École Polytechnique de l'Université de Tours
64, Avenue Jean Portalis
37200 TOURS, FRANCE
Tél. +33 (0)2 47 36 14 14
www.polytech.univ-tours.fr

Département Informatique

CAHIER DE SPECIFICATION & PLAN DE DEVELOPPEMENT

Projet :	Clustering interactif d'éléments de contenu (extraits de documents)	
Emetteur :	Jing CHEN	Coordonnées : EPU-DI
Date d'émission :	09/01/2014	

Validation

Nom	Date	Valide (O/N)	Commentaires
Jean-Yves Ramel		N	
Jean-Yves Ramel		N	
Jean-Yves Ramel			

Historique des modifications

Version	Date	Description de la modification
00	21/12/2014	Version initiale : synthèse de différents documents
01	28/12/2014	Version corrigée : correction des diagrammes et la partie de fonctionnalités
02	08/01/2015	Version finale :

TABLE DES MATIERES

CAHIER DE SPECIFICATION SYSTEME

1. Introduction

Ce document décrit les spécifications fonctionnelles et techniques nécessaires à la réalisation pour le Projet de fin d'études « Clustering interactif d'éléments de contenu (extraits de documents) » et à la mise en œuvre de logiciel Retro 2014. Il sert ensuite à organiser des tâches tout au long du projet comme un outil fondamental de communication.

Ce projet fait partie du projet PARADIIT (Pattern Redundancy Analysis for Document Image Indexation & Transcription) mis en place par l'équipe des Bibliothèques Virtuelles Humanistes (BVH) du Centre d'Études Supérieures de la Renaissance (CESR) et l'équipe Reconnaissance des Formes et d'Analyse d'Images (RFAI) du Laboratoire Informatique (LI) de l'université François Rabelais. PARADIIT a pour objectif la transcription de livres anciens en bibliothèques numériques.

Ce projet vise à poursuivre la conception et le développement des outils d'analyse de données et de classification permettant de comparer ces éléments les uns avec les autres pour regrouper les éléments , en clusters, puis pour étiquetés automatiquement(transcription).

2. Contexte de la réalisation

2.1. Contexte

Tous document est généralement constitué de motifs ou éléments de contenu répétitifs qu'il peut être intéressant d'exploiter pour mettre en place des mécanismes d'indexation et de recherche permettant de retrouver rapidement une informations spécifiques dans de grandes masses de documents (CF moteur de recherche GOOGLE).

Ce type de mécanismes est assez simple à mettre en place lorsque les documents sont sous forme textuelle (Pages WEB, ASCII) mais devient plus complexe lorsque l'on désire travailler sur des versions numérisées (images).

2.2. Objectifs

L'objectif essentiel du projet est l'amélioration la qualité de Clustering et la réalisation de transcription automatique.

Les principales tâches qui devront être effectuées durant ce PFE concernent :

- Etude de l'existant (architecture, dll et plug-ins existants)
- Etude des méthodes de Clustering et fouille visuelle de données utilisables pour comparer et caractériser les éléments sélectionnés
- Conception et développement des IHM et méthodes interactives permettant la visualisation et l'optimisation des clusters d'éléments de contenu
- Conception et développement d'une méthode de reconnaissance automatiquement de clusters
- Développement et test du système
- Parallèlement, rédaction de cahiers et rapports : document de conception et spécification, cahier de recettes et validation, rapport technique.

2.3. Bases méthodologiques

Afin de mener à bien ce projet, voici les outils qui seront utilisés :

- Logiciel de dessiner des interfaces graphiques : Balsamiq Mockups (Site web Balsamiq Mockups, s.d.)

- Générateur de documentation : SandCastle (documentation compilation pour gérer les bibliothèques de classes) (Site web Sandcastle, s.d.)
- IDE : Visual Studio 2012
- Langage de programmation : C#
- Design Patter : MVVM (Model – View – View Model)
- Framework : .Net 4.5
- Bibliothèques :
    - AForge : bibliothèque de traitement d'image et d'intelligence artificielle.
    - AvalonDock : un contrôleur de l'accueil fenêtre pour WPF
    - Accord.Net : est une extension de AForge.NET, un framework C # populaire pour vision par ordinateur et l'apprentissage. Il offre de nombreuses fonctions d'analyse et de traitement statistique (Site web Accord.Net, s.d.)
- Outil de gestion de projet :
    - Redmaine : gestionnaire de différentes version de source code du projet et gérer le processus de projet PFE.
    - TortoiseSVN : un logiciel de gestion de versions sous forme de Plugin pour Microsoft Windows

3. Description générale

3.1. Environnement du projet

3.1.1. PARADIIT

Ce projet s'inscrit dans le cadre du projet PAEADIIT (Pattern Redundancy Analysis for Document Image Indexation & Transcription) dont les membres sont :

- L'équipe des Bibliothèques Virtuelles Humanistes (BVH) du Centre d' Études Supérieures de la Renaissance (CESR).
- L'équipe Reconnaissance des Formes et Analyse d'Images (RFAI) du Laboratoire Informatique (LI) de l'université François Rabelais de Tours.

Le projet PARADIIT (Site Web PaRADIIT Projet, s.d.)est une solution de transcription de documents anciens, formée de plusieurs applications développées par l'équipe RFAI. Il permet de convertir des livres anciens en ressources numériques accessibles à tous. Il permet de traiter des images (livres ou documents scannés) comme étant des ensembles de caractères et de trouver la correspondance ASCII de ces caractères.

Les documents à numériser font partie de la bibliothèque du CESR et date set de la Renaissance jusqu'au début du XVIIe siècle. De ce fait, ils présentent des particularités telles que la présence de bruit dans l'image ou encore la présence de caractères anciens propres à l'époque de la Renaissance.

Depuis sa création, en 2004, PARADIIT a été sponsorisé par deux Google Digital Humanities Awards.

Figure 1: Logo de CESR, RFAI, LI, Avards Google

3.1.2. Retro Software

Les deux principales applications de PARADIIT permettant la transcription des documents sont : Agora et Retro.

- Agora : un outil d'analyse d'image qui permet, dans un premier temps, d'extraire les éléments contenus dans des images documents numérisés.
- Retro : réalisé par M. Jean-Yves Ramel, a pour but de visualiser et d'exploiter les résultats de segmentation fournis par AGORA.
- Un autre projet a également été développé dans le cadre d'un PFE pour l'année scolaire 2013/2014 par Samantha GEORGES : le projet PFE « Intégration et interfaçage de logiciels de clustering et de transcription ». L'application développée lors de ce PFE avait pour but d'intégrer deux logiciels de clustering dans RETRO.

Le processus d'extraire caractères est présenté en dessous. Les fonctionnalités à développer sont en rouges.

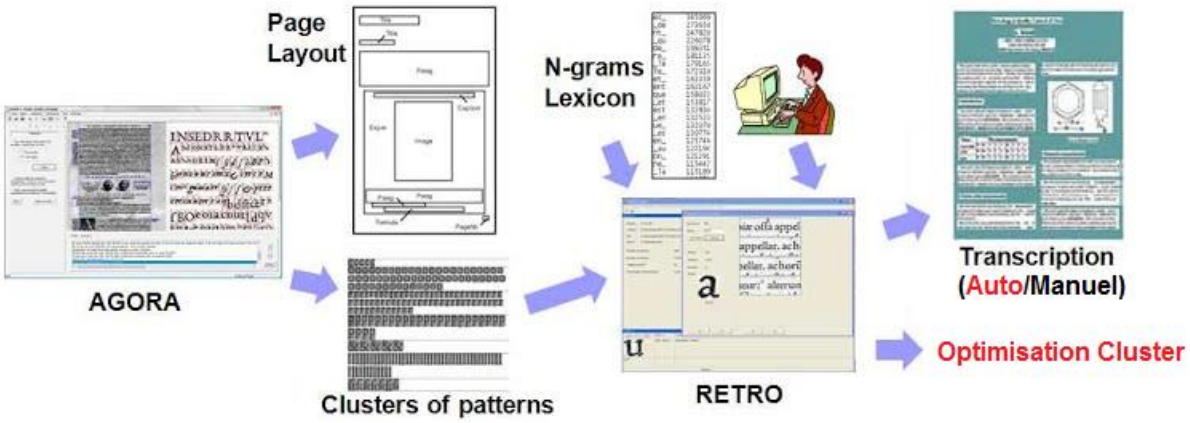


Figure 2: Processus d'extraire caractères

3.2. Caractéristiques des utilisateurs

Différents types d'utilisateurs pourront être amenés à utiliser l'application développée au cours de ce PFE. Parmi ces types, on distingue: les professionnels, les personnes du CESR et du LI.

