RAPPORT D'ALTERNANCE

Pour l'obtention du Master *Traitement automatique des langues*

Préparé au sein du Département de Sciences du langage
UFR Philosophie, Information-Communication, Langage, Littérature, Arts du Spectacle (PHILLIA), Université Paris Nanterre

L'analyse sémantique au service de l'industrie ferroviaire : Alternance chez SNCF Réseau du 29 octobre 2020 au 13 septembre 2021

Présenté et soutenu par
Martin Anthony SALUD

Rapport soutenu publiquement le 25 juin 2021 Sous la direction de		
TARAVELLA Iris	Professeure	Université Paris Nanterre
BENTCHIKOU KAMEL	Responsable Solutions 4.0	SNCF Réseau
GOULARD Énora	Adjointe Solutions 4.0	SNCF Réseau

Note liminaire

Les normes d'écriture inclusive seront appliquées tout au long de ce rapport. Ces normes sont inspirées de la thèse de Ranchon (2016) et du Manuel d'écriture inclusive (Haddad, 2016) :

1) Toute mention d'une personne qui peut être une femme ou un homme sera écrite avec un point médian « · ». Cela sera le cas pour le singulier et le pluriel.

   Par exemple : l'agent·e, les client·e·s, les voyageur·euse·s, les acteur·rice·s

2) Ce sera également le cas pour les adjectifs successifs.

   Par exemple : l'employé·e expérimenté·e

3) Dans le cas de deux objets pluriels de genres grammaticaux différents, nous appliquerons la règle de proximité.

   Par exemple : les enregistrements et les images incluses dans le projet

Remerciements

À mes professeur·e·s à l'Université de Nanterre, à l'Université Sorbonne Nouvelle et à INALCO, pour l'introduction rigoureuse mais très enrichissante au traitement automatique des langues.

À mon équipe au sein de SNCF Réseau, pour leur bienveillance, leur aide et leur envie de nous voir nous épanouir en équipe. Merci de m'avoir donné l'occasion de travailler dans le monde de la mobilité, ce qui est un de mes rêves depuis mon enfance.

À Guillaume, pour le soutien, la motivation et le réconfort tout au long de mon parcours universitaire.

À ma famille et à mes ami·e·s, pour leur soutien malgré les milliers de kilomètres et les multiples confinements nous séparant.

Para sa isáng lipunang malayà at mapagpalayà.

Pour une société libre et libératrice.

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteur·e·s/autrices figurent en citations et les auteur·e·s/autrices sont mentionné·e·s.
5. Les écrits sur lesquels je m'appuie dans ce rapport sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : SALUD

PRENOM : Martin Anthony

DATE : 18 juin 2021

SIGNATURE : 

Table des matières

Note liminaire 2
Remerciements 3
Table des matières 5
Introduction 6
Chapitre 1 – L'industrie ferroviaire et SNCF Réseau 7
1.1 Le groupe SNCF et Réseau 7
1.2 Smart Studio 11
Chapitre 2 – L'amélioration des enquêtes de satisfaction 13
2.1 Contexte et objectifs du projet 13
2.2 Premiers essais 13
2.3 Solution adaptée 15
2.4 Retour d'expériences et perspectives possibles 18
Chapitre 3 – Gestion des incidents caténaires 20
3.1 Contexte et objectifs du projet 20
3.2 Exploration du corpus 21
3.3 Solutions envisagées 23
3.3.1 Outil de prédiction statistique 23
3.3.2 Outil d'aide à la décision 24
3.4 Lot 1 : Définition des catégories avec le métier et Machine learning 25
3.5 Interface graphique 27
3.6 Limites et perspectives 29
Conclusion 31
Bibliographie 32
Tableau d'illustrations 34
Annexes 35
Annexe 1 : Exemple d'une Fiche d'incident (FI) 35
Annexe 2 : Exemple d'un Rapport d'analyse causale (RAC) 37
Annexe 3 : Arborescence de concepts concernant les incidents caténaires 40

Introduction

Dans le cadre de ma deuxième année de Master en Traitement automatique des langues, j'ai effectué une alternance en tant qu'analyste sémantique pour Smart Studio du 29 octobre 2020 au 13 septembre 2021. La première moitié de mon alternance s'est déroulée principalement en télétravail puisque mon contrat a débuté la veille du deuxième confinement. Nous sommes désormais en mode hybride avec 1-2 jours par semaine en présentiel au siège de SNCF Réseau au Campus Rimbaud, Saint-Denis.

L'objectif de ce rapport est double. Premièrement, j'essaierai de relater comment l'analyse sémantique a été employée pour répondre aux besoins de SNCF Réseau. Enfin, je détaillerai mon évolution personnelle et professionnelle au sein d'une grande structure.

Je présenterai dans un premier temps le secteur ferroviaire et comment Smart Studio y trouve sa place. Ensuite, je présenterai les deux projets auxquels j'ai pris part. Je terminerai par une partie sur mes acquis au sein de l'entreprise.

Chapitre 1 – L'industrie ferroviaire et SNCF Réseau

Mon alternance s'est déroulée du 29 octobre 2020 au 13 septembre 2021 au sein du Smart Studio chez SNCF Réseau. Cette équipe émergente s'appuie sur l'analyse de données afin d'accompagner la transformation digitale des métiers. Afin de comprendre le contexte de l'équipe, je détaillerai dans ce chapitre le placement du Smart Studio au sein de la SNCF et ses enjeux.

1.1 Le groupe SNCF et Réseau

La Société Société nationale des chemins de fer français a été créée à la suite de la nationalisation de plusieurs entreprises ferroviaires privées en 1938 (Décret-loi du 31 août 1937 portant réorganisation du régime des chemins de fer, 1937). Depuis sa naissance, la SNCF joue un rôle structurant dans le développement de la mobilité humaine et l'aménagement des territoires. La SNCF transporte 15 millions de voyageur·euse·s, non seulement en France, mais aussi à l'étranger dans de nombreux moyens de déplacement : train, tram, bus, etc. Elle ne s'intéresse pas uniquement à la mobilité des voyageur·euse·s, mais aussi au fret. Malgré son statut français, un tiers de ses revenus provient de l'étranger et ses 275 000 employé·e·s sont réparti·e·s dans 120 pays (Groupe SNCF, 2020).

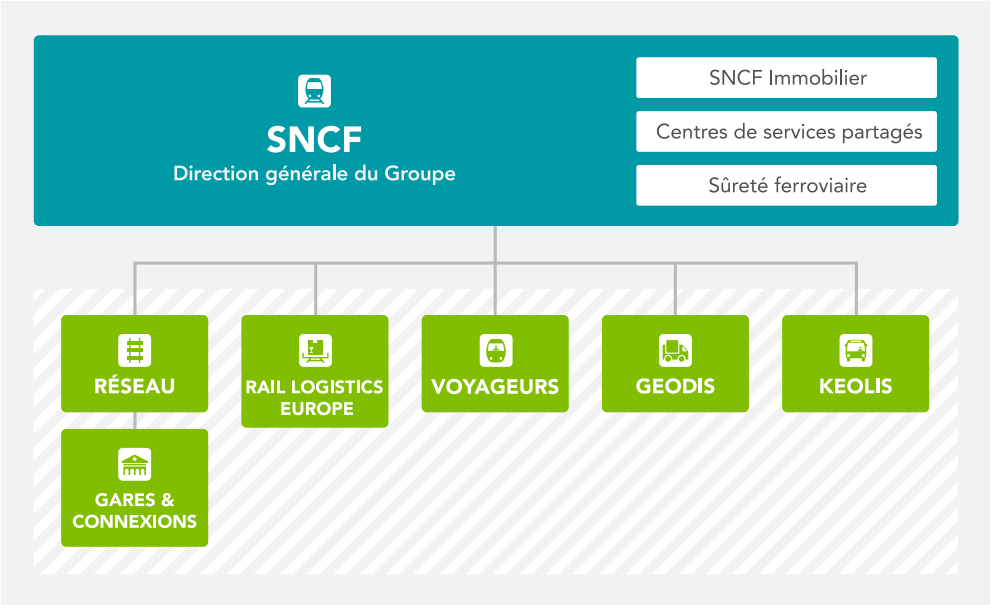
La nature internationale de la SNCF a poussé la refonte de son organisation en 2020 pour donner lieu à un groupe plus performant face aux défis de l'ouverture à la concurrence et de la mondialisation (LOI n° 2018-515 du 27 juin 2018 pour un nouveau pacte ferroviaire (1), 2018). Dans la Figure 1 ci-dessous, les différents composants du groupe sont détaillés.

Tout en haut, la société mère SNCF définit la stratégie globale. Elle détient également le contrôle sur tout ce qui concerne 1) le patrimoine commun du groupe (sauf les gares et infrastructures) avec SNCF Immobilier, 2) les services communs entre les filiales notamment sur les questions de rémunération et d'insertion sociale avec les Centres de services partagés et 3) la sécurité et la protection de l'ensemble des acteur·rice·s avec la Sûreté ferroviaire (Suge) (SNCF, s. d.).

D'un côté, Rail Logistics Europe et Geodis proposent le transport des marchandises avec ses nombreuses filiales ferroviaires. De l'autre côté, SNCF Voyageurs assume le transport des personnes en mode quotidien ou longue distance. Cela comprend notamment les trains à grande vitesse (TGV), les Transilien en Île-de-France et les Trains express régionaux (TER). Quant à Keolis, cette filiale de la SNCF se spécialise en transport du

quotidien tels que les déplacements par tramway, bus et vélos, pour n'en citer que quelques-uns (SNCF, s. d.).

Figure 1. Schéma simplifié : organisation de la SNCF depuis le 1er janvier 2020[1]



Le dernier des cinq groupes est Réseau. Si les quatre composants précédemment mentionnés sont orientés vers le transport de bout à bout de voyageurs ou de marchandises, Réseau s'intéresse aux quatre métiers suivants :

- « l'accès à l'infrastructure ferroviaire du réseau ferré national, comprenant la répartition des capacités et la tarification de cette infrastructure
- la gestion opérationnelle des circulations
- la maintenance, l'entretien et le renouvellement de l'infrastructure
- le développement, l'aménagement, la cohérence et la mise en valeur du réseau » (SNCF, s. d.)

Réseau assure donc que les rails sur lesquels les trains roulent sont en développement, amélioration et renouvellement constants. La Figure 2 ci-dessous donne un aperçu de l'étendue des chantiers gérés par le groupe en 2021. Enfin, Réseau gère sa filiale Gares et Connexions qui a pour vocation le développement et la gestion d'au moins 3 000 gares sur l'ensemble du territoire français (SNCF, s. d.).

[1] https://www.sncf.com/fr/groupe/profil-et-chiffres-cles/portrait-entreprise/qui-sommes-nous

Figure 2. Carte nationale des principaux chantiers en 2021 (SNCF Réseau, 2021)

Étant donné cette large étendue des domaines sur lesquels intervient Réseau, plusieurs directions sont nécessaires pour gérer ses compétences et ses métiers. La Figure 3 présente les différentes directions au sein du Réseau. Smart Studio, dans lequel mon alternance a été effectuée, fait partie de la Direction Générale Numérique.

