

# DOSSIER DE SPECIFICATION PROJET FIL ROUGE

Encadré par :

Julien PINQUIER

Etudiants du groupe :

BERNY Linda

RESSUGE William

PASSAMA Julien

ARCIN Gautier

MENAA Samir

# Introduction

Dans le cadre du projet fil Rouge que nous aurons à réaliser tout au long de notre première année d'école d'ingénieur, nous avons créé un groupe d'élèves provenant d'univers différents afin que nous ayons une meilleure complémentarité et que nous puissions nous en servir dans ce projet. En effet, parmi les élèves, deux proviennent de DUT Informatique, un de DUT GEII, un a suivi une formation EEA dans un institut étranger et enfin un élève provient de classe préparatoire. La diversité de ce groupe nous permettra d'aborder le projet sous tous ses angles sans omettre d'aspects et, en cas de difficulté, de faire face aux problèmes qui pourraient survenir par le biais des connaissances de chacun dans des domaines précis.

Le projet fil Rouge est un projet pluridisciplinaire composé de 3 parties. Ce document de spécification ne concerne que la première partie intitulée « Indexation de documents et moteur de recherche ». En effet, au vu de l'immense quantité de documents qu'il existe aujourd'hui, de différents types, il est nécessaire de trouver un moyen d'exploiter les informations qu'ils contiennent de façon rapide et efficace : ceci sera donc le but de la première partie de notre projet.

Tout d'abord, nous restituerons le cahier des charges qui a été mis à notre disposition, puis nous aborderons la représentation fonctionnelle complète du système et enfin, nous étudierons les processus de validation à mettre en place.

# Sommaire

## I. Restitution du cahier des charges

- a. Définitions
- b. Fonctionnalités

## II. Représentation fonctionnelle

- a. Diagrammes des cas d'utilisation
- b. Diagramme de séquence

## III. Processus de validation

- a. Contraintes
- b. Cas nominaux et dégradés

## Gestion de projet

# I. Restitution du cahier des charges

## a. Définition

Tout d'abord, nous avons jugé essentiel de définir tous les termes techniques que nous allons utiliser dans ce dossier de spécification afin que nous partions sur les mêmes bases et que nous soyons bien en accord avec notre client.

- Document : fichier support d'information. Dans notre cas, le document peut être de type texte, image, ou audio.
- Descripteur : Mot ou locution contribuant à caractériser l'information contenue dans un document et à en faciliter la recherche.
- Indexation : L'indexation consiste à donner accès aux documents à partir d'une indication concernant leur contenu et/ou leur nature (forme, type).
- Histogramme : L'histogramme est un moyen rapide pour étudier la répartition d'une variable (Par exemple, la couleur d'une image).
- Cas nominal : cela représente le cas où on réalise le processus habituel, avec uniquement des succès, sans être confronté à des problèmes.
- Cas dégradés : cela représente les cas critiques en imaginant toutes les possibilités où les tests unitaires sont susceptibles de ne pas aboutir.
- Proportion de similarité : La similarité repose sur le « calcul » d'une distance entre deux descripteurs. Si la distance est faible c'est-à-dire que les documents sont quasi-similaires. Si la distance est nulle, alors les documents sont identiques.
- Seuil : valeur à laquelle est comparée la distance pour établir la similarité de deux documents.
- Mode administrateur/mode utilisateur : l'administrateur est la personne qui détient les droits d'accès, elle est autorisée à donner des droits à son utilisateur ou bien de bloquer l'accès à certain contenu à l'aide d'un mot de passe.
- Test unitaire : il s'agit d'une procédure permettant de vérifier le bon fonctionnement d'une partie précise d'un programme.
- Structure de données : c'est une façon d'organiser les données pour les traiter plus facilement (exemple : matrice, tableau).
- Paramètre de configuration : modalité à spécifier au démarrage du logiciel.

- Quantification : étapes permettant le traitement informatique de tout type de signal, tels que le son ou l'image.
- Mot-clé : Dans le domaine du référencement dans les moteurs de recherche, le terme « mot-clé » désigne une expression, composée d'un ou de plusieurs mots, qui est tapée par l'internaute dans la zone de recherche du moteur de recherche.