I. Les professionnels

Il est compris toute personne ayant déjà travaillé sur du Clustering de manière professionnelle ou encore les personnes travaillant dans le domaine de la typographie ou des livres anciens.

Caractéristiques du « Professionnels » :

- Connaissance de l'informatique : Pas forcément
- Expérience de l'application : Non
- Fréquence d'utilisation : Utilisateurs réguliers ou occasionnels
- Droits d'accès utilisateurs : Tous les doits

II. les personnes du CESR et du LI

Les personnels du CESR et du LI à l'origine du projet PARADIIT peuvent être amenés à utiliser la nouvelle version du logiciel Retro qui sera créée lors de ce PFE.

Caractéristiques du « les personnes du CESR et du LI » :

- Connaissance de l'informatique : Oui
- Expérience de l'application : Oui
- Fréquence d'utilisation : Utilisateurs réguliers
- Droits d'accès utilisateurs : Tous les doits

3.3. Fonctionnalités et structure générale du système

L'objetif de mon PFE porte surtout sur les fonctionnalités du clustering et transcription automatique.

La description des fonctionnalités concernant ce projet sont présentés par un diagramme d'activités. Le cas d'activités ci-contre montre des scénarios principaux dans Retro2014 : les manipulations de projet Retro, le processus du Clustering, les manipulations de Cluster, les processus de la transcription. Des activités à développer sont en rouge.

Par rapport à la structure générale du système, je présente seulement des components concernant ce projet. Le diagramme de composants représente des interfaces requis et offerte du component. L'architecture de Retro2014 déjà existant présente en dessus :

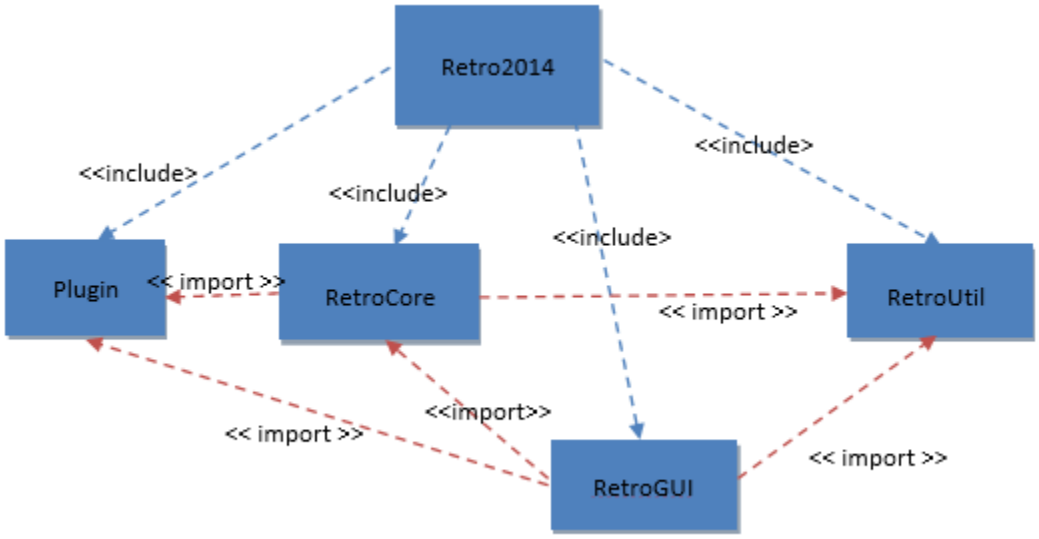


Figure 3: L'architecture de Retro2014

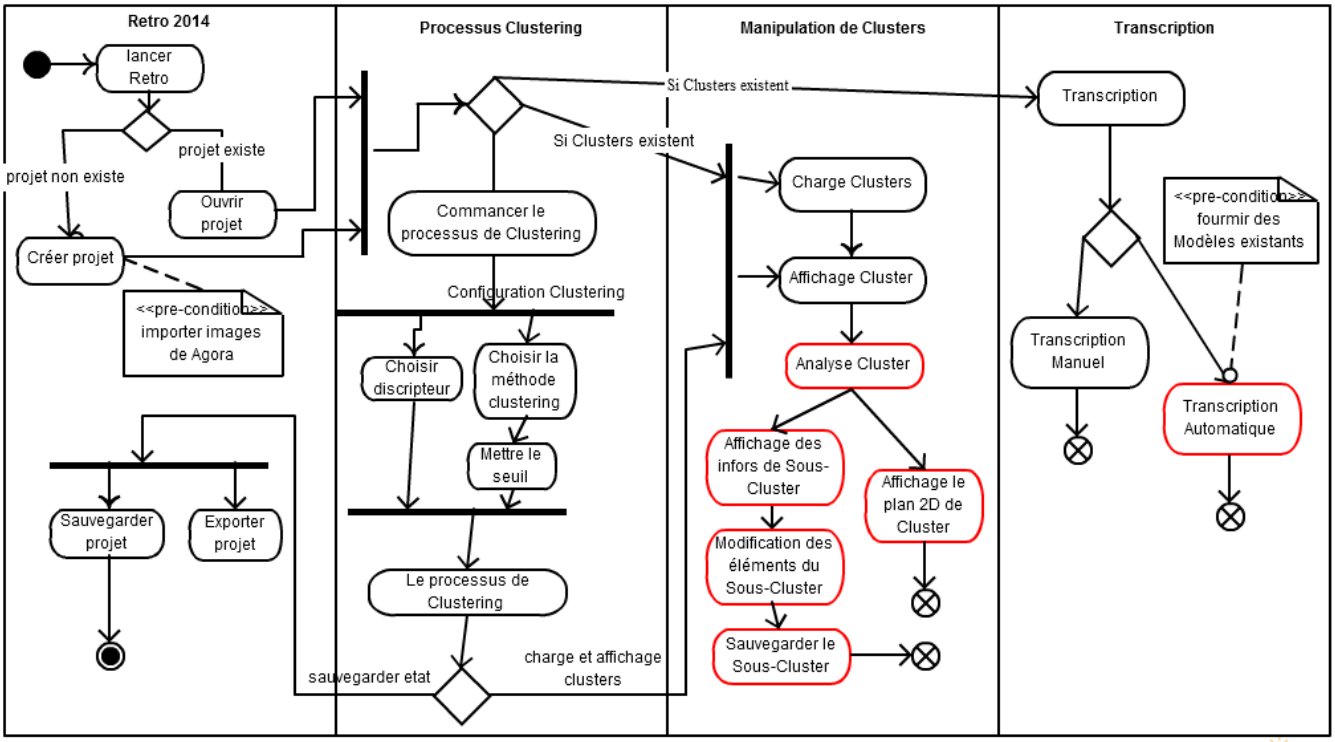


Figure 4 : Activités Diagramme

3.4. Contraintes de développement, d'exploitation et de maintenance

3.4.1. Contraintes de développement

Toutes les fonctionnalités de Retro ancien ne doivent pas être modifiées et doivent intégrer au développement du projet. Le projet développé doive être Open-Source constitue une autre contrainte car ceci implique un code parfaitement commenté (en anglais), l'utilisation de l'anglais pour les noms de variables et autres, et un code qui soit le plus modulaire possible.

4. Description des interfaces externes du logiciel

4.1. Interfaces matériel/logiciel

L'interface matériel/logiciel est seulement composée d'un ordinateur. Le logiciel n'est pas amené à communiquer en réseau avec d'autres machines.

4.2. Interfaces homme/machine

Le projet sera associé à trois interfaces homme/machine. La première étant l'interface permet de vérifier le processus de clustering. La deuxième étant l'interface de modification du cluster e. La troisième étant l'interface de transcription.

