le voir ci-dessus, la première tâche consistait à choisir un sujet. Les consignes du client étaient de trouver un sujet qui nous intéresse, qui présente un intérêt par rapport au contexte socio-économique dans lequel on vit et qui nous permettrait d'extraire suffisamment d'informations pour faire nos statistiques. Le client apparaît alors comme un

contrôle puisque, pour valider notre sujet, il a dû vérifier que ses exigences étaient respectées. L'ensemble de l'équipe de projet a bien évidemment participé à cette activité car un consensus sur le sujet était nécessaire. Pour ce faire, nous avons commencé par réaliser un brainstorming durant lequel chacun d'entre nous a pu lister deux ou trois idées de sujet qu'il souhaitait traiter. Une fois cette liste constituée, tous les membres de l'équipe ont voté pour ou contre chaque sujet. A l'issue de ce vote, seuls trois sujets faisaient l'unanimité :

- Les voitures électriques / à hydrogène
- Les élections américaines
- Impact des catastrophes météorologiques / du changement climatique sur les populations

Pour départager ces trois sujets, nous avons réfléchi sur les informations que l'on pourrait obtenir mais aussi regarder le nombre d'articles scientifiques disponibles sur les sites spécialisés. C'est finalement le thème des véhicules électriques et à hydrogène que nous avons retenu. Ce sujet a été validé par les clients, Madame Wahiba BAHSOUN et Monsieur Riad MOKKADEM. D'ailleurs suite à leurs conseils, nous avons restreint le sujet uniquement aux véhicules à hydrogènes (et non plus aux véhicules électriques) pour éviter l'éparpillement.

Une fois le sujet trouvé, la deuxième tâche était la formalisation du sujet. Pour cela, nous avons commencé par définir le sujet plus précisément. Effectivement, pour délimiter le cadre de réflexion et d'éviter un sujet trop large ou peu précis, l'objectif était d'essayer de répondre à cinq questions :

- Quoi : quels sont les composantes du sujet ?
- Qui : quels groupes, quelles personnes sont concernés par le sujet ?
- Où : quel contexte géographique, pays, région ?
- Quand : quelle période de temps, quelle durée ?
- Comment : sous quels angles aborder le sujet ?

Après avoir effectué ce travail, notre sujet était délimité, il portera sur l'évolution des technologies des véhicules hydrogènes dans le monde à travers un contexte de lutte contre le réchauffement climatique.

Ensuite, lors d'un nouveau brainstorming, nous avons cherché des termes (mots, expressions) qui caractérisent le mieux le sujet. Cette liste de termes devait contenir des mots-clés, des mots spécifiques, des synonymes, des sigles… Une fois le sujet précisé et défini par son vocabulaire spécifique, nous avons pu formuler notre problématique. Une nouvelle fois, plusieurs problématiques ont été proposées par les différents membres de l'équipe. Après concertation, la problématique retenue a été : ***Dans un contexte où l'écologie devient de plus en plus importante, comment évolue la conception et les technologies des véhicules à hydrogène à travers le monde ?***

L'étape suivante dans notre recherche documentaire, était de traduire chaque idée maîtresse en concept, puis associer chaque concept à des mots-clés pour élaborer l'équation de recherche. Ainsi, trois concepts ont été choisis : « véhicule », « hydrogène » et « écologie ».

La dernière étape fût donc de constituer l'équation logique en combinant les mots-clés avec des opérateurs (ET, OU, SAUF) et symboles (parenthèses, étoiles, point d'interrogation, etc.). Finalement, voici l'équation de recherche retenue :

(vehicle OR transport OR transports OR train OR trains OR tractor OR bike* OR bicycle* OR boat OR boats OR ship OR ships OR *plane OR *planes OR aircraft* OR car OR cars OR truck* OR lorry* OR bus OR automobile* OR motor* OR rocket*)*
AND hydrogen
AND (ecolog OR climat* OR pollution* OR environment* OR "renewable energy" OR emission* OR "carbon neutrality" OR "global warming" OR conservation* OR sustaina*)*

La première partie concerne les véhicules. On y retrouve donc la plupart des moyens de transports. Parfois, il est utile d'utiliser l'opérateur '*' pour remplacer des lettres manquantes (permet de prendre en compte les pluriels par exemple). Mais cet opérateur est à utiliser avec grande prudence, son abus peut entrainer la prise en compte de mots non souhaités. Par exemple, si pour le train on souhaite prendre en compte le pluriel et que l'on met « train* », cela entrainera la prise en compte de « trainspotter », « training », « trainspotting », …, des mots non adaptés au sujet. Il faut donc plutôt écrire « train » et « trains » pour éviter ce type de problème. En revanche, l'opérateur '*' est très utile pour « plane », puisque le fait d'utiliser « * plane » permet de prendre en compte les mots « airplane » (« avion » en anglais américain) et « aeroplane » (« avion » en anglais britannique).

La deuxième partie concerne l'hydrogène et la troisième partie l'aspect environnemental.

Enfin, une fois l'équation de recherche finalisée, la dernière étape pour la collecte de données était le scraping. Pour cela, il fallait choisir un site sur lequel il allait être effectué. Pour plus de facilité, il était préférable de ne sélectionner qu'une seule source. Nous étions d'abord partis sur le site « Science Direct ». Cependant, nous nous sommes rendus compte que ce site permettait l'usage d'uniquement huit booléens dans l'équation de recherche. Au vu de notre équation ci-dessus, nous avons très vite compris qu'il serait impossible d'utiliser ce site pour réaliser le scraping. Nous nous sommes donc rabattus sur le site « Core », qui nous permettait d'utiliser l'intégralité de notre équation tout en conservant une quantité d'articles suffisante.

C'est grâce à Python et plus précisément aux deux librairies « Requests » et « BeautifulSoup » que nous avons réalisé le scraping. Le scraping s'est déroulé en deux étapes.

Pour la première étape, nous utilisions l'équation de recherche qui nous permettait grâce au moteur de recherches de Core d'obtenir près de 11500 articles. Ces articles étaient rassemblés par 100 sur chaque page (115 pages au total). L'objectif de notre algorithme était de parcourir ces 115 pages et de récupérer à chaque fois les informations générales des 100 articles (identifiant, titre, auteurs, langue). Le tout était stocké dans un premier fichier (*df_infos.csv*). Le fait de stocker les informations dans un fichier permettait de ne pas réexécuter l'algorithme à chaque fois. En effet, son temps d'exécution était d'environ 20min, cela nous aurait fait perdre un temps précieux de le relancer à chaque début de séance de travail.

La deuxième étape avait pour but d'accéder plus en détail au contenu des articles et notamment d'obtenir le résumé et les éventuels mots-clés. Cela était rendu possible grâce aux identifiants récupérés lors de la partie 1. Effectivement, un identifiant nous permettait de constituer une URL permettant d'accéder directement à la page d'un article. Le programme chargé d'effectuer cette tâche nécessitait alors un temps d'exécution bien plus long puisqu'il fallait pour chaque article, envoyer une requête contenant son URL, attendre la réponse du serveur, scraper les informations souhaitées (après avoir vérifié qu'elles existaient) puis les stocker dans un fichier. Il a fallu entre 2 et 3h pour que l'exécution se termine, la nécessité de stocker le contenu du scraping dans un fichier (*df_content.csv*) était encore plus flagrante pour cette étape.

Pour finir, il est important de signaler un changement majeur lors du scraping. Initialement, le moteur de recherche du site « Core » sélectionne les articles qui respectent l'équation de recherche en se basant sur leur contenu en intégralité. Or, étant donné que nous nous souhaitions uniquement se baser sur les informations des résumés, nous nous sommes vite rendu compte que certains résultats étaient peu pertinents (les résumés ne parlaient pas forcément de véhicules hydrogènes). Ainsi, nous avons fait le choix de forcer le moteur à n'appliquer l'équation de recherche que sur les résumés et les titres (et non plus sur le contenu). Ce choix a entrainé une forte diminution du nombre d'articles (de 80 000 à 11 500) mais une augmentation sensible de la pertinence des résultats. 11 500 articles étant un nombre encore largement satisfaisant ce changement était clairement bénéfique pour la suite du projet.

Comme nous l'avons vu avec le diagramme A0, après l'étape de collecte de données, vient l'étape d'extraction.

IV.4 Processus A2 – Extraire les données

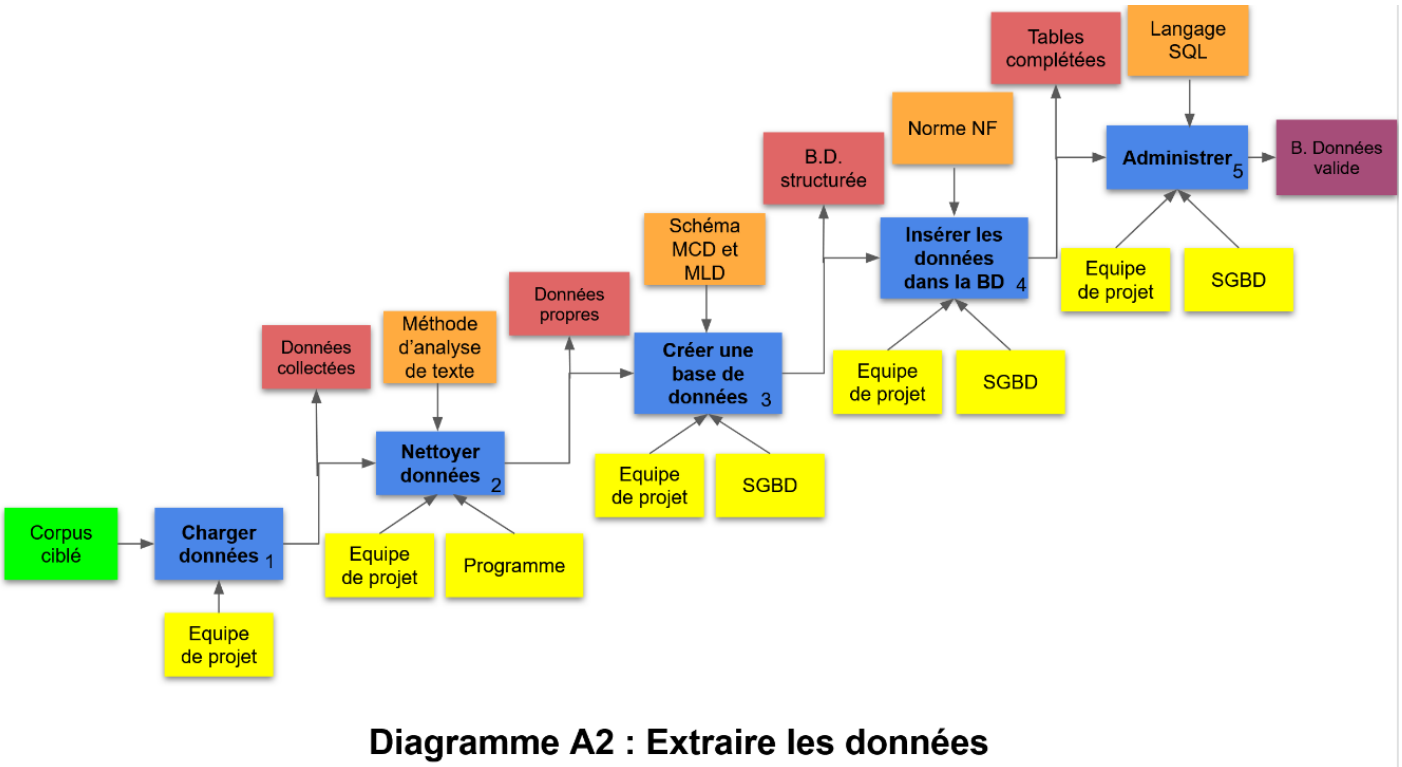


Figure 7 - Diagramme A2 - Extraire les données

Pour commencer cette nouvelle phase du projet, il a fallu charger dans des DataFrames les données scrapées qui étaient stockées dans des fichiers .csv. Ensuite, nous avons effectué une jointure entre les deux DataFrames pour en avoir plus qu'un (la jointure portait évidemment sur les identifiants d'articles). C'est également lors de cette étape que nous avons pu appliquer une fonction Python permettant d'éliminer les doublons. Finalement, il y a « seulement » un peu plus de 8000 articles distincts qui ont été récupérés. Cette quantité reste toutefois largement suffisante pour avoir des statistiques solides et en tirer des informations intéressantes.

Toutes les données issues du scraping étaient donc désormais contenues dans un seul et unique DataFrame. L'étape suivante consistait à nettoyer ces données. Ce nettoyage comprenait deux tâches distinctes : l'extraction d'informations depuis les résumés mais aussi le nettoyage à proprement parlé des données (définition d'un format uniforme pour les dates, nettoyage des données textuelles, etc.).

Les informations à extraire depuis les résumés portaient sur la citation de véhicules, de type de stockage de l'hydrogène, de type de moteur à hydrogène, de pays, de type de production, etc. Pour parvenir à extraire ces informations, il a fallu constituer des listes de vocabulaire à priori. Ensuite, le programme Python, grâce à la librairie « Re » notamment, se chargeait de vérifier si un mot du vocabulaire pour un type d'information était présent dans le résumé. Si c'était le cas alors l'information était présente dans l'article. Nous avons donc

répété cela pour chaque type d'informations et nous avons stocké les résultats dans un nouveau DataFrame puis un nouveau fichier .csv.

Ensuite, avant de procéder à l'insertion de ces données dans notre base de données, il a fallu les nettoyer et les uniformiser. Pour les dates, le format adopté fut « jour-mois-année ». Pour les informations textuelles (noms des auteurs, titres d'articles, mots-clés), il a fallu appliquer les méthodes d'analyse de texte. Par exemple, la suppression de certaines ponctuations et de caractères spéciaux, mise en minuscule, etc. Pour plus de précision, il est nécessaire de se reporter directement au Notebook qui est fortement commenté pour faciliter sa compréhension et qui détaille ce nettoyage.

Avant de pouvoir insérer toutes ces données dans une base de données, il faut bien évidemment la créer. La première étape lors d'une création de base de données est la conception du Modèle Conceptuel de Données (MCD), également appelé schéma entité-association. Après plusieurs versions et modifications (se reporter au chapitre 7 « Gestion de configuration » pour plus de détails), voici notre schéma entité-association validé par le client :

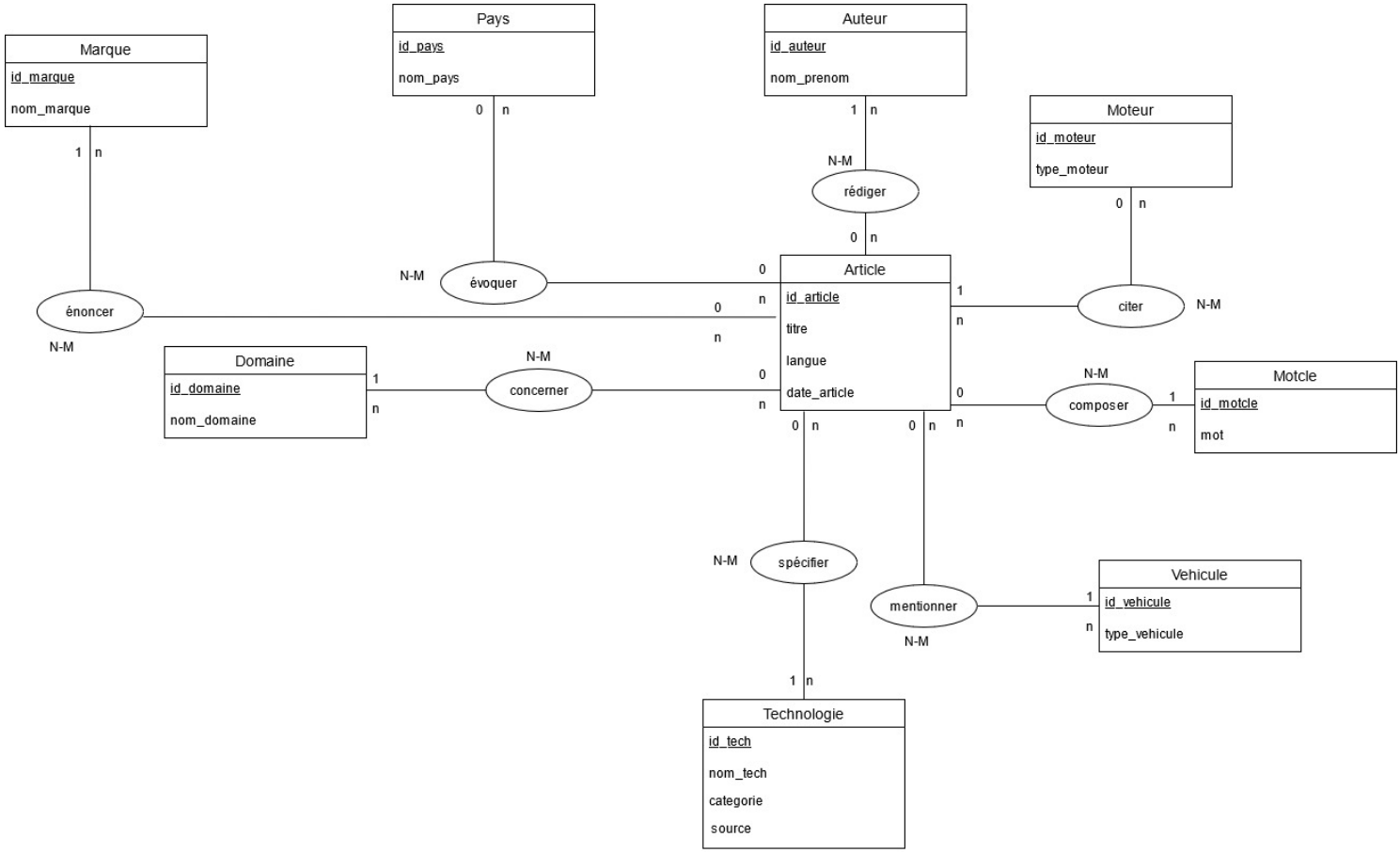


Figure 8 - Modèle Conceptuel de Données_V1.2

Pour chaque entité, nous avons choisi d'utiliser un identifiant comme clé primaire. Cet identifiant est un entier qui sera incrémenté à chaque fois que l'on insère une nouvelle ligne. Cela nous permet d'éviter les problèmes de duplication de clé primaire.

Rapidement, présentons l'intérêt de chaque entité.

- *Article* : entité permettant de stocker les différents articles récupérés. En plus de la clé primaire représentée par *id_article,* elle est composée des attributs suivants : *titre* correspondant au titre de l'article, *langue* pour connaitre la langue d'écriture de l'article (la grande majorité des résumés sont écrits en anglais mais cela n'empêche pas que les articles soient rédigés dans une autre langue) et *date_article* pour la date de publication de l'article.
- *Mot-clé* : entité permettant de stocker les mots-clés qui résume l'article dont ils sont issus. Elle est composée de *id_motcle* la clé primaire et de *mot* qui est un mot clé. Un mot peut être un mot simple ou un groupe de mots (mot composé, petite phrase, expression, etc.).
- *Vehicule* : entité permettant de regrouper les différents véhicules mentionnés dans les articles. *id_vehicule* est la clé primaire et *type_vehicule* correspond au type de véhicule : voiture, camion, bateau, fusée, etc.
- *Technologie* : entité stockant les différentes technologies mentionnées. Plusieurs types de technologie seront présentes : type de production de l'hydrogène et type de stockage de l'hydrogène. *id_tech* est la clé primaire, *nom_tech* est le nom de la technologie mentionnée, *catégorie* permet de différencier les technologies (stockage ou production), *source* renseigne l'origine de l'hydrogène lorsque la technologie est un stockage (hydrocarbures, eau, ou autre).
- *Domaine* : entité concernant le domaine d'un article. *id_domaine* est la clé primaire, *nom_domaine* est le nom du domaine : transport de passager, transport de marchandise, militaire, spatial, etc.
- *Marque* : entité concernant les marques citées dans les articles. *id_marque* est la clé primaire, *nom_marque* est le nom de la marque.
- *Pays* : entité concernant les pays mentionnés dans un article. *id_pays* est la clé primaire, *nom_pays* est le nom du pays rencontré.
- *Auteur* : entité concernant le ou les auteurs d'un article. *id_auteur* est la clé primaire, *nom_prenom* est la concaténation du nom et du prénom de l'auteur.
- *Moteur* : entité concernant le type de moteur évoqué dans un article. *id_moteur* est la clé primaire, *type_moteur* est le type de moteur mentionné : réaction, hybride, combustion interne ou pile à combustible.

