Dossier de Spécifications

Projet Fil Rouge

Conception d'un moteur de recherche

SeekFox

UPSSITECH SRI 1A

2019-2020

Gaël Gamba – Etienne Combelles

Raphaël Bizet – Oualid El Abdaoui – Clément Truillet

Client : Isabelle Ferrané

Dossier de Spécifications 1

I – Introduction

I.1 Objet du document

L'objet de ce document est de définir les spécifications fonctionnelles détaillées de la partie 1 du Projet Fil Rouge proposé à la promotion de 1ere année de la spécialité Systèmes Robotiques et Intelligents de l'UPSSITECH.
Les spécifications fonctionnelles détaillées ont pour but de décrire :

- l'ensemble des fonctionnalités de l'application;
- les données nécessaires à l'application;
- le fonctionnement interne de l'application.

II - Cadre

II.1 Contexte & Objectif

L'accès à l'information est un enjeu essentiel. Avec internet et les ordinateurs, de grandes masses de données sont accessibles à tout le monde. Pour exploiter ces masses de données, contenant des fichiers de diverse nature (fichiers textes, audio, images), des outils adaptés comme des moteurs de recherche sont indispensables. Notre objectif consiste en la création d'un tel outil, utilisable par tous, gérant à la fois l'indexation (création de descripteurs de documents utilisables par le moteur de recherche), la comparaison (des descripteurs des documents entre eux, permettant de déterminer la ressemblance entre les documents) et la recherche de documents.

II.2 Résumé

Notre moteur de recherche aura donc plusieurs tâches à remplir : indexation, comparaison, recherche, pour n'en citer que les principales. Il est à noter que seule la recherche sera disponible pour un utilisateur, le reste n'étant accessible qu'à l'administrateur.

Les documents traités étant de nature diverse (fichiers textes, audio, images) de formats prédéfinis (*XML*, *TXT*, *JPEG*, *WAV*…), nous devrons d'abord faire une description synthétique de chacun qui sera utilisée pour son traitement par le moteur de recherche : il s'agit là de la première fonctionnalité de celui-ci, l'indexation automatique. Ces descripteurs comprendront les informations essentielles du document, chaque document ne pouvant avoir qu'un seul descripteur dans la base de ceux-ci.

Le moteur de recherche utilisera ces descripteurs pour comparer les documents, afin de définir les similarités entre différents documents. Cette similarité se base sur le nombre de points communs entre deux descripteurs : si ces descripteurs sont identiques, alors les documents sont les mêmes, sinon le nombre de caractéristiques en commun définit un taux de similarité qui permettra de fixer un seuil à partir duquel on considérera que deux documents sont proches.

La fonction de recherche du moteur sera le résultat visible et utilisable des fonctionnalités décrites précédemment. Dans cette première phase du projet, l'interface utilisateur consistera en une entrée clavier et un affichage à l'écran des résultats ordonnés de la recherche.

La version graphique de l'interface sera alors développée dans la seconde partie du projet.

Le moteur de recherche aura deux modes d'accès, un mode administrateur qui permettra de configurer et de lancer l'indexation et de visualiser les descripteurs, et un mode utilisateur qui sera l'aspect tout public du moteur, à savoir la recherche de documents dans la base de documents depuis l'interface. La recherche fonctionnera en utilisant les fichiers de descripteurs créés à la phase d'indexation, et leurs comparaisons, et affichera une liste de documents répondant à la requête ordonnés selon leur taux de similarité.

II.3 Livrables

En plus du présent dossier de spécifications, le produit final de cette première partie sera fourni sous forme d'archive compressée, comprenant tout le code nécessaire au fonctionnement du moteur de recherche, ainsi qu'un mode d'emploi pour générer un exécutable (à partir d'un fichier Makefile) et un manuel d'utilisation du moteur de recherche.
Sera aussi fourni un rapport détaillant tout le processus de création de l'outil ainsi que son fonctionnement.