4.2.1. Description des interfaces existante

I- Interface du logiciel de Clustering

Cette première interface permettra la gestion du projet (chargement, sauvegarde, et export). Elle permettra aussi le processus de Clustering et l'affichage des Clusters.
Les différents menus seront suivants :
Le menu « Project »:

- New Project
- Open Project
- Save Project
- Close Projet
- Exit Retro

Le menu « EoC »:

- Compte Signatures
- View

Le menu « Clusters » :

- Generate
- View/Edit Clusters
- Generate Stats
- Load Clusters
- Manual Transcription
- Automatic Transcription

Le menu « Pages » :

- Browse Pages
- Browse Illustractions (deprecated)

Le menu « Books » :

- Export as Altro
- Export EoC Annotations

Le menu « Typograhy » :

- Model Selection
- BodyHeight
- Clusters To Models

A. Maquette du logiciel lors de son lancement

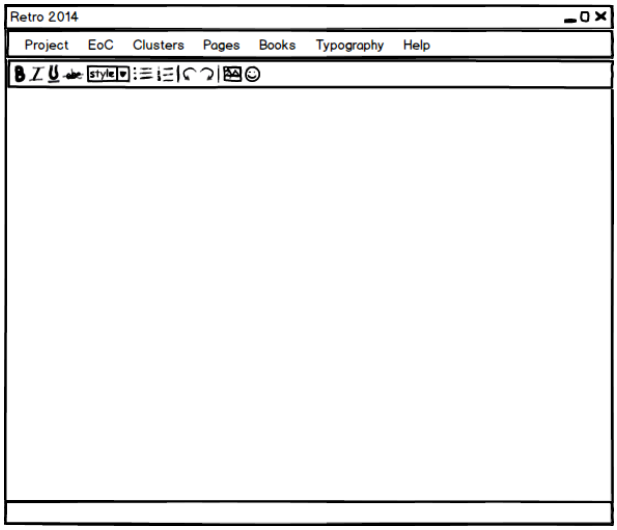


Figure 5: Accueil de Retro 2014

Lors de son lancement, le logiciel est épuré. Dans ce fait, l'utilisateur pourra effectuer ces actions :

- Ouvrir un Projet Retro.
- Créer un Projet Retro.
- Consulter les informations du logiciel.
- Quitter le logiciel.

Pour commencer d'utilisation, il faudra créer ou ouvrir un projet Retro.

B. Maquette du logiciel ouverture ou création d'un projet Retro

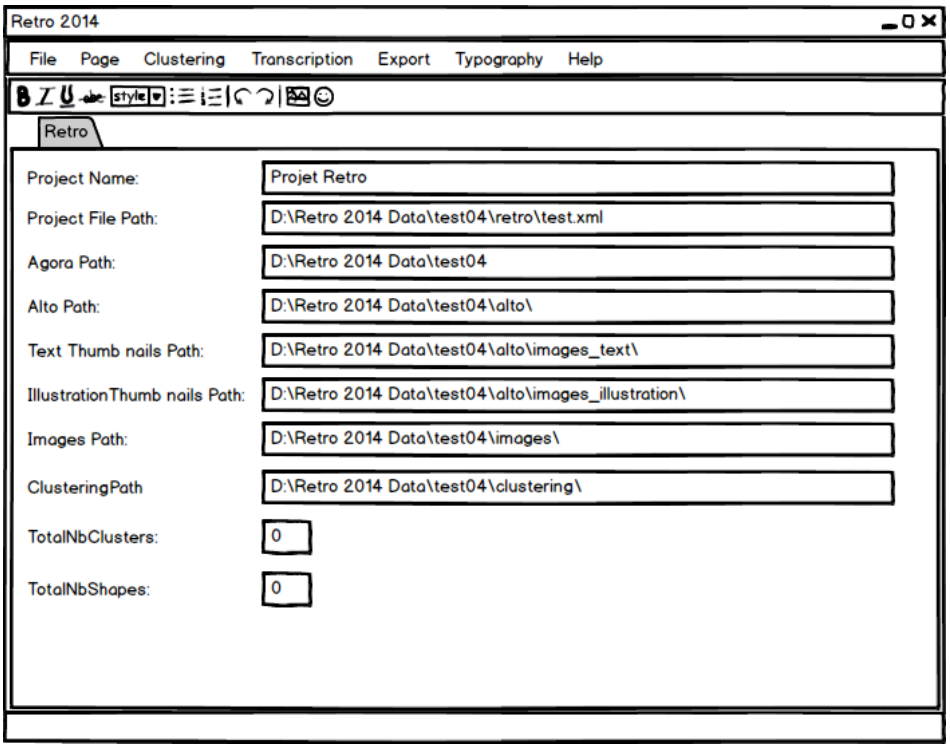


Figure 6: Ouvrir un Projet Retro2014

Après création ou ouverture d'un projet, on peut lancer le processus Clustering. A l'aide de menu « Clustering » -> « Process Clustering » permettant d'effectuer le traitement.

C. Maquette du processus Clustering

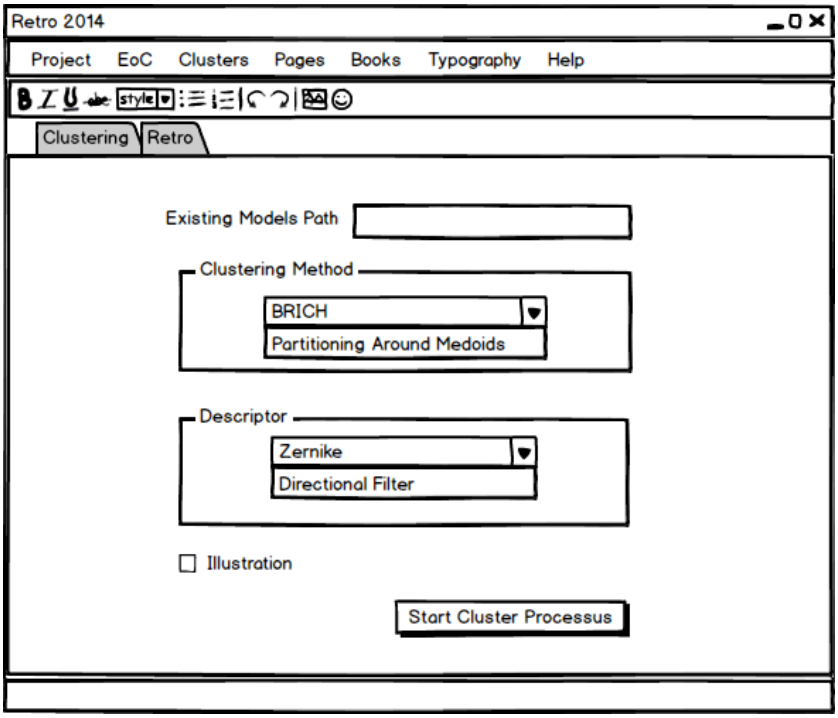


Figure 7 : La configuration du Clustering

Une fois que l'utilisateur lance le processus du Clustering, il faudra configurer des paramètres de Clustering :

- Le path de la base de modèles existant, s'il existe.
- Les paramètres du clustering méthode :
  - Le seuil
  - La manière du traitement de image : numérisation, normalisation, dénoies

En des paramètres valides, il peut démarrer le processus du Clustering. Cette fonction implémente l'interface « IClusteringPlugin »