## b. Fonctionnalités

### ❖ Indexation

L'indexation des différents fichiers se fera via la production de descripteurs, uniques à chaque type de document.

Nous distinguons trois différents types de documents : texte, image (noir et blanc/couleur) et audio.

Les processus d'indexation des trois types sont indépendants.

L'indexation est lancée de deux manières différentes :

- Soit en mode administrateur du logiciel, nous pouvons alors « forcer » l'indexation.
- Soit lors du lancement de l'application si les fichiers `base_descripteur_type`, `base_index_type` (ou `liste_base_texte` pour le texte) sont absents.

Si le temps nous le permet, nous pensons ajouter une possibilité afin que le programme vérifie que la base de descripteurs est bien à jour avant de lancer une recherche, et procéder à une ré-indexation dans le cas contraire.

A l'issue de l'indexation, deux à trois fichiers sont créés :

- Un fichier contient tous les descripteurs : `base_descripteur_type`.
- Un fichier contenant la liste des documents traités, avec un lien vers les descripteurs correspondant : `base_index_type`.
- (dans le cas de l'indexation texte uniquement) Un fichier répertoriant des mots-clés et leurs occurrences dans les documents textes traités, afin de faciliter la recherche : `liste_base_texte`.

Chaque descripteur est lié à son fichier, ce qui signifie que l'ajout d'un fichier doit créer un nouveau descripteur lors de l'indexation suivante, mais aussi que la suppression du fichier devra entraîner la suppression du descripteur correspondant.

Sur la première version du logiciel, un changement dans un répertoire indexé (ajout d'un fichier ou suppression d'un fichier existant) entraînera une ré-indexation totale de tous les fichiers. Si le temps nous le permet, nous espérons mettre en place une « indexation partielle », ne supprimant/ajoutant que le/les fichiers concerné(s).

L'indexation est configurable de multiples manières (cf. « Configuration »), qui changent le contenu des descripteurs. Nous forcerons alors une ré-indexation lorsque les paramètres seront modifiés.

Notons que pour les fichiers images et sons, les traitements ne sont pas faits sur les fichiers originaux mais sur des fichiers « bruts » contenant leurs données (.txt pour image, .bin pour audio).

## ❖ Recherche

Nous pourrions effectuer une recherche de deux manières :

- Par « critère de recherche » (1).
- Par similarité par rapport à un document de référence (2).

La recherche par critère de recherche (1) s'effectuera uniquement sur des documents textes.

L'utilisateur fournira alors un mot-clé au programme, qui retourne les fichiers contenant le mot-clé, par ordre décroissant de fréquence d'apparition.

La recherche par similarité (2) calcul une distance entre le descripteur du document de référence et tous les descripteurs correspondants à un fichier de même type. Si la distance est inférieure à un seuil (configurable, voir « Configuration »), le fichier sera affiché dans les résultats de la recherche.

Si la distance est nulle, les fichiers sont considérés comme strictement identiques.

Pour chaque recherche, il sera possible de sélectionner un fichier à ouvrir.

Si le temps nous le permet, nous afficherons le « taux de similarité » des fichiers retournés par rapport au fichier originel.

## ❖ Configuration

La configuration pourra s'effectuer lorsque nous ouvrirons le programme en mode administrateur.

Elle permettra de :

- Configurer le processus de création des différents descripteurs.  
(Exemple : A partir de combien d'occurrences un mot est considéré comme un mot-clé pour un fichier texte, le nombre de bits de quantification pour un fichier image, etc.).
- Définir les différents seuils pour la recherche.

Un fichier de configuration sera fourni avec le programme. Si ce fichier venait à manquer, le programme prendrait alors des valeurs « par défaut ».