De plus, les fonctionnalités du produit final seront présentées lors d'un entretien à la date convenue du 24 janvier 2020.

II.4 Planning

- Remise du dossier de spécifications : 13/11/2019
- Conception et développement du moteur de recherche : entre le 13/11/2019 et le 12/01/2020
- Tests et intégration : du 12/01/2020 au 20/01/2020
- Livraison du moteur de recherche : 24/01/2020
- Remise du rapport : 31/01/2020

Vous pourrez suivre l'avancée de notre travail ici :
https://trello.com/b/efDdpiIV/projet-fil-rouge-1a-sri-seekfox .

Afin d'avancer de manière efficace, nous avons décidé de prévoir les tâches et d'en estimer les durées de réalisation.
De manière graphique, ce planning se présente par un diagramme de Gantt (cf. figure 1).

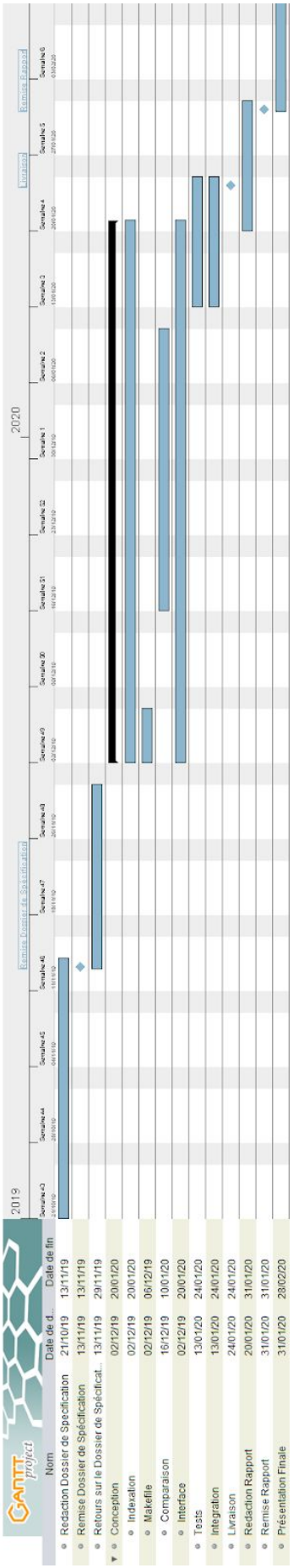

Figure 1 : Diagramme de Gantt

II.5 Présentation de l'équipe

Afin de mener à bien ce projet, notre équipe est constituée de 5 personnes réunies sous une équipe nommée SeekFox.

- Clément TRUILLET
- Etienne COMBELLES
- Gaël GAMBA
- Oualid EL ABDAOUI
- Raphaël BIZET

III - Spécifications Fonctionnelles

III.1 Fonctionnement Général

Le programme, un moteur de recherche, retourne les *n* fichiers les plus proches d'un fichier (texte, audio ou image) ou d'un mot-clé entré par l'utilisateur. Cette valeur *n*, fixée à 10 par défaut, pourra être modifiée via le mode d'accès *administrateur*.

Pour cela, il disposera d'un répertoire contenant les descripteurs[1] de chacun des fichiers présents dans le répertoire de recherche, ceux-ci étant créés par le programme lors de l'indexation des fichiers du répertoire.

Deux types d'utilisateurs pourront utiliser ce programme, via deux modes d'utilisation, à savoir le mode utilisateur et le mode administrateur[2]. Les deux auront accès à la recherche classique mais l'administrateur pourra faire des choses supplémentaires (voir la *Figure 3).*

[1] Descripteur : ligne de texte associée à un fichier (texte, image ou audio) et contenant les informations essentielles à la comparaison de ce fichier avec d'autres fichiers du même type. Un descripteur est stocké dans un fichier de descripteurs avec tous les autres descripteurs correspondant à un type de fichier (texte, audio ou image)