II - Interface du logiciel analyse Cluster

A. Maquette de l'affichage Clusters

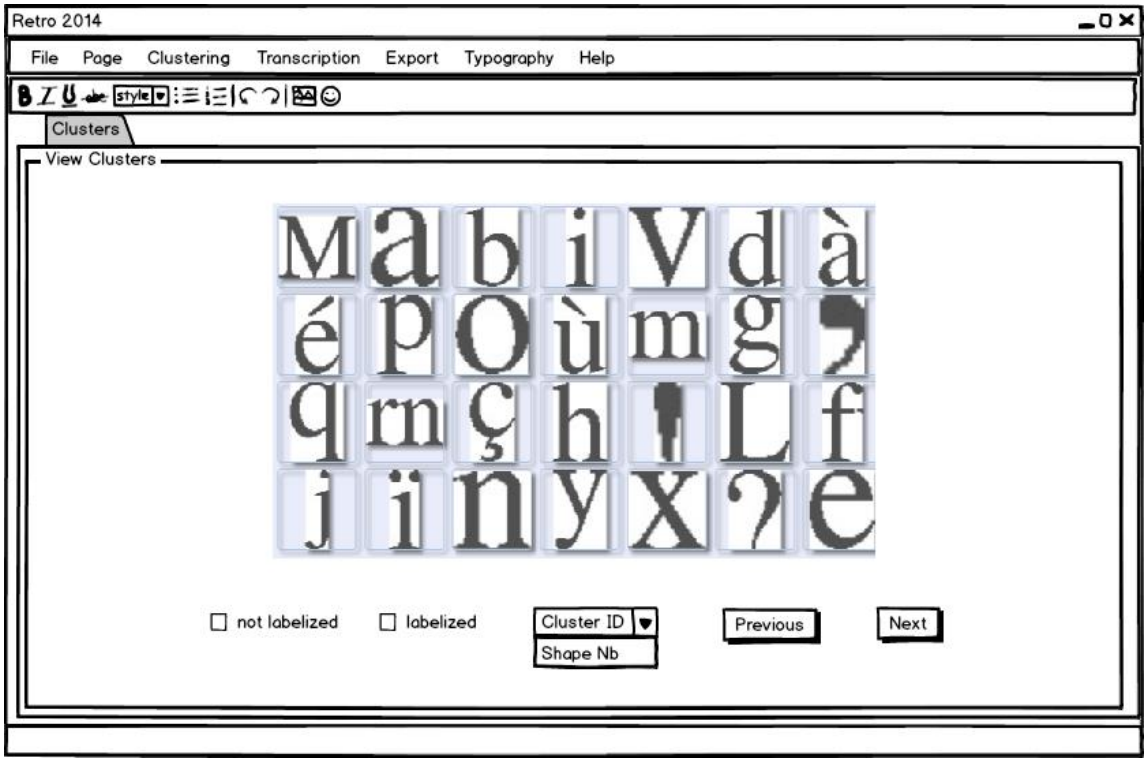


Figure 8 : Affichage de Clusters

Une fois que l'utilisateur a chargé des images de Clusters, voici à quoi il ressemblera. Désormais, les possibilités offertes seront :

- Affichage des caractères étiquetés/non-étiquetés
- Affichage des caractères par l'ordre de Cluster ID ou nombre de caractères
- Cliquer sur l'image d'un caractère pour consulter les informations

III - Interface du logiciel transcription

A. Maquette de transcription manuelle

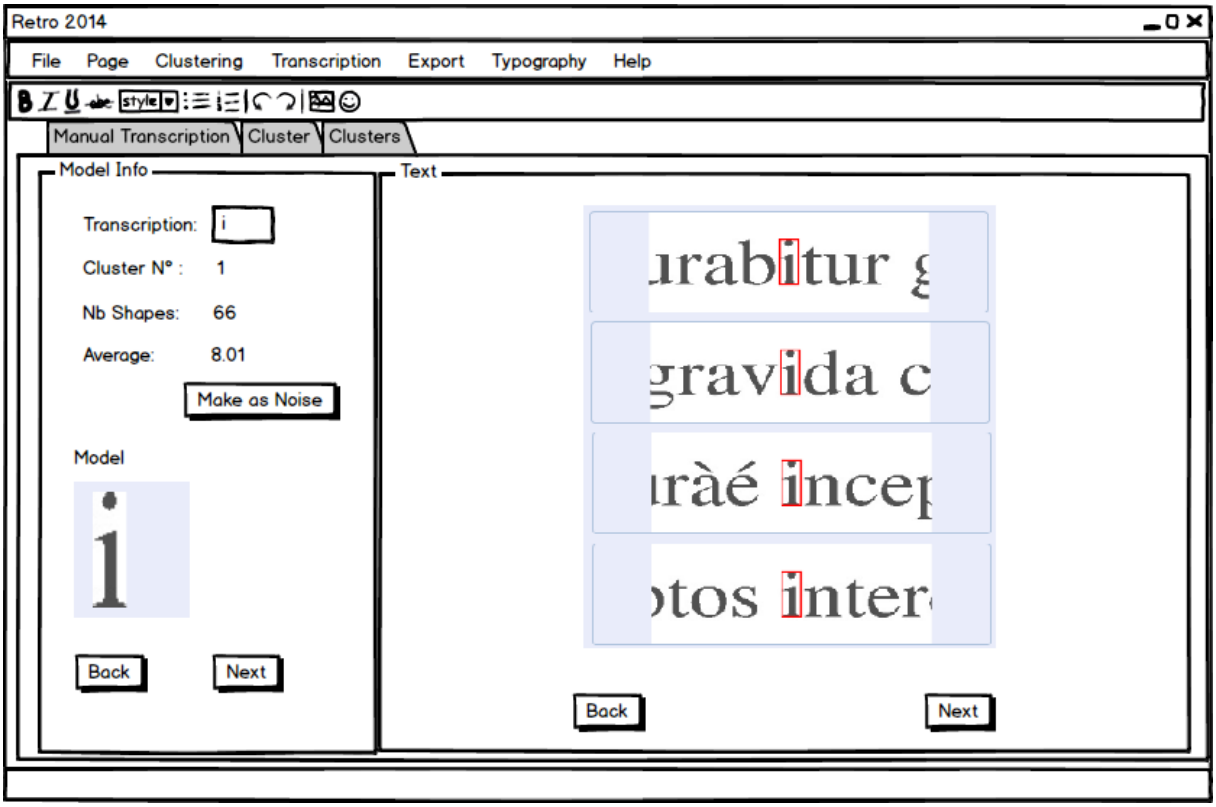


Figure 9: Transcription manuelle

Une fois que l'utilisateur a fait le processus du Clustering, il peut lancer Manual Transcription dans le menu. Désormais, les possibilités offertes seront :

- Mettre l'étiquette du Cluster et valide
- Marquer ce Cluster comme une noise
- Parcourir tous les Clusters
- Parcourir tous les éléments dans un Cluster

4.2.2. Description des interfaces à développer

A. Maquette de l'analyse d'un Cluster

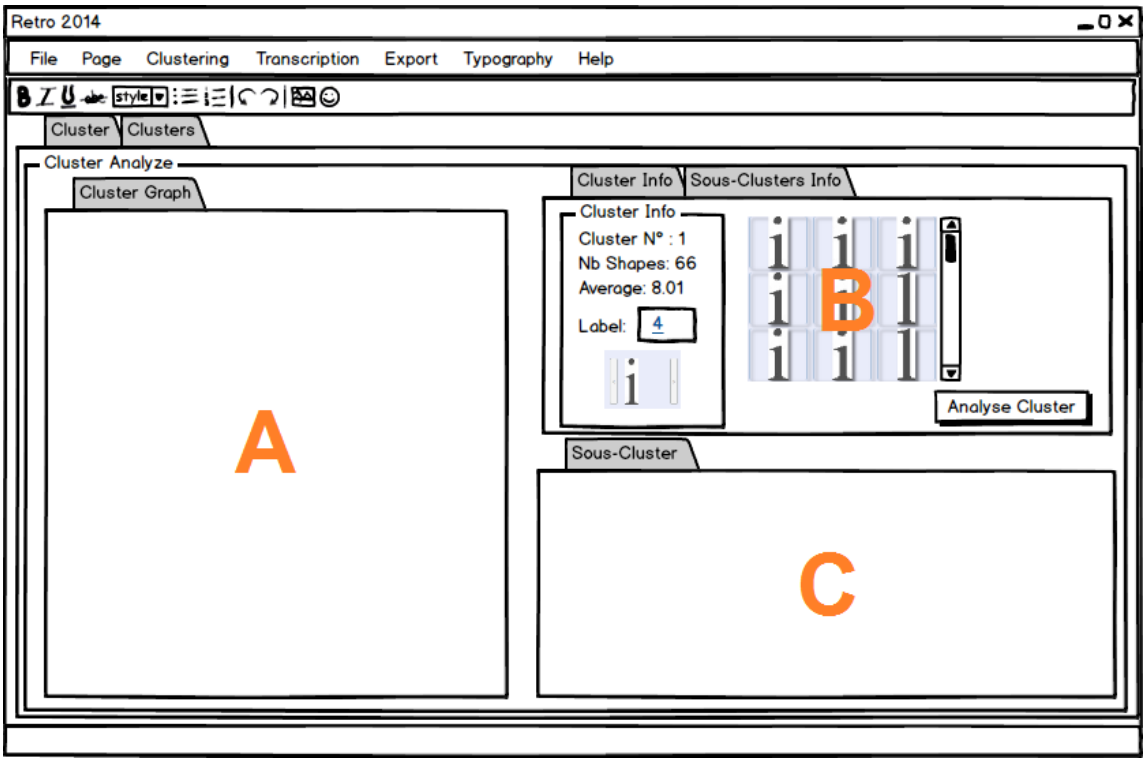


Figure 10: l'information d'un Cluster

Une fois que l'utilisateur clique une image de caractère (un Cluster), il affichera des informations du Cluster. L'interface propose une zone de « Info Cluster » pour afficher les informations:

- Toutes les images de caractères dans un Cluster
- Le numéro du Cluster
- Le nombre d'image dans le Cluster
- Le moyen de la signature du Cluster lequel est un signe du Cluster
- L'étiquette du Cluster
- L'image du représentant du Cluster

Pour simplifier l'analyse, on définit un ensemble appelé « Sous-Cluster » qui est un ensemble de caractères et consiste dans la sortie d'analyse. Il y a 3 parties principales dans la section Cluster Analyse :

- La partie A : Affichage le plan 2D/3D du Cluster ou un Sous-Cluster
- La partie B : les informations du Cluster et des informations de Sous-Clusters
- La partie C : les informations d'un Sous-Cluster

Le bouton « Analyse Cluster » permet de fournir des informations de Sous-Cluster dans la section B et le plan 2D/3D de Cluster dans la section A. L'interface ressemblera la maquette en dessous :

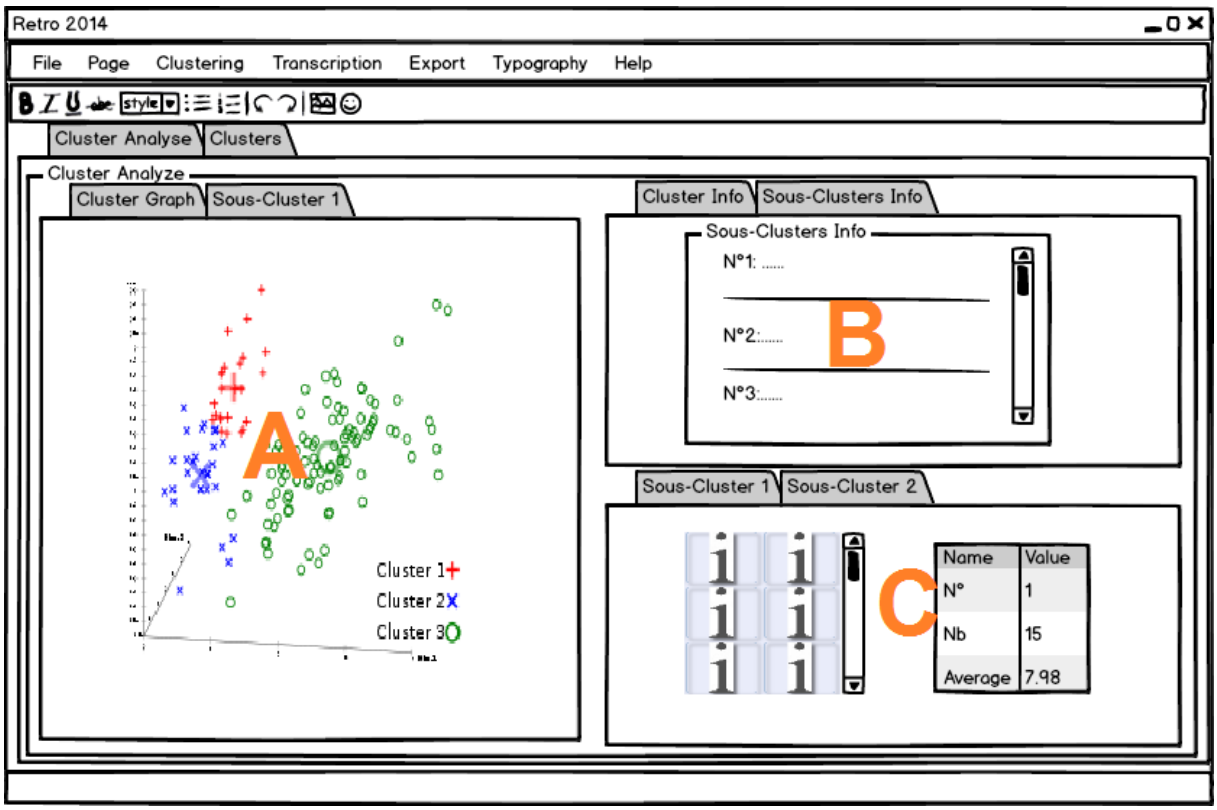


Figure 11: Analyse Cluster

La partie A affichera la distribution des points (un point par un élément/pattern). On peut distinguer les éléments anormaux et déployer des manipulations :

- En cliquant en droite sur un point, il peut envoyer l'élément dans un workspace en attendant de retravailler(le supprimer le déplacer etc…) et actualiser le fichier xml du Cluster.
- En flottant sur un point, il affichera ses informations (le numéro, des paramètres etc..) en une étiquette.

Quant à des manipulations de workspace, on fait des références à la fonctionnalité « Modifier Clusters » existant dans la version de PFE « Intégration et interfaçage de logiciels de clustering et de transcription » du Samantha GEORGES.
La partie B « Sous-Cluster Info » affichera les informations de tous sous-clusters :

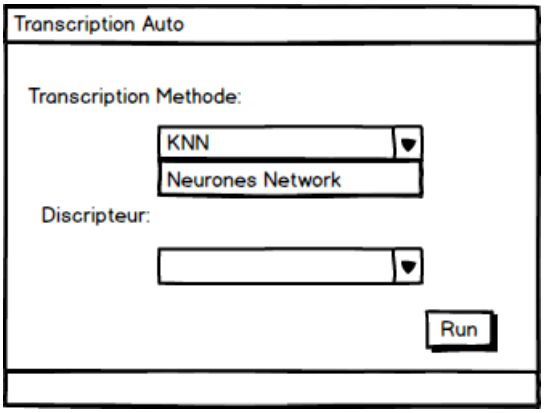
- Le numéro du Sous-Cluster
- Le nombre d'image dans le Sous-Cluster
- Le moyen de la signature du Sous-Cluster
- L'étiquette du Cluster

En cliquant sur un Sous-Cluster Info de la liste dans B, il affichera des informations du Sous-Cluster dans la partie C. La partie C « Sous-Cluster i » présente les informations de Sous-Cluster :

- L'image de caractère du Sous-Cluster
- Les informations du Sous-Cluster : numéro du Sous-Cluster, le nombre de caractères dans Sous-Cluster et le moyen de la signature du Sous-Cluster

Quand on ferme cette fenêtre, tous les informations de Sous-Cluster existent plus et on obtient un fichier .xml renouvelé du Cluster

B. Maquette de transcription automatique



Quand on lancer le processus de Auto Transcription un dialogue apparait pour choisir la méthode et le descripteur du transcription automatique. Après il y a une dialogue indique que la transcription automatique est terminé. On peut vérifier ou modifier le résultat dans l'interface de transcription manuelle.

4.3. Interfaces logiciel/logiciel

Le logiciel sera amené à traiter des images de Agora. Ces données seront stockées sur le répertoire où déposent des images de Agora. On peut justement configurer la description fichier .xml du projet Agora pour importer des images.

5. Architecture générale du système

Notre système est composé de 3 librairies et d'une application. L'application de l'Identifier les principaux composants/éléments du système ainsi que leurs relations. Sans être une analyse à part entière, cette partie doit montrer qu'une première réflexion sur la structure interne de Retro2014 qui complètera la vision sommaire donnée en 3.3. Un diagramme de package au sens large présentant les principales structures de données ainsi que les principaux composants du système sont fourni avec un ensemble d'explications.