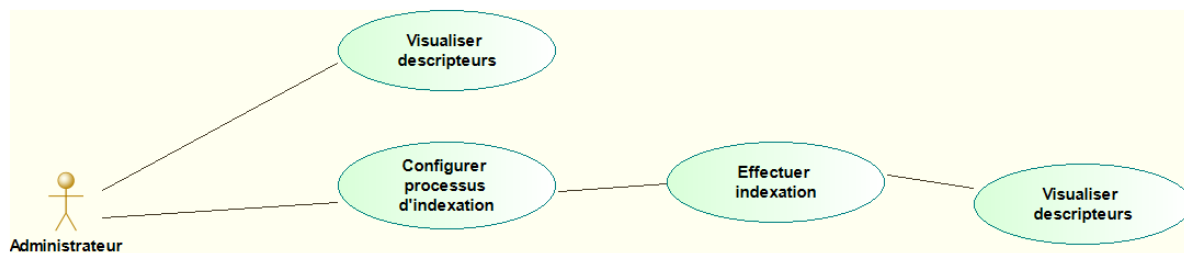
Un changement dans la configuration forcera la ré-indexation de tous les types impactés par la modification de paramètres.

## II. Représentation fonctionnelle

### a. Diagrammes de cas d'utilisation

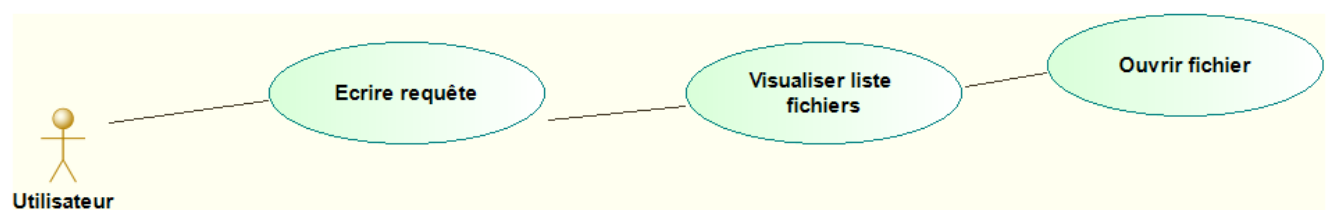
Les diagrammes ci-dessous ont pour but de donner une vision globale des fonctionnalités du logiciel. Ils représentent les interactions possibles entre les utilisateurs et le logiciel en fonction des différents niveaux d'utilisation.

