

# Observations : Automatisation

NOTE TECHNIQUE

GILBERT D. BRAULT

## Table des Matières

|   |    |
|---|----|
| Introduction .....  | 2  |
| Le problème à résoudre .....  | 2  |
| La solution proposée .....  | 2  |
| Etapes de traitement des observations .....   | 3  |
| La saisie.....  | 3  |
| L'analyse.....  | 3  |
| Le traitement.....  | 3  |
| Conclusion pour l'automatisation de la phase d'analyse.....                         | 4  |
| Nature des observations .....   | 5  |
| Définition du problème de groupement des observations similaires.....               | 7  |
| Quelles sont les limites des solutions classiques .....                             | 7  |
| Quelles sont les attentes pour une codification automatique des observations ?..... | 7  |
| Solution au problème de classification des observations .....                       | 8  |
| Les techniques nécessaires .....  | 8  |
| Technologies utilisées .....  | 9  |
| La méthode « Minimum Spaning Tree » (MST) .....                                     | 10 |
| Présentation de la solution Ilex .....  | 11 |
| Visualisation par arbre .....   | 11 |
| La recherche .....  | 15 |
| Choisir le niveau d'agglomération .....   | 17 |
| Conclusion .....  | 18 |
| Architecture de la solution .....   | 18 |
| Que doit-on retenir ? .....   | 18 |

# Introduction

## Le problème à résoudre

La maintenance des installations d'ascenseurs est régie par des lois et des décrets en France.

Les bureaux de contrôles, mandatés par le représentant du propriétaire, réalisent des visites générales périodiques (VGP) ou d'autres visites, une réception de travaux de modernisation par exemple, etc... Leurs auditeurs produisent alors des rapports de compte-rendu de visite d'inspection.

Ces rapports consignent, entre autres, des anomalies concernant les équipements visités. La réglementation, dans certains cas, oblige à régler ces problèmes. Ces anomalies sont dénommées observations ou réserves. Elles doivent faire l'objet d'un suivi de résolution des problèmes identifiés.

Ces rapports sont ensuite transmis aux sociétés de maintenance d'ascenseurs, mandatés par le représentant du propriétaire.

La société Ilex ascenseurs, opérant sur le territoire Français, dont le siège est à Antibes en France doit donc faire face à cette problématique.

Ilex ascenseurs reçoit régulièrement un flux de rapports de compte rendu de visite d'inspection, rédigés par différents bureaux de contrôles mandatés pour l'inspection d'appareils du parc dont Ilex ascenseurs est prestataire de maintenance.

L'objet de cette note est de décrire le processus de traitement des observations chez Ilex ascenseurs pour montrer l'architecture des technologies contribuant à l'automatisation du traitement de l'information, en se concentrant sur la phase analyse.

La phase analyse a pour but d'affecter une observation à une méthode de résolution de problème en vue de l'attribution des travaux à réaliser aux équipes d'exécutions d'Ilex ascenseur, voire parfois au client.

## La solution proposée

La solution proposée à Ilex est une intégration de la solution dans le système d'information existant avec la possibilité pour les utilisateurs :

- De visualiser l'arbre des observations sémantiquement similaires émises par un bureau de contrôle sur une période donnée. On peut choisir l'un des bureaux suivant le contenu de la base de données préexistante
- De rechercher les observations similaires sémantiquement en entrant un texte d'observation.
- D'interagir avec la représentation pour définir les groupements correspondants aux observations partageant la même sémantique en réglant la taille des groupes par une distance.
- De construire l'ensemble des données nécessaire à la visualisation de manière périodique pour mettre à jour la visualisation et la recherche et la définition des groupes par agglomération.

## Etapes de traitement des observations

Le traitement des observations comporte trois étapes

- La **saisie** : collecter l'ensemble des rapports et saisir les observations s'y trouvant
- L'**analyse** : identifier la nature des observations et les affecter à des méthodes de résolution
- Le **traitement** : suivre l'avancement du processus de résolution, quand l'observation a été attribué à une équipe Ilex ascenseur qui doit exécuter la résolution du problème.

Annuellement, Ilex ascenseurs doit traiter entre 20,000 et 30,000 observations. Des outils de traitements automatisés sont nécessaires pour obtenir le niveau de qualité requis dans le métier de la maintenance des ascenseurs afin d'obéir à la réglementation.

### La saisie

La saisie a les objectifs suivants

- Collecter tous les rapports reçus par Ilex ascenseurs au travers différents canaux
- Codifier l'information pour faciliter le traitement ultérieur
- Associer les observations à un appareil sous mandat Ilex ascenseur

Dans cette phase, toutes les observations traitées par Ilex ascenseur se retrouvent donc dans une base de données centralisée.

Il faut noter toutefois que l'observation en tant que telle, n'est pas codifiée : c'est un texte en langage naturel, tel qu'il a été saisi par l'auditeur lors de sa visite d'inspection.

### L'analyse

Le but de l'analyse, qui est réalisé par des experts de la société Ilex ascenseur (l'analyste), est d'analyser les observations afin de définir le travail correspondant à la résolution du problème posé.

L'analyste dispose devant lui d'un écran où toutes les observations pour un appareil donné sont représentées sous forme d'une table.

Il les prend une à une et remplit par observation une fiche d'affectation où il caractérise l'équipe en charge du traitement et la nature du travail à réaliser.

L'objectif de la présente note technique est de montrer comment ce travail d'affectation peut être automatisé.

### Le traitement

Le traitement fait l'objet d'un suivi de travaux classique permettant de suivre l'avancement des tâches affectés aux personnels des différentes équipes d'exécution. De manière ultime, la « levée de réserves » est réalisée – avec transmission au bureau de contrôle – quand les travaux sont terminés ou que leur planification certaine est acquise.

## Conclusion pour l'automatisation de la phase d'analyse

### La phase d'analyse

- Nécessite des ressources qualifiées quand elle est réalisée manuellement.
- Implique une certaine dispersion des résultats d'affectation du fait du travail manuel de classification lié à l'expertise de chacun et à la fatigue d'une activité routinière.
- Impose des délais de réalisation du fait du besoin en ressources humaines.
- Constitue un goulet d'étranglement dans le processus de traitement des observations.

Il est donc parfaitement naturel de rechercher les moyens d'automatiser cette tâche. Nous allons donc nous focaliser sur l'automatisation de la phase analyse, en particulier :

- Comprendre le problème de l'affectation des observations pour les attribuer à une méthode de résolution de problème. Une première étape c'est de regarder la nature des observations.
- Décrire l'architecture permettant d'automatiser ce processus.