L'étape suivante dans la conception d'une base de données est de transformer le MCD en MLD (Modèle Logique de Données).

En effet, le MCD ne peut pas être implanté dans une base de données sans modification. Il est obligatoire de transformer ce modèle. On dit qu'on effectue un passage du modèle conceptuel de données vers le modèle logique de données. Le Modèle Logique de Données est également appelé modèle (ou schéma) relationnel (lorsqu'on travaille avec une base de données relationnelle). Après ce passage, le MLD pourra être implanté dans une base de données relationnelle.

Pour réaliser ce changement, il faut suivre dans l'ordre ces trois règles d'or.

<u>- Règle 1 :</u>

A chaque entité est associé un schéma de relation composé de tous les attributs de l'entité. La clé de l'entité devient la clé de la relation.

<u>- Règle 2 :</u>

Si dans une association A, il existe une entité E pour laquelle la cardinalité du couple (E,A) est (0,1) ou (1,1) (autrement dit l'association est donc de type 1-1 ou 1-N), alors on ajoute dans le schéma de la relation R (celle qui traduit E) tous les attributs de l'association ainsi que la clé de chacune des autres entités de l'association.

<u>- Règle 3 :</u>

Si dans une association A, il n'existe pas une entité E pour laquelle la cardinalité du couple (E,A) est égale à (0,1) ou (1,1) (autrement dit l'association est de type N-M), alors on crée un nouveau schéma de relation contenant une clé de chaque entité participant à l'association ainsi que tous les attributs de l'association.

<u>Remarque</u> : Il n'y a aucune règle pour déterminer la clé de la nouvelle relation.

Une fois qu'on a appliqué ces règles à notre modèle conceptuel de données, on obtient le modèle relationnel suivant :

Modèle Logique de Données (MLD)

Le nom des associations dans le MCD devenant des relations dans le MLD ont été changés afin qu'ils soient plus significatifs et compréhensibles à la lecture et plus manipulable lors de l'implémentation.

Article(id_article, titre, langue, date_article)

Motcle(id_motcle, mot)

Vehicule(od_vehicule, type_vehicule)

Technologie(id_tech, nom_tech, categorie, source)

Domaine(id_domaine, nom_domaine)

Marque (id_marque, nom_marque)

Pays(id_pays, nom_pays)

Auteur(id_auteur, nom_prenom)

Moteur (id_moteur, type_moteur)

Motscles_Evoq(#id_article, #id_motcle)

Vehicules_Evoq(#id_article, #id_vehicule)

Technologies_Evoq(#id_article, #id_tech)

Domaines_Evoq(#id_article, #id_domaine)

Marques_Evoq(#id_article, #id_marque)

Pays_Evoq(#id_article, #id_pays)

Auteurs_Evoq(#id_article, #id_auteur)

Moteurs_Evoq(#id_article, #id_moteur)

Figure 9 - Modèle Logique de Données_V0.2

Comme indiqué au-dessus du schéma relationnel, lors du passage au modèle logique de données, les associations de types N:M sont devenues des relations. Leurs noms ont été changé pour qu'ils soient significatifs et compréhensifs. Pour avoir le détail de ces changements de noms, il faut se reporter à l'annexe 3 contenant le dictionnaire de données.

Une fois que le schéma relationnel fut créé, nous avons pu procéder à l'écriture du script de création de la base (fichier *create_base.sql*).

La base étant créée, l'étape qui vient est tout naturellement l'insertion des données dans celle-ci. Nous avons fait le choix d'insérer toutes les données via Python à partir d'une interface avec la base de données Oracle permis grâce au module « Cx_Oracle ». L'insertion via Python nécessitait tout de même l'usage de requête SQL et le parcours du DataFrame contenant toutes les informations. Pour plus de détails, il est une nouvelle fois judicieux de se reporter au Notebook correspondant. La présence de commentaires facilite sa compréhension et son fonctionnement.

La dernière étape du diagramme est l'activité « Administrer ». C'est en réalité une activité principale du diagramme A0. Comme vu en cours, étant donné qu'elle est difficilement divisible en un niveau plus détailler nous l'avons joint au diagramme A2 au lieu de créer un diagramme A3 qui serait composé de trop peu d'activité.

IV.5 Processus A3 – Administrer

Comme évoqué juste au-dessus, cette partie n'est pas divisible en un niveau plus fin. Nous n'avons donc pas fait de diagramme supplémentaire.

Globalement, le fait d'administrer une base de données signifie savoir :

- Anticiper les volumes de données à stocker
- Procéder au choix des outils pour construire la base de données
- Créer le dictionnaire de données (à retrouver en Annexe 3 – Dictionnaire de données)
- Etc.

Désormais que les données sont stockées dans une base de données, l'objectif est bien évidemment de les exploiter.

IV.6 Processus A4 – Exploiter les données

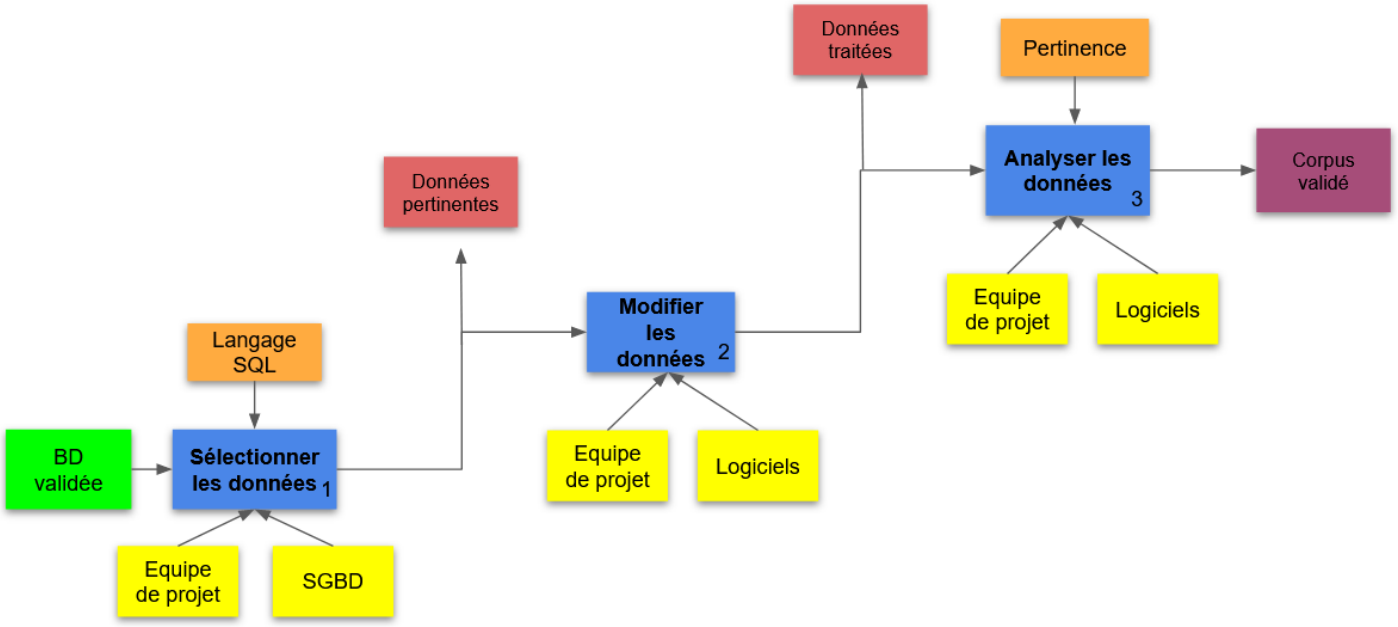


Figure 10 - Diagramme A4 - Exploiter les données

Pour la phase d'exploitation des données, il y a deux étapes essentielles, la sélection de données grâce à des requêtes SQL et l'analyse de données récupérées.

Pour sélectionner les données, nous avons réfléchi aux types d'informations que l'on voudrait mettre en lumière grâce aux éléments contenus dans notre base. C'est à partir de cela que nous avons tous ensemble créé des requêtes SQL permettant d'interroger la base de données. Au total, une vingtaine de requêtes ont été mises au point. Nous avons fait le choix de les regrouper en 4 catégories.

La première catégorie regroupe les informations basiques liées directement aux articles et leur contenu. Par exemple, quels sont les pays les plus évoqués dans les articles portant sur les véhicules hydrogènes, quelles sont les techniques de production les plus citées, les moteurs les plus mentionnés, etc. Cela permet pour commencer d'avoir une première idée du contenu des articles et d'avoir une vue d'ensemble.

Ensuite, la deuxième partie est plus axée sur les véhicules. On y souhaite notamment savoir pour chaque type de véhicules, quel est le type de moteur, les types de production d'hydrogène, de stockage, etc., les plus utilisés.

La troisième partie est axée sur l'évolution dans le temps. Premièrement, on souhaite montrer l'évolution du nombre d'articles chaque année depuis 1945. Deuxièmement, on rentre plus dans les détails, et on souhaite voir quelle est la technologie de production la plus

évoquée chaque année, s'il y a des changements depuis 1945 ou pas ? Même questions pour d'autres attributs.

Enfin, la dernière catégorie est portée sur les pays. Notre souhait est de déterminer pour chaque type de véhicules le top 3 des pays qui s'y intéressent le plus. Ou, dans l'autre sens, pour les 10 pays les plus cités, quels sont les types de véhicules qui les intéressent le plus.

Après avoir traduit toutes ces idées en requêtes SQL valides, nous les avons exécutées et observées les résultats. Cette observation des résultats a entrainé la deuxième étape : l'étape de modification des données. En effet, deux raisons expliquent ces modifications.

Premièrement, la modification est liée au fait que le résultat de certaine requête n'était pas aussi intéressant que ce que l'on avait imaginé. Il ne s'avérait donc plus utile de garder ce type de requête. Ensuite, toujours au vu des résultats il nous a semblé pertinent de modifier quelques éléments de requêtes (ajout de conditions sur les dates ou autre).

Deuxièmement, la modification de données est nécessaire pour certains résultats qui n'ont pas un format adéquat à la visualisation. Par exemple, pour certains auteurs, notamment les auteurs asiatiques, leurs noms et prénoms étant composés de caractères spéciaux ils ne pouvaient pas être stockés en tant que tel dans la base de données. En effet, avant leur insertion nous avons nettoyé mais surtout encodé ces caractères spéciaux pour qu'il n'y ait pas de problème d'insertion. Cependant, les résultats des requêtes sont toujours sous la forme encodée, il faut donc les décoder pour avoir un affichage convenable.

Pour terminer ce processus l'activité finale était l'analyse de données en prévision de l'activité A5 sur la valorisation des données. Effectivement, l'objectif était d'observer les résultats des requêtes pour déterminer sous quelles formes nous pourrons les présenter pour qu'ils soient le plus parlant possible. Autrement dit, il fallait réfléchir aux types de graphiques/diagrammes à mettre en place pour chaque résultat. Cette phase était importante avant de débuter la phase de valorisation des données car elle nous a permis de mieux nous préparer et donc de nous faire gagner du temps.

IV.7 Processus A5 – Valorisation des données

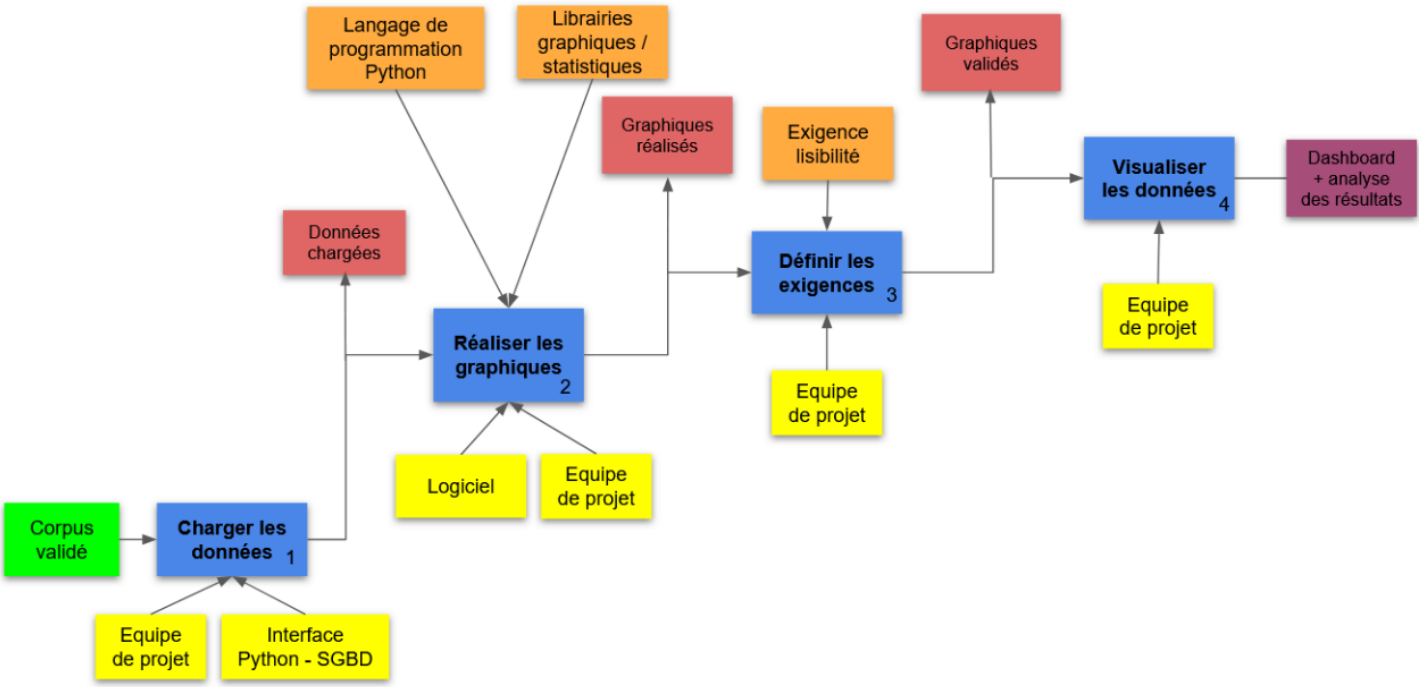


Diagramme A5 : Valorisation les données

Figure 11 - Diagramme A5 - Valorisation des données

Ultime étape du projet, la valorisation des données est primordiale car c'est durant celle-ci que nous allons pouvoir visualiser nos résultats et surtout les analyser pour mettre en place un système d'aide à la décision.

Premièrement, nous avons chargé sur Python les données grâce aux requêtes SQL créées et validées à l'étape précédente.

Une fois ces données récupérées, nous avons pu commencer la réalisation des graphiques. Comme expliqué précédemment, cette étape avait déjà été préparée en amont lors de l'étape 4 puisque nous savions quels types graphiques nous utiliserons pour tel ou tel type de requêtes. Les types de graphiques que nous avons sélectionné étaient très variés. Ils peuvent être très classiques : histogrammes, diagrammes circulaires, courbes en fonction du temps ; un peu plus poussés : diagrammes empilés 100% ; voir originaux : carte du monde pour représenter les pays les plus impliqués dans la recherche scientifique sur les véhicules hydrogènes.

La lisibilité d'un graphique est un aspect auquel nous avons accordé une importance particulière. Effectivement, pour qu'un graphique soit validé, nous veillons à ce qu'il soit lisible, compréhensible, pas trop chargé. Par exemple, les diagrammes circulaires étaient employés uniquement si le nombre de modalités n'excédaient pas 4. Si c'était le cas, nous privilégions les histogrammes.

Après avoir réalisé et validé tous les graphiques, la dernière étape de ce projet était leur analyse dans l'objectif de notre aide à la décision sur les véhicules à hydrogène.

Cette analyse des résultats s'est faite en quatre parties (les mêmes parties qu'évoquées pour les requêtes).

IV.7.1 Partie 1 – Analyses générales sur les articles

Pour rappel, cette première catégorie regroupe les informations basiques liées directement aux articles et leur contenu. Voici nos principales observations basées sur des histogrammes, des diagrammes circulaires et un nuage de mots.

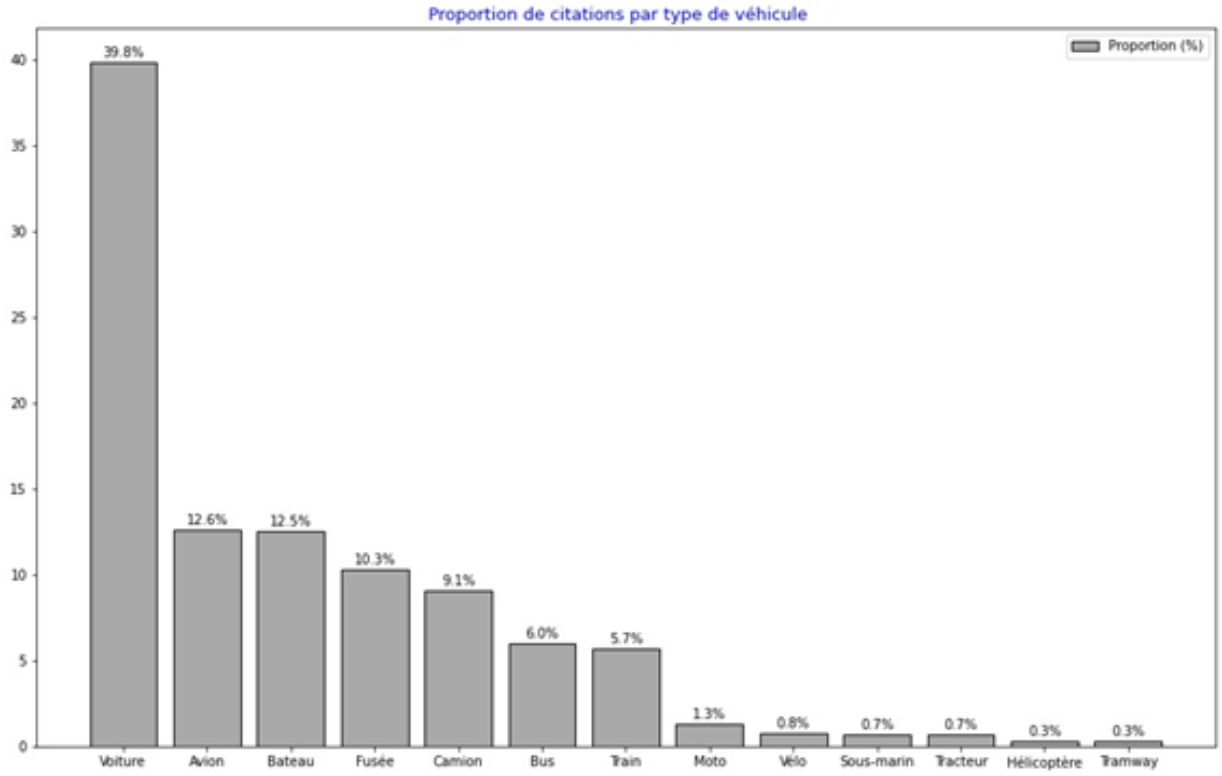


Figure 12 – Proportion de citations par type de véhicule

A l'aide de ce premier graphique, on constate que la voiture est clairement le type de véhicule le plus représenté dans les articles sur les véhicules hydrogènes. En effet, lorsqu'un véhicule est mentionné dans un article, dans 39.8% des cas il s'agit de la voiture. Attention, il ne faut pas interpréter ce graphique comme étant la proportion d'articles parlant des voitures hydrogènes (même si forcément ces deux résultats sont corrélés positivement).