Figure 3. SNCF Réseau et ses directions générales

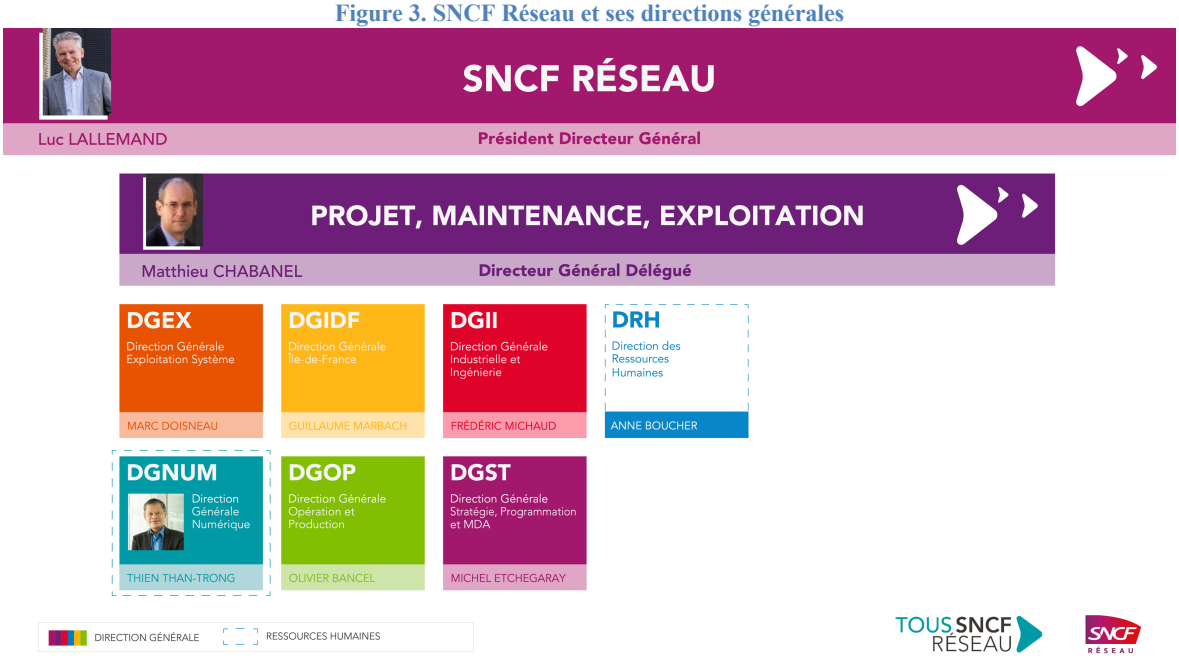
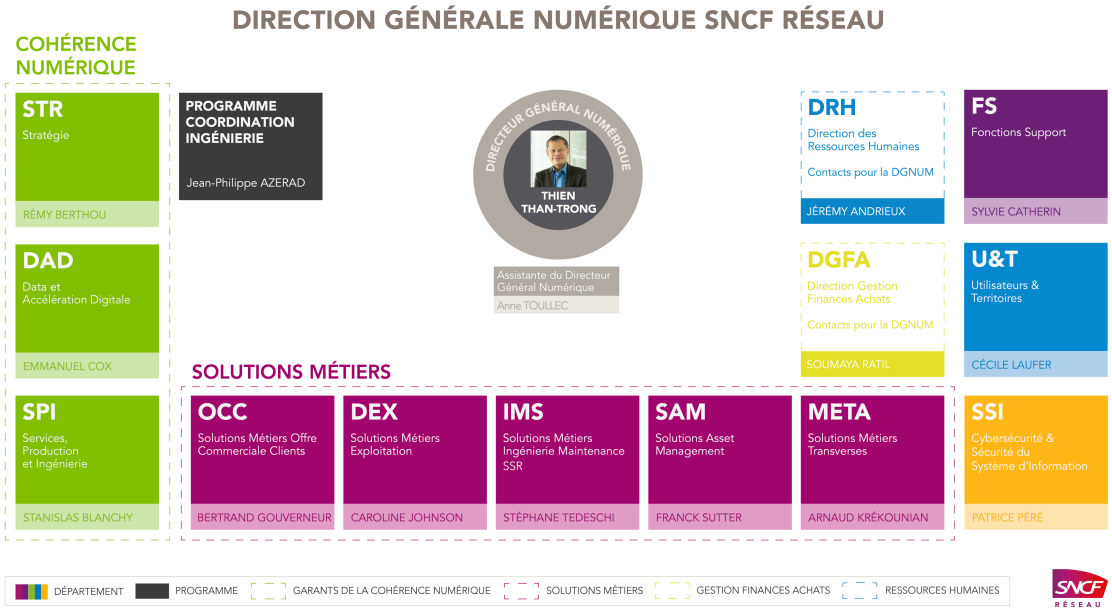


Figure 4. Les départements composant la DGNUM



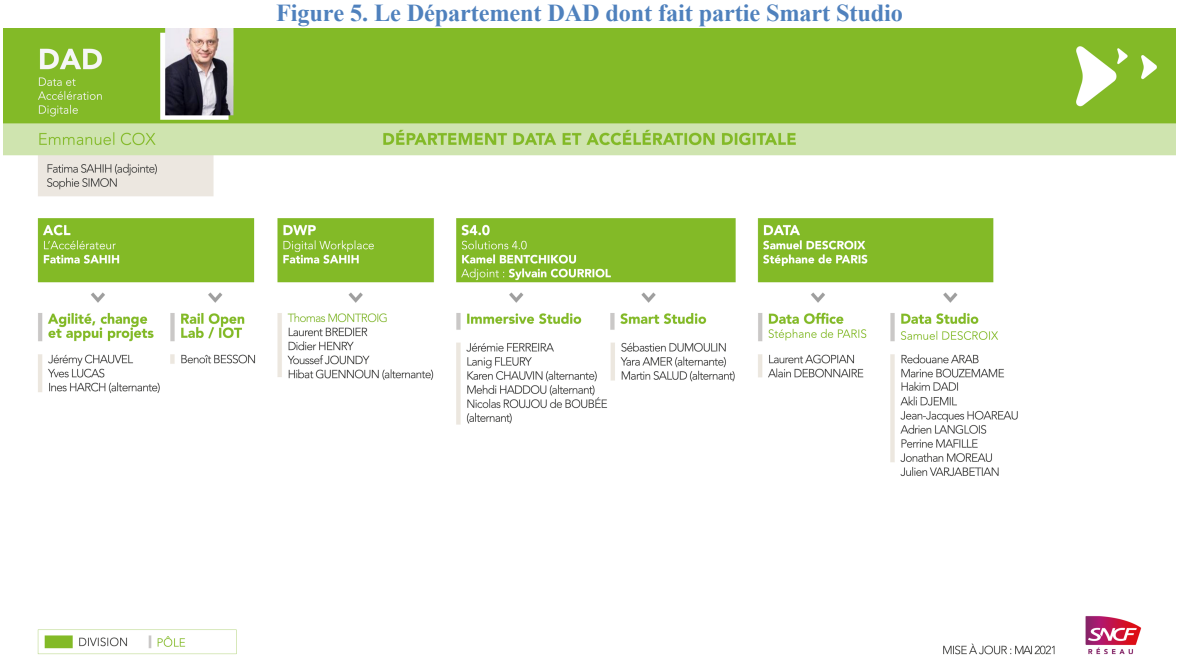
La DGNUM est composée de trois sections, onze départements et un programme de coordination (Figure 4). Elle assure la transformation digito-numérique de Réseau et fournit un « haut niveau de service aux métiers et aux clients et utilisateurs » (SNCF Réseau, 2021). Ses quatre missions sont les suivantes :

- « simplifier la gouvernance du numérique,
- installer une relation équilibrée entre numérique et métier,
- rassembler les forces qui [concourent] au pilotage du numérique,
- améliorer [son] delivery et [son] pilotage économique » (SNCF Réseau, 2021)

En son sein se trouve le Département Data et Accélération Digitale (Figure 5) chapeauté par Emmanuel COX. Ce département abrite quatre divisions, chacune responsable de plusieurs domaines. Son but est d'accompagner le métier dans la co-construction des solutions digitales, de les développer, et d'accélérer leur déploiement.

Les deux premières divisions, l'Accélérateur et le Digital Workplace, s'occupent entre autres de l'industrialisation et de la conduite du changement au sein de l'équipe. Ensuite, la division Solutions 4.0 assure le développement et le déploiement de nouvelles technologies pour des usages variés et concrets pour Réseau et éventuellement à destination externe. Enfin, la Division Data met en place des outils innovants pour mieux appréhender les besoins de Réseau en termes de données.

Figure 5. Le Département DAD dont fait partie Smart Studio



Les projets proposés par la division Solutions 4.0 s'appuient sur plusieurs technologies émergentes telles que l'intelligence artificielle, la blockchain et la réalité augmentée. L'Immersive Studio produit notamment des formations en réalité virtuelle pour mieux simuler les situations réelles de travail pour les agent·e·s de maintenance, les gestionnaires de crises, etc.

1.2 Smart Studio

Sous la division Solutions 4.0 se trouve également Smart Studio. Notre équipe travaille en lien étroit avec les autres services au sein de la DGNUM et est composée de :

- Kamel BENTCHIKOU : responsable de Solutions 4.0
- Énora GOULARD : adjointe de Solutions 4.0
- Sébastien DUMOULIN : responsable du Smart Studio
- Yara AMER et Martin SALUD : alternant·e·s analystes sémantiques

Victor LEUTARD qui est chef de projet Digital & Innovation de la division Accélérateur vient également en appui à l'élaboration de nos projets depuis le début. Yara et moi travaillions en binôme sur toutes les missions qui nous ont été confiées. Ceci nous a permis de diviser la charge de travail et de mettre en commun nos compétences et nos acquis.

Le studio a plusieurs hubs dont Smart Maintenance et Smart Language. Le premier s'appuie sur les data sciences et la blockchain avec pour objectif la maintenance prédictive,

l’anticipation et la limitation des perturbations, la gestion du patrimoine et la détection des anomalies.

Quant au Smart Language, nous utilisons le traitement automatique des langues pour réaliser l’analyse des enquêtes de satisfaction et améliorer la gestion des incidents caténaires et pour explorer d’autres projets en sciences cognitives et en traduction automatique.

Chapitre 2 – L'amélioration des enquêtes de satisfaction

Le premier projet auquel j'ai été assigné concerne l'utilisation de l'analyse sémantique dans le but d'améliorer l'efficacité des enquêtes de satisfaction.

2.1 Contexte et objectifs du projet

Une enquête de satisfaction est lancée au sein de Réseau tous les trois ou quatre mois pour recueillir les ressentis des collaborateur·rice·s et identifier des pistes d'amélioration pour les services proposés. Étant donné le nombre élevé des personnes enquêtées (3 000 à 6 000), ce processus est automatisé. Les réponses aux questions fermées et aux questions sur échelle (de type « très satisfaisant » versus « pas du tout satisfaisant ») sont facilement traitées par le biais des logiciels déjà existants dans la plupart des suites bureautiques.