#### Use case administrateur



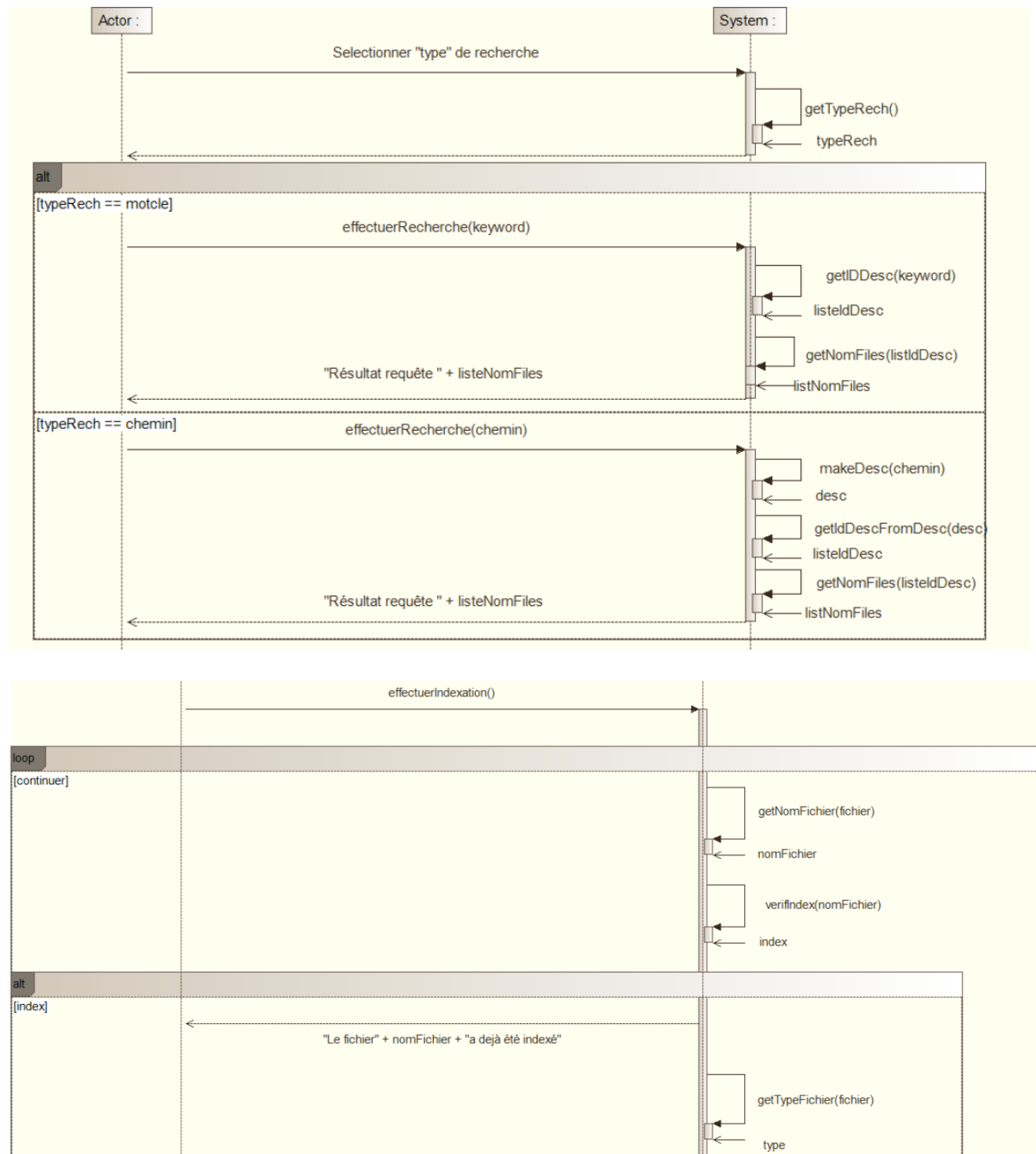
Les différentes utilisations que pourra effectuer l'administrateur du logiciel sont principalement liées au paramétrage. Il pourra **configurer** les différents processus d'indexation en fonctions des besoins utilisateur. Il pourra également **lancer l'indexation** manuellement afin de pouvoir mettre à jour les différentes bases de données. De plus il lui sera possible de **visualiser** tous les descripteurs afin de vérifier que cela correspond à ce qui est attendu. Bien évidemment, l'administrateur pourra effectuer les mêmes tâches que l'utilisateur, i.e. une recherche par mot clé ou par fichier.