Nettement derrière la voiture, l'avion, le bateau, la fusée et le camion sont les quatre autres types de véhicules pour lesquels la recherche d'un fonctionnement à l'hydrogène est la plus avancée.

En prenant du recul, il est tout à fait normal de voir la voiture si loin devant. Effectivement, c'est le mode de transport le plus utilisé dans le monde (et de loin). Ainsi, dans la lutte contre les émissions de gaz à effet de serre, le plus urgent est de s'attaquer aux voitures. Même analyse pour les avions qui eux représentent un volume d'utilisation moins important, mais tout aussi polluant.

La présence de la fusée si haut dans le classement s'explique par le fait que l'hydrogène est un composant quasiment incontournable dans l'usage des moteurs-fusées.

Le bus n'arrive lui qu'en 6e position, bien que depuis 2019 ils soient de plus en plus présents dans les grandes villes françaises notamment (Paris, Pau, Versailles, etc.).

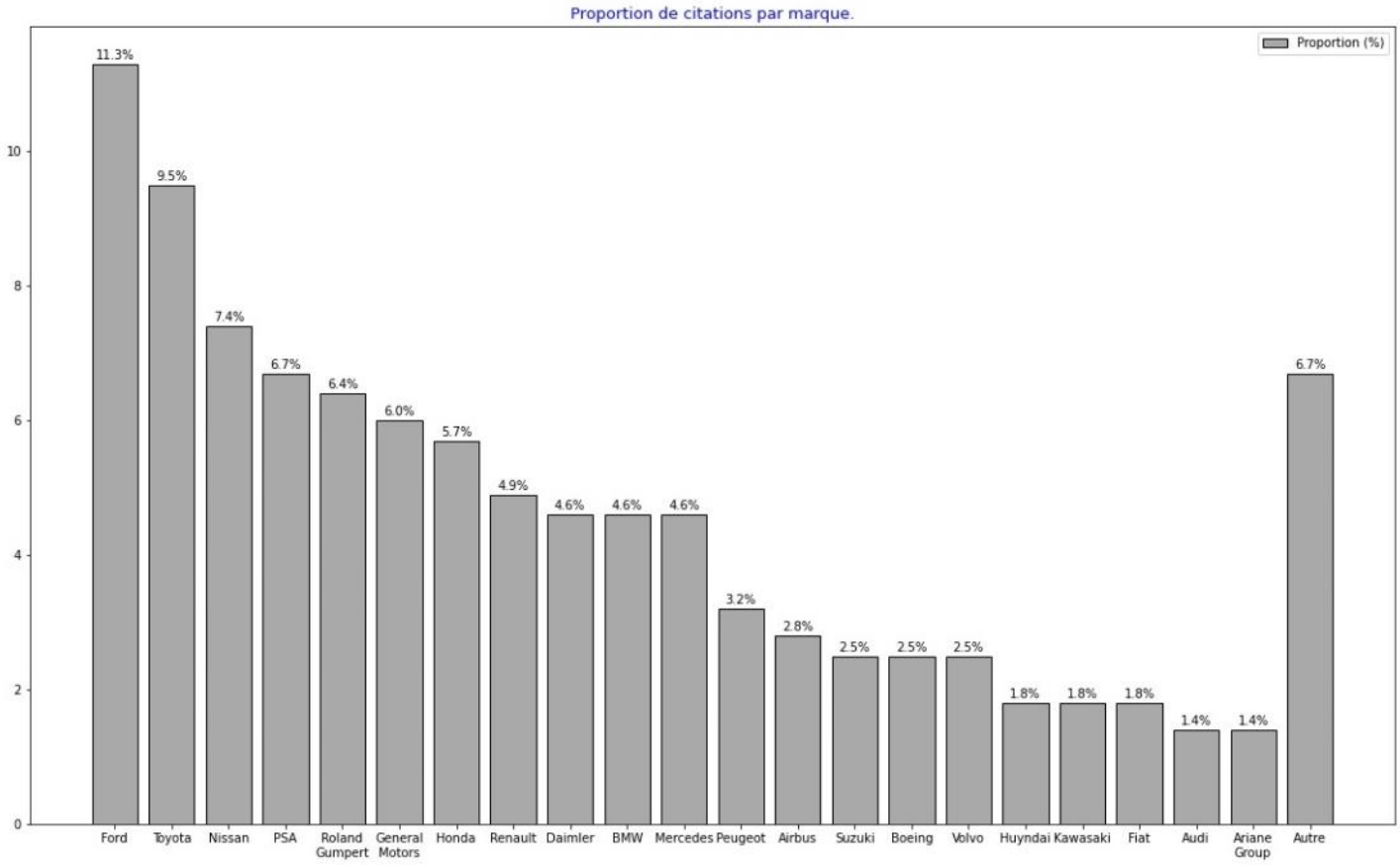


Figure 13 – Proportion de citations par marque de véhicules.

Pour compléter cette partie, il nous a semblé intéressant de voir quelles étaient les marques les plus citées dans les articles.

On retrouve le fait que la voiture est le véhicule le plus représenté dans les articles puisque la majorité des marques plus citées sont des marques automobiles. C'est le constructeur automobile Ford qui semble être le plus en avance sur la recherche à propos des véhicules hydrogènes. Il est évoqué dans 11.3% des cas où une marque de véhicules est citée.

A noter, que la marque Hopium Française crée en 2019, qui ambitionne de créer une berline à hydrogène haut de gamme n'apparait pas dans ce classement (ce qui peut s'expliquer par son apparition très récente par rapport aux autres constructeurs).

Comme marque non-automobile, on retrouve entre autres Daimler (camion) en 9e position et Airbus et Boeing (avion et spatial) respectivement au 12e rang et 14e rang.

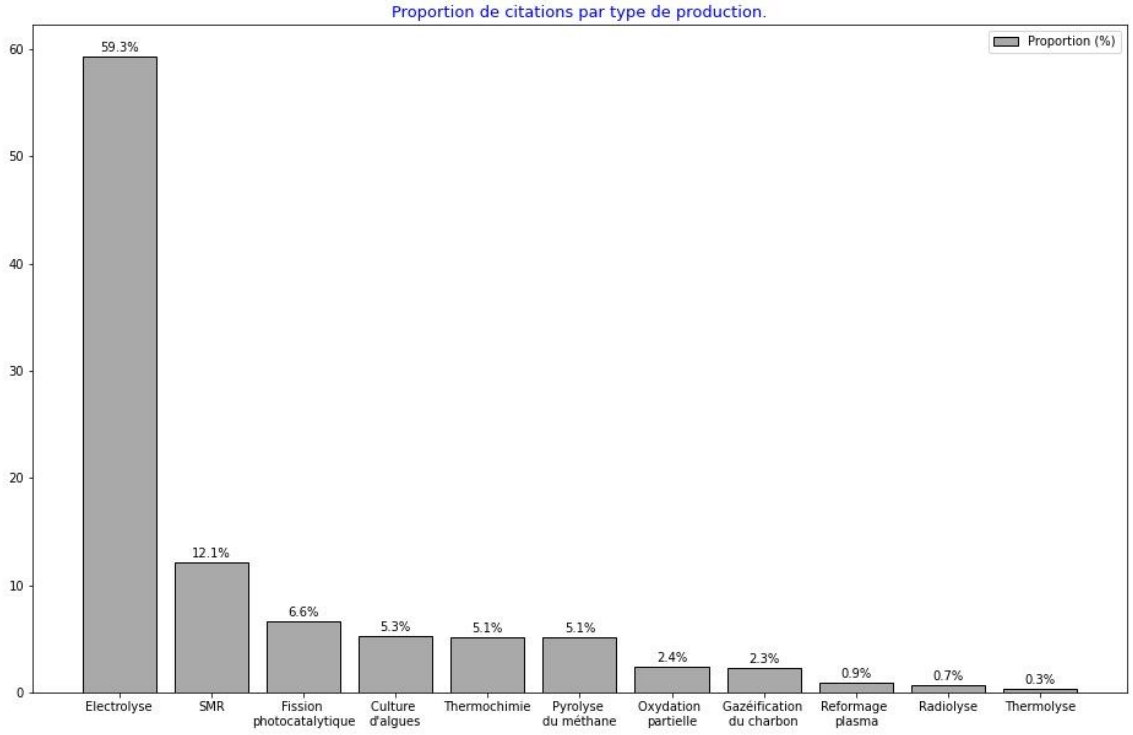


Figure 14 - Proportion de citations par type de production

En ce qui concerne le type de production de l'hydrogène, on voit que l'électrolyse est clairement le mode de production le plus discuté dans les articles de recherches avec 59.3% des citations. C'est procédé électrolytique qui décompose l'eau en dioxygène et dihydrogène grâce à un courant électrique. Un type de production qui n'émet pas de gaz à effet de serre.

La méthode de SMR (pour Steam Methane Reforming en anglais qui signifie le vaporeformage du méthane) arrive loin dernière avec 12.1% des mentions. C'est une méthode de conversion des hydrocarbures qui cause des rejets de gaz à effet de serre. Les autres méthodes ne dépassent pas les 10%.

Ces résultats sont très intéressants si on les compare à la réparation « réelle » du type de production dans l'industrie. Effectivement, en 2015, 49% de la production de dihydrogène à l'échelle mondiale venait de vaporeformage du méthane (SMR) (contre seulement 12.1% de citations dans les articles de recherche), 29 % de l'oxydation partielle du pétrole (seulement 2.4% dans les articles de recherches), 18 % de la gazéification du charbon (2.3% dans la recherche) et seulement 4 % de l'électrolyse de l'eau (plus de 59% de citations dans les articles scientifiques). Ces résultats s'expliquent par le rendement plus faible de l'électrolyse mais surtout un prix de production beaucoup plus élevé ce qui repoussent les industriels à opter

pour celle-ci. Cependant, on voit que la recherche scientifique accorde une importance particulière à l'électrolyse de l'eau, étant donné qu'elle est « verte » c'est surement le mode de production de l'hydrogène du futur au vu du contexte actuel (à condition qu'on règle son problème de rendement et de coût). En effet, le fait que la production de l'hydrogène soit à 95 % basée sur les hydrocarbures, nuance fortement le qualificatif de véhicule « décarboné » qu'on lui accorde parfois.

Pour compléter cette observation nous avons regroupé les types de productions en fonction de leur source :

- soit à partir d'hydrocarbures (SMR, oxydation partielle du pétrole, gazéification du charbon, etc.) et donc polluant (on parle d'hydrogène gris)

- soit à partir d'eau (on parle d'hydrogène vert).

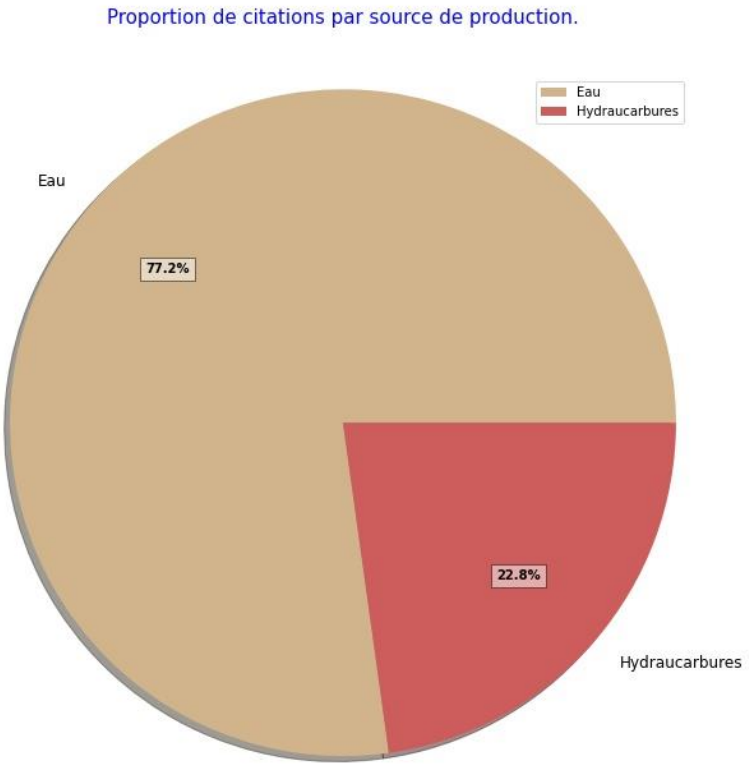


Figure 15 – Proportion de citations par source de production

Nous avons vu précédemment que la production de dihydrogène dans le monde était à 96% basée sur des hydrocarbures et seulement 4% à partir de l'eau. Or, dans les articles de recherches, c'est littéralement l'opposé qui se produit, 77.2% des types de productions évoqués sont des méthodes de production à partir de l'eau contre seulement 22.8% pour les hydrocarbures. Cette différence s'explique forcément par le fait que depuis 30 ans, la lutte face aux émissions de gaz à effet de serre est de plus en plus importante. Il est donc naturel que la science s'intéresse à des types de production « décarbonés » plutôt que polluants.

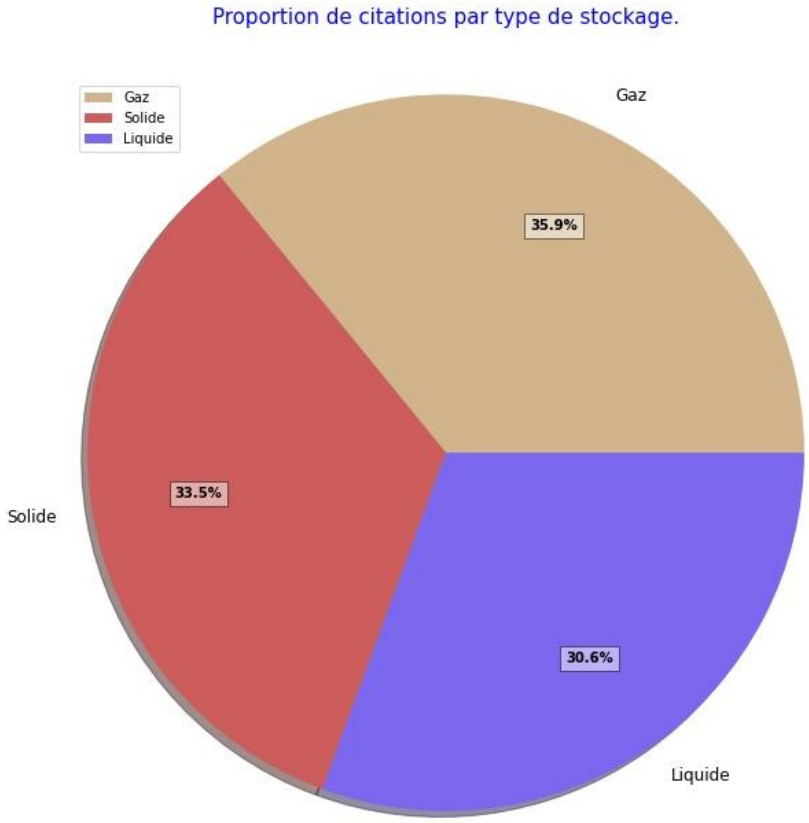


Figure 16 – Proportion de citations par type de stockage

Ce nouveau graphique, présente la répartition des types de stockage de l'hydrogène. On constate une quasi parfaite égalité entre les trois modes de stockages.

L'hydrogène est un gaz extrêmement léger qui occupe un volume important dans les conditions de pression standard, c'est-à-dire en pression atmosphérique. Pour le stocker et le transporter efficacement, il faut fortement réduire ce volume (et donc d'augmenter sa densité).

La méthode la plus simple permettant de diminuer le volume d'un gaz, à température constante, est d'augmenter sa pression, c'est le stockage sous pression (sous forme de gaz).

Une autre technique de pointe pour stocker un maximum d'hydrogène dans un volume restreint consiste à transformer de l'hydrogène gazeux en hydrogène liquide en le refroidissant à très basse température. L'hydrogène se liquéfie lorsqu'on le refroidit à une température inférieure de -252,87°C. C'est le stockage sous forme liquide.

Enfin, les méthodes de stockage de l'hydrogène sous forme solide sont des techniques mettant en jeu des mécanismes d'absorption ou d'adsorption de l'hydrogène par un matériau. En pratique, c'est une technique de stockage un peu moins utilisé que les autres. Cependant, comme l'affirme l'entreprise Air Liquide (groupe industriel français d'envergure internationale spécialiste des gaz industriels) le stockage sous forme solide est « une piste de recherche prometteuse »[1], ce qui explique qu'elle apparaisse si souvent dans les articles de recherches scientifiques.

[1]Citation à retrouver paragraphe « Sous forme solide » à l'adresse suivante : Citation Air Liquide

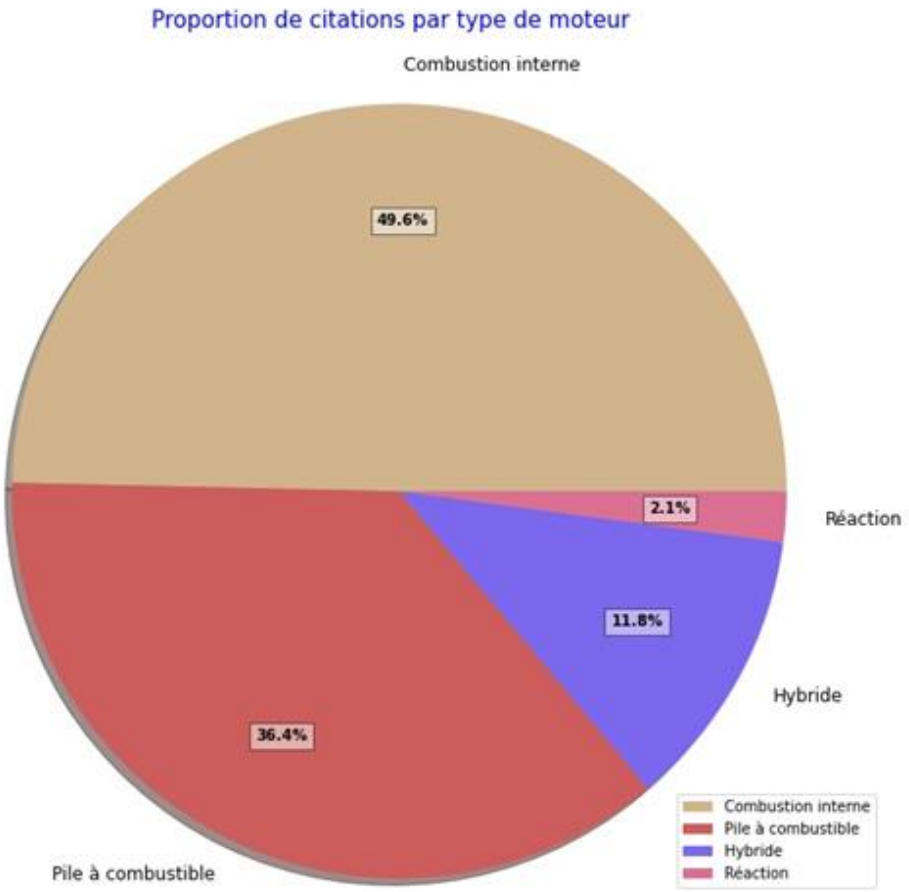


Figure 17 – Proportion de citations par type de moteur

Ce graphique nous permet de connaitre la répartition des moteurs évoqués dans les articles. Lorsqu'un moteur est évoqué dans un article, près d'une fois sur deux c'est pour parler d'un moteur à combustion interne. C'est type de moteur à combustion dans lequel l'énergie thermique dégagée par la combustion est convertie en énergie mécanique à l'intérieur du moteur. A noter, que ce type de moteur est si bien représenté dû au fait que dans certains cas les moteurs à réaction peuvent être classés comme moteur à combustion interne. Par ailleurs, le rendement d'un tel moteur est de 30 à 40%. Ce rendement faible explique que la recherche scientifique cherche à l'améliorer. Cela ne veut pas forcément signifier que c'est le moteur le plus employé dans les véhicules hydrogènes aujourd'hui.