[2] Administrateur : compte utilisateur protégé avec un login et un mot de passe et ayant accès à des fonctionnalités supplémentaires (menu de configuration, lancer l'indexation et visualiser des descripteurs)

<table>
<tr><th></th><th>Actions disponibles</th><th>Données recherchables</th><th>Type de recherche</th></tr>
<tr><td rowspan="4">Mode Utilisateur</td><td rowspan="5">Rechercher</td><td rowspan="2">Texte</td><td>Par mot-clé</td></tr>
<tr><td>Par Fichier requête</td></tr>
<tr><td>Audio</td><td>Par Fichier requête</td></tr>
<tr><td rowspan="2">Image</td><td>Par Fichier requête</td></tr>
<tr><td rowspan="4">Mode Administrateur</td><td>Par Couleur Dominante</td></tr>
<tr><td>Lancer l’indexation manuellement</td><td colspan="2" rowspan="3"></td></tr>
<tr><td>Visualiser un descripteur</td></tr>
<tr><td>Configurer le programme</td></tr>
</table>

Figure 3 : Visualisation des choix possibles suivant le type d'utilisateur

Le menu de configuration permettra d’afficher la liste de réglages du moteur de recherche qu’il sera possible de modifier, comprenant :

- Le chemin du répertoire de la base de documents
- Le chemin du répertoire de la base de descripteurs
- Le nombre de fichiers affichés par page de recherche
- Le pourcentage de ressemblance minimum requis pour qu’un fichier soit affiché

Dans ce menu seront aussi présentes deux actions disponibles uniquement pour l’administrateur :

- Lancer une indexation manuelle
- Afficher la liste des descripteurs ainsi que leur contenu

Les changements seront enregistrés dans un fichier ce qui permettra de les conserver à chaque redémarrage du programme.

Le scénario ci-dessous, illustré par la *figure 4*, donne le schéma de fonctionnement du moteur de recherche.

Scénario :

1. Au lancement du programme un premier menu demande à l'utilisateur de choisir entre le mode administrateur ou le mode utilisateur de base
2. Si le mode choisi est le mode utilisateur normal
    - Le programme propose à l'utilisateur via un menu de choisir entre la recherche par mot clé, par couleur dominante, ou par fichier requête.
        - Si l'utilisateur choisit la recherche par mot clé, il peut alors écrire le mot à rechercher.
            1. Si le mot n'est pas valide (caractère non alphanumérique), un message s'affiche indiquant à l'utilisateur que sa requête est invalide. Cela peut se produire 3 fois, après quoi il sera renvoyé au point 2
        - Si l'utilisateur choisit la recherche par couleur dominante, le programme lui propose un set de couleurs à choisir.
            1. Si la couleur choisie n'est pas valide (réponse ne correspondant à aucune proposition), un message s'affiche indiquant à l'utilisateur que sa requête est invalide. Cela peut se produire 3 fois, après quoi il sera renvoyé au point 2
        - Si l'utilisateur choisit un fichier parmi les fichiers requête, la recherche est effectuée.
            1. Si le fichier choisi n'est pas valide (inexistant, format non reconnu), un message s'affiche indiquant à l'utilisateur que sa requête est invalide. Cela peut se produire 3 fois, après quoi il sera renvoyé au point 2
    - Une fois la recherche configurée, le programme va chercher dans le corpus de fichiers ceux correspondant le plus aux critères demandés. Il les affichera sous forme de pages de réponses comprenant un nombre fixé par l'administrateur de fichiers par page, et rangés dans un ordre décroissant de correspondance. Le fichier le plus ressemblant s'ouvrira.
    - Un menu va alors apparaître dans lequel l'utilisateur peut:
        - Relancer une recherche (Retour au point 2)
        - Afficher d'autres fichiers (Afficher une nouvelle page de fichiers s'il y en a d'autres qui ont été trouvés)
        - Quitter le programme