## Nature des observations

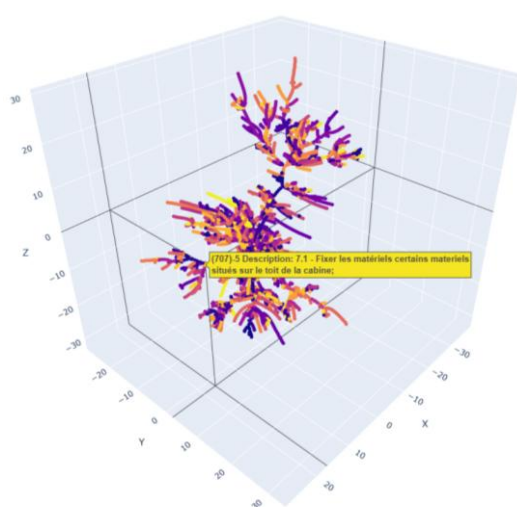
Prenons par exemple les observations qui sont émises par le bureau de contrôle AASA.

AASA a émis sur les deux dernières années 4721 observations dans les rapports concernant des appareils dont Ilex est mandaté pour la maintenance.

Sur ces 4721 observations, 2755 font l'objet d'une description unique. C'est-à-dire qu'il y a 2755 phrases en langage naturel, qui décrivent des observations de manière unique : ce sont des phrases qui ont une syntaxe différente les unes des autres.

Plongeons maintenant sur le détail d'une observation pour comprendre ce que veut dire cette variété, liée à l'utilisation du langage naturel.

Nombres -- Observations(4721|2755)|groupes=(743)



Dans l'arbre des observations, comprenant 743 groupements, sélectionnons une observation au hasard :

7.1 - Fixer les matériels certains matériels situés sur le toit de la cabine

Et regardons maintenant les observations qui sont voisines

|      | index | Enregistre_le       | societe_rattachement | Agence_Id | Appareil_Id | Description  |
|------|-------|---------------------|----------------------|-----------|-------------|--|
| 1231 | 1231  | 2022-12-23 14:39:39 | ILEX-06              | 601       | 22491       | 7.1 - Fixer les matériels certains matériels situés sur le toit de la cabine;        |
| 1389 | 1389  | 2023-01-04 15:35:24 | ILEX-06              | 601       | 24990       | 7.1 - Fixer les matériels, situés sur le toit de la cabine. Prestation non réalisée  |
| 2444 | 2444  | 2023-02-19 23:34:49 | ILEX-06              | 601       | 24502       | 7.1 - Fixer les matériels qui le nécessitent situés sur le toit de la cabine.        |
| 2894 | 2894  | 2023-03-02 09:52:20 | ILEX-06              | 601       | 2142        | 7.1 - Fixer les matériels rajoutés sur le toit de la cabine. Prestation non réalisée |
| 3389 | 3389  | 2023-03-28 15:26:52 | ILEX-06              | 601       | 25102       | 7.1 - Fixer les matériels qui le nécessitent situés sur le toit de la cabine.        |

Avec « la distance » qu'on a imposée, un groupe de 5 d'observations a été trouvé.

Parmi ces 5, 4 sont uniques :

7.1 - Fixer les matériels certains matériels situés sur le toit de la cabine;  
7.1 - Fixer les matériels, situés sur le toit de la cabine. Prestation non réalisée  
7.1 - Fixer les matériels qui le nécessitent situés sur le toit de la cabine.  
7.1 - Fixer les matériels rajoutés sur le toit de la cabine. Prestation non réalisée

On voit donc qu'un groupe contient des observations présentant des petites variations de langage, ou des compléments.

Augmentons la distance, on voit que le nombre de groupes chute à 241. Regardons l'évolution du groupe de notre observation précédente.

On a changé la distance de 1 à 1.8 et on a recalculé les groupes (re coloriage).

| index | Enregistre_le       | societe_rattachement | Agence_Id | Appareil_Id | Description   |
|-------|---------------------|----------------------|-----------|-------------|---|
| 428   | 2022-05-31 14:40:32 | ILEX-06              | 601       | 23829       | 7.1 - Fixer les différents éléments situés sur le toit de la cabine, notamment les alimentations des vigiks et les différents téléphones. |
| 824   | 2022-07-20 15:59:07 | ILEX-06              | 601       | 25513       | 7.1 - Fixer les matériels certains matériels situés sur le toit de la cabine, notamment le téléphone....                                  |
| 844   | 2022-07-20 15:59:21 | ILEX-06              | 601       | 25514       | 7.1 - Fixer les matériels certains matériels situés sur le toit de la cabine, notamment le téléphone....                                  |
| 1231  | 2022-12-23 14:39:39 | ILEX-06              | 601       | 22491       | 7.1 - Fixer les matériels certains matériels situés sur le toit de la cabine;   |
| 1355  | 2023-01-04 11:19:12 | ILEX-06              | 601       | 24989       | 7.1 - Fixer les matériels , notamment le téléphone, situés sur le toit de la cabine. Prestation réalisée                                  |
| 1389  | 2023-01-04 15:35:24 | ILEX-06              | 601       | 24990       | 7.1 - Fixer les matériels, situés sur le toit de la cabine. Prestation non réalisée   |
| 1426  | 2023-01-04 15:35:30 | ILEX-06              | 601       | 24991       | 7.1 - Fixer certains matériels situés sur le toit de la cabine, notamment le téléphone.... Prestation réalisée                            |
| 1687  | 2023-02-06 09:49:14 | ILEX-06              | 601       | 24448       | 7.1 - Fixer les matériels certains matériels situés sur le toit de la cabine, notamment le téléphone.... Prestation non réalisée          |
| 2444  | 2023-02-19 23:34:49 | ILEX-06              | 601       | 24502       | 7.1 - Fixer les matériels qui le nécessitent situés sur le toit de la cabine.   |
| 2758  | 2023-02-28 15:32:49 | ILEX-06              | 601       | 25371       | 7.1 - Fixer les matériels qui le nécessitent situés sur la cabine, notamment le téléphone..   |
| 2894  | 2023-03-02 09:52:20 | ILEX-06              | 601       | 2142        | 7.1 - Fixer les matériels rajoutés sur le toit de la cabine. Prestation non réalisée  |
| 3389  | 2023-03-28 15:26:52 | ILEX-06              | 601       | 25102       | 7.1 - Fixer les matériels qui le nécessitent situés sur le toit de la cabine.   |

On passe alors à un groupe de 12 observations dont 10 sont uniques :

```
7.1 - Fixer les différents éléments situés sur le toit de la cabine, notamment les
alimentations des vigiks et les différents téléphones.
7.1 - Fixer les matériels certains matériels situés sur le toit de la cabine, notamment le
téléphone.....
7.1 - Fixer les matériels certains matériels situés sur le toit de la cabine;
7.1 - Fixer les matériels , notamment le téléphone, situés sur le toit de la cabine.
Prestation réalisée
7.1 - Fixer les matériels, situés sur le toit de la cabine. Prestation non réalisée
7.1 - Fixer certains matériels situés sur le toit de la cabine, notamment le téléphone.....
Prestation réalisée
7.1 - Fixer les matériels certains matériels situés sur le toit de la cabine, notamment le
téléphone..... Prestation non réalisée
7.1 - Fixer les matériels qui le nécessitent situés sur le toit de la cabine.
7.1 - Fixer les matériels qui le nécessitent situés sur la cabine, notamment le téléphone..
7.1 - Fixer les matériels rajoutés sur le toit de la cabine. Prestation non réalisée
```

De nouveaux éléments se sont glissés dans les phrases, toutefois, le travail à réaliser reste certainement identique.