5.1. Architecture générale des projets existants

Concernant l'architecture du code source, Retro2014 est composé de 4 sous-projets : « Plugin », « RetroCore », « RetroGUI » et « RetroGUI ». On détaille des descriptions et fonctions de sous-projets, après il y a une diagramme pour présenter les relations entre eux.

5.1.1. Plugin

Ce sous-projet qui est un paquet contient des modules d'extension d'application pour compléter le logiciel hôte pour lui apporter de nouvelles fonctionnalités. Le type de sortie est une librairie Plugin. Il est séparé en 3 parties (répertoires) :

- *DatabaseObjects* : Les objets fondamentaux pour stocker des données du projet.
  - « Database » comme une base de données (dataset) pour sauvegarder tous les données à traiter
  - « Document » contient tous les patterns (EoC) d'un document
  - « Cluster » après le traitement de Clustering, tous les informations peuvent sauvgarder dans un objet de la classe Cluster. Un Cluster est composé d'une liste de Partterns et ses paramètres de priorités
  - « Pattern » représentant un élément générique qui possède une liste de signatures
  - « Signature » est un pointeur de caractéristiques représentant un Pattern.
- *Interfaces* : Les interfaces et des classes abstraites de Clustering permettant de développer un module de Clustering, de descripteur ou de la lecture du document.
  - « IClusteringPlugin » : une interface pour le processus de Clustering
  - « IConfig » : s'intègre au plugin configuration avec la méthode de sauvegarder
  - « IDescriptorPlugin » : une interface pour ajouter le descripteur.
  - « IDocumentReaderPlugin » : une interface pour charger la base de données
- *PluginTools* : il est un outil de traitement image qui contient les fonctionnalités de débruitage de l'image et normalisation l'image.

5.1.2. RetroCore

Sous-projet principal de l'application qui fait appel aux autres sous-projets. Le type de sortie est une librairie RetroCore. . Il est séparé en 3 parties :

- *Model* : Les éléments fondamentaux de projet Retro : des paramètres de configuration, l'énumération d'exceptions, l'information sur logiciel.
- *OcrTypo* : il comprend les fonctionnalités liées à l'algorithme TemplateMatching, à l'export et à l'emplacement des données. Il sert à réaliser la transcription automatique.
  - « FontModel » : définir l'objet Font. Et on considère un Font comme un modèle
  - « IOCR » : l'interface de OCR moteur (Optical Character Recognition) est la reconnaissance optique de caractères. Celui-ci permet de récupérer le texte dans l'image d'un texte imprimé et de le sauvegarder dans un fichier pouvant être exploité dans un traitement de texte pour enrichissement
  - « TemplateMatchingOCREngine » : héritage « IOCR ». Il est un OCR moteur utilisant TemplateMatching pour la transcription automatique.
- *ViewModel* : il define the ViewModel éléments du Retro2014 en l'architectural pattern -- MVVM.

5.1.3. RetroGUI

Sous-projet de l'application est composé de IHM d'application. Le type de sortie est l'Application Windows.

5.1.4. RetroUtil

Ce sous-projet incorpore les outils liés à l'affichage de l'écran dynamique de démarrage et à la conversion d'une image.

5.2. Diagramme de Package

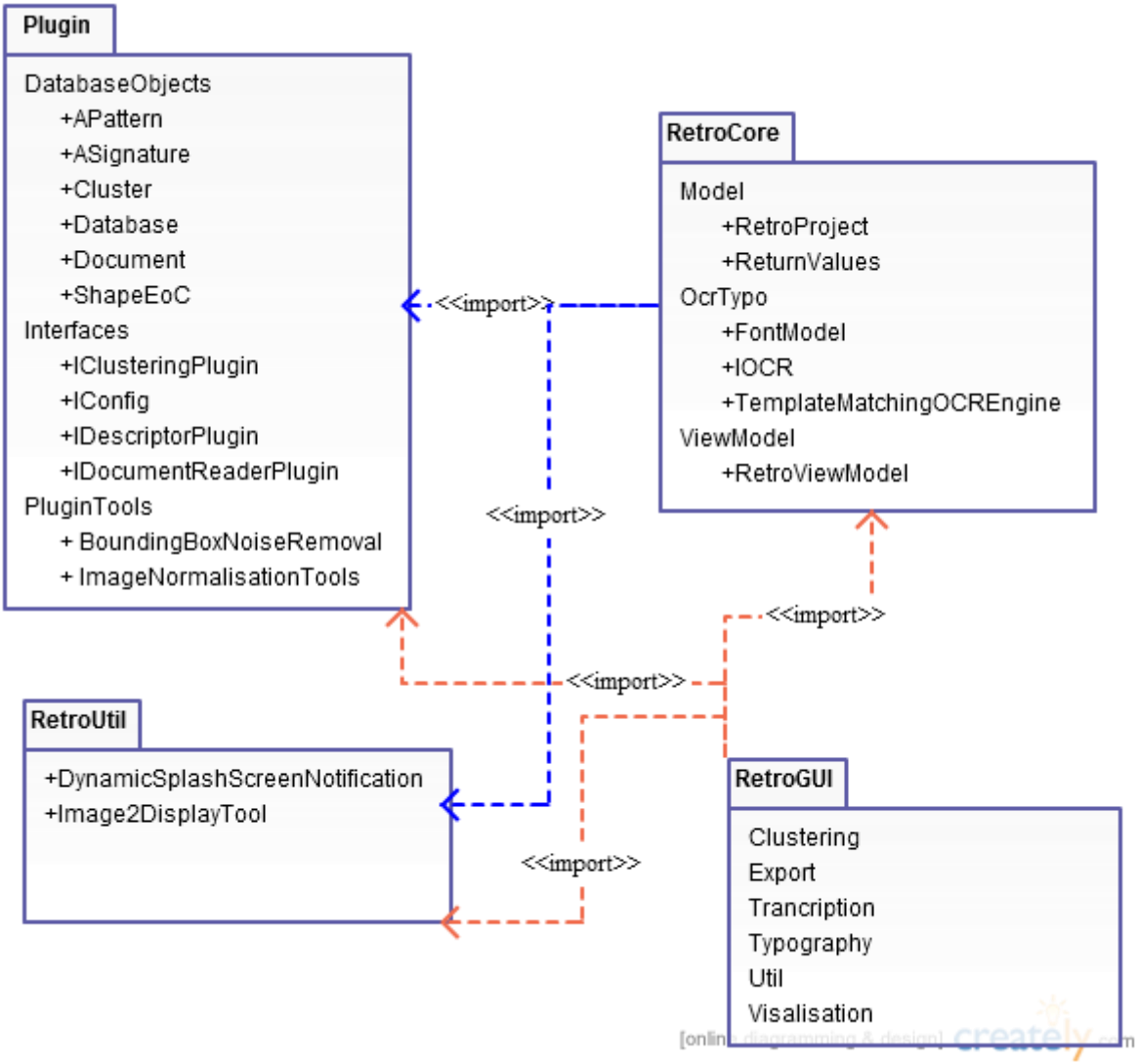


Figure 12: Package Diagramme

5.3. Proposition d'amélioration

Les propositions qui ont été suggérées et qui seront certainement mises en place dans la nouvelle version de RETRO2014 :

- Ajout des nouvelles IHM dans le sous-projet « RetroGUI »:
    - IHM et des méthodes correspondantes de « Analyse Cluster » sont met dans le répertoire « Clustering »
- Ajout des méthodes abstrait implémentant des différentes analyses du « Analyse Cluster » à la classe « IClusteringPligin » dans répertoire « Interface » du sous-projet « Plugin ». Création un nouveau répertoire « Traitement » dans le sous-projet « RetroCore » et ajout une nouvelle classe « AnalyseCluster » celui-ci implémenté l'interface « IClusteringPligin ».
- Ajout des méthodes de « TranscripAuto » dans la classe « TemplateMatchingOCREngine » du sous-projet « RetroCore » .

6. Description des fonctionnalités

L'objectif de ce PFE concerne la conception et le développement des IHMs et des méthodes correspondants, l'optimisation des clusters d'éléments de contenu. Mon travail s'agit de mettre en place analyser Cluster, l'optimisation de transcription manuelle et effectuer la transcription automatique.
Pour donner une vision globale, je fournirai un diagramme de l'utilisation. Des utilisations à développer sont met en rouge. En outre, la fonction « Analyse Cluster », la fonction « Transcription Manuelle » et la fonction « Transcription Automatique » seront décrite précisément.