3. Si le mode choisi est le mode administrateur
    - L'utilisateur saisi les identifiants de l'administrateur
        - S'ils sont incorrect, l'utilisateur peut les retaper jusqu'a 3 fois. Au bout de la 3ème fois le programme repart à l'étape 1.
    - L'utilisateur choisit une action à accomplir parmi une liste comprise dans un menu
        - Si l'utilisateur choisit de changer les paramètres du moteur de recherche, un menu s'affiche lui permettant de le configurer:

changer les processus d'indexation, avec une option lui permettant de remonter au menu précédent

1. Si l'utilisateur entre des configurations qui ne sont pas valides (par exemple s'il écrit qu'il veut afficher -1 fichiers après une recherche), le programme réitère sa requête d'information un maximum de 3 fois avant de retourner au point 3.

- Si l'utilisateur choisit de lancer l'indexation manuellement, l'indexation se lance et le menu d'actions se réaffiche
- Si l'utilisateur choisit d'afficher un descripteur, il choisi ensuite le descripteur à afficher, et l'affiche.
    1. Si son choix n'est pas valide (descripteur inexistant par exemple) il peut le réitérer un total de 3 fois avant de retourner au point 3.
- Si l'utilisateur choisit de faire une recherche, il se passe exactement la même chose pour une recherche par l'utilisateur classique