C'est donc bien le problème de ces observations en langage naturel :

Beaucoup d'observations peuvent être regroupées car elles correspondent à une méthode de résolution de problèmes identique du fait qu'elles ont la même sémantique.

Le problème technique posé au système de traitement de l'information, c'est donc de regrouper ensemble des observations qui sont exprimées en langage naturel lorsqu'elles ont la même sémantique.

# Définition du problème de groupement des observations similaires

## Quelles sont les limites des solutions classiques

Les systèmes de base de données relationnels sont les outils traditionnels du traitement de données en entreprise.

Lorsqu'on utilise les systèmes de base de données relationnels, le traitement de l'information est très efficace, dès lors que l'information a été codifiée.

La codification de l'information, c'est non seulement la traduction de l'information dans une forme assimilable par une machine, mais il y a une étape supplémentaire, c'est de définir l'information de telle manière que la machine hérite d'une partie de la connaissance contenue dans cette information.

Si les observations avaient fait l'objet, en amont, d'un travail de codification suffisamment fin pour définir de manière univoque le travail à réaliser, la machine, avec un système de base de données relationnel ne rencontrerait aucunes difficultés pour grouper les observations !

Mais ce n'est pas le cas, du fait de la multitude des intervenants et de la relative complexité des problèmes à traiter, le codage en langage naturel des observations est une nécessité.

Les systèmes conventionnels de gestion de base de données sont donc impuissants aujourd'hui pour résoudre ce problème de classification des observations !

Il était donc naturel que cette codification soit faite manuellement : c'est l'étape d'analyse.

Un système de gestion de bases de données classique ne sait pas regrouper des phrases en langage naturel qui ont la même sémantique

## Quelles sont les attentes pour une codification automatique des observations ?

Le problème qui doit être résolu, c'est donc de codifier automatiquement les observations pour résoudre la question de l'automatisation de la phase analyse.

Avant de passer à la résolution effective du problème, il faut se poser la question de savoir quels sont les besoins pour arriver à une solution viable.

Les besoins sont les suivants :

1. Trouver un moyen automatique pour grouper les observations de même sémantique.
2. Représenter les observations pour qu'un expert métier puisse confirmer que la classification proposée est conforme à l'état de l'art du métier et aussi procéder à des ajustements sur les groupes.
3. Rechercher un groupe en donnant le texte d'une observation et de la repérer sur la représentation

Le premier point coule de source et le deuxième, c'est la capacité pour les hommes de terrain de juger de la pertinence de la classification et de se l'attribuer de manière opérationnelle.



# Solution au problème de classification des observations

## Les techniques nécessaires

Pour résoudre le problème de classification automatique des observations, plusieurs techniques sont nécessaires pour répondre aux besoins analysés dans l'introduction.

Voici le tableau donnant les techniques, leurs descriptions et leurs contributions aux besoins

| Technique  | Description  | Contribution aux besoins  |
|--|--|---|
| Vectorisation Recherche                            | Les phrases en langage naturel doivent être transformées en vecteurs ( $\mathbb{R}^n$ ). La projection des phrases dans cet espace vectoriel doit faire en sorte que <ul style="list-style-type: none"><li>• Des phrases de même sémantique doivent être voisines (distance entre deux vecteurs)</li><li>• On doit pouvoir retrouver une phrase qui a été vectorisée en présentant un texte ayant une sémantique voisine</li></ul> | <ul style="list-style-type: none"><li>• Préalable au groupement automatique</li><li>• Recherche des groupes en donnant le texte d'une observation</li></ul> |
| Réduction de dimensions                            | Les vecteurs dans des espaces de grandes dimensions ne sont pas représentable de manière ergonomique et interprétable facilement par l'homme <ul style="list-style-type: none"><li>• Il faut donc une technique de réduction de dimensions pour projeter les vecteurs en 2D ou 3D</li><li>• Il est souhaitable que cette réduction permette de montrer des observations ayant la même sémantique l'une à côté de l'autre</li></ul> | <ul style="list-style-type: none"><li>• Représentation des observations pour utilisation par l'expert métier</li></ul>                                      |
| Création de groupes par agglomération (clustering) | On a des vecteurs dans l'espace à n dimensions qui représentent des observations, comment grouper automatiquement les observations des sémantiques voisines <ul style="list-style-type: none"><li>• Possibilités de grouper les observations à une distance donnée par connexité</li><li>• Coloriage des groupes</li></ul>   | <ul style="list-style-type: none"><li>• Création des groupements d'observation de même sémantique</li></ul>   |
| Représentation de graphes en 3D                    | Après avoir réduit la dimension des vecteurs, il faut le représenter sur l'écran   | <ul style="list-style-type: none"><li>• Visualisation des graphes d'observations</li></ul>  |

## Technologies utilisées

Les techniques requises sont supportées par des logiciels open source disponibles et bien entretenus.

Voici les composants les plus significatifs qui sont utilisés pour la solution

| #   | Composant        | Technique                                 | Description   |
|---|------------------|---|---|
| Création des données pour en vue de la visualisation et de la recherche | all-MiniLM-L6-v2 | Vectorisation                             | Origine : Microsoft<br>« Sentence Transformer » qui a la capacité de transformer une liste de phrases en vecteurs (dimensions : 384).<br>Les vecteurs voisins correspondent à des phrases de sémantique voisine |
|   | FAISS            | Recherche                                 | Origine : META (Facebook)<br>Base de données de vecteurs permettant la recherche par distance ou par voisins les plus proches   |
|   | UMAP             | Réduction de dimension                    | Origine : Leland McInnes, Tutte Institute for Mathematics and Computing (TIMC) in Ottawa, Canada<br>Uniform Manifold Approximation and Projection<br>Méthode de réduction des dimension                         |
|   | igraph           | Réduction de dimension                    | Origine : igraph<br>Bibliothèque de traitement de graphe en langage C++ avec proxy Python<br>Graphe de la matrice des distances, Technique MST et calcul 3D de la représentation d'un graphe en 3D              |
| Interaction 3D  | Plotly           | Visualisation et interaction graphique 3D | Origine : Plotly<br>Bibliothèque de visualisation et d'interaction graphique 3D   |
|   | Sklearn          | Groupeement par agglomération             | Origine : Sklearn<br>Bibliothèque de machine learning<br>On utilise l'algorithme de clustering agglomératif   |

La solution repose sur de solides bases :

- Une logique de résolution parfaitement explicable et relativement simple d'un point de vue algorithmique.
- Des technologies existantes, correspondant à des hommes-années de recherches et développement, qui sont fiables car généralement bien supporté par des institutions ou des groupes de développeurs de renom.