En revanche, il est plus complexe de tirer des conclusions à partir des réponses aux questions ouvertes et les commentaires laissés par les employé·e·s. Ces verbatim étaient, au moment de la conception du projet, analysés de façon manuelle et individuelle, voire pas du tout exploités. À défaut d'une équipe dédiée à cette fin, une analyse automatisée a été conçue pour rendre le processus plus fluide et efficace.

Le projet a donc trois objectifs :

1. Exploiter pleinement les verbatim dans les enquêtes ;
2. Identifier les pistes d'amélioration proposés dans les verbatim ;
3. Faire remonter ces axes d'amélioration aux services concernés.

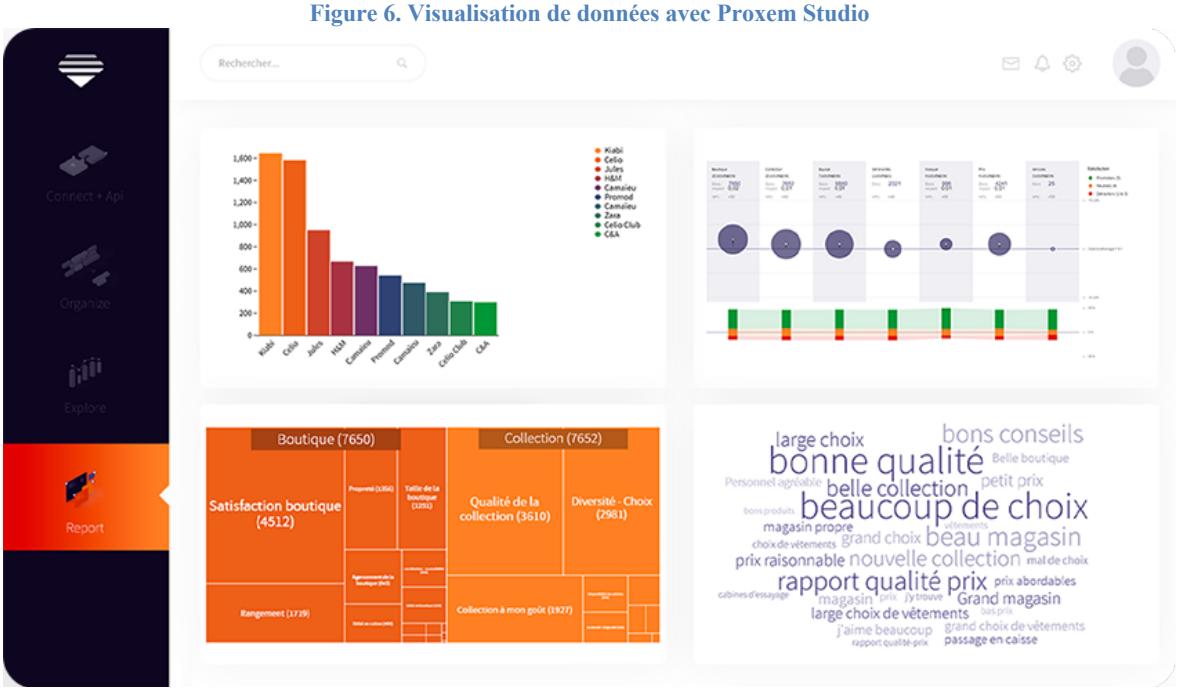
2.2 Premiers essais

Le premier corpus de réponses aux enquêtes provient de l'année 2019. Cette base a été constituée par une entreprise tierce et comptait environ 6 000 réponses. La codification des réponses, qu'elles soient fermées ou en texte libre, se faisait sur un seul fichier Excel. Chaque colonne correspondait à une question. Au total, il y avait 145 000 cellules à traiter, ce qui constituait une tâche conséquente au niveau de l'exploration.

Proxem Studio est l'outil choisi pour mener à bien le projet. Il s'agit un outil d'analyse sémantique en mode SaaS (*software as a service* ou logiciel en tant que service) à partir de son site web (Dassault Systèmes, 2021). Cela permet à l'utilisateur·rice d'appréhender l'outil peu importe son système d'exploitation. Le fait que ce soit sur le web

permet également de procéder de manière collaborative. À part l'analyse TAL/NLP qui s'effectue en arrière-plan, les données peuvent être explorées directement sur le site avec des fonctionnalités de visualisation de données incluses (Figure 6).

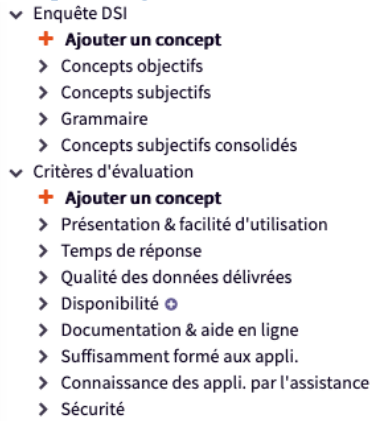
Figure 6. Visualisation de données avec Proxem Studio



Proxem Studio est conçu pour traiter des fichiers Excel et les lit par colonne. Il est possible de définir chaque colonne en fonction de sa nature : identifiants, métadonnées, texte libre, etc. Cela simplifie la tâche d'analyse puisque les recherches sémantiques peuvent être filtrées sur une colonne donnée. Pour ce projet en particulier, cette fonctionnalité est très pertinente puisqu'une colonne correspond à une question.

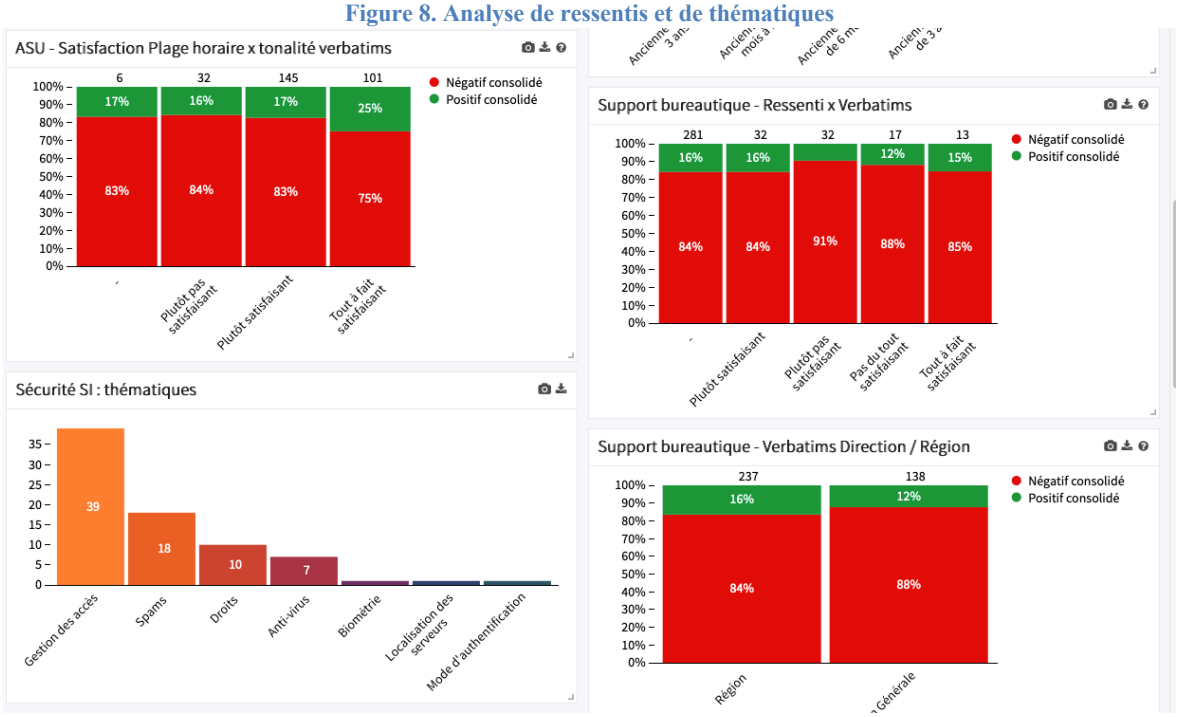
Pour les premiers essais, l'équipe a balayé le questionnaire pour identifier les questions ouvertes et les réponses possibles données par les personnes répondantes. Cette étape a donc permis de définir d'une part un premier thésaurus de mots-clés pertinents et d'autre part une base de mots répartis par son ressenti positif ou négatif (Figure 7). À la suite de cela, des catégories ont été établies pour bien trier les verbatim en fonction de leur thématique et ressenti.

Figure 7. Quelques catégories établies sur Proxem Studio



Les graphiques produits à la fin de ces essais démontrent bien les ressentis des agent·e·s et permettent de déterminer les thématiques de leur réponses (Figure 8). Cependant, il n'était pas possible de croiser ces deux données et de creuser plus. Cela veut dire que les ressentis par thématique globale étaient identifiés, mais les soucis plus fins et les axes d'améliorations liés ne l'étaient pas. Les conclusions sont donc faites de manière globale sans proposer des actions à prendre et des étapes par lesquelles commencer.

Figure 8. Analyse de ressentis et de thématiques



2.3 Solution adaptée

L'entreprise tierce qui gère les enquêtes a été changée au lancement de la nouvelle enquête pour l'année 2021. Les questions, la manière dont elles sont posées, la présentation et sa codification sur Excel ont donc été modifiées par la suite. Les questions ont été simplifiées et organisées dans des sections plus claires et unifiées. Un changement

notamment très adapté à notre analyse était la division en deux du champ des commentaires qui se trouve dans chaque section. Au lieu d'avoir toutes les réponses, qu'elles soient positives ou négatives, dans une seule colonne sur Excel, les questions ouvertes ont été divisées en deux, afin de répartir les réponses positives et négatives dans deux colonnes différentes (Figure 9).