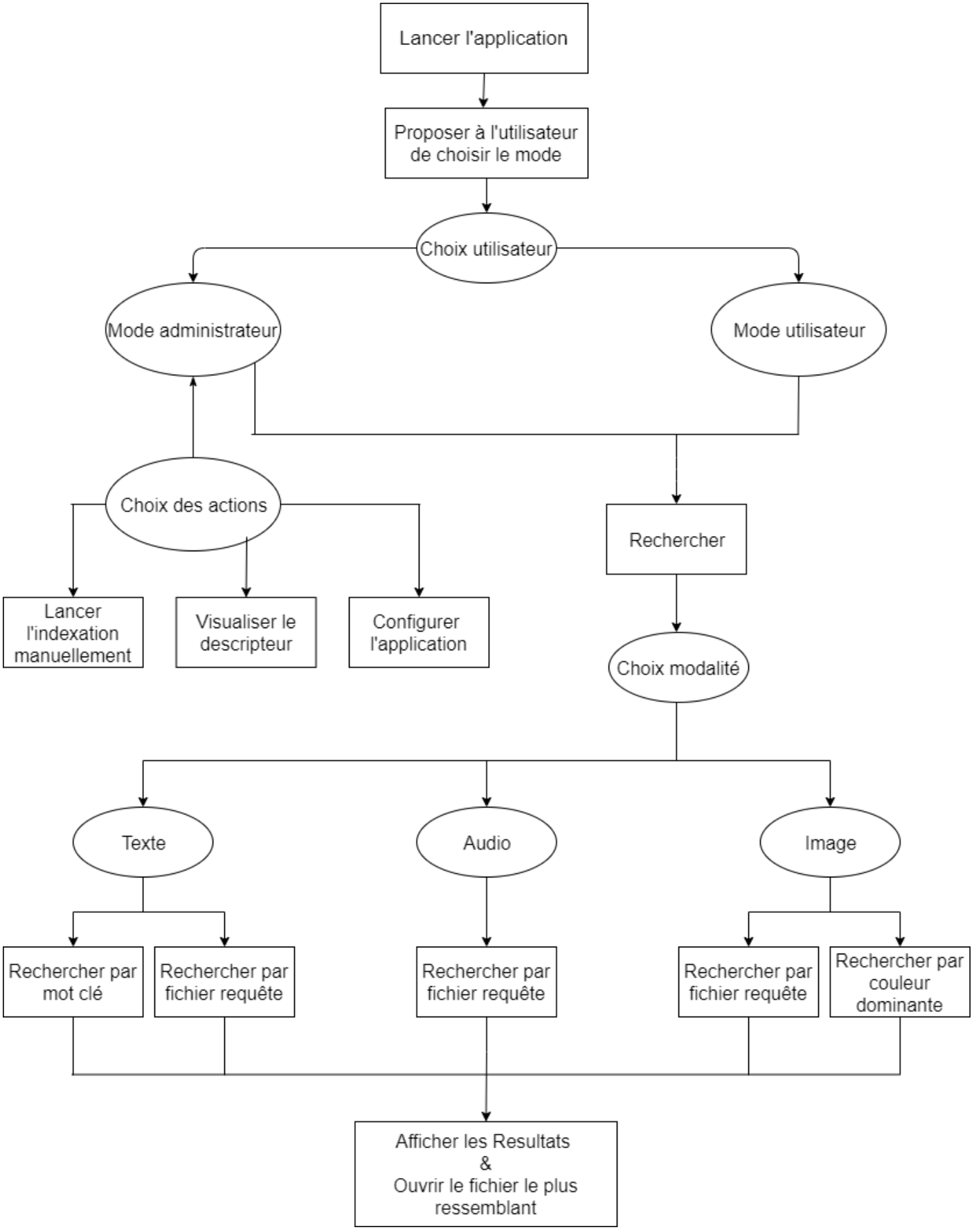

Figure 4: Schéma de fonctionnement

III.2 Indexation

L'indexation consiste en la création des descripteurs pour pouvoir représenter les fichiers de manière plus synthétique, afin qu'ils soient utilisable par le moteur de recherche.

Ces descripteurs seront sous forme de texte et pourront représenter des fichiers de différents types : texte, image, audio; le processus et les critères de création sont détaillés plus bas.

L'indexation aura deux modes de lancement, un automatique et un manuel.

III.2.1 Lancement Manuel

Le lancement manuel de l'indexation fera partie des options disponibles pour l'administrateur: il sera accessible depuis son menu, où il sera présenté sous la forme d'une option "mettre à jour la base de descripteurs".

Cette indexation commencera par vérifier si une base de descripteurs est déjà présente. Si c'est le cas, l'indexeur continuera en vérifiant quels descripteurs sont "orphelins", c'est-à-dire quand leur fichier d'origine n'existe plus ou a été modifié ; ces descripteurs seront supprimés.

Ensuite, il passera en revue l'ensemble des fichiers présents dans la base de documents, et créera des descripteurs pour ceux qui n'en ont pas déjà.
Dans le cas où il n'y a pas de base de descripteurs, l'indexeur va simplement en créer un avec tous les fichiers présents dans la base de documents.
Une fois ces opérations faites, le moteur renverra le message "Base de descripteurs à jour".

Si l'indexeur a rencontré un obstacle l'empêchant de correctement s'exécuter, il renverra un message d'erreur décrivant au mieux l'erreur rencontrée, et proposant une solution pour la réparer dans les cas les plus courants (par exemple lorsque la base de documents comprend des fichiers dans des formats non pris en compte).

III.2.2 Lancement Automatique

Le mode d'indexation automatique s'exécutera à chaque lancement du moteur de recherche en vérifiant la présence d'une base de descripteurs. Pour ne pas ralentir le lancement du moteur de recherche à chaque fois, il ne fera l'indexation que s'il voit qu'il n'y a pas d'index de descripteurs (donc au premier lancement, ou si l'index a été supprimé manuellement).

En cas d'indexation, elle fonctionnera comme pour le mode manuel dans le cas où il n'y a pas d'index de descripteurs, renvoyant les mêmes messages de réussite ou d'erreur. Lorsqu'il y a déjà un index de descripteurs, l'indexation ne se fera tout simplement pas, elle devra être faite manuellement par l'administrateur si la base de données est modifiée.