## La méthode « Minimum Spaning Tree » (MST)

C'est finalement la méthode qu'on a retenu pour la représentation des observations. Ce n'est pas la plus performante, UMAP est réputée plus rapide, mais c'est celle qui s'explique le plus facilement et qui offre un rendu qui correspond au problème :

1. Dans l'espace à  $n$  dimensions, on a la collection des observations qui ont été transformées en vecteurs.
2. On peut donc calculer la matrice des distances entre toutes les observations grâce aux vecteurs.
3. Mais une matrice de distance entre vecteurs, c'est aussi la représentation d'un graphe dont les nœuds sont les observations et les relations sont les distances. Il existe des codes pour créer un graphe à partir de la matrice des distances.
4. Ensuite, on calcule le graphe correspondant à l'arbre couvrant de poids minimum ou MST de ce graphe.
5. Un arbre couvrant de poids minimum, c'est comme un réseau de routes qui relie plusieurs villes de la façon la plus économique possible. Imaginons qu'on veut relier toutes les villes d'une région avec des routes, mais qu'on veut dépenser le moins d'argent possible pour construire ces routes. Un arbre couvrant de poids minimum montre exactement quelles routes construire pour relier toutes les villes sans faire de détours inutiles, et en dépensant le moins d'argent possible. La distance d'une ville à une autre est la plus petite, d'où le terme poids minimum.
6. Ensuite, on utilise un code qui permet de générer les coordonnées 3D des nœuds du MST : c'est ce qu'on utilise pour représenter l'arbre à l'écran.
7. On ne montre pas les arrêtes car il y en a trop

Cette méthode permet de représenter les observations avec une vue 3D, chaque point ou nœud du graphe correspond à une observation.

Les observations qui sont sémantiquement similaires sont représentées par de longs fils, ce qui permet de les repérer très rapidement.

Ce graphe est extrêmement pratique pour visualiser les observations et leurs proximités.

Le graphe présenté au paragraphe « Nature des Observations » est un exemple illustrant l'ensemble des 4,721 observations émises par le bureau de contrôle AASA. Nous avons l'expérience avec des graphes de plus de 15,000 nœuds qui fonctionnent parfaitement.

La création des MST et de leurs coordonnées 3D est la plus gourmande en ressource et temps CPU.

Pour des petits graphes, par exemple pour 2725 nœuds, il faut environ 95 secondes, mais pour 16061 nœuds, il faut 3974 secondes. Ceci sur un CPU puissant (I7, voir I9 ou bien Xeon, bien qu'un seul core soit utilisé) et avec au moins 32GO de mémoire RAM.

# Présentation de la solution Ilex

Cette solution s'inscrit dans l'intégration à un système d'information existant : la base de données des observations est préexistante.

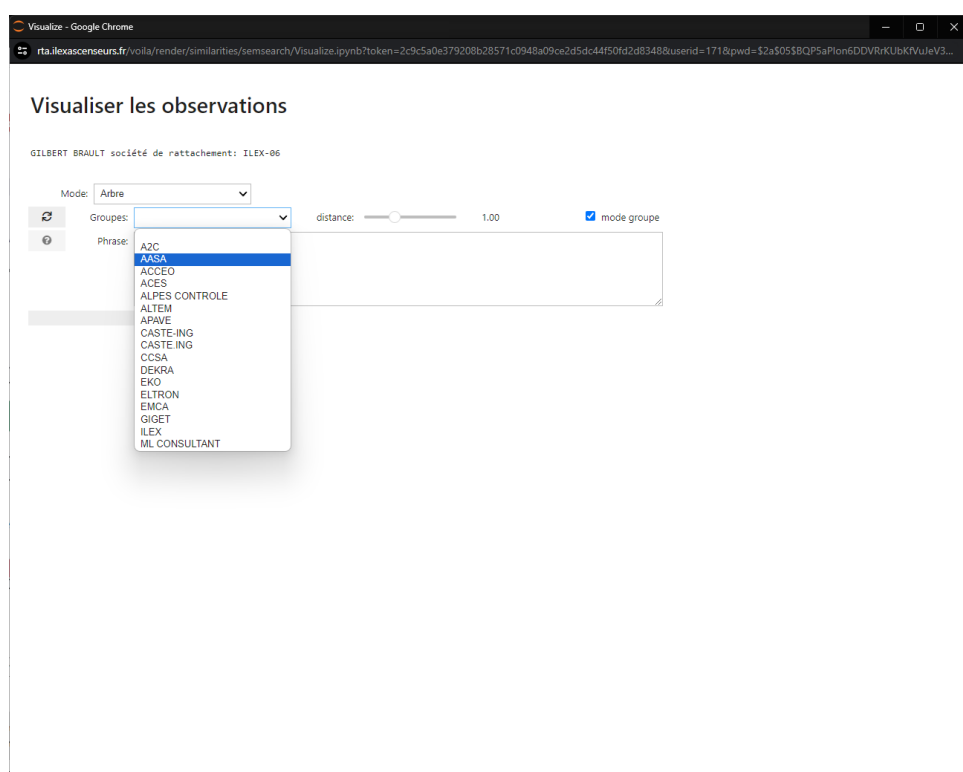
La solution met en œuvre un serveur dédié qui va chercher les données dans le serveur SQL de la base de données.