#### Use case utilisateur

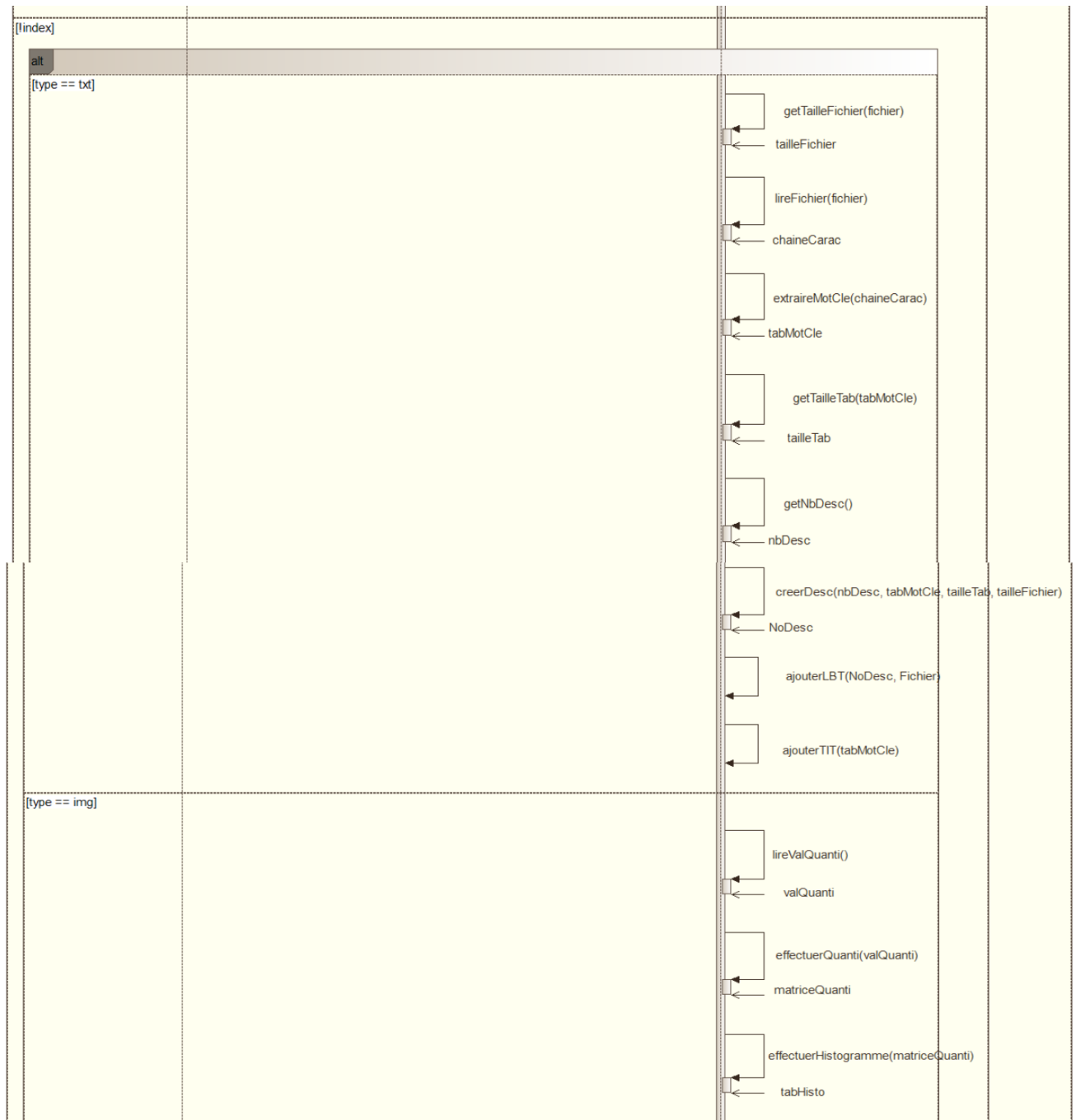


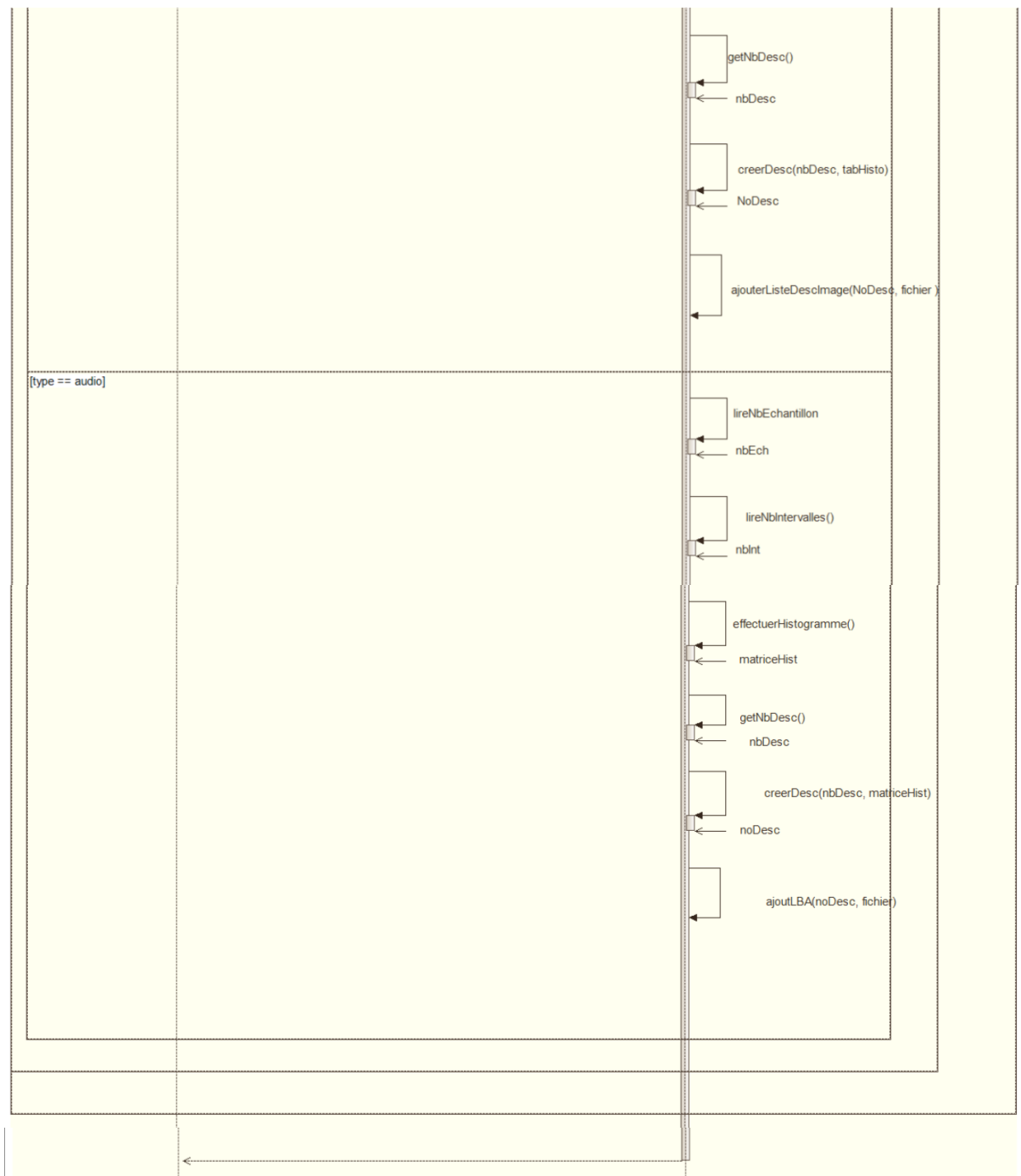
Pour l'utilisateur, il sera possible d'effectuer une recherche en tapant un **mot-clé** ce qui lui affichera la liste des fichiers dont le **mot-clé** apparaît dans les descripteurs ou ressemble fortement. Il lui est également possible de mettre dans sa requête un **chemin d'accès à un fichier**, auquel cas le descripteur de ce fichier sera comparé à ceux existant, les fichiers ayant les descripteurs les plus similaires seront affichés. Il sera également possible d'ouvrir les fichiers qui lui sont affichés.

## b. Diagramme de séquence









Le diagramme de séquence a pour but de modéliser les interactions entre les différents acteurs et le système.

Ici, si un utilisateur souhaite faire une recherche, il doit spécifier si elle se fait par mot-clé ou par fichier. Si elle se fait par fichier, un descripteur correspondant est temporairement créé et il est par la suite comparé aux autres. Ensuite, lorsque l'administrateur lance l'indexation, chaque fichier est indexé (s'il ne l'est pas déjà) en fonction de son type (texte, image, audio).