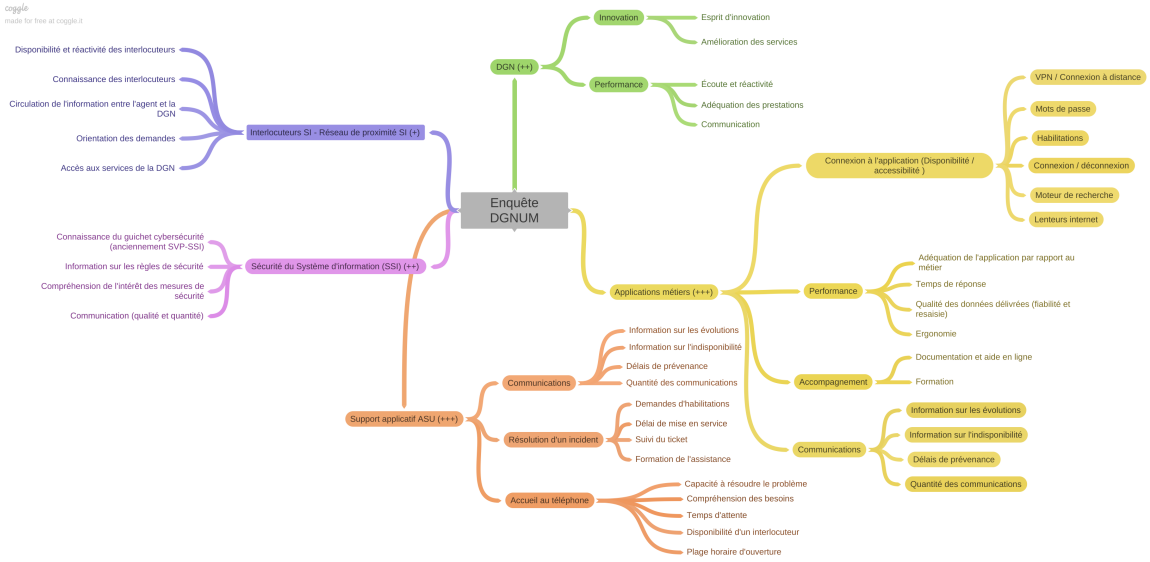
Figure 9. Séparation des champs textuels correspondant aux réponses positives et négatives

PROG : B6/B7 sur le même écran
SI B1 <> 99 (utilise aucune applications)
B6. Qu'est-ce que vous appréciez le plus dans cette application ?
Dites-le avec vos mots
NR autorisé

SI B1 <> 99 (utilise aucune applications)
B7. Qu'est-ce que vous appréciez moins dans cette application ?
Dites-le avec vos mots
NR autorisé

Partant sur un nouveau corpus provenant des enquêtes de cette année avec 3 337 réponses, nous avons décidé de procéder tout d'abord par une définition du besoin de nos collaboratrices de l'Activité « Expérience et connaissance de l'utilisateur » au sein du Département Utilisateurs & Territoires (voir Figure 4 pour l'organigramme). En co-construisant une ontologie de connaissances avec les clientes en amont (Figure 10), le processus a été raccourci et les mots-clés pertinents pour chaque thématique ont été plus facilement repérés lors de la phase d'exploration.

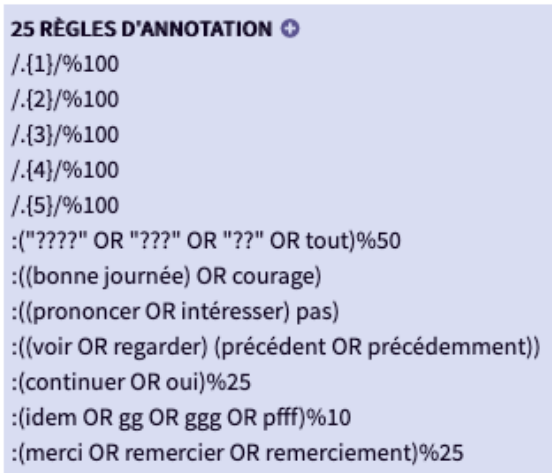
Figure 10. Arborescence de thématiques et de concepts définis avec les client·e·s



La phase d'exploration effectuée a donné lieu à une catégorie supplémentaire non définie dans l'arborescence. Comme déjà mentionné en haut, chaque question avait sa propre colonne en fonction de son ressenti négatif ou positif. En revanche, quelques commentaires négatifs se trouvaient tout de même dans la colonne pour le texte positif et beaucoup de commentaires ne portaient pas d'intérêt dans l'analyse sémantique globale. C'est pour cela qu'une catégorie « Verbatims à écarter » a été créée pour enlever tous les textes comportant la mention « merci », « ras », « à voir », et ainsi de suite (Figure 11).

Au niveau d'évaluation, nous avions établi un seuil de 40% de reconnaissance de verbatim avec les clientes pour les premières expérimentations avec le nouveau corpus. Sur 3 976 verbatim au total, seulement 2 942 étaient exploitables en supprimant les verbatim à écarter. 1 572 réponses ont pu être triées, ce qui nous donne un taux de fiabilité à 53,43%, soit 13 points de mieux que le seuil décidé avec le métier.

Figure 11. Exemples de verbatim à écarter



Les graphiques créés sont aussi plus pertinents par rapport aux premiers essais. Grâce à l'interactivité de l'outil Proxem, nous pouvons non seulement voir les chiffres par thématique, mais aussi explorer plus en détail ces réponses individuelles. Dans la Figure 12, nous avons un aperçu en haut à gauche des verbatim trouvés dans le champ destiné aux commentaires positifs sur les applications métiers. Il est possible de filtrer ces réponses en cliquant sur les trois autres graphiques.

Figure 12. Graphiques générés après l'analyse

2.4 Retour d'expériences et perspectives possibles

Tout d'abord, le taux de reconnaissance à 53,43% reste à améliorer. L'amélioration de l'efficacité du système repose sur l'exploration de données et sur l'extraction de mots-clés. À mesure que les thématiques seront définies plus précisément, ce score évoluera. En effet, le temps passé à filtrer une à une les réponses sera conséquent faute de précisions sur l'arborescence.

Le thésaurus établi pendant ce projet reste assez restreint. Malgré quelques données déjà existantes sur la base de Proxem, le champ lexical ferroviaire est très spécialisé et requiert l'entraînement du système sur plus de volumétrie. En effet, à l'arrivée de plus de réponses au fil du temps, plus de mots-clés pourront être recueillis.

Nous pouvons aussi aller plus loin en nous focalisant également sur les bases lexicales déjà recueillies en amont par la SNCF. Ces bases pourront nourrir le système avec non seulement les mots-clés, mais aussi des expressions multi-mots, acronymes et abréviations dont nous n'avions pas connaissance.

L'utilisation de Proxem pour un premier lancement dans l'analyse sémantique en entreprise s'est montrée formatrice. Plusieurs étapes de prétraitement du corpus ont pu être écartées, notamment la lemmatisation, le nettoyage des fautes d'orthographes, et ainsi de suite. En effet, ces fonctionnalités existent déjà dans l'outil Proxem et les filtrages lexicaux peuvent se faire à base d'expressions régulières, de règles sémantiques, de recherches lexicales lemmatisées et même en intégrant la distance du Levenshtein pour retrouver les

éventuelles fautes d'orthographes. Ce gain de temps permet finalement de se consacrer à des tâches plus orientées vers l'analyse sémantique pure.

Chapitre 3 – Gestion des incidents caténaires

Le deuxième projet auquel j'ai pris part et qui est toujours en cours concerne le traitement des rapports d'incidents caténaires. Bien que très formateur, le premier projet détaillé dans le chapitre précédent a une portée limitée puisqu'il s'agit d'une analyse sur des documents internes et très ponctuels. Ce deuxième projet a une plus grande ampleur car il est le fruit de la collaboration entre plusieurs services au sein de Réseau, et son objectif est d'être utilisé par de nombreux·ses agent·e·s sur le terrain.

3.1 Contexte et objectifs du projet

Comme évoqué auparavant, SNCF Réseau gère le développement et la maintenance du réseau ferroviaire en France. Les aléas sont nombreux et les défauts d'infrastructure liés à plusieurs facteurs peuvent empêcher le bon roulement des trains. Un des composants les plus importants, et qui peut subir des dommages pour maintes raisons est le système pantographe-caténaire. Ce système de captage électrique assure en effet l'alimentation des trains en électricité et permet son fonctionnement. Aujourd'hui, une enquête est lancée à la suite de chaque incident caténaire.

Ces enquêtes suivent un parcours long et chronophage impliquant beaucoup d'acteur·rice·s aux niveaux local, régional et national. Chaque rapport est traité de manière manuelle et individuelle, rallongeant ainsi le processus. Plusieurs incidents peuvent être classés sans suite au bout d'un certain temps. Un incident classé sans suite signifie que les causes et les responsables n'ont pas pu être identifié·e·s, ce qui oblige Réseau à dédommager l'État ou l'entité en question. De fait, Réseau y perd non seulement du temps, mais aussi des ressources financières.

L'analyse des éléments d'incident nécessite des savoir-faire techniques et particuliers qui ne sont pas forcément transposables à partir des référentiels métiers. Les agent·e·s impliqué·e·s dans les enquêtes développent leurs compétences au fil d'une longue expérience au sein du Réseau. Il est donc difficile de former des agent·e·s débutant·e·s en un temps raccourci et il vaut mieux par la suite pérenniser ces savoir-faire et les rendre plus accessibles.

Quatre objectifs ont été définis pour ce projet :

1) Obtenir des statistiques concernant les incidents caténaires en trouvant des régularités dans un corpus de rapports d'incident

2) Déterminer et ensuite limiter les causes racines en se fondant sur les statistiques et sur des prédictions
3) Limiter les perturbations sur le réseau pour minimiser les nuisances aux transporteur·euse·s et voyageur·euse·s
4) Réduire le temps et les coûts d'interventions en maitrisant plus rapidement chaque incident

3.2 Exploration du corpus

Notre corpus est composé de documents datant de 2015 à 2019 et provenant de la base Incidents de Sécurité Conservés et Historisés Informatiquement pour Analyse (ISCHIA) de la SNCF. Ces documents correspondent à deux types de rapport que les responsables rédigent quand une enquête est lancée.

Le premier type de document est une Fiche d'incident (FI) qui fournit les premiers constats concernant un incident. Elle relate les faits techniques, la localisation de chaque incident, le modèle du train et les informations immédiates disponibles telles que le contexte raconté par un·e conducteur·rice ou les mesures qu'il·elle a prises. 1 125 documents PDF de ce type ont été récoltés.

Quant au second, il s'agit d'un Rapport d'analyse causale (RAC). Lors des premiers essais, 959 rapports avaient été rassemblés, mais cette base s'agrandit au fur et à mesure que de nouveaux en sont rédigés. Contrairement aux FI, les RAC représentent l'avis final des enquêteur·euse·s. Ils font foi légalement et sont utilisés pour l'arbitrage. Par conséquent, ils sont plus complets, à savoir que les causes, les risques et les conséquences de chaque incident y sont bien décrites. Les documents sont principalement en format PDF sauf quelques-uns qui sont en fichier Word.

Étant donné la structuration différente des deux types de document, deux scripts Python ont été créés pour aspirer les données. Les seules différences entre ces deux scripts sont le nombre de catégories à rechercher ainsi que la manière dont ces recherches est effectuée. Comme le montrent les Figures 13 et 14, la façon dont les données se présentent sont différentes (cf. Annexes 1 et 2 pour un exemple d'une FI et d'un RAC). Les expressions régulières utilisées pour les deux types de documents sont donc différentes. À cela s'ajoute la complexité d'emploi des fichiers PDF. Pour aspirer le texte de chaque PDF, nous avons utilisé la bibliothèque Tika (Mattmann, 2020). Grâce à pandas (Reback et al., 2021), nous

avons ensuite exporté les données aspirées en fichier CSV avec le contenu textuel divisé en plusieurs colonnes. Chaque colonne correspond à une information importante : nature, lieu, ligne, type d'évènement, etc.

Figure 13. L'en-tête d'une FI

Evènement

Date : 05/09/08 17:10

Nature : Rupture de caténaire

Lieu : Voie M2 à Paris Montparnasse

Transmis par : Adjoint PSSR PR

Figure 14. L'en-tête d'un RAC

RAPPORT D'ANALYSE CAUSALE n °14982

Identifiant de l'événement : 11 février 2014 / 18:12 / 420000 / Viroflay-Rive-Gauche (78) / Le conducteur du train 165452 (EF SNCF) constate un pendule caténaire décroché lors de son arrêt sur la voie 2 bis en gare de Viroflay (repris au CRJ).