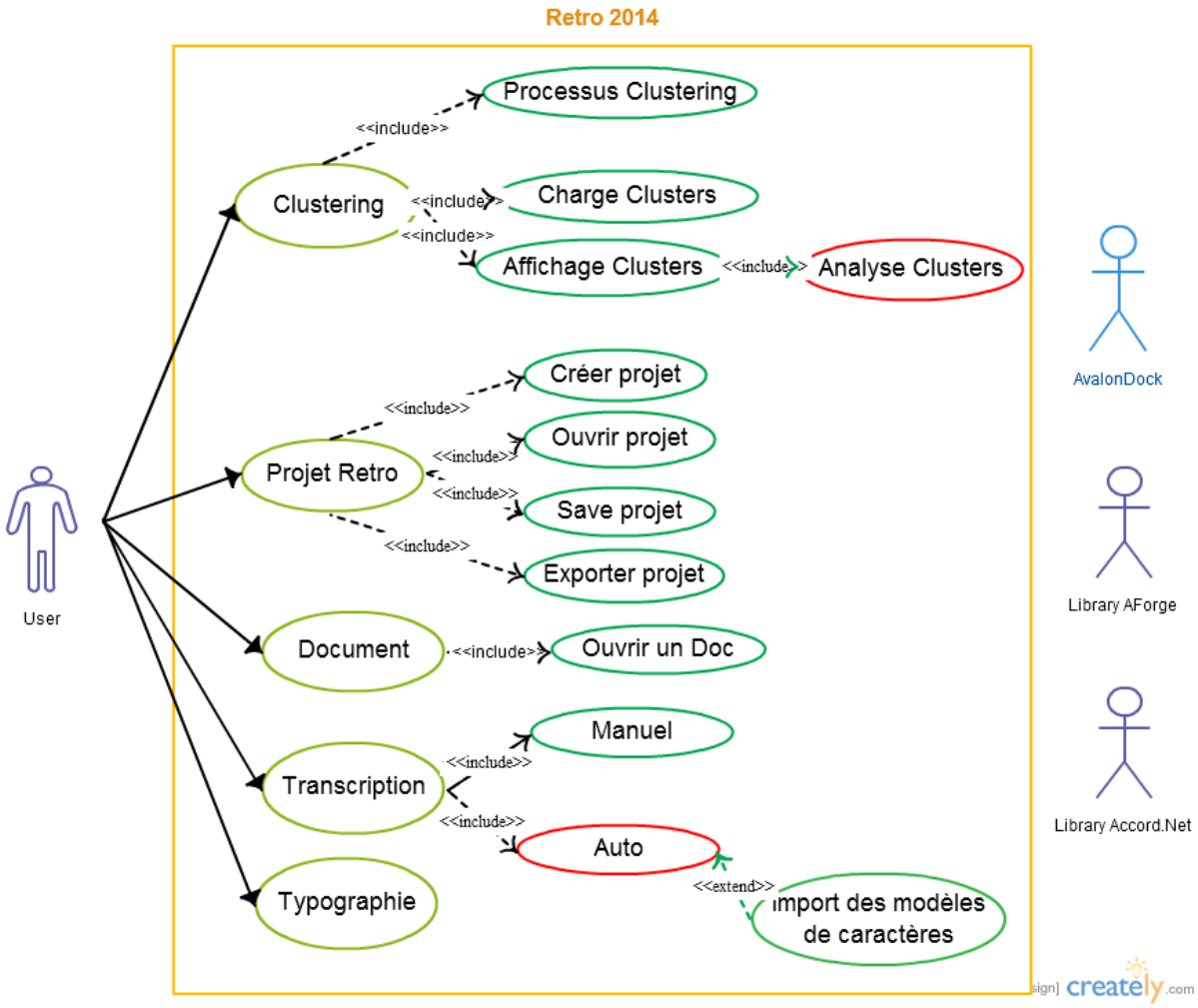


Figure 13: User Case Diagramme

6.1. Définition de la fonction « Analyse Cluster »

Une fois le Clustering effectué, l'utilisateur pourra consulter les Clusters créés et analyser le Cluster. Cette fonctionnalité liée à l'interface « IClusteringPlugin » et le libraire Accord.Net.
« Analyse Cluster » c'est-à-dire, découverte d'informations intéressantes dans un Cluster. On peut explorer des liens entre patterns dans un Cluster et signifier les objets homogènes et l'objet hétérogène. L'entrée d'analyse est le fichier .xml d'un Cluster. La sortie d'analyse : un plan de visuel Cluster ; une liste de Clusters qui regroupe des objets dans le Cluster et on les appelle Sous-Cluster. Le type de Sous-Cluster est Cluster. Il possède tous les propriétés de Cluster.

Voici un diagramme d'utilisation présentant les fonctions de « Analyse Cluster ».

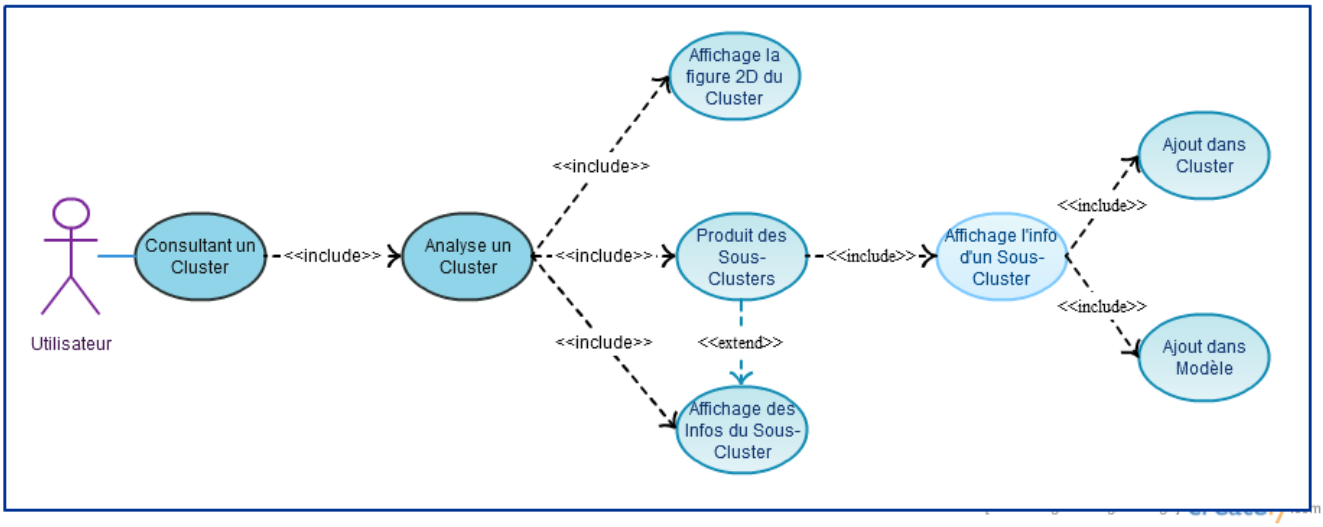


Figure 14: Use Case Analyse Cluster

6.1.1. Fonction « Analyse Cluster »

Nom : Analyse Cluster
Rôle : Permet de découvrir les liens entre les patterns dans un Cluster et refait Clustering sur un Cluster.
Priorité : Primordiale
Description : La méthode soit *PerformClustering* dans l'interface *IClusteringPlugin*, soit la méthode KMeans dans Accord.NET. Un répertoire sera créé pour le Cluster analysé dans le répertoire « *analyse cluster/* » . Dans le répertoire on sauvegarde le fichier xml de chaque Sous-Cluster. Et un fichier xml pour sauvegarder des informations d'analyse : le numéro de Cluster, le nombre de Sous-Cluster, les numéros de Sous-Cluster, le nombre de patterns dans le Sous-Cluster.
Entrée: Un Cluster
Sortie: Une liste de Sous-Clusters.