Voici maintenant une présentation des fonctionnalités utilisateurs principales

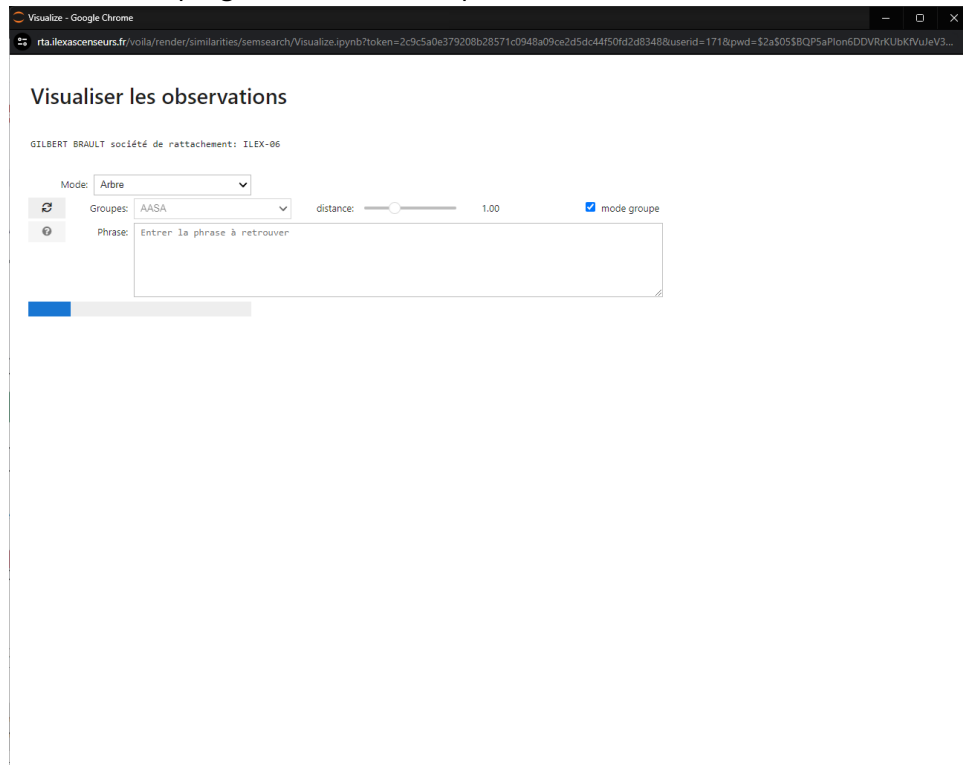
## Visualisation par arbre

Les bureaux de contrôles ont chacun leur arbre de représentation des observations

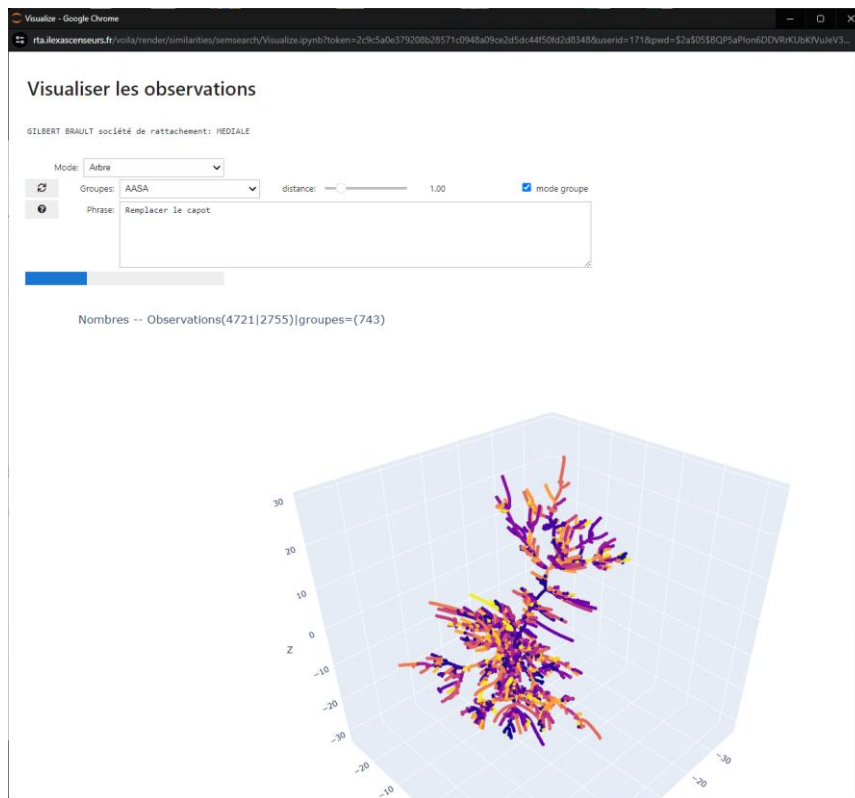
La première chose est de choisir un groupe :



Une barre de progression se met en place



Le temps que le calcul de la coloration et la représentation 3D de l'arbre soit effectués.

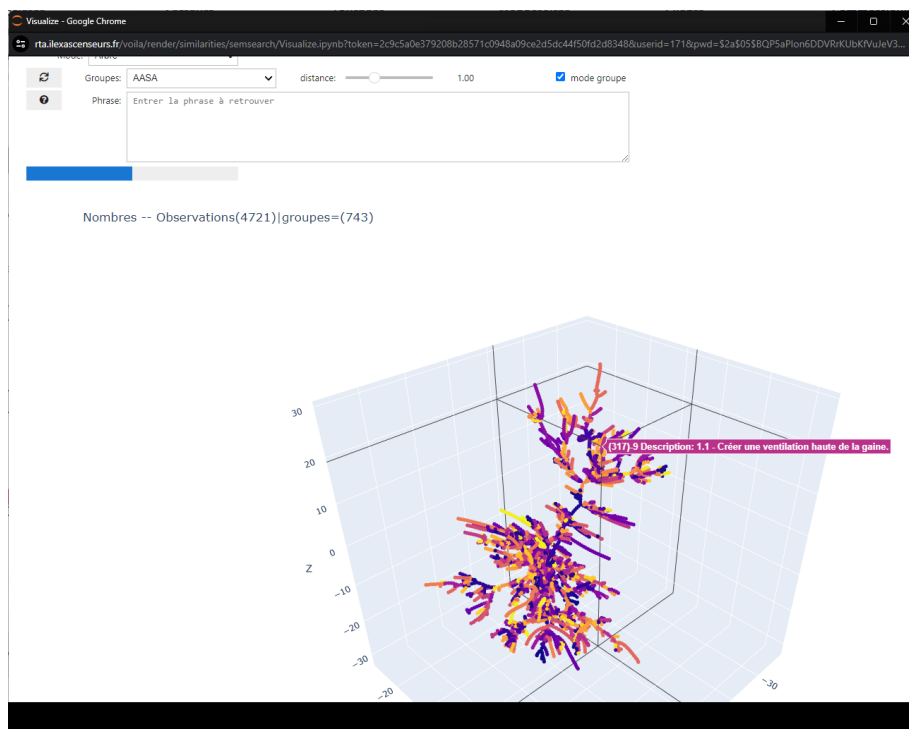


Quand c'est terminé (quelques dizaines de secondes au maximum), l'arbre apparaît en 3D.