### III. Processus de validation

#### a. Contraintes

La première contrainte que nous impose le cahier des charges réside dans la composition du groupe. Nous devons être 5 étudiants, ni plus, ni moins. Cela permettra d'être suffisamment nombreux afin d'aller au bout de projet et suffisamment nombreux pour confronter nos différents points de vue.

D'une part, il nous est également imposé le langage informatique. Tout au long du projet, les développements devront être implémentés sous langage C et sur LINUX seulement. Il est recommandé d'utiliser des commandes ou script Unix pour simplifier certains traitements.

D'autre part, nous devons veiller attentivement à la portabilité du code ainsi qu'à l'utilisation de commentaires afin que chacun de nous puisse comprendre le travail de l'autre.

De plus, tous les programmes et fonctions devront être testés dans leurs cas limites ; cela fait l'objet de la partie suivante portant sur les cas nominaux et les cas dégradés.

#### b. Cas nominaux et dégradés

Admin → Visualiser descripteurs② → texte

Admin → Visualiser descripteurs② → image

Admin → Visualiser descripteurs② → audio

Admin → Configurer le processus d'indexation → texte → nombre d'occurrences

Admin → Configurer le processus d'indexation → texte → taille des mots clés

Admin → Configurer le processus d'indexation → image → ?

Admin → Configurer le processus d'indexation → image → ?