.

1. DESCRIPTIF

Lieu : Viroflay-Rive-Gauche (78) Km : 13+900
Voie : 2 bis (VP) Régime d'exploitation : Double voie Block Automatique Lumineux
Electrification : 1500 volts Continu
Train n°: 165452 Parcours de Rambouillet (17h22) à Paris Montparnasse (18h22)
Activité : Voyageur EF : SNCF
Engin moteur : 7608 avec Rame VB2N C 06

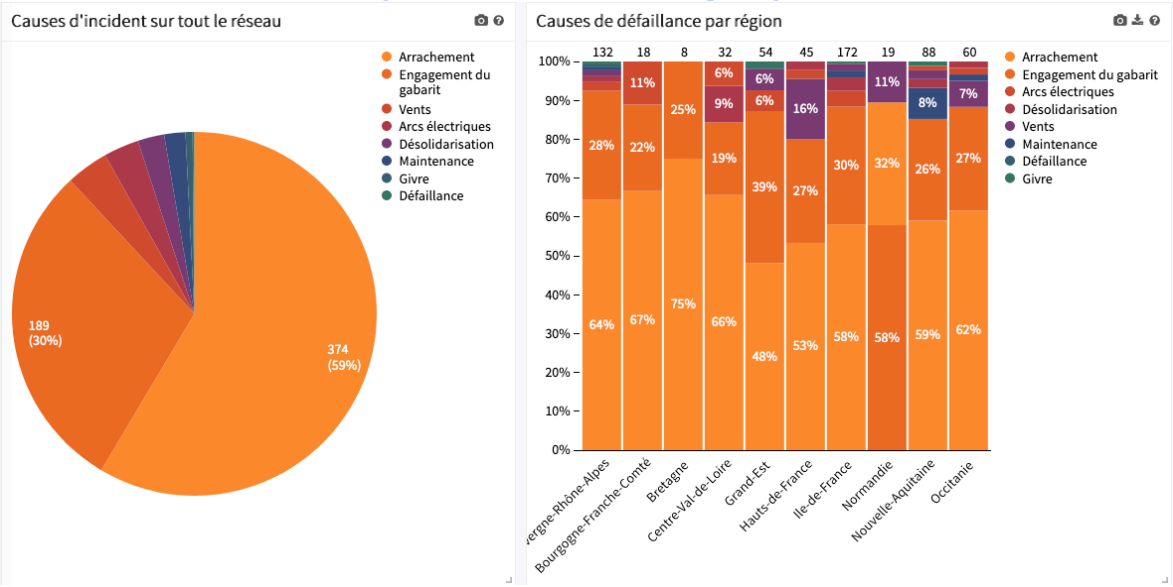
La Figure 15 donne un aperçu de la structure des données aspirées depuis les RAC. Pour donner suite à la demande du métier, nous avons écarté tous les fichiers FI de notre analyse. Comme nous l'avons déjà établi, les FI ne présentent que les premières constatations et ne font pas foi en cas d'arbitrage. Malgré cela, il nous reste tout de même près de mille documents à analyser datant de 2014 jusqu'à aujourd'hui.

Figure 15. Structure du premier fichier CSV

	Fichier	Lieu	Région Administrative	Circonstances	Conséquences
0	Cause 178/RAC/RAC 20170421 45 PSE Ischia 23181 Gien-V.pd	Gien (45)	Centre-Val-de-Loire	Contexte : Le train FRET 75049 de l'EF SI	Gravité des dommages humains : Personnes transportées : Sans c
1	Cause 178/RAC/RAC 20170619 75 PRG Ischia 23575 Javel-V.p	Javel (75)	Ile-de-France	Contexte : Les installations de sécurité de	Gravité des dommages humains : Personnes transportées : Néant.
2	Cause 178/RAC/Rapport d'enqu ├-κte d ├-ñfinitif 20111109 PRC	Entre Béziers et Vias (34)	Occitanie	Le 09112011 vers 23h05 le train de voyag	Gravité des dommages humains : Personnes transportées : Aucune
3	Cause 178/RAC/RAC 20160607 34 LR Ischia 20984 entre B ├-ñ	Ecole Valentin (25)	Bourgogne-Franche-Comté	Contexte : Pendant le mouvement social	Gravité des dommages humains : Personnes transportées : néant
4	Cause 178/RAC/RAC 20140331 PDL Ischia 15315 Tierc ├-ñ (49)	Arlac (33)	Nouvelle-Aquitaine	A 19h28, le conducteur du TER 857340 E	Gravité des dommages humains :Personnes transportées : aucune,
5	Cause 178/RAC/RAC 20170328 25 BFC Ischia 23009 Ecole Va	Nîmes (30)	Occitanie	Contexte : Le 28 mars 2017, une demand	Gravité des dommages humains : Personnes transportées : Sans c
6	Cause 178/RAC/RAC 20171213 33 APC Ischia 24836 ARLAC-V	Juvisy (91)	Ile-de-France	Contexte : Le train n°7214223 de IEF SN	Gravité des dommages humains : Personnes transportées :Néant A
7	Cause 178/RAC/RAC 20150813 30 LR Ischia 18786 Nimes-V.pc	Bailleul-Sir-Berthoult (62)	Hauts-de-France	Contexte : Le conducteur du train SNCF M	Gravité des dommages humains : Personnes transportées : Néant.
8	Cause 178/RAC/RAC 20170121 91 PRG Ischia 22577 Juvisy-V.p	Libourne (33)	Nouvelle-Aquitaine	Contexte : Des travaux sont en cours en g	Gravité des dommages humains : Personnes transportées : Néant.
9	Cause 178/RAC/RAC 20150902 62 NPC 18922 Bailleul Sir Bertl	Entre Magenta et Hausmann-S	Ile-de-France	Contexte : Section de ligne de double voi	Gravité des dommages humains : Personnes transportées : Néant
10	Cause 178/RAC/RAC 20140719 APC Ischia 16019 Km 29,530 e	Entre Gravigny-Balizy et Chilly	Ile-de-France	Contexte : Le train autotrain 24599 de IEF	Gravité des dommages humains : Personnes transportées : Aucune

Nous avons ensuite analysé ce fichier CSV avec Proxem Studio pour explorer les concepts qui s'y trouvent. Cette étape nous a permis de créer une première ontologie de catégories que nous avons tirée des référentiels métier. Grâce à cela, nous avons pu créer des graphiques qui résument la totalité du contenu textuel de notre corpus. Dans la Figure 16, nous voyons les causes trouvées dans les rapports qui sont ensuite croisées par région.

Figure 16. Causes sur le réseau et par région



Nous avons pu tirer des conclusions concernant la volumétrie des données en fonction des types d'incident, des causes racines, de mesures d'urgences prises, et des composants et pièces impliquées. En revanche, le croisement de données sur Proxem est limité à deux : une métadonnée et un champ textuel. Cela ne nous a donc pas permis de croiser plusieurs catégories. De plus, Proxem ne permet pas de faire de prédictions. Au vu des limitations de cet outil, il fallait donc développer une solution en interne.

3.3 Solutions envisagées

Étant donné l'ampleur du projet, il a été découpé en deux lots afin de fournir un produit fonctionnel et de le tester sur le terrain le plus rapidement possible.

3.3.1 Outil de prédiction statistique

La question à laquelle tentera répondre l'outil de prédiction statistique est « Qu'est-ce qui s'est passé ? ». Pour ce faire, nous procédons en quatre étapes :

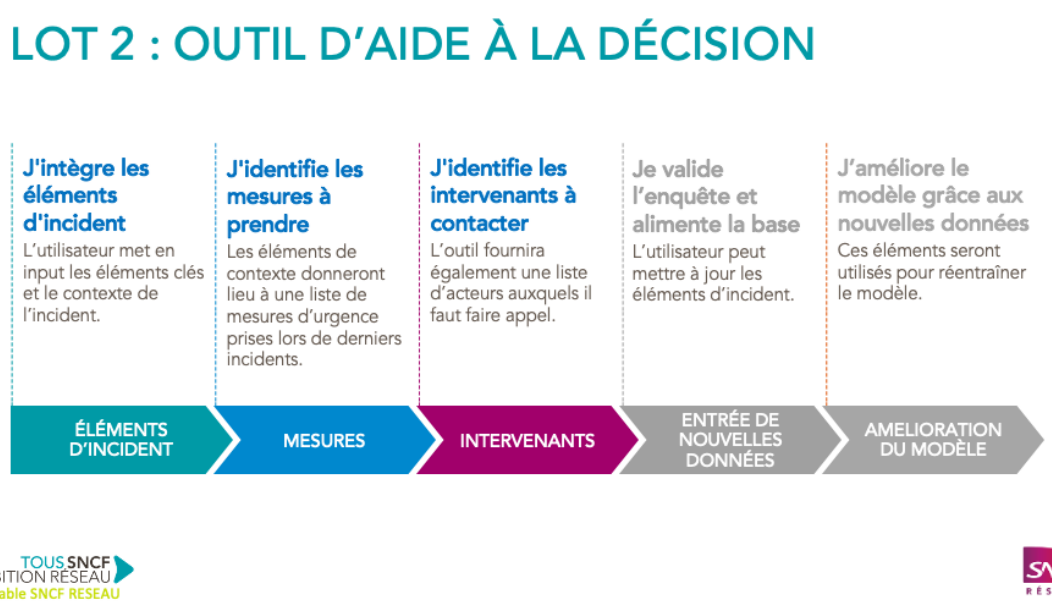
1) Automatisation de l'analyse des documents,
2) Définition des catégories à prédire selon les besoins du métier,
3) Création d'une interface graphique qui permettra à l'utilisateur·rice d'entrer des éléments d'incident,
4) Génération et affichage des statistiques liées à ces éléments grâce à une classification multilabel.

Le but de ce lot n'est pas de donner une liste d'actions à l'agent·e, mais plutôt de lui donner un aperçu global de probables causes racines étant donné des éléments de contexte. À titre d'exemple, il·elle peut rechercher quelle cause de défaillance est la plus probable pour une chute de fil de contact en Bretagne.

3.3.2 Outil d'aide à la décision

Le deuxième lot fonctionne en sens inverse. Il tentera de répondre à la question : « Que puis-je faire en sachant tous ces éléments ? ». Les étapes spécifiques de ce lot restent à définir, mais la vision globale est présentée dans la Figure 17.

Figure 17. Perspectives pour l'outil d'aide à la décision



L'utilisateur·rice a donc accès à l'interface précédemment développée où il·elle peut entrer les faits constatés d'un incident. Grâce à ces inputs, il·elle recevra une liste d'actions à prendre et d'intervenant·e·s à contacter. Prenons à titre d'exemple un·e employé·e sur le terrain qui observe un incident concernant un train spécifique. Il·elle pourra mettre en input la végétation comme cause de l'incident. L'outil lui fournira ensuite les prochaines actions à suivre : déclencher une enquête régionale, couper le courant électrique d'une zone géographique plus large, etc.