Chaque point de l'arbre est une observation émise par un bureau dans un rapport.

Le nombre total d'observations, le nombre des observations uniques et le nombre de groupes sont indiqués.

Quand on passe la souris sur l'arbre, le texte de l'observation s'affiche



Les données en début d'observations sont – un numéro de groupe de couleur – le nombre d'élément dans le groupe

#### **(317)-9 Description: 1.1 - Créer une ventilation haute de la gaine.**

Il faut aussi maximiser la fenêtre et quand on passe la souris sur l'arbre pour voir une barre d'outils apparaître quand la souris passe sur l'arbre.



Pour activer un outil, il suffit de cliquer dessus et il devient gris un peu plus intense



Le « zoom » : il suffit d'actionner la roulette de la souris pour agrandir l'arbre autour du point de la souris



Le « pan » : il permet de translater l'arbre. Choisir un point de l'arbre et, tout en maintenant appuyé le bouton gauche de la souris, bougez-là. L'arbre se translate du déplacement de la souris



La « rotation orbital » : choisir un point, maintenir le bouton gauche de la souris enfoncé, bouger la souris. L'arbre tourne autour du point

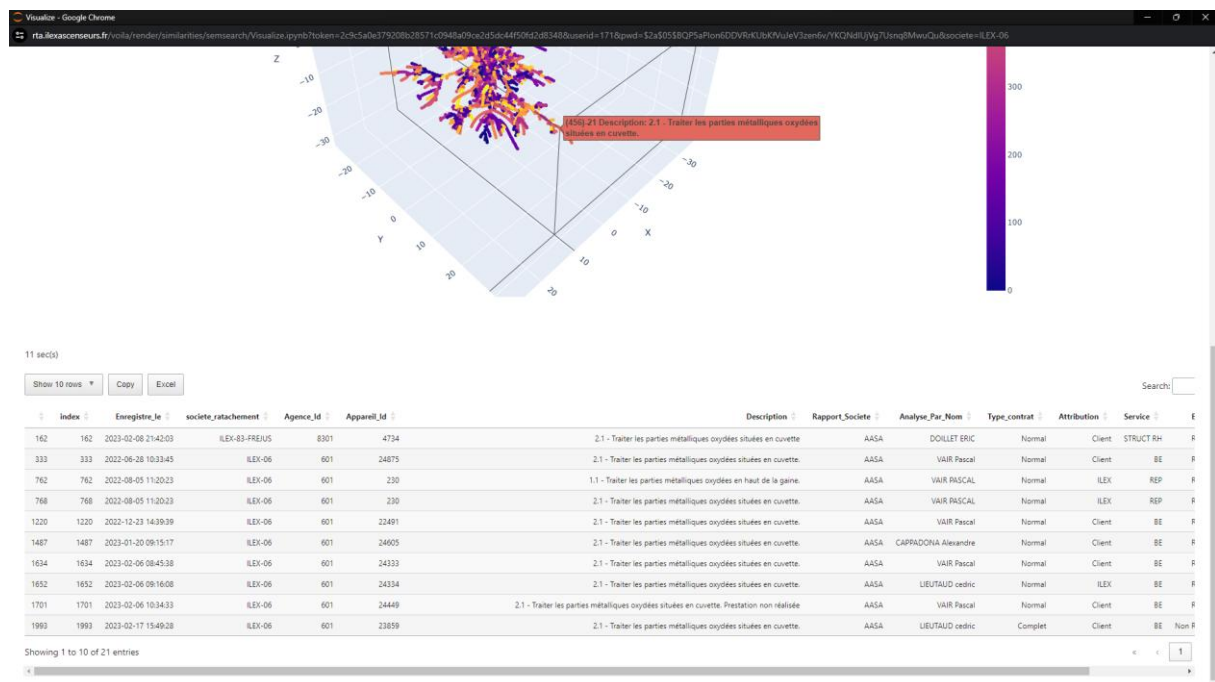


La « rotation 'table tournante' » : idem à précédemment, mais au lieu de tourner autour d'un point, c'est autour d'un axe.



La « maison » : retour au positionnement initial

Quand on clique sur un point, on peut voir la liste des observations du groupe coloré



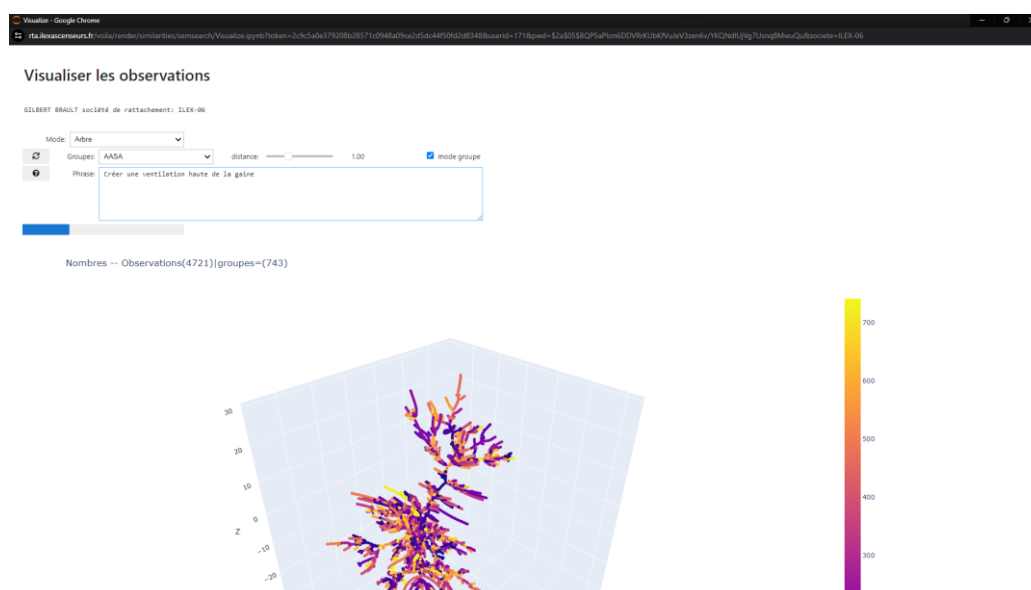
On peut

- Agrandir la liste avec l'outil « Show xx rows »
- Copier dans le presse-papier les données du tableau
- Exporter au format Excel la liste des observations du groupe
- Utiliser la pagination pour voir les autres pages du groupe