Le deuxième moteur le plus représenté est le moteur qui utilise une pile à combustible. La pile à combustible produit de l'électricité et non pas un mouvement mécanique.

Les moteurs hydrides correspondent à des moteurs où l'on additionne le dihydrogène aux hydrocarbures utilisés classiquement comme carburants.

Le moteur à réaction ne représente que 2.1% des citations. Nous verrons dans la partie 2 pourquoi il apparait si peu.

Intéressons-nous désormais aux domaines les plus souvent présents dans ces articles.

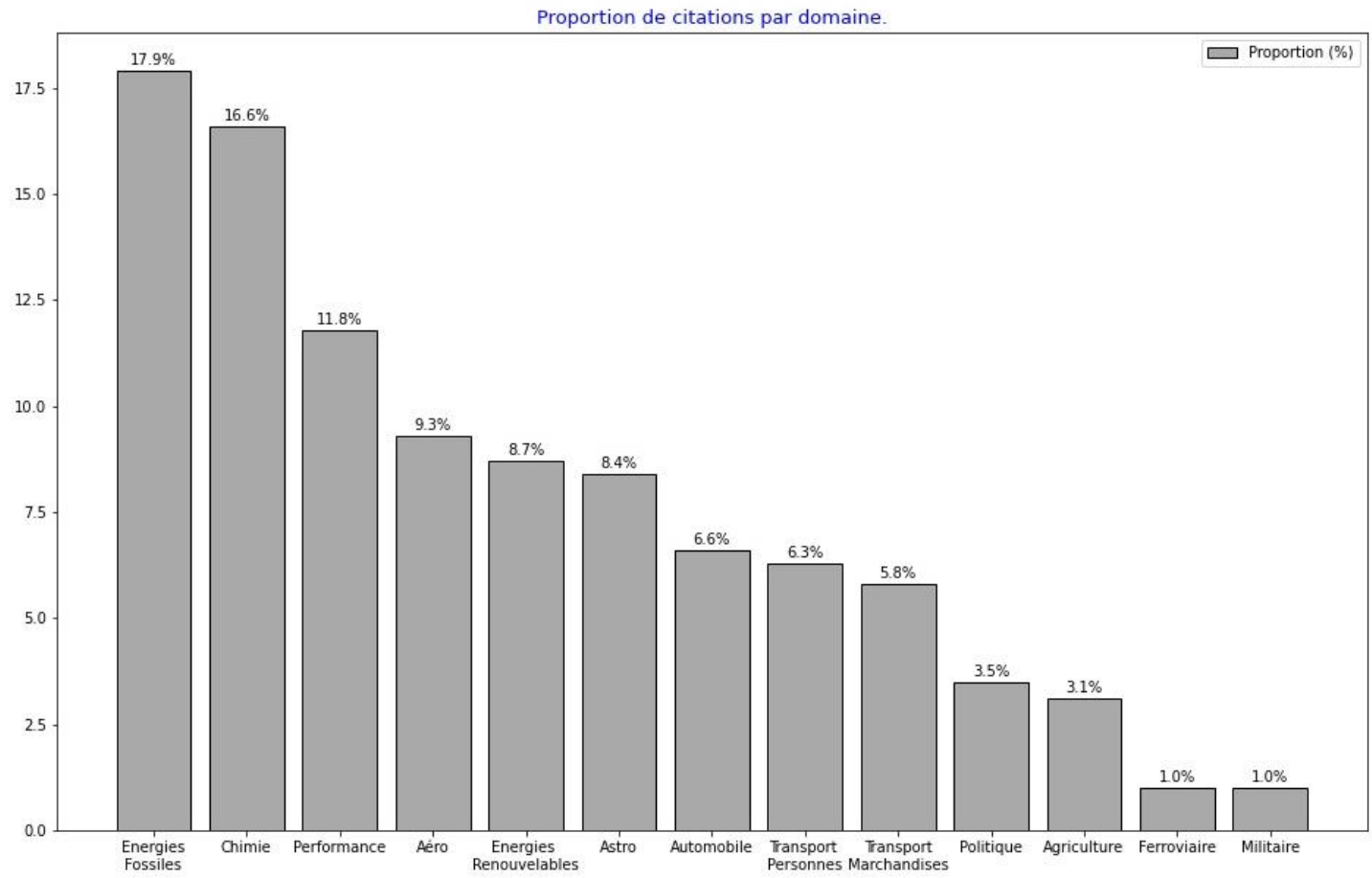


Figure 18 – Proportion de citations par domaine.

Le domaine des énergies fossiles arrive en tête, nous nous doutons alors que pour mettre en avant le côté « décarboné » des véhicules hydrogènes, les auteurs « critiquent » les véhicules classiques fonctionnant aux énergies fossiles. On rappelle également, qu'une partie de la production d'hydrogène est basée sur les hydrocarbures.

Le coté performance est également un domaine qui apparait régulièrement. On imagine alors que les auteurs mettent en avant le rendement, la puissance, l'autonomie de leurs véhicules par rapport à la normale grâce à leurs travaux de recherches.

Paradoxalement, les domaines de l'aéronautique et l'astronomique arrivent devant le domaine automobile alors qu'on a vu en début de partie que la voiture était de loin le véhicule le plus cité dans les articles. Cette anomalie s'explique par le fait que certains mots de vocabulaire sont à la fois associés à l'aéronautique et à l'astronautique, ce qui « dope » leur fréquence d'apparition.

Enfin, on s'aperçoit que les véhicules hydrogènes sont aussi bien destinés au transport de marchandises qu'au transport de personnes.

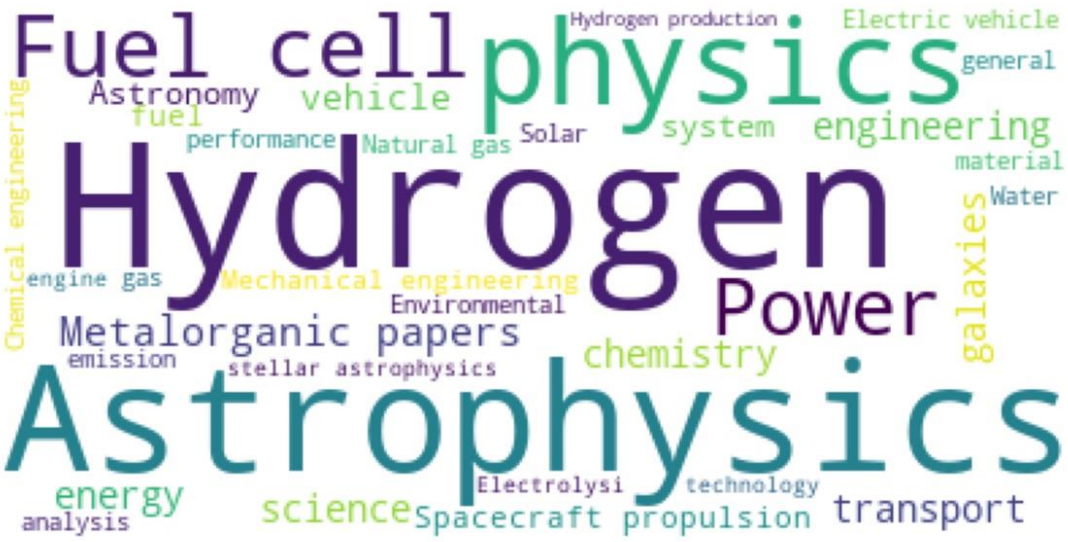


Figure 19 – Nuage de mots représentant la fréquence d'apparition des mots-clés.

Pour conclure cette partie, nous avons réalisé un nuage de mots basé sur la fréquence d'apparition de mots-clés dans les articles.

Sans surprise, on constate que le mot hydrogène est le plus présent, accompagné des mots « vehicle » et « transport », preuve que notre sujet à bien été ciblé et que les informations scrappées sont en adéquation avec ce que l'on cherchait.

Ensuite, on retrouve des informations très similaires que celles issues du diagramme sur les domaines. Effectivement, on retrouve les mots « Astrophysics », « Spacecraft propulsion » et « Galaxies » pour le domaine aérospatial, les mots « Chemistry, « Chemical enginering » et « Electrolysi » pour le côté chimie. Enfin les domaines des énergies fossiles, de la performance et des énergies renouvelables sont représentés respectivement par les mots « gas », « performance » et « solar ».

Voici ce qui conclut la première partie de nos observations. Pour résumer, nous avons donc vu que la voiture qui est notre mode de transport principal est également le véhicule pour lequel la technologie semble être la plus avancée. Cela explique que les marques les plus citées soient des constructeurs de voitures. Par ailleurs, nous avons constaté que la recherche se tournent vers des technologies plus "vertes" afin de protéger la planète.

Désormais, nous pouvons passer à l'analyse plus spécifique des véhicules.

IV.7.2 Partie 2 – Analyses sur les véhicules

Cette partie est donc plus axée sur les véhicules. On y souhaite notamment savoir pour chaque type de véhicule, quel est le type de moteur, le type de production d'hydrogène, le type de stockage, ..., les plus utilisés.

Commençons par confronter les types de véhicules au type de stockage.

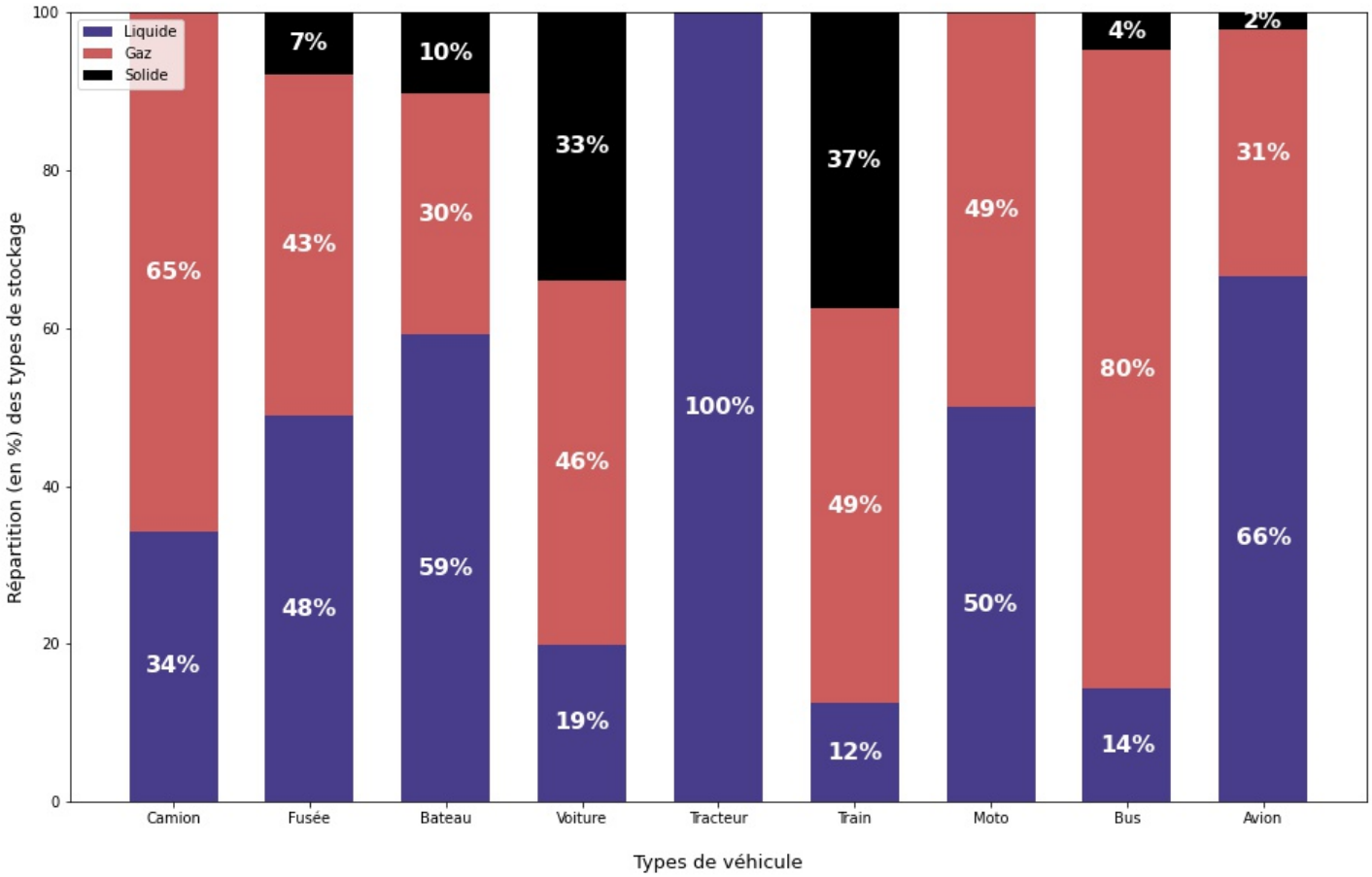


Figure 20 - Répartition des types de stockage en fonction des types de véhicule.

On a vu dans la partie 1, que les trois types de stockages avaient quasiment une proportion de répartition identique dans les articles. Or, en séparant en fonction du type de véhicule, on s'aperçoit que ce n'est plus le cas.

Tout d'abord, première analyse qui saute aux yeux c'est que le stockage sous forme solide n'est majoritaire pour aucun des véhicules. Pourtant nous avons vu précédemment qu'il est cité dans 33.5% des cas où l'on évoque un type de stockage. Comment expliquer cela ? Tout simplement en se rappelant que ce type de stockage n'est encore qu'expérimental et peu développé. Ainsi, lorsqu'on parle de ce type de stockage dans les articles, il n'est pas forcément associé à un type de véhicule. Toutefois, le stockage sous forme solide représente 33% des modes de stockage pour les voitures, le véhicule le plus évoqué dans les articles, une preuve de son intérêt grandissant.

Ensuite, on constate que le stockage de l'hydrogène liquide semble être destiné aux véhicules aériens ou maritime puisqu'il représente 66% des modes de stockage pour l'avion,

59% pour les bateaux et 48% pour les fusées. A noter que pour le tracteur, le 100% est à relativiser du fait qu'il y a très peu d'articles mentionnant les tracteurs (cf. partie 1).

Enfin, le transport routier semble lui privilégier le stockage sous forme gazeuse puisque 80% des bus utilisent ce type de stockage, 65% des camions et 46% des voitures.

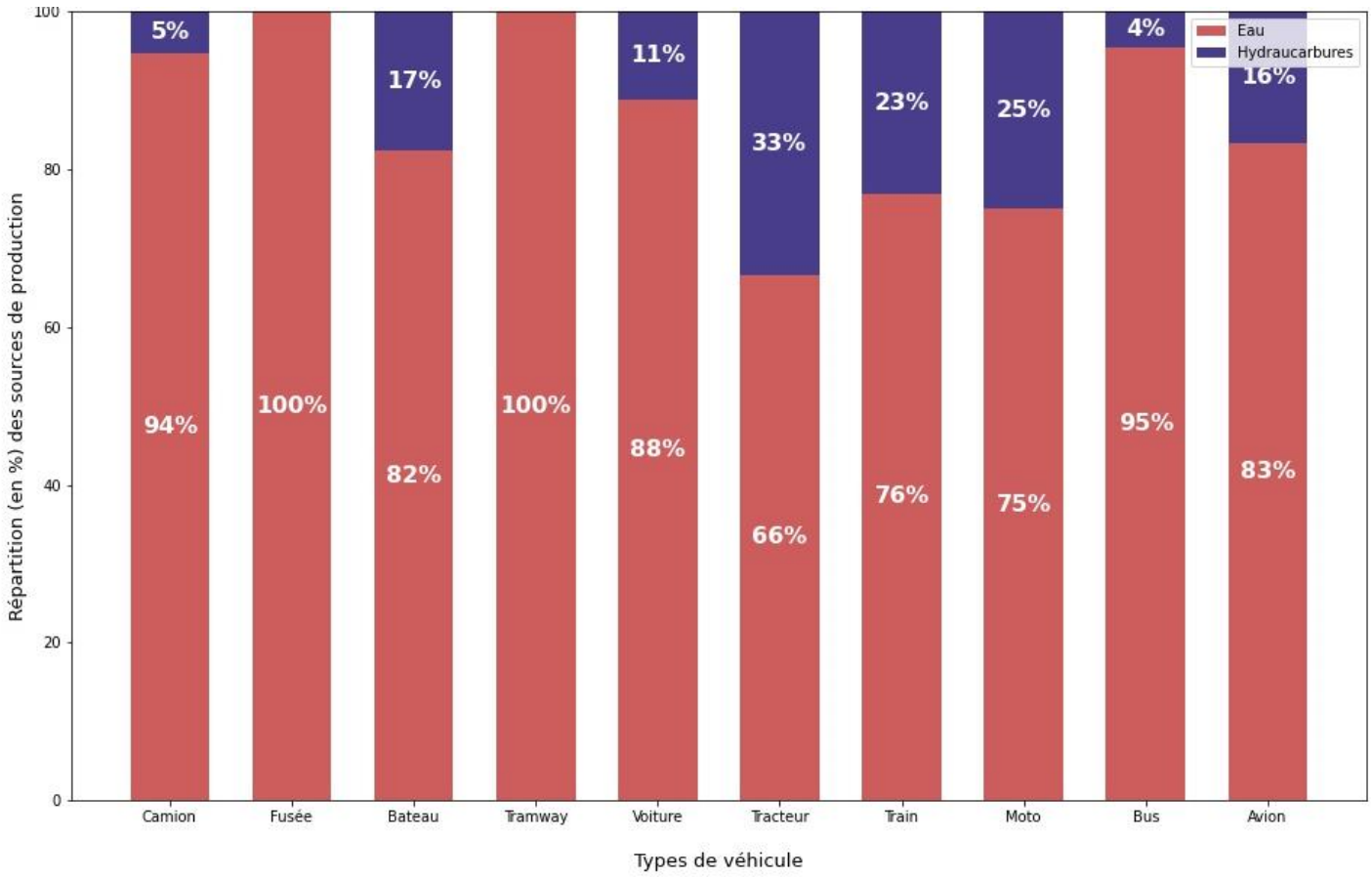


Figure 21 - Répartition des sources de production en fonction des types de véhicule.

Pour les sources de productions de dihydrogènes, il n'y a quasiment aucune différence entre les véhicules, tous accordent une grande importance à une source qui n'émet pas de gaz à effet de serre. On rappelle, que cela n'est valable que dans la « théorie », c'est-à-dire à travers la recherche. En pratique l'énorme majorité de la production de l'hydrogène se fait à partir d'hydrocarbures, en tout cas, pour l'instant…

Pour finir, nous avons fait le choix de faire une analyse croisée entre les véhicules et les types de moteur.