Admin → Configurer le processus d'indexation → image → ?

Admin → Configurer le processus d'indexation → audio → ?

Admin → Configurer le processus d'indexation → audio → ?

Admin → Configurer le processus d'indexation → audio → ?

Admin → Lancer l'indexation③

Utilisateur → Recherche avec mot clef → Visualiser les fichiers trouvés par choix④

Utilisateur → Rechercher avec un chemin d'accès à un fichier⑥⑤ → Visualiser les fichiers trouvés par choix④

- ① La saisie est incorrecte (pas un numéro existant dans le menu) : la personne devra à nouveau ressaisir. Ce cas est présent à chaque fois que l'application demande une saisie d'un choix.
- ② Cas dégradé : si le fichier n'existe pas alors affichage du message d'erreur « configurer le fichier .config et/ou lancer l'indexation ».
- ③ S'il n'y a pas de fichier de configuration, il sera créé automatiquement avec des valeurs par défaut.
- ④ Sélection d'un fichier par son numéro : ouverture de celui-ci dans l'application correspondante.
- ⑤ Si le fichier donné n'a pas de descripteur il est alors créé.
- ⑥ Aucun fichier trouvé message d'information.

? Paramètres éventuellement configurables.

### Explications :

Les numéros ci-dessus, représentent tous les cas dégradés qui peuvent se produire lors du lancement ou lors de l'exécution de l'application.

La première partie qui représente l'arborescence des choix possibles se lit de la manière suivante (exemple non contractuel) :

l'utilisateur saisit « 1 » pour le mode administrateur, puis il a 3 choix possibles :

- 1) Visualiser les descripteurs : → **Visualiser descripteurs**
- 2) Configurer le processus d'indexation : → **Configurer le processus d'indexation**
- 3) Lancer l'indexation : → **Lancer l'indexation**

...

Les flèches représentent les relations de précedence entre les différents choix possibles. De ce fait la fin de la ligne représente la profondeur maximale accessible grâce aux choix précédents. Par exemple la ligne suivante : **Admin → Visualiser descripteur② → texte**  
Se lit : *J'ai effectué le choix du mode « Administrateur » puis j'ai effectué le choix de « visualiser les descripteurs » enfin j'ai effectué le choix du média « texte » (descripteurs des documents texte).*

# Gestion de projet

Il est nécessaire que nous nous accordions, suite à ce document de spécification, sur le travail qu'il faudra fournir dans les délais qui suivent. Nous allons entamer une phase de conception, dans laquelle nous allons concevoir l'application.

Il est assez difficile pour nous à ce stade de prévoir la répartition exacte des tâches qui vont suivre, néanmoins, en s'appuyant sur ce que nous avons prévu de faire, énoncé dans le présent dossier de spécification, nous pensons attribuer à trois élèves la phase d'indexation, chacun se chargera d'un type de fichier (texte, image, audio). Les deux autres étudiants se chargeront de la recherche et de la partie interface. Nous n'avons pas encore réparti les tâches.

Outils utilisés : Plusieurs outils seront utilisés dans le cadre de ce projet.

*Git* permettra la mise en commun et la mise à jour du code.

Un groupe de travail *messenger* a d'ores et déjà été créé pour faciliter les échanges dans le groupe, et les adresses e-mails constitutionnelles seront utilisées pour échanger avec les clients.

Le planning sera disponible en ligne sur *Trello*, et sera mis à jour en fonction de l'avancée du projet (le planning n'est pas disponible à ce jour, mais la création du compte commun a été effectuée).

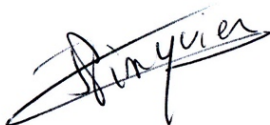
Planning prévisionnel : Nous prévoyons une réunion vendredi 24 novembre à 15h30 pour répartir les tâches de la conception.

Nous rappelons que la présentation finale de la première partie du projet est le 18 janvier.

Nous fixons donc la fin de l'élaboration de la conception au 22 décembre et l'élaboration d'une première version prototype au 4 janvier.

## **Accord du maître d'ouvrage**

Bon pour accord le 25 novembre 2017

A handwritten signature in black ink, appearing to read 'D. Guyer', written over a horizontal line.