À la fin de l'enquête, l'application permettra le téléchargement des rapports validés pour alimenter la base et le modèle, et ainsi produire des résultats améliorés.

3.4 Lot 1 : Définition des catégories avec le métier et Machine learning

Nous avons transféré la totalité des catégories élaborées sur Proxem dans un script Python. Ces dernières, étant plutôt des produits de notre compréhension des référentiels métiers, ont été présentées et modifiées en travaillant avec le métier. Une arborescence détaillée (voir Annexe 3) a donc été co-construite avec les clients et cela nous a permis d'identifier les termes les plus pertinents de chaque catégorie.

Ensuite, nous avons repris le fichier CSV créé afin de lemmatiser le contenu textuel et de nettoyer les stopwords avec la bibliothèque Stanza (Manning et al., 2014). Cette étape permet de simplifier la recherche avec des expressions régulières sur Python. La Figure 18 fournit un exemple du traitement effectué pour chaque colonne. Ici, nous faisons la recherche sur la colonne 'Circonstances'. Puis, la fonction 'analyze_column()' effectue la lemmatisation et le nettoyage grâce à Stanza. Pour tous les lemmes trouvés, nous vérifions s'ils peuvent être rangés dans les catégories 'collision', 'arrachement', etc.

Figure 18. Expressions régulières pour rechercher les thématiques qui ressortent dans les rapports

```
def contexte():
  contexte = df_RAC["Circonstances"]
  analyze_column(contexte)

  x = [[] for i in range(len(all_lemmes))]
  for item in all_lemmes:
    collision = re.search(r'collision', item)
    if collision:
      x[all_lemmes.index(item)].append("collision")
    else:
      x[all_lemmes.index(item)].append("None")
    arrachement =
     re.search(r'rupture|casser|arracher|arrachemment|rompre', item)
    if arrachement:
      x[all_lemmes.index(item)].append("arrachement")
    else:
      x[all_lemmes.index(item)].append("None")
    chute = re.search(r'chute.*fil de contact', item)
    if chute:
      x[all_lemmes.index(item)].append("chute fil de contact")
    else:
      x[all_lemmes.index(item)].append("None")
    perte = re.search(r'perdre|perte', item)
    if perte:
      x[all_lemmes.index(item)].append("perte")
    else:
      x[all_lemmes.index(item)].append("None")
    retournement = re.search(r'couverture.*obstacle', item)
    if retournement:
      x[all_lemmes.index(item)].append("retournement")
    else:
      x[all_lemmes.index(item)].append("None")
    enchevêtrement = re.search(r'enchevêtrement', item)
    if  enchevêtrement:
      x[all_lemmes.index(item)].append("enchevêtrement")
    else:
      x[all_lemmes.index(item)].append("None")
  global contexte_list
  contexte_list = x
```

Une fois les recherches faites sur la totalité du corpus et des catégories choisies, nous les rangeons dans un autre fichier CSV ne contenant que des valeurs binaires (Figure 19). Ces valeurs correspondent à la présence d'une catégorie dans un rapport donné. Au moment de la rédaction de ce rapport, nos recherches sont limitées au croisement de 'Région', 'Contexte' et 'Causes'. À terme, l'objectif est que l'utilisateur·rice puisse choisir les éléments qu'il·elle trouve pertinent à croiser, de type 'Mesures d'urgence', 'Mesures conservatoires', etc.

Figure 19. Valeurs binaires indiquant la région où un incident est produit selon les rapports

Bourgogne-Franche-Comté	Bretagne	Centre-Val-de-Loire	Corse	Grand-Est	Hauts-de-France	Ile-de-France
0	0	1	0	0	0	0
0	0	0	0	0	0	1
0	0	0	0	0	0	0
1	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	1
0	0	0	0	0	1	0

Après, nous avons utilisé NumPy (Harris et al., 2020) pour manipuler les données dans les CSV et SciKit-Learn (Pedregosa et al., 2011) pour créer et implémenter notre modèle pour le machine learning. Étant donné qu'un incident peut être causé à la fois, à titre d'exemple, par un problème sur la voie et une défaillance caténaire, nous ne pouvons pas nous contenter à une seule étiquette pour chaque incident. La classification que nous avons appliquée ici est donc la classification multilabel.

Le Tableau 1 présente les résultats de nos tests sur plusieurs classifieurs. Nous avons gardé une seule mesure d'évaluation pour les premiers tests. Le Hamming Loss, adapté à la classification multilabel, détermine le taux d'erreur de prédiction sur les étiquettes. Pour cette mesure d'évaluation, un score moins élevé signifie un meilleur résultat.

Tableau 1. Résultats de classification multilabel

Classifieur	Hamming Loss
Multilabel k Nearest Neighbors	0.166
Multi-layer Perceptron	0.168
Binary Relevance k Nearest Neighbors	0.153
Label Powerset x Random Forest	0.189

Binary Relevance x Random Forest	0.161

Les classifieurs donnent des résultats assez similaires, mais *Binary Relevance kNN* présente les meilleurs. Nous avons sauvegardé le modèle de ce classifieur en local avec Joblib (Joblib Development Team, 2020). Cela nous permet de charger le modèle à partir de l'interface graphique.

3.5 Interface graphique

Une interface graphique était nécessaire pour permettre aux utilisateur·rice·s d'accéder à l'application. Nous en avons développé quatre au vu des besoins qui ont évolué entre temps. Tout d'abord, nous avons créé un simple formulaire sur Python qui se lance en même temps que la partie machine learning. Cela nous a servi à tester s'il était éventuellement possible d'intégrer des choix dans le script.

En revanche, cette solution ne permet pas à une personne sans compétence en Python de l'appréhender. Il fallait donc créer une application ou une interface qui puisse être lancée par un·e utilisateur·rice sans connaissances spécifiques. Google Colab, sur lequel nous avons effectué la plupart de notre développement, ne permettait pas la création de ce genre d'interface. Nous avons donc sauvegardé nos scripts en local et essayé la bibliothèque Tkinter (Van Rossum, 2020). Tkinter permet d'avoir un GUI simple avec des boutons faciles à développer.

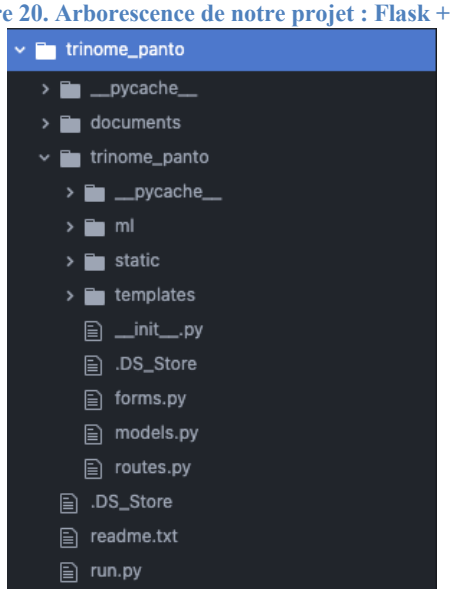
Cette solution ne s'est pas avérée suffisante puisqu'il faut s'assurer que les futur·e·s utilisateur·rice·s de l'application ou de l'interface ne soient pas trop contraint·e·s techniquement. En effet, l'utilisation de Tkinter semble assez compliquée et demande une installation. Nous avons par la suite décidé de tout simplifier est de rester sur une page Web qui est accessible à tous·tes depuis n'importe quel navigateur.

Pour ce faire, il fallait que la page web soit hébergée sur les serveurs de la SNCF. Notre premier essai était une page en HTML-PHP, mais son intégration avec Python était trop complexe. Nous avons donc cherché une autre solution qui permette à la fois d'exécuter des scripts Python à travers un serveur et d'afficher une page HTML. La solution qui a été trouvée est Flask (Grinberg, 2018) avec Bootstrap 5.0 comme framework d'affichage.

L'architecture du projet se trouve dans la Figure 20. L'exécution de l'interface se fait par le biais de 'run.py' tandis que la navigation et l'affichage des résultats se trouvent dans

‘routes.py’. Tous les scripts Python concernant le machine learning se trouvent dans le dossier ‘ml’ et les pages HTML sont stockées dans ‘static’.

Figure 20. Arborescence de notre projet : Flask + ML



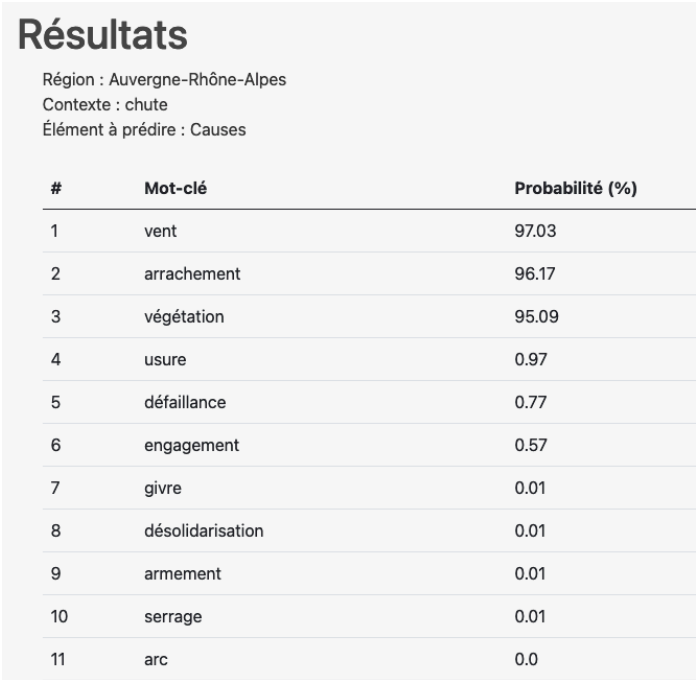
L’utilisateur·rice a le choix entre une région (toutes les régions administratives sauf la Corse y sont incluses), un contexte et un élément à prédire (Figure 21). Le choix du contexte se réfère à l’élément de l’incident connu. Ces choix sont limités à ‘Collision’, ‘Arrachement caténaire’, ‘Chute de fil du contact’, ‘Perte d’un archet ou d’un pantographe’, ‘Enchevêtrement’ et ‘Retournement pantographe’. Quant à l’élément à prédire, nous l’avons limité à ‘Causes’ pour cette première étape de test.

Figure 21. Page d'accueil de l'interface graphique

Ensuite, la page de résultats s’affiche pour donner la probabilité de chaque label étant donné les choix cochés. Dans la Figure 22, l’agent·e veut connaître les possibles causes racines pour une chute de fil de contact en Auvergne-Rhône-Alpes. D’après le modèle

entraîné sur les RAC analysés, nous pouvons dire que ‘vent’, ‘arrachement’ et ‘végétation’ sont les causes les plus probables.