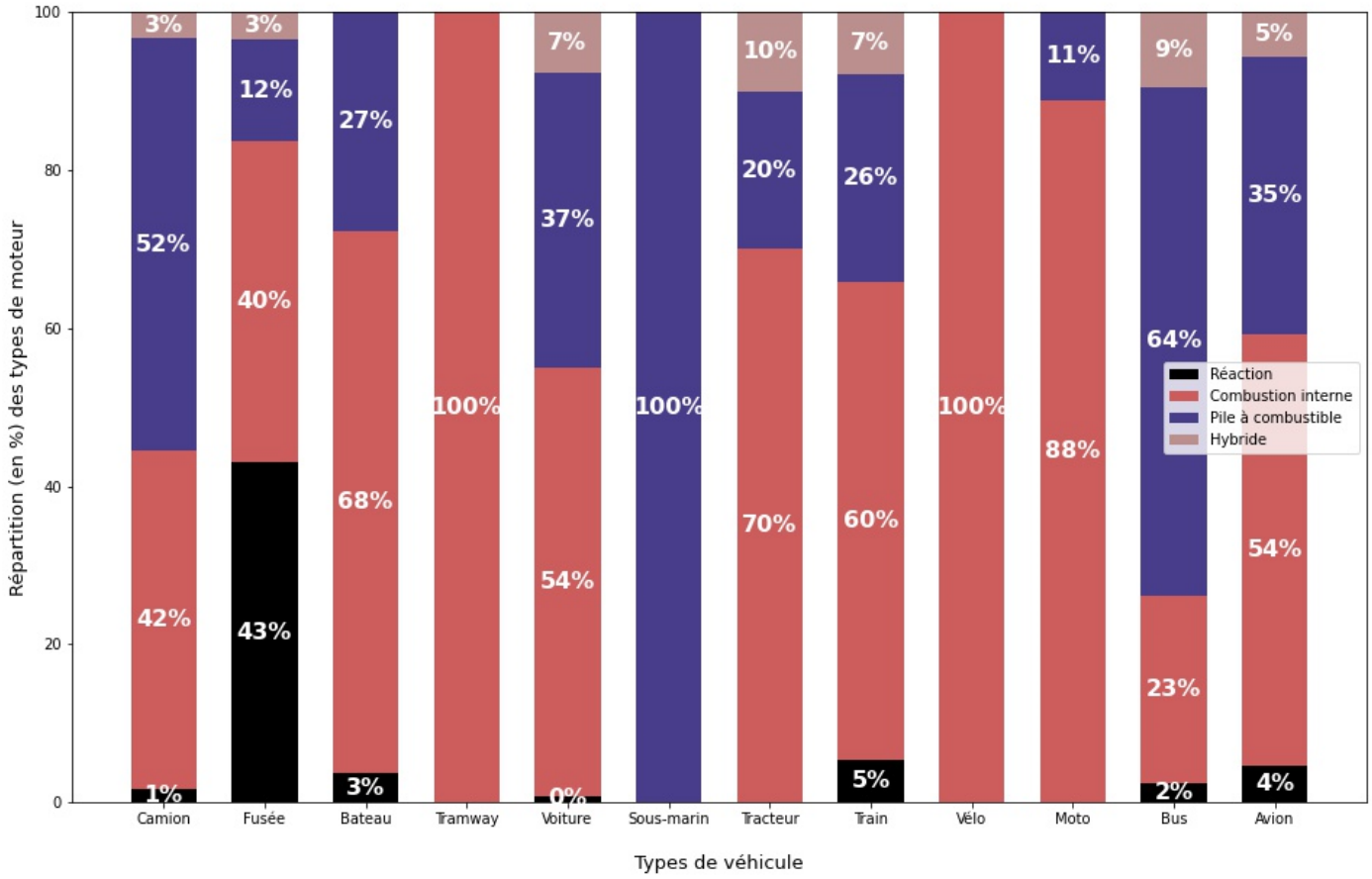


Figure 22 - Répartition des types de moteur par types de véhicule.

Sans surprise, la quasi-totalité des moteurs à réaction se trouvent au niveau des fusées. A noter qu'il peut sembler perturbant de ne pas avoir 100% des moteurs à réaction pour les fusées. Cela s'explique par deux raisons. Premièrement, dans certains cas, les moteurs à réaction peuvent être considérés comme des moteurs à combustion interne (d'où le 40%). Deuxièmement, le scraping ne peut être parfait, il y a forcément quelques approximations, qui explique pourquoi certaines fusées sont associées aux moteurs qui couple une pile à combustible (ce qui est en pratique impossible).

Ensuite, il semble y avoir une nouvelle fois une même tendance qui se dégage pour les véhicules routiers : la pile à combustible. Effectivement, 64% des moteurs associés aux bus sont des moteurs utilisant une pile à combustible, 52% pour les camions et 37% pour les voitures.

Le transport aérien et maritime est lui plutôt tourné vers les moteurs à combustion interne.

Pour finir, les véhicules hybrides représentent pour chaque type de véhicules moins de 10% des solutions envisagées.

Pour conclure, grâce à cette partie, on constate qu'il y a une tendance qui semble se dégager dans la recherche scientifique sur les véhicules à hydrogène. En effet, les véhicules routiers (voiture, camion, bus) semblent être plus enclins à fonctionner avec un moteur de type pile à combustible nécessitant un stockage de l'hydrogène sous forme gazeuse.

Au contraire, les véhicules aériens et maritimes semblent eux mieux supporter le stockage de l'hydrogène sous forme liquide et d'un moteur à réaction ou à combustion interne.

Enfin, pour tous les véhicules, la priorité est donnée à une production de l'hydrogène à partir de l'eau plutôt qu'à partir des hydrocarbures.

IV.7.3 Partie 3 - Analyses temporelles

A travers cette nouvelle partie, nous allons nous intéresser à l'évolution des véhicules à hydrogène au fil des années depuis 1945.

Pour commencer, nous avons décidé de regarder l'évolution du nombre d'articles.

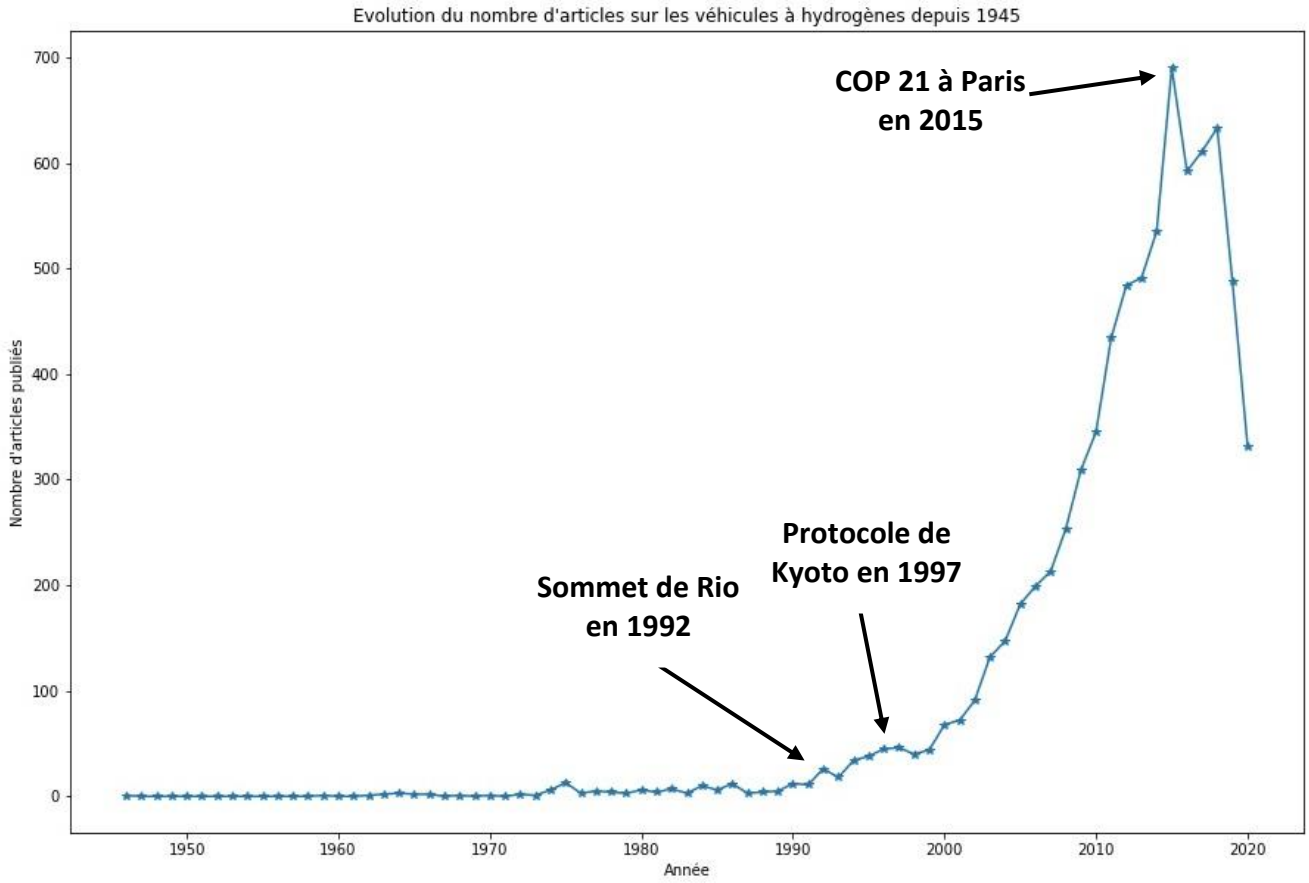


Figure 23 - Evolution du nombre d'articles sur les véhicules hydrogènes depuis 1945.

On voit clairement que le nombre d'articles sur les véhicules hydrogènes connait une augmentation exponentielle depuis les années 1990/2000. Cette augmentation a atteint son apogée en 2015 avant de stagner les années suivantes. La baisse en 2020 est à relativiser étant donné la crise sanitaire liée au Covid-19. Cependant, le nombre d'articles avait déjà connu une importante baisse en 2019. L'année 2021 sera une année importante pour juger si cette diminution va se confirmer ou pas.

La brusque augmentation à partir des années 1990 peut s'expliquer par le développement d'Internet et donc la multiplication des publications scientifiques en ligne mais cela n'explique pas tout.

Effectivement, comme indiqué sur le graphique, cette augmentation est surement aussi liée à deux événements majeurs dans la lutte contre le réchauffement climatique témoignant de l'émergence d'une prise de conscience au niveau mondial.

Premièrement, la convention-cadre des Nations unies sur les changements climatiques (CCNUCC) est signée en 1992 lors du sommet de la Terre à Rio de Janeiro. Les signataires de cette convention, entrée en vigueur en 1994, se fixent comme objectif de stabiliser la concentration des gaz à effet de serre dans l'atmosphère à « un niveau qui empêche toute perturbation anthropique dangereuse du climat »

Deuxièmement, les signataires de la CCNUCC adoptent en 1995 lors de la troisième conférence des parties à la convention (COP 3) à Kyoto, au Japon, le protocole de Kyoto. Cet accord international vise à la réduction des émissions de gaz à effet de serre (dioxyde de carbone, méthane, protoxyde d'azote et trois substituts des chlorofluorocarbones) et vient s'ajouter à la CCNUCC.

Face à cette prise de conscience de la dangerosité du réchauffement climatique et de la nécessité de diminuer les émissions de gaz à effet de serre au niveau mondial de manière urgente, trouver des moyens de transports « décarbonés » est l'une des solutions à exploiter (le transport étant un secteur particulièrement responsable de ces émissions). C'est donc l'une des raisons qui explique sans doute l'augmentation brutal de la recherche sur les véhicules hydrogènes à partir des années 1990.

L'apogée de cette croissance est atteinte en 2015, ce qui coïncide avec l'arrivée sur le marché des premiers véhicules hydrogènes (Toyota Mirai notamment). Une période qui coïncide également, de manière plus anecdotique certes, avec la date de la Conférence des parties 21 (COP 21) à Paris, autre événement clé dans la lutte contre le réchauffement climatique.

Désormais, poursuivons en regardant pour chaque année, quelle technologie, quel véhicule ou quel pays arrivait en tête des citations.

A noter que nous commençons ces analyses depuis 1990 car avant le nombre d'articles est trop faible pour avoir des analyses fiables.

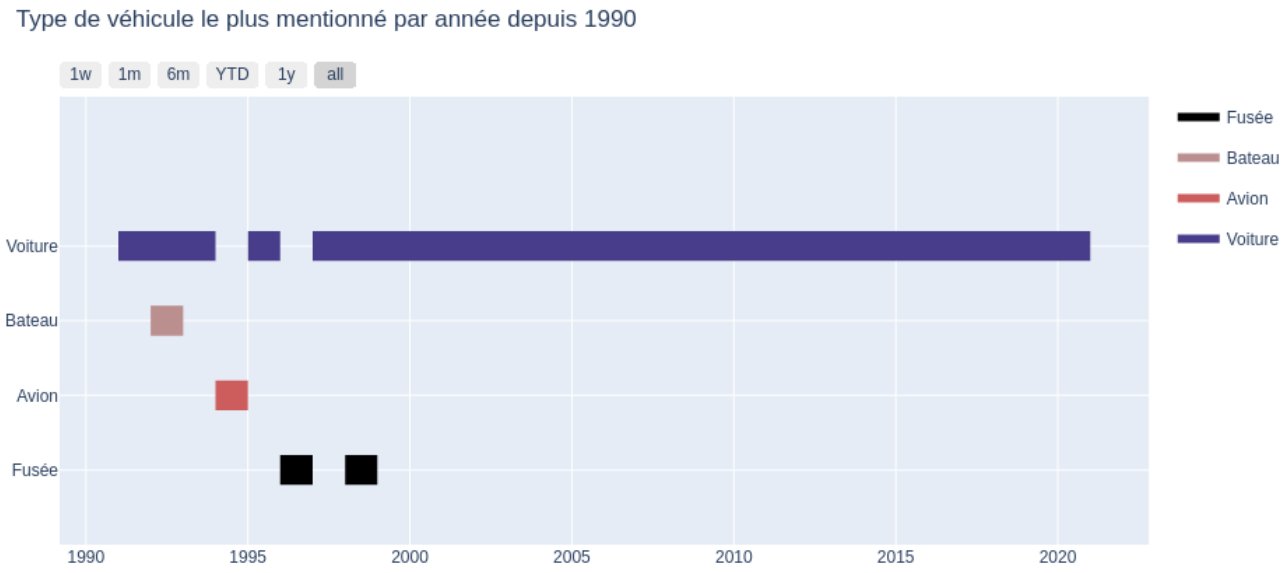


Figure 25 - Type de véhicule le plus mentionné chaque année depuis 1990

On observe clairement que la voiture est largement devant, notamment depuis les années 2000. Toutefois, dans les années 1990, il y avait d'autres types de véhicule qui occupaient les premiers rôles aux côtés de la voiture. Voir la fusée dans les premières positions à la fin des années 1990 est plutôt naturel étant donné qu'à cette période la conquête spatiale était en pleine expansion (par exemple, le premier module de la station spatiale internationale (ISS) a été mis en place en 1998).

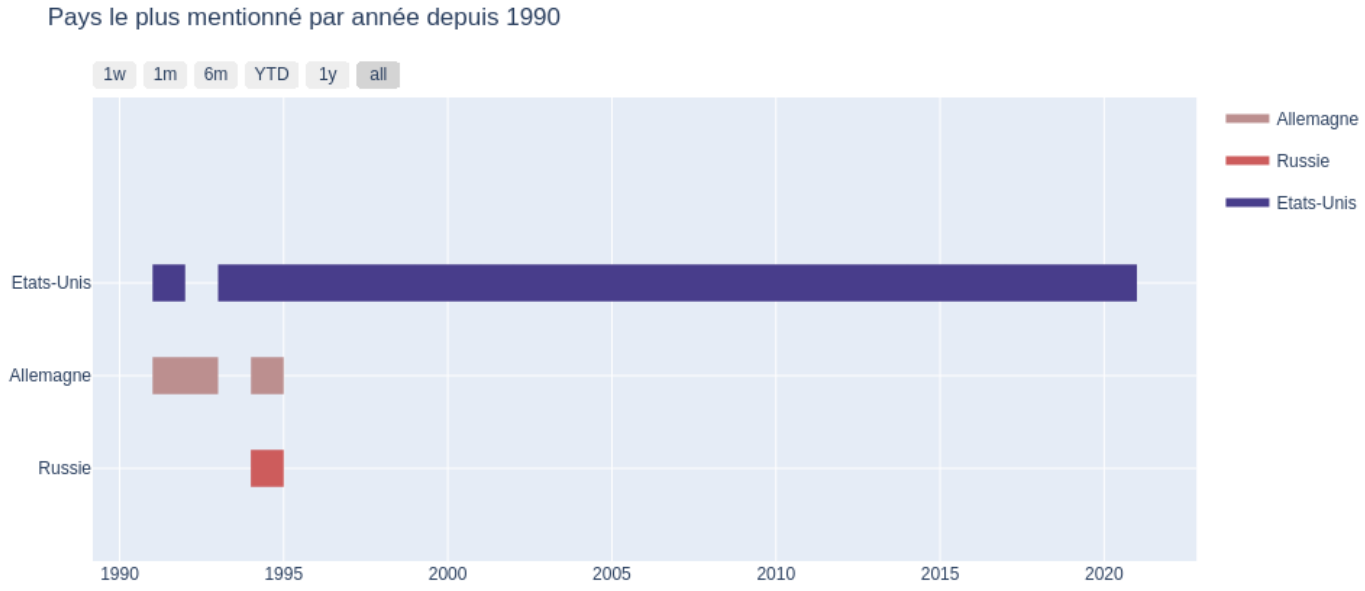


Figure 24 - Pays le plus mentionné chaque année depuis 1990

L'analyse est plutôt similaire avec les pays, en effet, on constate que ce sont les Etats-Unis qui caracolent largement en tête depuis les années 1995. Cependant, au début des années 1990, l'Allemagne et la Russie faisaient jeu égal avec les Etats-Unis. Nous reviendrons plus en détail sur les pays dans la partie 4.

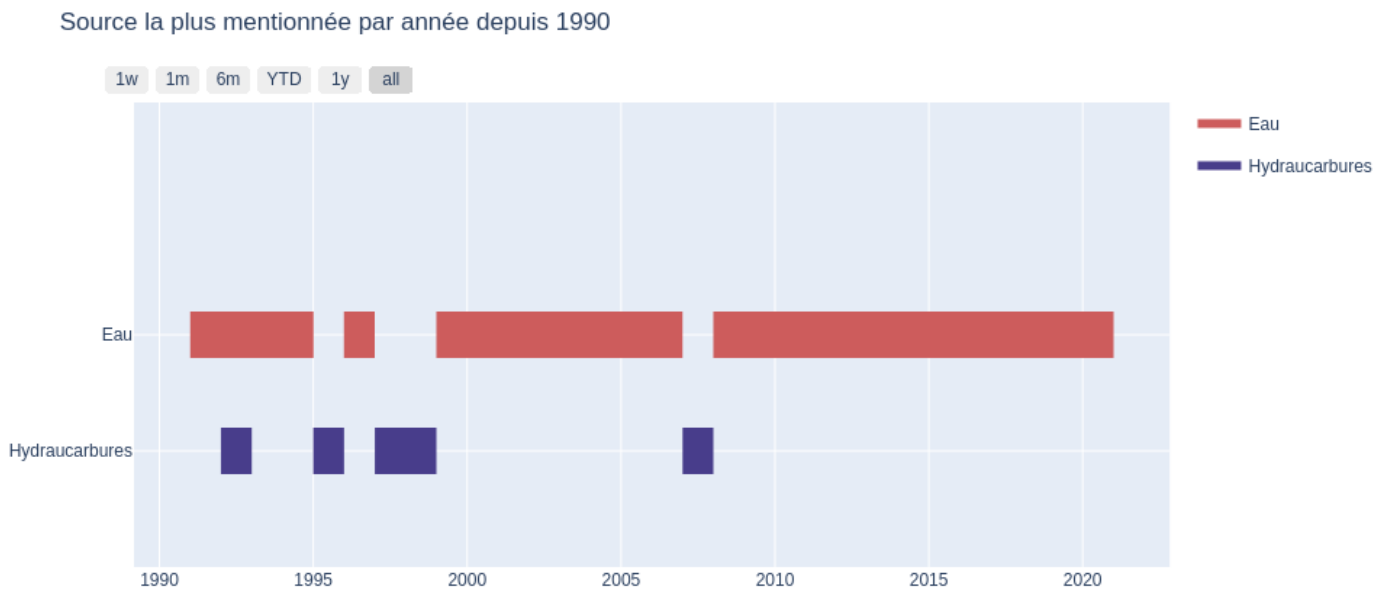


Figure 26 – Source d'hydrogène la plus évoquée chaque année depuis 1990

Nous nous sommes intéressés ensuite aux types de source de production de l'hydrogène. L'observation que l'on peut faire à partir du graphique ci-dessus est très claire, il y a eu une réelle prise de conscience dans les années 1990 au sujet du danger des émissions de gaz à effet de serre liées à l'exploitation des hydrocarbures. Ainsi, les articles de recherche sur la production d'hydrogène à partir des hydrocarbures n'est plus majoritaire au profit de la production à partir de l'eau depuis cette prise de conscience.