III.3 Recherche

Chaque recherche par l'utilisateur ou l'administrateur se déroule selon un scénario global décrit ci-dessous :

Scénario :

1. L'utilisateur saisit le chemin d'accès au fichier ou choisit le fichier (stocké dans le répertoire requête) dans une liste affichée. (Note : Une recherche de texte par mot-clé ou d'image par couleur dominante est aussi possible.)
2. Le système crée un descripteur du fichier requête et effectue une comparaison avec les descripteurs du même type déjà indexés. Dans le cas d'une recherche par mot-clef ou par couleur, la recherche s'effectue directement dans la base de descripteurs sans en créer un.
3. Le système affiche la liste des fichiers les plus proches du fichier requête (ou dont les mot-clefs ou la couleur correspondent le mieux à ceux rentrés) et ouvre le fichier le plus pertinent avec un éditeur de texte, lecteur d'image ou audio suivant le type du fichier.

Exceptions :

1. Le fichier d'entrée est inexploitable.
    - Le fichier n'est pas un fichier texte, image ou audio
    - Le fichier n'est pas lisible par le système
    - Le fichier est dans un format incompatible avec le système
    - Le chemin vers le fichier est non valide
2. Il n'y a aucun fichier indexé.

Pour chaque exception, un message d'erreur la décrivant sera affiché, permettant à l'utilisateur de savoir quoi changer à sa requête pour qu'elle fonctionne.

Il est important de noter que lors d'une recherche à partir d'un fichier, les fichiers affichés le seront selon un taux de similarité défini par la comparaison (voir chapitre suivant).

Le seuil d'affichage (soit le taux minimal à partir duquel on considère deux fichiers comme similaires) ainsi que le nombre de résultats fera partie des paramètres configurables en mode administrateur.

Le type de recherche le plus commun est la recherche par fichier, celle-ci a pour but de trouver des fichiers ressemblant le plus possible au fichier que l'on fournit et le programme affichera tous les fichiers qui correspondent à la recherche et ouvrira le fichier le plus ressemblant. La manière dont nous comparerons ces fichiers est expliquée plus bas.

Cependant, chaque recherche d'un fichier de format différent a ses spécificités, elles sont décrites dans les sous-parties suivantes.

III.3.1 Recherche d'un texte

En plus de la recherche par fichier, nous pouvons également faire une recherche par mot-clé, dont le but est de trouver les descripteurs de fichiers ayant le plus d'occurrence du mot-clé en question, cela va se faire à partir du descripteur généré pour chaque fichier.

III.3.2 Recherche d'image

Il y a également une deuxième option pour la recherche d'image, il s'agit de la recherche par couleur dominante ayant pour but d'afficher les fichiers contenant le plus la couleur recherchée.

III.3.3 Recherche de son

La recherche audio est orientée sur la recherche de l'occurrence d'un jingle audio (ou motif audio) dans les fichiers sons. Ici, l'utilisateur va nous fournir un fichier sonore court et le programme va afficher les fichiers sonores contenant ce son, donc ce motif, ainsi que le time code de l'apparition dudit son.

III.4 Comparaison

La comparaison entre deux fichiers est traitée de façon interne par l'application, sans que l'utilisateur n'ait à intervenir. Pour n'importe quelle comparaison, le scénario global est le même :

Scénario :

1. Le système reçoit un fichier requête.
2. Le système génère le descripteur associé.
3. Le système compare tour à tour ce descripteur avec ceux déjà stockés en prenant en compte le type de comparaison (ex:par mot clé).
4. Le système renvoie la valeur correspondante à un score de similarité ou une distance

Exceptions :

1. Le fichier d'entrée est inexploitable, pour les mêmes raisons que précédemment.
2. Il n'y a aucun fichier du même type indexé.

III.4.1 Comparaison de son

La comparaison entre fichiers audio est un peu différente étant donné qu'on ne cherche pas à comparer la ressemblance entre 2 descripteurs mais à savoir si l'un des descripteurs (qui est une représentation du fichier) est compris dans l'autre. La comparaison va donc retourner plus d'informations que pour les autres, à savoir le nombre d'occurrences du jingle ainsi que le time code lié à ces occurrences.

III.5 Interface utilisateur

L'interface sera pour l'instant seulement composée du terminal, toutes les interactions se feront à l'aide de messages textuels.