Figure 22. Exemple d’une page de résultats



Résultats

Région : Auvergne-Rhône-Alpes
Contexte : chute
Élément à prédire : Causes

#	Mot-clé	Probabilité (%)
1	vent	97.03
2	arrachement	96.17
3	végétation	95.09
4	usure	0.97
5	défaillance	0.77
6	engagement	0.57
7	givre	0.01
8	désolidarisation	0.01
9	armement	0.01
10	serrage	0.01
11	arc	0.0

3.6 Limites et perspectives

Si nous reprenons les quatre objectifs définis en amont pour le Lot 1 du projet, nous pouvons dire que le Lot 1 est réussi. Nous avons pu 1) automatiser l’analyse des documents à travers une aspiration des rapports PDF, leur lemmatisation et leur nettoyage, 2) définir les catégories avec le métier, 3) créer une interface graphique fonctionnelle qui prend en compte les choix d’utilisateur·rice et 4) générer et afficher les statistiques depuis notre modèle. Toutefois, il nous reste plusieurs axes d’amélioration à traiter :

1) Il existe des rapports dont une grande partie des données n’est pas captée. C’est notamment le cas pour le captage des régions pour une partie conséquente des rapports. Il faut soit revoir les expressions régulières utilisées, soit écarter les documents problématiques et refaire un script traitant ces cas spécifiques.

2) Certains documents sont en format Word et ne sont donc pas exploités pour le moment.

3) En ce qui concerne l’ontologie développée avec le métier, grâce à une extraction des mots-clés, nous avons pu repérer les mots les plus fréquents qui ressortent de

chaque section du rapport. Nous sommes toujours en train de classer ces mots en fonction des catégories avec le métier.

4) Ces mots et leurs classifications ne sont stockés que dans nos expressions régulières dans le script. Il faudrait créer un thésaurus de ces mots-clés pour pérenniser ces informations.

5) L'interface graphique ne permet que le choix d'une région et d'un élément de contexte. À terme, l'objectif est de permettre de cocher plus d'éléments significatifs tels que le type d'engin, le composant impacté et ainsi de suite.

Quant aux perspectives de ce projet, le Lot 2 sera lancé prochainement et aura des enjeux plus larges. Ensuite, le projet en général pourra être étendu à d'autres utilisations par de multiples utlisateur·rice·s. À titre d'exemple, les statistiques générées peuvent aider les enquêteur·euse·s dans leur travail d'analyse, mais aussi les responsables au niveau national pour le reporting ou la définition de plans d'actions. De plus, les technologies employées lors du projet pourraient très bien être appliquées à d'autres types d'incidents, que ce soit en interne dans le groupe SNCF ou pour d'autres sociétés.

Conclusion

Ce rapport d'alternance a présenté les deux projets auxquels j'ai pris part. J'ai pu mettre en lumière comment notre équipe a mis en place l'analyse sémantique au profit de SNCF Réseau. Les deux projets portaient sur des enjeux et des thématiques différentes. Notre projet sur l'amélioration des enquêtes de satisfaction permettra à nos client·e·s d'exploiter un grand volume de données qui ne sont pas exploitées depuis un long moment. Le modèle de machine learning que nous avons développé pour l'incidentologie caténaire sera possiblement démultiplié à un large panel d'utilisateur·rice·s. Globalement, des améliorations restent à effectuer sur l'ensemble de nos travaux, mais les premiers jalons sont déjà posés.

Hormis l'application des méthodes TAL dans le secteur, mon alternance m'a aidé à m'améliorer personnellement et professionnellement. La communication était un des aspects clés pour le bon déroulement de nos projets et c'est l'une des compétences sur laquelle je me suis largement développé et m'améliore encore. En effet, c'était à travers des ateliers de cadrage et des présentations que l'on nous a présenté les enjeux, les risques et les perspectives non seulement de nos projets, mais aussi de nos collaborateur·rice·s. Il fallait apprendre à être synthétique sans omettre les détails les plus importants et à rester toujours bienveillant·e à l'égard de tous·tes.

De plus, au vu de la taille de notre équipe, il était facile de se perdre dans les tâches à accomplir. Les deux projets se sont déroulés en même temps, ce qui nous a appris à prioriser nos tâches efficacement afin de faire avancer le projet au maximum tout en conservant du temps pour préparer nos présentations au reste de l'équipe. J'ai également appris comment avoir une vision plus claire et détaillée à partir des méthodes de gestion de projets qui nous ont été enseignées.

Enfin, j'ai appris à être autonome et à avoir l'initiative d'apprendre plus sur le terrain. Mon expérience à la SNCF m'a permis de croiser mes nouvelles connaissances en TAL avec ma passion pour le transport ferroviaire et la mobilité en général. J'espère ainsi pouvoir continuer de travailler dans un cadre stimulant dans lequel je pourrai m'épanouir tout en devant relever quotidiennement de nouveaux défis.

Bibliographie

Dassault Systèmes. (2021). Logiciel d'analyse sémantique des données textuelles pour l'entreprise. *Proxem NLP for Business*. https://www.proxem.com/a-propos/

Décret-loi du 31 août 1937 portant réorganisation du régime des chemins de fer, (1937). https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000848872

Grinberg, M. (2018). *Flask web development : Developing web applications with python*. O'Reilly Media, Inc.

Groupe SNCF. (2020). *SNCF, Votre partenaire de tous les services de mobilités*. SNCF International. https://medias.sncf.com/sncfcom/finances/Publications_Groupe/SNCF_Votre_Partenaire_Service_toutes_Mobilites_2020.pdf

Haddad, R. (2016). *Manuel d'écriture inclusive* (Mots-Clés, Éd.). Mots-Clés. https://www.univ-tlse3.fr/medias/fichier/manuel-decriture_1482308453426-pdf

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

Joblib Development Team. (2020). *Joblib : Running Python functions as pipeline jobs*. https://joblib.readthedocs.io/

LOI n° 2018-515 du 27 juin 2018 pour un nouveau pacte ferroviaire (1), 2018-515 (2018).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Association for Computational Linguistics (ACL) System Demonstrations*, 55-60. http://www.aclweb.org/anthology/P/P14/P14-5010

Mattmann, C. (2020). *Tika-Python*. https://github.com/chrismattmann/tika-python

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn : Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Ranchon, G. (2016). *Une didactique de la langue, de la culture et du genre : Le manuel FLE, discours et réalisations*. Université de Lyon.

Reback, J., McKinney, W., Jbrockmendel, Van Den Bossche, J., Augspurger, T., Cloud, P., Hawkins, S., Gfyoung, Sinhrks, Roeschke, M., Klein, A., Terji Petersen, Tratner, J.,

She, C., Ayd, W., Naveh, S., Patrick, Garcia, M., Schendel, J., … H-Vetinari. (2021). *pandas-dev/pandas : Pandas 1.2.4* (v1.2.4) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3509134

SNCF. (s. d.). *Qui sommes-nous ?* SNCF. Consulté 1 juin 2021, à l'adresse https://www.sncf.com/fr/groupe/profil-et-chiffres-cles/portrait-entreprise/qui-sommes-nous

SNCF Réseau. (2021). *Carte nationale des principaux chantiers en 2021*. https://www.sncf-reseau.com/fr/carte/carte-nationale-principaux-chantiers-en-2021

SNCF Réseau. (2021). *Présentation de la Direction Générale Numérique SNCF Réseau*. https://sncf.sharepoint.com/sites/SNCF-Reseau-Direction-Numerique/SitePages/Presentation-Direction-Numerique.aspx

Van Rossum, G. (2020). *The Python Library Reference, release 3.8.2*. Python Software Foundation.

Tableau d’illustrations

Figure 1. Schéma simplifié : organisation de la SNCF depuis le 1er janvier 2020 8
Figure 2. Carte nationale des principaux chantiers en 2021 (SNCF Réseau, 2021) 9
Figure 3. SNCF Réseau et ses directions générales 9
Figure 4. Les départements composant la DGNUM 10
Figure 5. Le Département DAD dont fait partie Smart Studio 11
Figure 6. Visualisation de données avec Proxem Studio 14
Figure 7. Quelques catégories établies sur Proxem Studio 15
Figure 8. Analyse de ressentis et de thématiques 15
Figure 9. Séparation des champs textuels correspondant aux réponses positives et négatives 16
Figure 10. Arborescence de thématiques et de concepts définis avec les client·e·s 16
Figure 11. Exemples de verbatim à écarter 17
Figure 12. Graphiques générés après l'analyse 18
Figure 13. L'en-tête d'une FI 22
Figure 14. L'en-tête d'un RAC 22
Figure 15. Structure du premier fichier CSV 22
Figure 16. Causes sur le réseau et par région 23
Figure 17. Perspectives pour l'outil d'aide à la décision 24
Figure 18. Expressions régulières pour rechercher les thématiques qui ressortent dans les rapports 25
Figure 19. Valeurs binaires indiquant la région où un incident est produit selon les rapports 26
Figure 20. Arborescence de notre projet : Flask + ML 28
Figure 21. Page d'accueil de l'interface graphique 28
Figure 22. Exemple d’une page de résultats 29

Tableau 1. Résultats de classification multilabel 26

Annexes

Annexe 1 : Exemple d'une Fiche d'incident (FI)

Fiche Evènement **N° 1 881**

Evènement

Date : 05/09/08 17:10
Nature : Rupture de caténaire
Lieu : Voie M2 à Paris Montparnasse
Transmis par : Adjoint PSSR PR

Date de prise de connaissance :

Par le GI :
Par le RLS :
Amorçage :

Ligne

Ligne : 553000 - Ligne d'Ouest-Ceinture à Chartres
Tronçon : du pk 1 + 102 au pk 84 + 948

Région : PRG - Paris-Rive-Gauche
Pk : 2 + 400

RT : 3041 - Ensemble Ferroviaire de Paris-Montparnasse

Département administratif: Paris
PN inscrit au PSN:
Type de bloc : BAL - Block automatique lumineux
Régime d'expl. : DV - Double voie

Catégorie ligne : NORMAL - Ligne du réseau conventionnel à écartement normal
Electrification : 1500V - 1500 volts continu
Groupe UIC : 5 - 5
Présence KVB :
Nb. voies : 2 voie(s)
N° voie : V2 - VPL -

Particularités d'exploitation : BANAL

Type évènement

Evt. EC :
Evt. ES :
Evt. ESR :
Affectation :
Annexe 3 : N
Evt. Sig. : N

RFN : O

VP : VP

Eng. pt.pro : N
Prés. DAAT:
DAAT serv.:
KVB serv. : O
Vit. dérail. > 30 km/h :
Int. circ. > 6 h : N

Trains et Véhicules concernés

N° Cir. Fer.	Mvt.	Circ	Titulaire CS	Act.	Sous Traitant	Act.	N° engin	Vit.	Nb. véh.	Ton-nage	Lon-geur	Nb. véh.	KVB pres.	KVB serv.	DAAT prés.	DAAT serv.	Train mvt	Train voy
8844	T	01	SNCF MOB.	VOYAGEUR									O	O				

N° véhicule	Type (B/E)	Position	AZ	Chargement	Particularité	C. danger
TGV	B					

Conséquences

0 Sans conséquence

Conséq. humaines

	Agents		Voyageurs		Personnes non autorisées sur les installations ferroviaires	Usagers des PN	Autres	TOTAL
	ds le train	hors train	ds le train	hors train				
Tués	0	0	0	0	0	0	0	0
Blessés	0	0	0	0	0	0	0	0
Dont blessés graves	0	0	0	0	0	0	0	0

Version du : 06/01/2020 Page 1 sur 2

CS	Tit. CS	Périmetre	ACT / DOM	ST	ACT / DOM	ENTITE CONCERNEE	ACT / DOM
178	SNCF RES.		M et T-FIX			SNCF RES.	M et T-FIX

Libellé automatique

Incidents caténaires de Infrastructure Équipement Sans accident

Entité GI en cause concernée

PRG - REGION Paris-Rive-Gauche

Fonctions

MI - Mainteneur de l'infrastructure

Gravité / risque / classification

Gravité dommage matériel : 2 - Dommages matériels mineurs

Gravité dommage personnel : 1 - Aucune conséquence

Risque : 2 - Risque faible

Classification : O - Incident inclassable niv. de pond. : 0.00000

Nomenclature EPSF :

Nomenclature secondaire EPSF :

Gravité EPSF :

Production du rapport GI : Pas de rapport

Information immédiate

Description de l'évènement :

A 17H10, le TGV 8844 arrache la caténaire sur la voie M2 en gare de Paris MOntparnasse du km 2 au km 2,4. Le gabarit de la voie TGV1 se trouve engagée par la caténaire arrachée sur la voie M2.