Pour rappel, nous avions vu lors de la partie 1 que la production de dihydrogène avait pour source des hydrocarbures dans 22.8% des cas. D'après ce graphique, on suppose que cette proportion est bien moins élevée depuis les années 2000 (même si en pratique dans le monde industriel, c'est toujours la source de production la plus employée à l'heure actuelle).

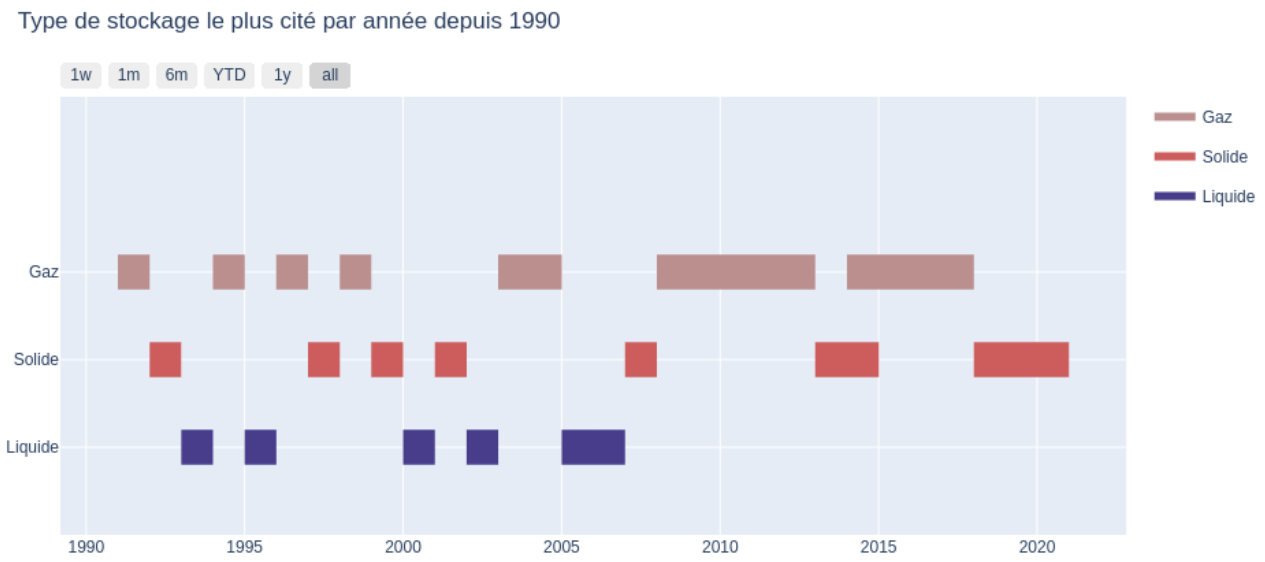


Figure 27 - Type de stockage le plus cité chaque année depuis 1990

En ce qui concerne les types de stockage, comme ce que nous avons observé dans la partie 1, il y a une forte homogénéité. Cependant, depuis les années 2010 il y a plutôt une tendance pour le stockage sous forme gazeuse, puis sous forme solide plus récemment. Retrouver le stockage sous forme solide en tête depuis 2017/2018 n'est pas une surprise. Effectivement, rappelons-nous que dans la partie 1 nous évoquions que, selon l'entreprise Air Liquide, le stockage sous cette forme est « une piste de recherche prometteuse », ce graphique confirme et valide cette information.

Le stockage sous forme liquide semble lui avoir atteint son maximum en termes de recherche à la fin des années 2000. Cependant, cela ne veut pas dire que ce mode de stockage n'est plus utilisé, au contraire, les performances attendues ont dû être atteintes entrainant la diminution des articles le spécifiant.

Voici ce qui conclut cette troisième et avant-dernière partie. La dernière catégorie d'analyse que nous avons réalisée était basée sur les pays.

IV.7.4 Partie 4 – Analyses sur les pays

Pour commencer, il nous a semblé pertinent de regarder quels étaient les pays les plus cités dans les articles parlant d'hydrogènes.

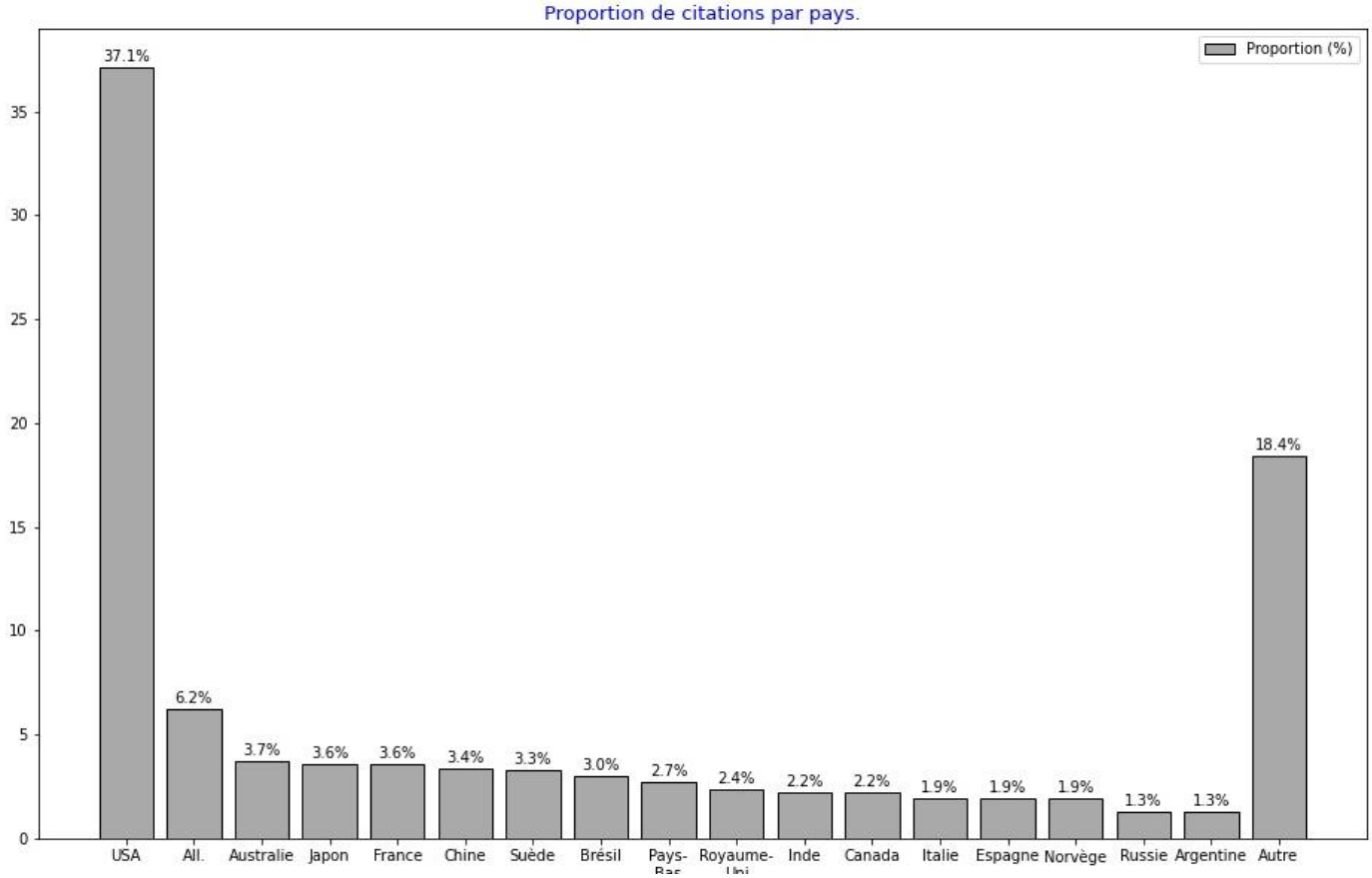


Figure 28 - Proportion de citations par pays

L'histogramme ci-dessus est très clair, les Etats-Unis sont de loin le pays le plus mentionné dans les articles parlant de véhicules à hydrogène. En effet, à chaque fois qu'un pays est mentionné dans un article, dans plus de 37% des cas, ce sont les Etats-Unis dont il s'agit. L'Allemagne arrive loin derrière en deuxième position avec 6.2% des mentions. A noter que la Russie ne se trouve qu'au 16e rang alors que nous avons vu dans la partie précédente qu'elle était dans les pays plus cités dans les années 1995. Elle semble donc avoir un peu délaissée la recherche sur les véhicules hydrogènes depuis.

On constate sans grande surprise la présence des 7 nations membres du Groupe des sept (G7) dans les 13 premières places. Le G7 est un groupe de discussion et de partenariat économique des sept plus grandes puissances avancées du monde qui détiennent environ 45% de la richesse nette mondiale en 2019 : Allemagne, Canada, États-Unis, France, Italie, Japon et Royaume-Uni. A leurs côtés sont présents d'autres puissances dites « développées » (l'Espagne, l'Australie, les Pays-Bas, etc.) mais aussi plusieurs pays émergents. Effectivement, on remarque la présence de la Chine, du Brésil, de l'Inde et de la Russie, pays faisant partie des BRICS (acronyme anglais pour désigner un groupe de cinq pays qui se réunissent depuis 2011 en sommets annuels : Brésil, Russie, Inde, Chine et Afrique du Sud) et de l'Argentine membre des Jaguars (groupe de pays d'Amérique Latine émergents : Mexique, Chili, Colombie et Argentine). Plus généralement, tous ces pays font partie des nouveaux pays industrialisés (NPI).

Nous avons fait le choix de représenter cette information également sous forme de carte :

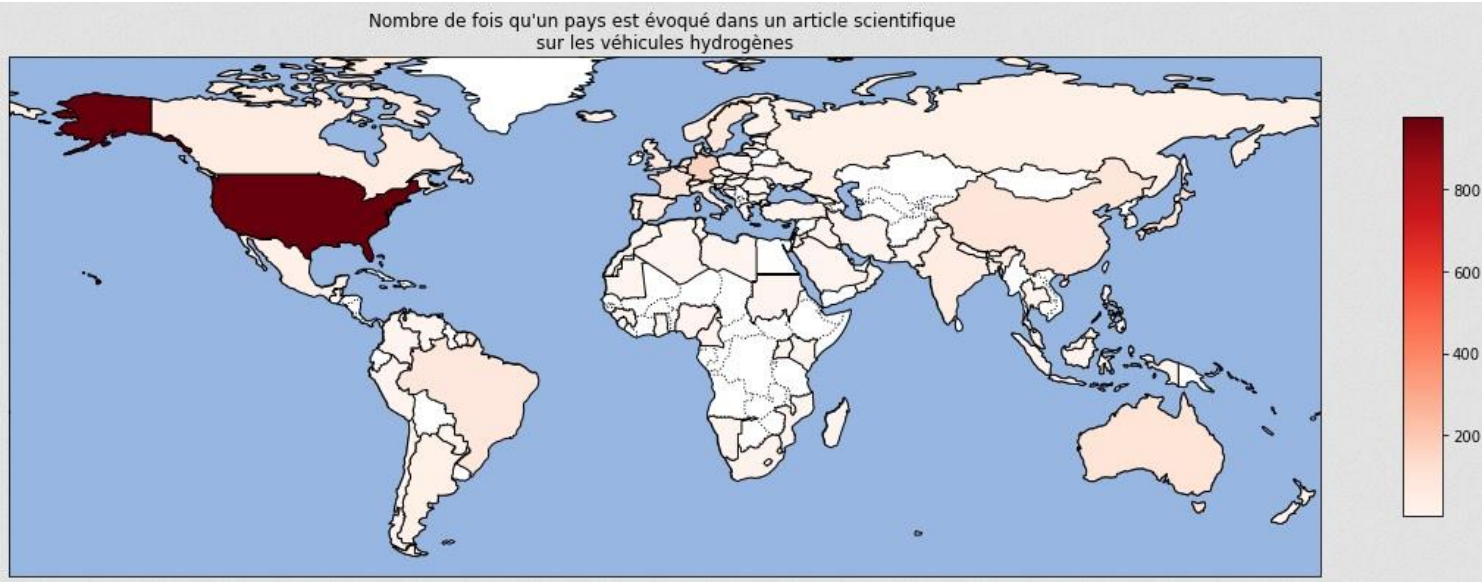


Figure 29 - Nombre de fois qu'un pays est évoqué dans un article sur les véhicules hydrogènes.

Cette carte représente donc le nombre de fois qu'un pays est évoqué dans les articles de recherches scientifiques portant sur les véhicules à hydrogènes. A cause de l'hégémonie des Etats-Unis, la carte n'est pas forcément très lisible, nous avons donc réalisé une deuxième version en retirant les Etats-Unis.

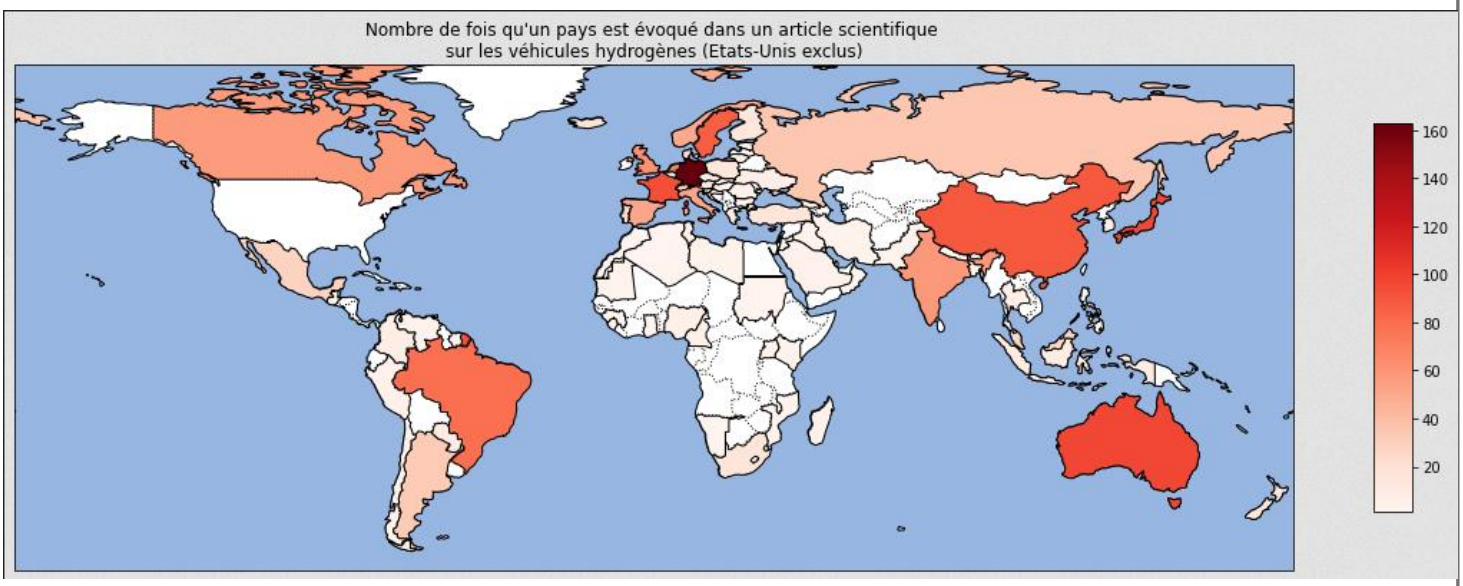


Figure 30 - Nombre de fois qu'un pays est évoqué dans un article scientifique sur les véhicules hydrogènes (USA exclus).

Cette carte est donc beaucoup plus parlante. On constate d'ailleurs que certains pays d'Afrique Subsaharienne ou du Moyen-Orient sont quelques fois évoqués dans les articles scientifiques, information que l'on ne pouvait pas voir avec l'histogramme précédent.

Pour conclure cette partie et l'ensemble de nos analyses, nous avons opté pour la recherche d'informations entre les pays et les types de véhicules.

Premièrement, nous souhaitions observer quel est le top 3 des types de véhicules par pays (nous avons pris que les 10 premiers pays en termes de citations pour limiter le graphique).

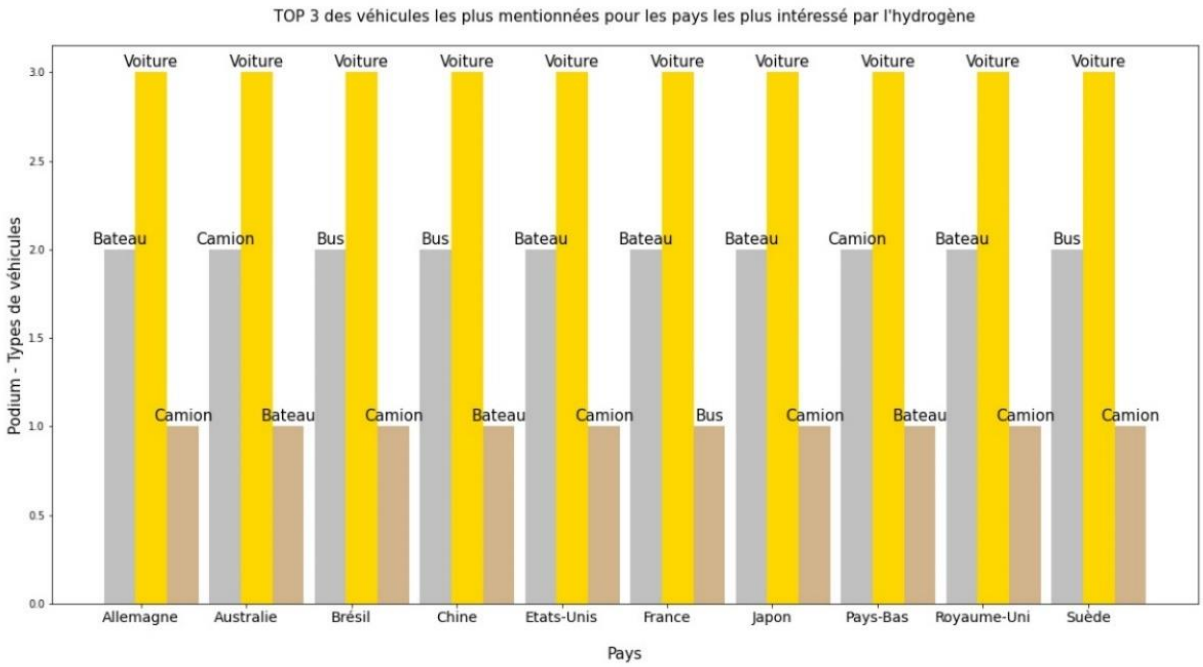


Figure 31 - Top 3 des véhicules les plus mentionnées pour les 10 pays les plus évoqués

Sans surprise, on observe que la voiture est en première position pour chaque pays. En revanche, il y a des différences au niveau de la deuxième place. Majoritairement les pays s'intéressent plutôt aux bateaux, mais pour certains pays ce sont les bus ou les camions qui se classent en deuxième position.