## La recherche

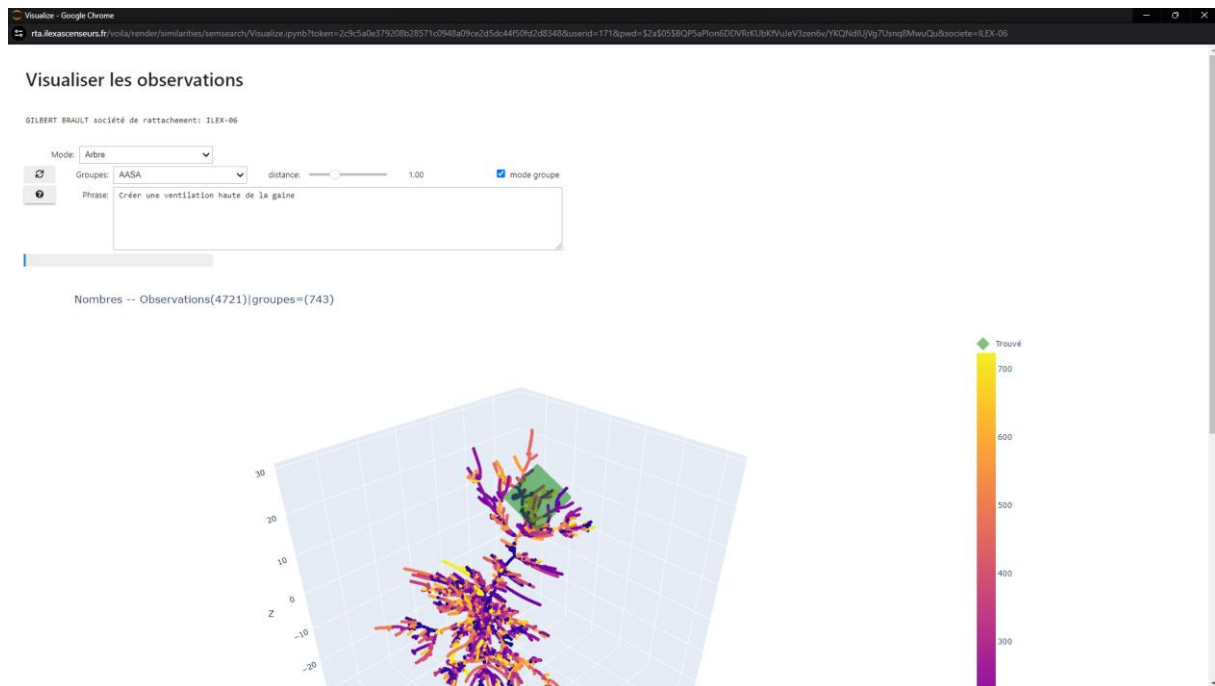
Pour trouver, à partir d'un texte d'une observation, l'endroit où elle est située dans l'arbre

Entrer un texte dans la boîte associée



Puis cliquer sur le bouton ? sur la gauche.






Un pavé vert, centré sur l'observation existante la plus proche devient visible.

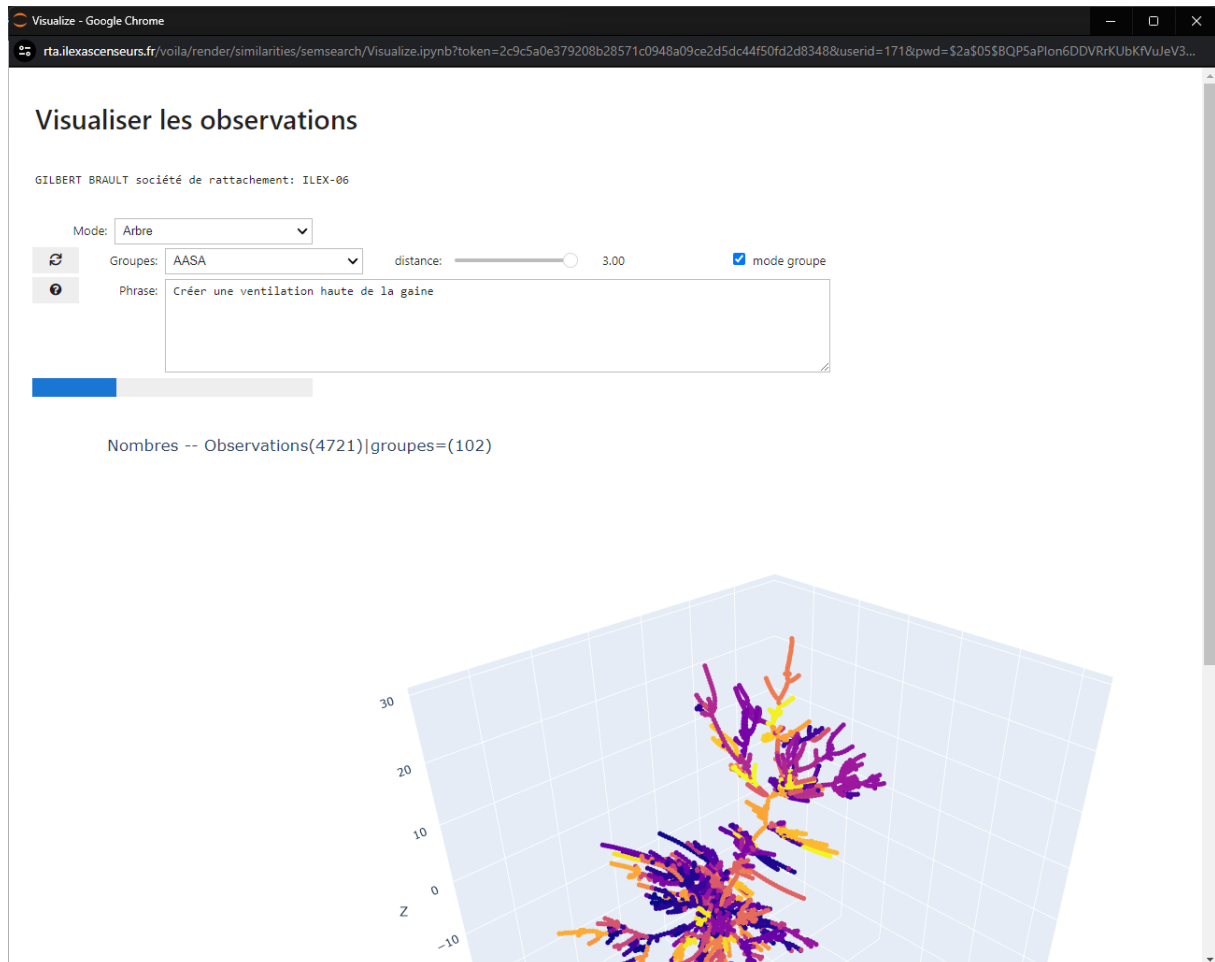
Il suffit de zoomer sur la zone pour trouver le point.