Actions urgentes mises en oeuvre par les exploitants :

Coupure caténaire suite à disjonction des SEL, puis demande de consignation C.

Mesures conservatoires prises par le GI :

Transbordement des voyageurs en gare de Ouest Ceinture. Dégagement de la rame du 8844 de la voie M2 par des engins diesel.

Autres commentaires / mesures autres mises en oeuvre :

Mesures conservatoires :

Sans objet

Entité origine de l'écart :

Suivi DSSR/RLS

Demande de complément

Date de dem. :

Date de réc. :

Emetteur :

Commentaires :

Demande decret 2006-1279

Date de dem. :

Date de réc. :

Emetteur :

Commentaires :

Demande d'enquete nationale

Date de dem. :

Date de réc. :

Commentaires :

Visibilité

SNCF MOB.
SNCF RES.
DC
DSSR
RFF
EPSF
BEA

Pièces jointes

Titre	Date de création	Date de mise à jour	Utilisateur	Conf. DSSR
SQ3_1237_PRG_050908.pdf	11/09/2008	11/09/2008	CONTENU ANONYMISÉ	N

Annexe 2 : Exemple d'un Rapport d'analyse causale (RAC)

DIRECTION DE LA CIRCULATION FERROVIAIRE

DIRECTION DE L'EXPLOITATION
Département Sécurité et Veille Opérationnelle
Division Veille Opérationnelle (IOS-V)

CONTENU ANONYMISÉ

CONTENU ANONYMISÉ

Avertissement : Le présent rapport est réalisé dans le cadre du décret 2006-1279 du 19/10/2006. Ce rapport décrit les circonstances, les conséquences et les causes directes et indirectes d'un évènement de sécurité. Il ne vise pas à déterminer des responsabilités à la suite de cet évènement.

RAPPORT D'ANALYSE CAUSALE n °14982

Identifiant de l'événement : 11 février 2014 / 18:12 / 420000 / Viroflay-Rive-Gauche (78) / Le conducteur du train 165452 (EF SNCF) constate un pendule caténaire décroché lors de son arrêt sur la voie 2 bis en gare de Viroflay (repris au CRJ).

.

1. DESCRIPTIF

Lieu : Viroflay-Rive-Gauche (78) Km : 13+900
Voie : 2 bis (VP) Régime d'exploitation : Double voie Block Automatique Lumineux
Electrification : 1500 volts Continu
Train n°: 165452 Parcours de Rambouillet (17h22) à Paris Montparnasse (18h22)
Activité : Voyageur EF : SNCF
Engin moteur : 7608 avec Rame VB2N C 06

2. CIRCONSTANCES

Le conducteur du train 165452 (EF SNCF) est reçu à 18h03 sur la voie 2 Bis en gare de Viroflay-Rive-Gauche pour assurer la desserte commerciale.
Lorsqu'il se remet en marche, il constate un flash accompagné d'une disjonction.
Après immobilisation de son train et accord d'une protection personnel sur la voie 1 IN à 18h33, il effectue la visite de son train et constate la présence d'un pendule décroché sur sa voie (2bis).
Il en informe le régulateur, restitue la protection voie 1 IN qui n'est pas impactée par l'incident (circulation rétablie à 18h40) et effectue une demande de secours.
Les voyageurs présents dans la rame sont transbordés.
Les astreintes concernées sont avisées (EIC, Caténaire, Voyageur,…)
Les installations sont remises en état par les agents de maintenance caténaire et après annulation de la Demande de Secours par le conducteur, le train 165452 est acheminé hors service commercial à Paris Montparnasse.

3. CONSEQUENCES

Gravité des dommages humains : Personnes transportées : néant
Autres : néant

afaq ISO 9001 Qualité AFNOR CERTIFICATION

Gravité des dommages Matériels : Matériel : néant
Infrastructures : un pendule rond décroché et pendules adjacent ligaturés.
Circulation : 38 trains impactés et 638 minutes de perdues.
Autres : néant

4. MESURES PRISES

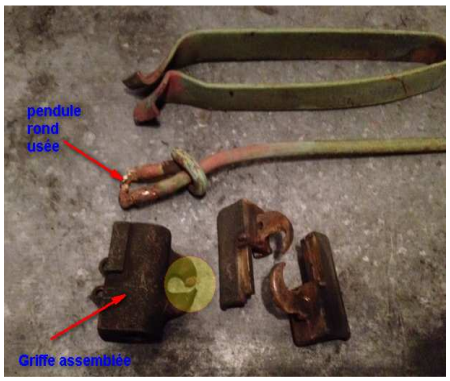
Par le GID :
Interruption de la circulation sur la voie 2 bis.
Appel des astreintes concernées pour intervention.
Accord protection personnel au conducteur Voie 1N pour visite de sa rame.
Accord consignation C et DFV sur la zone de l'incident au service maintenance pour réparation.

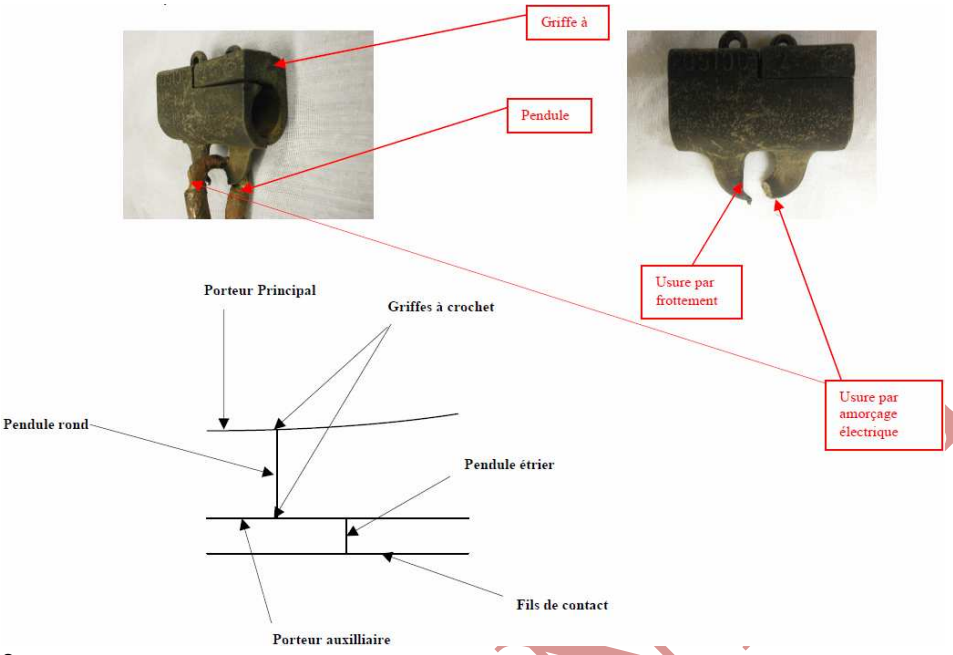
Par l'EF :
Avis au SGTC de l'incident
Demande d'une protection Personnel voie 1N pour visite de sa rame.
Avis à la clientèle et évacuation de la rame.
Adaptation du plan de transport.
Etablissement d'une Demande de Secours.

LEVEE DES MESURES CONSERVATOIRES :
La reprise de la circulation sur la voie 1N à 18h40 après visite de la rame par le conducteur.
Le rétablissement de la circulation sans restriction sur la voie 2 bis à lieu à 23h19 à l'issue de la remise en état des installations caténaire par la maintenance.

5. CAUSES

Analyse :
Le conducteur du train 165452 alors qu'il démarre de la gare de Viroflay circule à la vitesse de 25 km/h lorsqu'il constate les indices suivant : Flash + disjonction. Il s'arrête d'urgence en abaissant ses pantographes (machine positionnée en queue du train), la tête de la rame est au Km 13+750. Il observe la caténaire depuis la cabine de conduite et ne constate rien d'anormal. Un train croiseur lui signale qu'un pendule est décroché au dessus de la deuxième voiture de tête, Il informe le SGTC par la Radio Sol Train et effectue une demande de Secours.
Les agents caténaires constatent : un pendule rond cassé et les deux pendules ronds encadrant qui ne sont plus en portance. La voie 2Bis est équipée d'une caténaire Normale en 1,5 kV Courant Continu, elle est constituée de deux fils de contact de 107 mm², d'un porteur principal bronze de 116 mm², d'un porteur auxiliaire de 104mm² et d'un feeder de 262mm².
La réparation de la caténaire est terminée à 22h44 et après annulation de la Demande De secours la rame est acheminée en autonome sans voyageur vers Paris Montparnasse.
Lors de l'incident il n'y a pas eu de choc avec le pantographe de la circulation, quelques impacts en toiture seront observés en atelier ne nécessitant pas d'intervention. La rame a été remise à disposition de l'EF.





Cause :
Le décrochage du pendule rond est dû à une rupture mécanique de la griffe suite aux mouvements d'oscillations de la caténaire qui sont provoqué par les circulations et à un phénomène électrique (observation de traces d'amorçages) lié à une sollicitation plus importante de la zone correspondant au point de stationnement des rames (intensités plus importantes et pression du panto).

6. Risques

Risques afférents à l'évènement :
Aucun, L'arrêt du train est obtenu avant que le pantographe ne vienne toucher le pendule décroché.
La rame étant encore à quai l'évacuation des voyageurs n'a pas posé de difficultés.

Risques génériques :
Ce type d'événement peut provoquer une détérioration du pantographe de l'engin moteur suite à choc, voir un arrachement de la caténaire au passage d'une circulation avec toutes conséquences que cela peut entrainer vis à vis des personnes qui pourrait se trouver dans l'environnement d'un tel incident.

7. Conclusions du GID
Enquête conclue : A la suite de l'enquête et sur la base des éléments fournis par l'Infrapôle, le GID conclu à une défaillance des installations suite au décrochage d'un pendule rond de sa griffe, liée à une usure mécanique et électrique.

CONTENU ANONYMISÉ

Copie à : DET EIC – DSIN – EF SNCF

Annexe 3 : Arborescence de concepts concernant les incidents caténaires