Double petite surprise : l'absence à la fois de l'avion et de la fusée sur chacun des podiums pour ces dix pays. Pourtant, nous avons vu dans la partie 1 que les avions se classaient deuxièmes en termes de citations de véhicules (12.6%) et la fusée 4e (10.3%).

Le deuxième graphique est l'inverse du précédent, il recense pour chaque type de véhicules, le top 3 des pays qui s'y intéressent le plus.

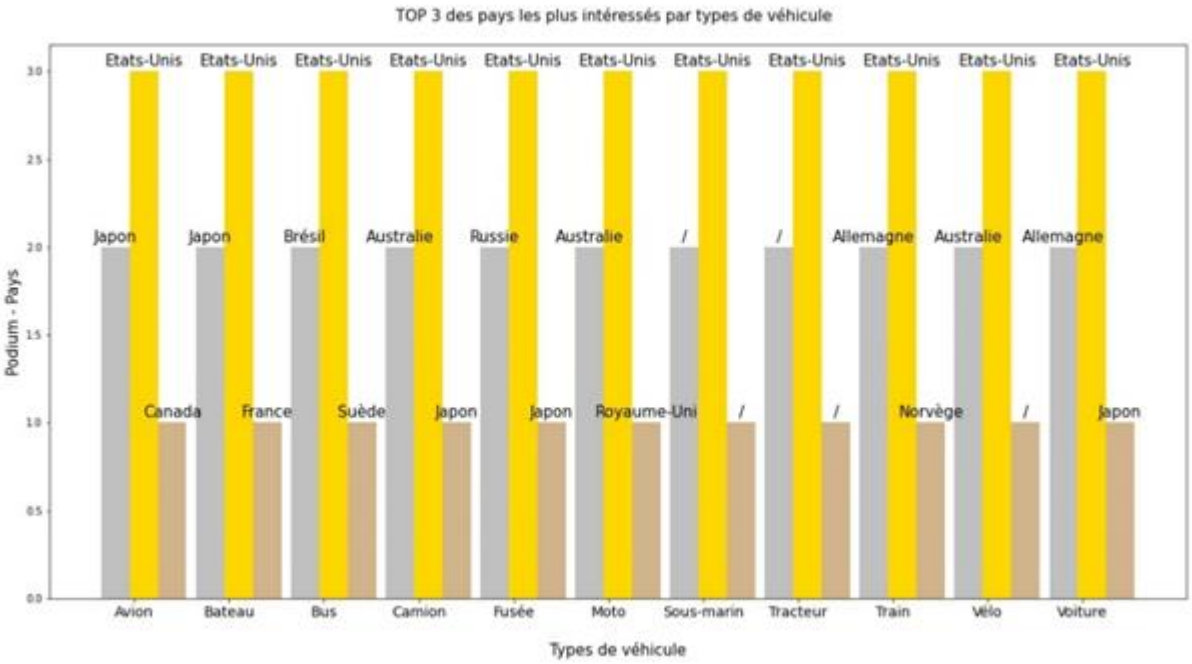


Figure 32 - Top 3 des pays les plus intéressés par types de véhicule.

Une nouvelle fois, les Etats-Unis sont en première position pour chaque type de véhicules. Derrière les Etats-Unis, ce sont souvent le Japon, l'Allemagne ou l'Australie qui s'y trouvent.

A noter la présence de « / » pour indiquer l'absence de pays car le nombre de mentions étaient soit insignifiant ou soit inexistant.

Enfin, l'une des principales observations que l'on peut faire sur ce graphique, est la présence de la Russie en 2e place pour la fusée alors qu'elle ne représente que 1.3% des pays évoqués (voir le début de cette partie). Cette présence n'est en fin de compte pas si étrange puisqu'on sait que le programme spatial russe a été, notamment durant la Guerre Froide avec les Etats-Unis, l'un des programmes les plus avancés au monde.

Grâce à toutes les informations scrappées dans les articles et stockées dans notre base de données, nous avons pu réaliser une quantité d'analyse pléthorique.

Pour résumer rapidement les principaux points à retenir de toutes ces observations, nous pouvons commencer par dire que les articles de recherches portant sur les véhicules à hydrogène sont principalement axés sur la voiture. Etant le moyen de transport le plus utilisé dans le monde, et donc responsable d'une part importante des émissions de gaz à effet de serre, la nécessité de développer une voiture « décarbonée » est urgent pour limiter les dégâts liés au réchauffement climatique. La voiture à hydrogène peut alors être une solution.

Cependant, pour être considérée comme véritablement « décarbonée », la voiture à hydrogène doit surmonter un défi de taille : utiliser un type de production à partir de l'eau et non plus à partir d'hydrocarbures comme c'est actuellement le cas. La production à partir de l'eau étant plus coûteuse et moins rentable, il est nécessaire d'améliorer cette technologie. On voit à travers ces articles de recherches les efforts fournis par la science dans ce sens-là : plus de trois articles sur quatre qui évoquent un type de production, mentionnent un type de production à partir de l'eau.

Le nombre d'articles sur les véhicules à hydrogène a d'ailleurs connu une croissance exponentielle à partir des années 1990, coïncidant avec une prise de conscience générale sur les dangers des émissions de gaz à effet de serre (sommet de Rio en 1992, protocole de Kyoto en 1995). Depuis 2015, on constate une stagnation du nombre de publications d'articles par année, voire même d'une diminution depuis 2 ans, il sera intéressant d'observer si cette tendance va se confirmer dans les années futures.

Par ailleurs, on a constaté qu'il y a une tendance qui semble se dégager dans la recherche scientifique pour les véhicules routiers (voiture, camion, bus) qui semblent être plus enclins à fonctionner avec un moteur de type pile à combustible nécessitant un stockage de l'hydrogène sous forme gazeuse.

Au contraire, les véhicules aériens et maritimes semblent eux mieux supporter le stockage de l'hydrogène sous forme liquide et d'un moteur à réaction ou à combustion interne. Le moteur à combustion interne est justement le type de moteur le plus évoqué dans les articles, cela s'explique par le fait qu'il possède un rendement de seulement 30% à 40%, il faut donc trouver des solutions pour améliorer son rendement.

Enfin, malgré l'hégémonie des Etats-Unis, tous les pays développés ou émergents sont impliqués dans cette recherche sur le véhicule à hydrogène. Reste à savoir si ce sera le véhicule du futur ? Seul l'avenir nous le dira.

V. Gestion de configuration

Nous dévoilons dans cette partie la description technique des divers composants que l'on a développé pour notre système d'aide à la décision. Cette partie a également l'objectif de gérer l'ensemble des modifications apportées au cours de l'évolution du système.

En d'autres termes, grâce à la gestion de configuration, nous sommes à mesure de référencer tous les éléments qui composent le logiciel quel que soit leur état, tout en étant capable de donner les différences entre deux états.

Pour plus de lisibilité nous séparons en plusieurs parties.

- Partie création de la base de données :

Nom du fichier	Version	Date	Modification	Commentaire
MCD_V0	V0.0	08/02/2021	Début de réflexion sur le MCD 1ere version réalisée	
	V0.1	17/02/2021	Ajout d'entités	
MCD_V1	V1.0	19/02/2021	Ajout d'entités et d'attributs plus axé sur notre sujet + suppression des éléments inutiles	Suite à la revue du 17/02/2021 avec le client
	V1.1	03/03/2021	Suppression de l'association triple + ajout de nouvelles entités ('Marque' et 'Moteur')	
	V1.2	10/03/2021	Ajout de nouveaux attributs dans la relation 'Technologie' ('catégorie' et 'source')	
MLD_V0	V0.0	24/02/2021	Création du MLD à partir du MCD validé	Suite à la validation (sous réserve de quelques modifications) du MCD par le client le 24/02/2021 par mail.
	V0.1	03/03/2021	Modifications entrainées par la modification du MCD.	
	V0.2	10/03/2021	Modifications entrainées par la modification du MCD.	
createbase_V0.sql	V0	03/03/2021	Création de la base de données à partir du MLD validé.	Suite à la validation du MLD lors de la revue du 03/03/2021 avec le client.
createbase_V1.sql	V1	10/03/21	Modification suite à la modification du schéma entité association (MCD) et donc du MLD	
insertdata_V0.sql	V0	03/03/2021	Création d'un fichier permettant de peupler la base.	Fichier devenu obsolète le 10/03/21 => Nous privilégions finalement le peuplement via la connexion Python.

- Partie Scraping :

Nom du fichier	Version	Date	Modification	Commentaire
Scraping_V0.ipynb	V0.0	08/02/2021	Création du fichier et des premières fonctions	
	V0.1	15/02/2021	Modification de la partie 1 – Scraping sur les informations générales	
Scraping_V1.ipynb	V1.0	18/02/2021	Modification de l'équation de recherche	Au lieu d'appliquer cette équation de recherche sur le contenu des articles, nous l'appliquons uniquement sur les résumés et les titres => diminution du nombre d'articles mais augmentation sensible de la pertinence des résultats
	V1.1	24/02/2021	Fin de la partie 1 – Scraping sur les informations générales + Début de la partie 2 – Scraping sur les informations détaillées	
	V1.2	03/03/2021	Fin de la partie 2 – Scraping sur les informations détaillées + Début de la partie 3 – Informations à partir des résumés	
	V1.3	10/03/2021	Fin de la partie 3 – Informations à partir des résumés	
	V1.4	20/03/2021	Relecture + Vérification de la qualité	
df_infos.csv	V0	24/02/2021	Fichier permettant de stocker les résultats de scraping de la Partie 1.	
df_content.csv	V0	03/03/2021	Fichier permettant de stocker les résultats de scraping de la Partie 2.	
df_scraping.csv	V0	10/03/2021	Fichier permettant de stocker toutes les informations issues du scraping qui seront insérées dans la base de données.	

- Partie insertion des données dans la base via Python :

Nom du fichier	Version	Date	Modification	Commentaire
insert_database_V0.ipynb	V0.0	10/03/2021	Création du fichier	
	V0.1	10/03/2021	Début partie 1 : Ecriture des requêtes et des fonctions permettant d'insérer les informations « de base »	
	V0.2	13/03/2021	Fin partie 1 + Début partie 2 : Ecriture des requêtes et des fonctions permettant d'insérer les informations liées au scraping	
	V0.3	15/03/2021	Fin Partie 2	
	V0.4	20/03/2021	Relecture + Vérification de la qualité	

- Partie interrogation des données :

Nom du fichier	Version	Date	Modification	Commentaire
requetes_V0.sql	V0.0	10/03/2021	Création du fichier	
	V0.1	10/03/2021	Mise en forme des premières requêtes	
	V0.2	15/03/2021	Création de nouvelles requêtes	
	V0.3	17/03/2021	Correction de certaines requêtes	
	V0.4	20/03/2021	Relecture + Vérification de la qualité	

- Partie visualisation et valorisation des données :

Nom du fichier	Version	Date	Modification	Commentaire
data_visualization_V0.ipynb	V0.0	15/03/2021	Création du fichier	
	V0.1	17/03/2021	Réalisation des premiers graphiques (histogrammes, diagrammes circulaires)	A partir des résultats des requêtes issues du fichiers *requetes_V0*.
	V0.2	18/03/2021	Réalisation de nouveaux graphiques (diagrammes empilés à 100%, cartes, courbes)	
	V0.3	20/03/2021	Relecture + Vérification de la qualité	

VI. Assurance et contrôle qualité

Dans cette partie, nous revenons sur les principales revues qui se sont déroulées tout au long du projet. Nous y détaillons pour chacune d'elles, leur date, leurs participants, ce qui a été évoqué et les validations éventuelles de la part du client. Cette partie permet donc d'avoir une traçabilité des réunions lors du projet.

- Revue du 01/02 :

Première véritable revue entre les clients et notre équipe de projet. L'ensemble des parties prenantes était présent que ce soit du côté des clients (Madame Wahiba BAHSOUN et Monsieur Riad MOKADEM) que du côté de l'équipe de projet (Vincent BARUDIO, Damien CHALIGNE, Théo SACCAREAU, Stévan STRICOT).

L'objectif de cette réunion était de valider le sujet. Pour rappel, nous devions nous orienter vers un sujet qui nous attirait et qui présentait un intérêt par rapport au contexte socio-économique actuel.

Pour ce faire, nous nous étions retrouvés pour que chacun puisse proposer aux autres membres du groupe différents sujets. Après avoir collecté une dizaine de sujets, nous avons procédé à un vote permettant de retenir trois sujets appréciés par tous les membres du groupe. Enfin, en amont de la revue, nous avons commencé pour ces trois sujets à regarder si le volume de données était suffisant.

Lors de la revue, nous avons donc présenté ces trois sujets aux clients ainsi que le nombre d'articles faisant référence à chacun d'eux. Suite à un échange avec les clients, nous nous sommes mis d'accord sur le sujet qui paraissait le plus intéressant, autant pour eux que pour nous. C'est donc sur les véhicules hydrogènes que nous avons décidé de travailler.

Suite à cette réunion, nous devions formaliser plus précisément le sujet (notamment la problématique) et construire l'équation de recherche.

- Revue du 08/02

Lors de la deuxième revue qui s'est déroulée une semaine plus tard, le 8 Février, seul Monsieur Riad MOKADEM était présent pour le client, ainsi que l'ensemble de l'équipe de projet.

Pour commencer, nous avons présenté plus en détail notre sujet en révélant notre problématique que nous chercherons à répondre via notre système d'aide à la décision. L'objectif du projet est de comprendre comment évolue la conception et les technologies des véhicules à hydrogène à travers le monde dans un contexte où l'écologie devient de plus en plus importante. Nous chercherons notamment à savoir quelles technologies, quels types de moteurs, quels types de véhicules sont aujourd'hui les plus développés (et donc les plus évoqués dans les articles scientifiques).

Par ailleurs, nous avons également présenté notre méthode qui a permis de définir l'équation de recherche. C'est cette équation qui sera utilisée pour récupérer que des articles pertinents vis à vis de notre sujet lors du scraping.

L'ensemble de nos présentations ont été validées par le client. En prévision de la prochaine revue prévue la semaine suivante, nous devions nous séparer en deux groupes. L'un pour travailler sur l'élaboration du schéma conceptuel de données (MCD), aussi appelé schéma Entité-Association. Le deuxième groupe devait lui commencer à réfléchir au scraping.

- Revue du 17/02 :

Cette troisième revue, en présence de Monsieur Riad MOKADEM, avait pour but de faire le point sur le schéma conceptuel de données.

Nous avions donc présenté une première version de notre schéma. Cette première proposition n'a pas totalement convaincu le client nous faisant notamment remarquer qu'avec ce schéma nous ne mettions pas assez en avant notre sujet dans le choix des entités et des attributs.

Le schéma n'étant pas validé, nous avions conclu qu'il faudrait qu'on échange par mail durant la semaine de vacances pour apporter les modifications attendues d'ici la rentrée.

- Revue du 03/03 :

Pour cette revue, Monsieur Riad MOKADEM représentait le client. L'ensemble de l'équipe de projet était présent.

L'objectif était de valider définitivement le schéma Entité-Association. Bien aidé par les échanges par mail en amont de la réunion, la validation de notre deuxième version du schéma n'a pas posé de problème. Les remarques indiquées par le client lors de la revue du 17 février ont été entendues et les modifications qu'elles impliquées ont été apportées.

Nous avions également profité de cette réunion pour présenter nos avancées sur le scraping, la création de la base de données et le rapport.

Suite à cette revue, nous devions poursuivre ces différentes tâches en parallèle.

- Revue du 10/03 et 15/03 :

Ces deux revues se sont déroulées en présence de Madame Wahiba BAHSOUN et l'ensemble de l'équipe de projet. Elles portaient sur l'aspect modélisation SADT.

Durant ces deux revues, nous avions présenté au client nos schémas de la méthode SADT. Nous avions également pris le temps d'apporter les quelques modifications suggérées par Madame BAHSOUN.

A l'issue de la deuxième revue, tous nos schémas de modélisation étaient réalisés et validés par le client.

- Revue du 17/03 :

La dernière revue avec le client avant la livraison du projet s'est déroulée en présence de tous les membres du projet, que ce soit au niveau du client avec la participation de Madame Wahiba BAHSOUN et de Monsieur Riad MOKADEM qu'au niveau de l'équipe de projet.

L'objectif était de faire un ultime bilan sur l'avancée du projet. Nous leur avions expliqué que le scraping et le peuplement de la base de données étaient finis et que l'écriture de requêtes pour l'interrogation des données arrivait elle aussi à son terme.

Ainsi, durant la dernière semaine du projet, notre objectif était de valoriser les données en réalisant le tableau de bord à l'aide de plusieurs types de diagrammes. En parallèle, nous devions finaliser le rapport.

VII. Bilan du projet

Ce projet nous a permis d'apprendre à travailler dans une situation similaire à celle que nous allons probablement rencontrer dans notre vie professionnelle, notamment ce qui concerne la relation client-fournisseur. Par ailleurs, entre le fait de travailler en groupe et le fait de devoir modifier sans cesse notre travail suite aux remarques du client ou à des problèmes rencontrés, il nous a fallu une bonne capacité d'adaptation et de communication.

A titre d'exemple, l'une de nos principales difficultés a été la mise au point du schéma entité-association. Effectivement, notre première version présentée au client ne répondait pas suffisamment à ses attentes, elle n'a donc pas été validée et le client nous a demandé de retravailler dessus. Nous avons donc redoublé de travail pour parvenir à satisfaire ce qu'il attendait. Cette expérience a été enrichissante puisque c'est une situation à laquelle nous serons confrontés dans quelques mois lors de notre entrée dans le monde du travail.

En ce qui concerne les problèmes techniques rencontrés, le premier obstacle a été le choix de la source pour le scraping. En effet, nous avions initialement pensé à scaper nos données via le site « ScienceDirect », cependant suite à des problèmes d'accès au site mais aussi un choix limité dans le nombre de connecteurs logiques à utiliser pour l'équation de recherche, nous nous sommes retrouvés dans une impasse. Il a alors fallu trouver très vite (pour ne pas prendre du retard sur notre planning prévisionnel) une solution de replis. Cette solution est venue du site « Core » qui n'avaient pas les inconvénients de « ScienceDirect ».

Le point le plus dur du projet a sûrement été la communication. Effectivement, dû à la Covid-19, les cours, et par conséquent notre projet, se sont passés essentiellement déroulés par visioconférence. Ainsi, pour pallier au fait de ne pas pouvoir se réunir de manière physique, afin de garder une bonne communication au sein du groupe et de faire circuler les informations de la bonne manière, nous avons mis en place plusieurs règles. Premièrement, nous partagions tous nos documents et informations via un Google Drive commun afin que chacun puisse avoir accès et avoir connaissance de ce qu'il se passe à tout moment du projet. Ensuite, nous nous sommes réunis sur l'application « Discord », afin d'avoir un endroit où nous pouvions communiquer par écrit mais aussi par oral. La répartition en sous-groupe était également possible. Enfin, nous avons mis en place, au début de chaque séance, une réunion à propos de l'avancement du projet afin de savoir ce qui était déjà fait et ce que nous devions faire durant la séance. De plus, on organisait à la fin de chaque séance une nouvelle réunion afin de faire le point sur l'avancement effectué pendant la session. Cela permettait que chacun d'entre nous soit au courant de l'avancement de chaque partie du projet, mais aussi de préparer ce que l'on aura potentiellement besoin pour la prochaine séance. Toutes ces règles en place nous ont permis, selon nous, d'avoir une communication fluide et un travail efficace malgré cette forte contrainte du travail à distance.