On voit aussi au-dessus de la barre colorée le carré vert jouxtant le mot Trouvé. Si vous cliquez dessus avec la souris, la zone verte disparaît et si on clique à nouveau, elle réapparaît.

On peut ainsi découvrir la zone correspondant à la recherche

## Choisir le niveau d'agglomération

On déplace le curseur distance (dans notre cas, il est placé à 3). On clique sur le bouton « Synchronisation » . Le calcul des groupes de couleur s'effectue et le résultat s'affiche en donnant le nombre de groupe et en ajustant les couleurs de points sur le graphe



Bien sûr, plus la distance est grande, moins il y a de groupes, mais aussi la vraisemblance métier n'est plus là : c'est juste un artéfact mathématique.

Note : la distance<sup>1</sup> affichée sur le curseur n'est pas la même que celle utilisée pour la recherche. On peut estimer expérimentalement un rapport de l'ordre de 3.

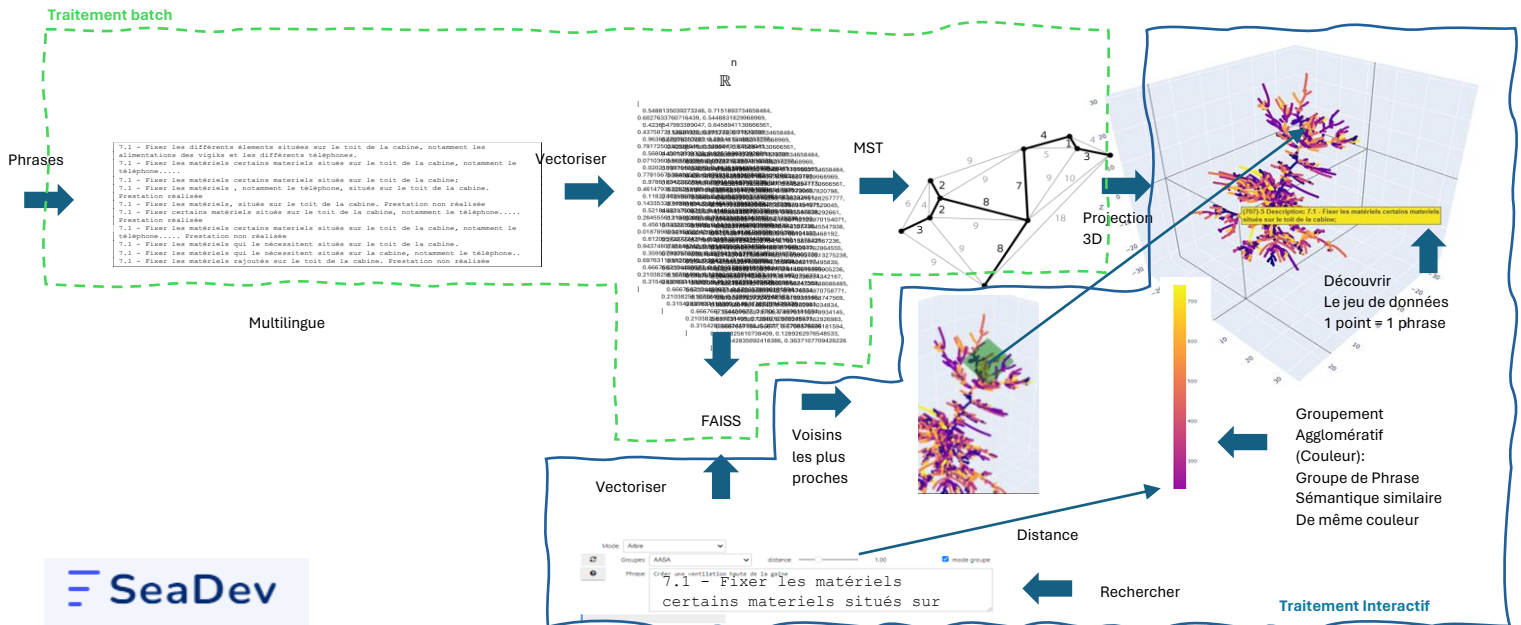
On peut vérifier en utilisant le mode voisin et en regardant les listes qui affiche la distance de recherche.

---

<sup>1</sup> La distance de recherche est calculée par le module FAISS alors que la distance des couleurs est calculée par la fonction AgglomerativeClustering.

# Conclusion

## Architecture de la solution



- Dans la partie du traitement effectué en traitement par lot
  - Les observations sont les phrases utilisées en entrée du système
  - Ces phrases sont vectorisées et les vecteurs sont
    - Soumis à FAISS pour créer une base de vecteurs en vue de la recherche
    - Transformés en MST grâce à la matrice des distances (une matrice de distance, c'est un graphe)
    - Les MST sont transformés en coordonnées 3D (un point par observation) pour affichage ultérieur
- Dans la partie interactive
  - Découvrir les observations
    - L'arbre MST est présenté en 3D interactif avec une observation par point
    - Pour voir le texte de l'observation, il suffit de passer la souris sur le point
    - Quand un point est sélectionné, on obtient la liste du groupe obtenu par agglomération
  - Rechercher une observation
    - On entre un texte pour rechercher les observations sémantiquement similaires
    - Grâce à FAISS, on effectue le calcul de recherche
    - On peut ensuite indiquer la zone correspondante sur le graphe 3D
  - Définition des groupement agglomératifs
    - On fixe une distance
    - On utilise alors un algorithme de calcul de groupes par agglomération qui associe tous les points à moins de cette distance par connexité.

## Que doit-on retenir ?

- La classification et la recherche sémantique de phrases permet de grouper les observations qui ont une même sémantique
- Ces groupes d'observations, si le critère de distance pris pour les définir est correctement positionné, correspondent à des travaux identiques ou de même nature.
- L'outillage mathématique utilisé est assez complexe, mais on a pu partager avec les hommes de terrains les résultats grâce à une représentation qui permet interactivement
  - De définir la « taille » des groupements
  - De vérifier si leur contenu sémantique est effectivement équivalent pour l'homme de métier en produisant la liste des observations sémantiquement similaire
    - Soit à partir d'un point qu'on sélectionne sur le graphique 3D et on produit l'ensemble des observations du groupe par groupement agglomératif.
    - Soit en entrant un texte d'observation et la recherche des voisins les plus proches sur le graphique 3D donne les observations qui sont sémantiquement similaires du texte entré pour la recherche.
- On a donc un outil qui permet de découvrir et de standardiser la liste des tâches correspondants aux observations émises par un bureau de contrôle.