Les utilisateurs devront faire des choix, comme par exemple quel type de recherche ils veulent faire.

Pour cela le programme affichera des chiffres devant chaque choix et l'utilisateur devra écrire le chiffre correspondant à l'action voulue.
Cela nous permettra en plus de tester la validité des entrées de l'utilisateur afin qu'il n'écrive pas des caractères indésirables ou hors sujets.

Dans le cas où l'utilisateur ferait trop d'erreurs lors de cette saisie de caractère, le programme retournerait automatiquement au menu précédent, ou au premier menu affichable le cas échéant.

IV – Annexes

IV.1 Tâches

Dans une envie de réaliser ce projet de manière efficace, il est cohérent que chacun des membres de l'équipe SeekFox se répartissent les tâches de manière équitable.
L'équipe étant composée de 5 membres, nous avons décomposé les tâches en 5 grandes parties.

Fonction	Fonctionnalité	Responsable
Texte Intégration globale	Comparaison et génération d'un descripteur de texte Intégration des différentes parties	Raphaël Bizet
Image	Comparaison et génération d'un descripteur d'image	El Abdaoui Oualid
Audio	Comparaison et génération d'un descripteur de fichier audio	Gaël Gamba
Indexation	Génération de la base de descripteurs	Etienne Combelles
Interface et Recherche	Mode administrateur Gestion du fichier de configuration Interface et interaction utilisateur Gestion de la mise à jour de la base de documents	Clément Truillet

IV.2 Scénarios de test

Afin de produire un moteur de recherche fonctionnel, il est primordial d'effectuer des tests.
Ces tests peuvent être unitaires (sur une fonction seulement), ou sur une fonctionnalité, nous allons alors utiliser ces scénarios de tests.

Tests de l'interface
Pour chaque menu ou chaque aspect de l'interface, des tests unitaires seront fait au fur et à mesure du développement du moteur de recherche pour vérifier leur bon fonctionnement. Le test d'un menu lambda suivra le modèle suivant :

1. Test de l'erreur en entrant un caractère qui ne correspond pas aux choix possibles
2. Test de chaque option du menu

Tests en mode administrateur uniquement
Test de connexion :
Par connexion on entend l'entrée par l'administrateur du mot de passe pour pouvoir accéder à ses menus.

1. Test avec un mot de passe erroné
2. Test avec le bon mot de passe

Test du menu des paramètres
Le scénario suivant sera appliqué pour chaque paramètre du menu :

1. Changement du paramètre
2. Ouverture manuelle pour constater la prise en compte de ces changements
3. Lancement de quelques recherches pour montrer que le moteur utilise correctement ce fichier de configuration

Indexation

1. Ajouter un nouveau fichier (texte, image ou audio) dans le dossier des documents
2. Lancer le moteur de recherche ou l'indexation
3. Visualiser le nouveau descripteur

1. Supprimer le descripteur d'un fichier
2. Visualiser le nouveau descripteur

Tests de recherche (mode utilisateur ou administrateur)
Pour des tests optimaux, chaque série de test sera effectuée pour chaque option de recherche (donc pour chacun des types de fichier et modes de recherche).

1. Lancer le moteur de recherche sur une base de descripteurs vide
2. Effectuer une recherche à partir d'un fichier déjà indexé
3. Constater que le fichier le plus pertinent est bien le fichier entré

1. Lancer le moteur de recherche
2. Effectuer une recherche à partir d'un fichier non indexé
3. Constater la validité de la recherche (test subjectif)

1. Lancer le moteur de recherche
2. Effectuer une recherche à partir d'un fichier non pris en charge[3]
3. Constater l'affichage d'une erreur

1. Lancer le moteur de recherche
2. Effectuer une recherche en entrant un chemin invalide
3. Constater l'affichage d'une erreur

.

[3] non pris en charge : le type du fichier n'apparaît pas dans le cahier des charges et implique sa non-indexation par le moteur de recherche.