De plus, nous avons dû réussir à nous mettre d'accord sur des plages horaires, en dehors des cours prévus afin de pouvoir avancer le projet tous ensemble.

Une autre difficulté que nous avons rencontrée était liée à notre sujet. En effet, nous sommes partis sur le thème des voitures hydrogènes, bien que celui-ci nous plaisait bien, aucun d'entre nous n'avait de connaissances réelles sur ce domaine. Nous avons donc dû faire des recherches sur le sujet afin de nous renseigner sur les technologies utilisées et les nouveautés dans le domaine.

En ce qui concerne les réussites et les points positifs du sujet, on peut commencer par le fait que le codage de la base de données et des requêtes d'interrogation associées aux graphiques n'a posé aucun problème étant donné l'attirance de l'ensemble du groupe pour les bases de données et le langage SQL.

Plus généralement, ce projet a été l'occasion de revoir, de consolider et d'approfondir des connaissances et des notions qui nous ont été transmises aux cours de notre Licence 3 ou de notre Master 1. Cela concerne aussi bien les cours de base de données, que les cours de génie logiciel et de gestion de projet ou encore que les cours d'extraction d'informations. Cet ensemble de notions nous sera forcément d'une aide précieuse dans la suite de notre formation et surtout dans l'optique de notre vie professionnelle.

Par ailleurs, le scraping de données sur un site Web était quelque chose de nouveau pour nous. Nous avons donc dû faire des recherches afin de scraper correctement les informations souhaitées. Cela nous a permis d'approfondir nos connaissances dans le domaine de l'extraction d'informations mais surtout de découvrir de nouvelles méthodes de codage.

Enfin, d'un point de vue professionnel, ce projet nous a permis de renforcer notre manière de travailler en groupe mais également d'améliorer notre capacité d'adaptation. En effet, faire le projet en distanciel dans sa totalité, nous à imposer une stricte rigueur afin de rester sur notre ligne directrice. De plus, au vu du contexte sanitaire, il nous a fallu nous adapter, afin de pouvoir travailler efficacement et ce, tous ensemble, en dépit du distanciel. Pour finir, nous avons, selon nous, plutôt bien réussi à nous répartir sur les différentes tâches afin d'avancer vite mais surtout de manière efficace. Nous avons réussi à toujours rester à l'écoute des autres et des nouvelles idées apportées, au gré de quelques compromis, afin de toujours rester en groupe et de tous travailler dans la même direction.

A travers ce projet nous avons été confrontés à plusieurs obstacles proches de ceux que nous rencontrerons sûrement dans notre futur métier. Ces obstacles ont été dépassés et nous en sommes ressortis plus compétent et grandi dans notre manière de travailler**.**

Nous considérons que l'objectif fixé par le client a été atteint, nous avons réussi à effectuer des analyses statistiques basés sur des informations issues d'articles portant sur les véhicules à hydrogène. Cependant, un point pourrait être améliorer. Effectivement, même si ce n'était pas une exigence du client, présenter les résultats de nos graphiques sous forme d'un véritable tableau de bord via une page Web pourrait être un aspect intéressant que par marque de temps nous n'avons pu développer. C'est la principale piste d'amélioration de notre système d'aide à la décision.

Pour conclure, réussir à mener, en groupe, un projet de bout en bout dans le but de résoudre un problème bien précis en faisant intervenir plusieurs domaines est bien évidemment une situation que l'on rencontrera souvent à travers notre futur métier et nous sommes fiers d'avoir accompli ce travail.

Annexes

- **Annexe 1 – Cahier des charges**

Université Paul Sabatier
Formation SID

Année Universitaire 2020 - 2021

Projet 'Tableau de Bord'
SID Master 1
Cahier des charges
V 5

HISTORIQUE DES MODIFICATIONS

Référence	SID Master 1 /Tableau de bord			
Version	**Objet de la modification**	**Date**	**Auteur(s)**	**Relecteur**
V 0	Création du document	6/12/2016	W.Bahsoun, R Mokadem	-
V 1	Mise à jour du cahier des charges	11/01/2017	W.Bahsoun, R Mokadem	
V2	Mise à jour du cahier des charges	17/01/2018	W.Bahsoun, R Mokadem	-
V3	Mise à jour du cahier des charges	16/01/2019	W.Bahsoun, R Mokadem	-
V4	Mise à jour du cahier des charges	15/01/2020	W.Bahsoun, R Mokadem	-
V5	Mise à jour du cahier des charges	22/01/2021	W.Bahsoun, R Mokadem	-

SOMMAIRE

1. INTRODUCTION 3
1.1 Avant Propos 3
1.2 Objectifs 3
2. DOCUMENTS APPLICABLES ET DE REFERENCE 5
2.1 Documents applicables 5
2.2 Documents de référence 5
2.3 Terminologie
3. EXIGENCES SUR LE PROCESSUS DE DEVELOPPEMENT 6
4.1 Organisation 6
4.2 Livrables 6
4.3 Planification 6
4. EXIGENCES SUR LE PRODUIT LOGICIEL 7
4.1 Plan de développement 7
5. EXEMPLES DE PROJETS 7

1. Introduction

1.1 Avant Propos

Le présent document constitue le Cahier des Charges de l'application à développer dans le cadre du projet 'Tableau de bord' en SID Master 1ère année. Il contribue à la création d'un cadre formel de type Client-Fournisseur où les enseignants jouent le rôle du client et les étudiants, celui du fournisseur.

Ce document précise l'organisation de l'équipe de développement et le fonctionnement tout au long du projet.

Le projet proposé dans le cadre de ce bureau d'études couvre les principaux thèmes suivants :

- Les entrepôts de données (data waherouse) [EDD],
- Les bases de données [BDD],
- L'Analyse statistique,
- La visualisation de données
- Le Génie logiciel [GL],
- La Gestion de projet [GPRO],
- La Gestion de Configuration
- …et la programmation (Java, Perl, Apache, Python…).

1.2 Objectifs

L'objectif de ce projet est de développer un système d'aide à la décision élaboré par des analyses détaillées sur des points d'intérêts à partir de la visualisation de données textuelles issues des BDD en ligne.

Ce système va recouvrir plusieurs étapes du processus décisionnel telles que :

1. Recherche et Collecte de données,
2. Préparation de données,
3. Valorisation de données.
4. Diffusion des connaissances.

Nous regroupons ces étapes en trois grandes parties à savoir :

- Recherche et préparation de données.
- Valorisation des données : Data Warehouse,
- Visualisation.des données

Ces activités seront supportées par les méthodes de génie logiciel notamment :

- Processus de développement,
- Gestion de projet,
- Ainsi que les règles d'assurance de qualité.

Partie Recherche d'informations et Préparation de données

Recherche et Collecte des données. Cette étape permet de rechercher l'information et d'identifier celle qui est utile dans les sources sélectionnées pour alimenter l'analyse. Nous définirons un domaine d'analyse pour chaque groupe.

La collecte de données se fera sur un périmetre de sujet bien défini à l'avance entre le fournisseur et le client. Cela est raffiné suivant une équation de recherche (pré-traitement de données). Enfin, le client validera par la suite la pertinence des données collectées (extraction de données pertinentes).

Partie Valorisation de données

Stockage de données,

Dans cette partie, un filtrage est appliqué sur la base de données afin d'obtenir un sous ensemble qui répond aux besoins des clients. Les sous ensembles constituent une vue restreinte de la base. Ces sous ensembles sont utilisés comme point d'entrée pour chaque groupe afin que les étudiants puissent appliquer leurs requêtes.

A cause de la crise sanitaire et du contexte d'enseignement en distanciel, le système de gestion de bases de données utlisé sera Oracle. Les étudiants pourront implémenter leur base de données à distance via SQL Developer. Oracle constitue un SGBD très répandu et très puiissant dans le milieu socio-économique.

Interrogation de données

Cette partie consiste à générer des structures de données pour chaque groupe. Cela est possible en appliquant des requêtes sur les sous ensembles de la base de données produite lors de l'étape précédente. En fonction des données manipulées par chaque groupe, une méthodologie sera proposée par l'enseignant pour exporter les données afin de peupler l'entrepôt de données. Par la suite, les requêtes doivent être écrites. Ces requêtes concernent les données présentes dans les différentes tables obtenues. Suivant le type des requêtes, les sorties (résultats) sont classifiées en quatre (4) types (dimensions):

- Une variable.
- Un vecteur.
- Une matrice.
- Un cube.

Partie Visualisation de Données

Les fonctions relatives à la visualisation de données sont essentielles pour réussir la présentation d'un travail de veille et pour convaincre les décideurs à travers une présentation lisible, pertinente et concise.

La visualisation de données consiste à représenter graphiquement (icônes, carte, arbres…) une information (données, processus, relations, concepts) souvent abstraite et/ou à très grande volumétrie.

Pour chaque groupe, nous définissons différents types de sorties adaptés à chaque type de requête tels que les histogrammes d'évolution 2D et 3D, les cartes géographiques, les nuages de mots,…etc.

2. Documents applicables et de références

2.1 Documents applicables

[CONT] Contrat de developpement.

2.2 Exemples de documents de référence

[GL] Cours « Génie Logiciel»
Wahiba Bahsoun

[DAWA] Cours « Data Waherouse»
Gilles Hubert

[CONC] Cours « Concepts Fond. BD »
Franck Morvan

[GPRO1] Cours et TD « Gestion de Projet »
Wahiba Bahsoun

2.3 Terminologie

Tous les termes utilisées dans le contexte de ce projet doivent etres expliquées par l'équipe de projet.

3. Exigences sur le processus de développement

3.1 Organisation

[ORG-00] La réalisation intégrale du projet est à la charge du Fournisseur constitué d'un groupe de 4 étudiants de Master M1 de la formation SID. La promotion actuelle est d'un effectif de 36 étudiants. Cela conduit alors à la formation de 9 groupes.

[ORG-01] Le Fournisseur doit assigner, pour toute la durée du projet :
- un chef de projet,
- un responsable de gestion de configuration,
- un responsable Assurance et Contrôle Qualité.

Ces fonctions doivent être assurées par des ressources distinctes.

[ORG-03] L'organisation de l'équipe de développement mise en place par le Fournisseur sera présentée au Client.

[ORG-04] Chaque réunion d'avancement se fera en présence d'au moins un représentant du Client, du chef de projet, du responsable de gestion de configuration et du responsable Assurance et Contrôle Qualité.

[ORG-05] Le Fournisseur doit être en mesure de présenter au Client l'état d'avancement du projet pendant toute sa durée.

3.2 Livrables

[LIV-00] Le Fournisseur doit livrer au Client :
- Un rapport du projet,
- Le cahier de Recettes,
- le code source et les données associées.

Les exigences relatives à ces éléments sont décrites au chapitre suivant.

[LIV-01] Les livrables mentionnés en [LIV-00] seront fournis par voie électronique (wahiba.bahsoun@irit.fr et riad.mokadem@irit.fr).

[LIV-02] Les livrables mentionnés en [LIV-00] seront remis au Client lors de la soutenance du projet (date prévisionelle : lundi 22 mars 2021).

3.3 Planification (Application Scrum)

[PLAN-00] Le projet sera réalisé dans le cadre de séances de cours et TP d'une durée de 2 heures ou de 4 heures. Les dates et heures précises sont communiquées dans l'emploi du temps.

[PLAN-02] Les réunions d'avancement seront hebdomadaires.

[[PLAN-04] La soutenance aura lieu vers la fin du mois de mars 2021 (date prévisionnelle : mercredi 24 mars 2021), ceci correspond à la présentation des résultats aux clients.

4. Exigences sur le produit logiciel

4.1 Plan de développement

[PD-Q-01] [GPRO] est applicable au plan de développement.

[PD-Q-02] Le rapport du Projet doit présenter les dispositifs d'assurance Qualité mise en place par le fournisseur, et plus précisément :

- En Introduction, la description du sujet d'analyse
- Les méthodes, techniques et outils utilisés,
- Les règles de développement appliquées sur toutes les phases de développement. en conception, en codage,
- Les dispositifs de suivi du projet
- Et une interprétation rigoureuse du sujet d'analyse.

[PD-Q-03] Le plan de développement doit présenter les dispositifs de gestion de configuration mise place pendant toute la durée du projet.

5. Exemples de Projets sur les sujets suivants :

Exemples de sujets traités 2015-2016:

Bitcoin, COP21, Harcélement sexuel, Euro 2016, Les addictions

Exemples de sujets traités en 2016-2017:

Internet of things (maisons connectées) , La thérapie par réalité virtuelle, Impression 3D des organes humains, Inégalité homme / femme, Impact des écrans sur les enfants, Données abérantes.

Exemples de sujets traités en 2017-2018:

Blockchains, L'impact des attentats, Les catastrophes naturelles, Les cyber attaques, L'évolution du e-sport, e-commerce, SmartGrid, Légalisation du cannabis, Ted Talks.

Exemples de sujets traités en 2018-2019:

Pollution due au plastique, Véhicules autonomes, Catastrophes naturelles, Brexit, La déforestation, Le flux migratoire, La coupe du monde au Qatar (Corruption dans le sport), Gilets jaunes.

Exemples de sujets traités en 2019-2020:

Dopage dans le sport, feux de foret, fuite des cerveaux, impact des écrans, impact des OGM sur la santé, impact du plastique sur l'environement, impact de l'organisation des jeux olympique, PMA, thérapie sonore, le tourisme spatial

- **Annexe 2 – Charte qualité**

- **Partie SQL :**

1- Nommage entités, attributs, contraintes :

- Nom d'une entité / table : court, significatif et première lettre en majuscules.
- Nom d'un attribut : court, significatif, en minuscules, underscore pour séparer les mots.
- Nom d'un attribut qui est un identifiant : id_attribut.
- Nom des contraintes de clé primaire : pk_Table (exemple : pk_Auteur)
- Nom des contraintes de clé étrangère : fk_attributTable (exemple : fk_auteurAuteur_Evoq)

2- Règles de codage :

- Mots clés et instructions en majuscules.
- Retour à la ligne à chaque nouvelle instruction.
- Déclaration des attributs : retour à la ligne et indentation pour chaque attribut.
- Sélection : 3 sélections d'attributs par ligne maximum, si plus : retour à la ligne et indentation.
- Pour les requêtes, les tables sont renommées avec un alias (exemple : Article A)
- Pour les requêtes, les attributs sont appelés de manière *alias_table.attribut* (exemple : a.id_article)

3- Autre:

- Les associations qui deviennent des tables sont renommées de manière suivante Table_Evoq (exemple : Auteur_Evoq)

- **Partie Python**

- **Import :** Une ligne par import. L'utilisation de 'from' est tolérée si on importe quelques fonctions spécifiques.
- **Fonctions / Variables** : En anglais et en minsicule avec éventuellement des underscore. Ex : my_function(), my_variable
- **Arguments de fonction / méthode** : minsicule et underscore si besoin (snake case). Ex: unit_price
- **Indentation** : 4 espaces pour bien séparer les blocs et avoir un code lisible.
- **Longueur de ligne** : Pas plus de 79 caractères par ligne
- **Ligne vides** : utile pour séparer les différentes parties du code. Il est recommandé de :
  - Laisser 2 lignes vides avant la définition d'une classe.
  - Laisser 1 seule ligne vide avant la définition d'une méthode dans une classe
  - On peut aussi laisser une ligne vide dans le corps d'une fonction pour séparer les sections logiques de la fonction, mais cela est à utiliser avec parcimonie.
- **Commentaires** : Commencent par le symbole #. Phrase complète en anglais de préférence. Eviter les commentaires sur la même ligne du code. Toutes les fonctions doivent avoir une docstring (triples doubles guillemets) renseignant son but, ses paramètres et sa sortie.

Figure 33 - Charte qualité

- **Annexe 3 – Dictionnaire de données**

Table cible	Attribut cible	Fichier d'extraction (sur Python après manipulation)	Colonne d'extraction	Type
Article	id_article	df_infos	Id	Number
	titre	df_infos	Titre	Varchar(500)
	langue	df_infos	Langue	Varchar(20)
	date_article	df_infos	Date	Date
MotCle	id_motcle	/	auto-incrémenté	Number
	mot	df_content	MotCle	Varchar(200)
Vehicule	id_vehicule	/	auto-incrémenté	Number
	type_vehicule	df_scraping	Vehicule	Varchar(20)
Technologie	id_tech	/	auto-incrémenté	Number
	nom_tech	df_scraping	Technologie	Varchar(50)
	categorie	df_scraping	Technologie	Varchar(20)
	source	df_scraping	Technologie	Varchar(20)
Domaine	id_domaine	/	auto-incrémenté	Number
	nom_domaine	df_scraping	Domaine	Varchar(25)
Marque	id_marque	/	auto-incrémenté	Number
	nom_marque	df_scraping	Marque	Varchar(50)
Pays	id_pays	/	auto-incrémenté	Number
	nom_pays	df_scraping	Pays	Varchar(50)
Auteur	id_auteur	/	auto-incrémenté	Number
	nom_prenom	df_infos	Auteurs	Varchar(350)
Moteur	id_moteur	/	auto-incrémenté	Number
	type_moteur	df_scraping	Moteur	Varchar(20)
MotCle_Evoq = composer	id_article	/	clé étrangère	Number
	id_motcle	/	clé étrangère	Number
Vehicule_Evoq = mentionner	id_article	/	clé étrangère	Number
	id_vehicule	/	clé étrangère	Number
Technologie_Evoq = Specifier	id_article	/	clé étrangère	Number
	id_tech	/	clé étrangère	Number
Domaine_Evoq = Concerner	id_article	/	clé étrangère	Number
	id_domaine	/	clé étrangère	Number
Marque_Evoq = Enoncer	id_article	/	clé étrangère	Number
	id_marque	/	clé étrangère	Number
Pays_Evoq = Evoquer	id_article	/	clé étrangère	Number
	id_pays	/	clé étrangère	Number
Auteur_Evoq = Rediger	id_article	/	clé étrangère	Number
	id_auteur	/	clé étrangère	Number
Moteur_Evoq = Citer	id_article	/	clé étrangère	Number
	id_moteur	/	clé étrangère	Number

Figure 34 - Dictionnaire de données

- **Annexe 4 – Détail diagramme de Gantt**

Tache	Date début	Date fin	Ressources
Choix du sujet	25/01/21	01/02/21	Tout le monde
Formalisation du sujet	01/02/21	08/02/21	Tout le monde
Choix de l'équation de recherche	01/02/21	08/02/21	Tout le monde
Choix des sources pour le scrapping	08/02/21	15/02/21	Tout le monde
Scrapping et nettoyage des données	08/02/21	10/03/21	Saccareau Théo
Elaboration du schéma entité / association	15/02/21	03/03/21	Tout le monde
Codage de la création de la base de données	28/02/21	10/03/21	Stricot Stévan
SADT	03/03/21	17/03/21	Barudio Vincent / Chaligné Damien
Insertion des données scrappées et nettoyées	10/03/21	15/03/21	Saccareau Théo
Création du dashbord (dont élaboration des requêtes)	10/03/21	20/03/21	Tout le monde
Ecriture du rapport	15/02/21	22/03/21	Tout le monde