6.1.2. Fonction « Visual Cluster »

Nom : Visual Cluster
Rôle : L'analyse visuelle d'un Cluster, représenter le nuage de points du Cluster dans le plan (un point par objet)
Priorité : Primordiale
Description : En utilisant la méthode PCA (Principal Component Analysis) dans Accord.NET pour présenter la distribution de points du Cluster dans le plan.
Entrée: Une liste de Patterns (un Cluster)
Sortie : La plan 2D du Clusters

6.1.3. Fonction « Add To WorkSpace»

Nom : Add To WorkSpace
Rôle : un pattern du Cluster.
Priorité : Primordiale

Description : Ajouter un pattern du Cluster dans un workspace en attendant de le retravailler(le supprimer, déplacer etc..),
Entrée: Un pattern du Cluster
Sortie: Le fichier .xml du Cluster mise à jour et le « workspace » mis à jour

6.1.4. Fonction « View Hover Info »

Nom : View Hover Info
Rôle : Affichage des informations d'un pattern dans le nuage de points sur le plan visuelle d'un Cluster.
Priorité : Primordiale
Description : En flottant sur un point dans le plan, il affichera ses informations (le numéro, des paramètres etc..) en une étiquette
Entrée: Un pattern
Sortie: /

6.1.5. Fonction « List Sous-Cluster Info »

Nom : List Sous-Cluster Info
Rôle : Afficher des informations de tous les sous-clusters dans le panneau
Priorité : Secondaire
Description : Après l'opération d'analyse Cluster, un nouvel onglet s'ouvre et permet de naviguer des informations brèves liées à chaque Sous-Cluster : numéro du sous-cluster, le nombre de patterns, le moyen de sous-cluster etc.
Entrée: Résultat de Analyse Cluster
Sortie: Affichage des informations de sous-clusters

6.1.6. Fonction « View One Sous-Cluster »

Nom: View One Sous-Cluster
Rôle : Afficher un Sous-Cluster dans un mode plus précis.
Priorité : Secondaire
Description : Lorsqu'on clique sur un Sous-Cluster dans la liste de Sous-Cluster Info, un nouvel onglet s'ouvre et permet de naviguer dans le Sous-Cluster et de visualiser les formes qui lui ont été assignées. On fait référence à la fonction « View One Cluster ».
Entrée: Un Sous-Cluster
Sortie: Affichage des informations et du contenu du sous-cluster choisi en entrée.

6.2. Définition de la fonction « Transcription Automatique »

Une fois le Clustering effectué, l'utilisateur pourra lancer la transcription. Cette fonctionnalité liée à la classe « TemplateMatchingOCREngine » et le libraire Accord.Net. Il faut fournir des sources de modèles caractères pour démarrer la transcription automatique.

Voici un diagramme d'utilisation présentant les fonctions de « Transcription »

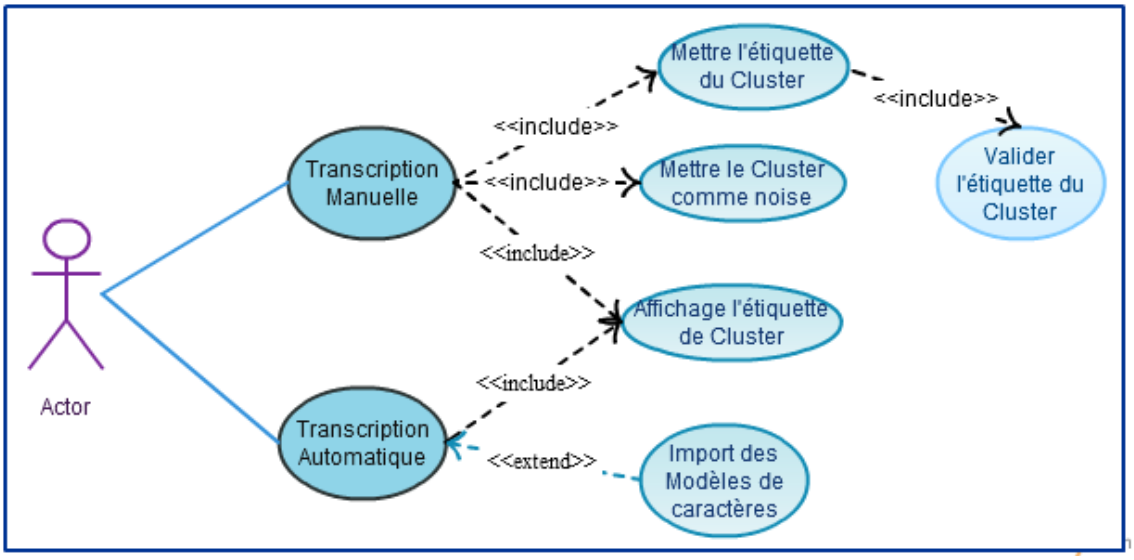


Figure 15: Use Case Transcription

6.2.1. Fonction « Transcription Automatique »

Rôle : Transcription Automatique
Priorité : Primordiale
Description : Transcrire le Cluster automatiquement avec des modèles existant. La méthode de transcription, soit Neurones Network, soit KNN (k plus proches voisins) en utilisant la librairie Accord.Net
Entrée: La liste de Clusters et la liste de Modèle
Sortie: Les Clusters étiquetés, mis à jours les fichiers .xml des Clusters

7. Conditions de fonctionnement

7.1. Performances

Les performances sont liées aux types d'algorithmes utilisés Préciser en termes mesurables, les spécifications temps réel liées à l'utilisation du système :

- du point de vue de l'utilisateur : temps de réponse souhaité, fréquence d'utilisation, temps d'indisponibilité acceptable, etc. ;
- du point de vue de l'environnement : fréquence moyenne d'acquisition d'états ou de mesures, fréquence maximale d'E/S, etc.

7.2. Capacités

En termes de temps de traitement, associés au nombre de pages d'un livre. Toutefois pour être utilisé dans des temps acceptables. Néanmoins il faudra s'assurer que le temps d'exécution soit dans des temps acceptables.

7.3. Sécurité

Au niveau de confidentialité du système, il n'y a aucun contrôle d'accès des utilisateurs, mots clefs, mots de passe, etc.

7.4. Conformité aux standards

Les conventions de nommage liés au langage C# seront respectés, ainsi que l'utilisation de la langue anglaise à l'intérieur du code.

7.5. Facteurs de Qualité

Les facteurs de qualité liés à l'amélioration de RETRO2014 sont :

- la qualité du code
- la documentation
- En effet, puisque le code se veut modulaire et que l'application pourra être extensible, il est nécessaire de respecter le modèle MVVM (Model View View Model) et de fournir une documentation développement. Il sera également important de mettre à jour le guide utilisateur de RETRO une fois l'application terminée. Il faudra également le mettre en ligne avec le code source, l'éxécutable et la documentation produisant automatiquement (Doxygen) grâce aux commentaires mis dans le code

.

PLAN DE DEVELOPPEMENT

8. Découpage du projet en tâches

Le projet sera découpé en 12 tâches principales, réunis en 5 parties différentes.

1. Prise en main du projet
   - ✓ Etude du contexte et des logiciels existants (architecture, dll et plug-ins existants)
   - ✓ Etude des méthodes de clustering et fouille visuelle de données utilisables pour comparer et caractériser les éléments sélectionnés
2. Analyse le sujet
   - ✓ Conception des IHM et méthodes interactives permettant la création, la visualisation et l'optimisation des clusters d'éléments de contenu
   - ✓ Rédaction du cahier de spécification
3. Développement du Retro2014
   - ✓ Rédaction du document de conception et spécification et rapport technique.
   - ✓ Correction de l'existant + phase de test
   - ✓ Création et développement de IHM de Analyse Cluster + phase de test
   - ✓ Implémentation de transcription automatiquement + phase de test
4. Phase de test du système
   - ✓ Test complets
   - ✓ Rédaction du cahier de recettes et validation
5. Livraison finale
   - ✓ Rédaction du rapport et préparation de la soutenance

On indiquera également ici les tâches relatives à la gestion de projet (prise en mains de l'existant, bibliographie, rédaction du cahier de spécification, du rapport, de manuels techniques ou utilisateurs, mise en production et recette globale, etc.
Chaque tâche doit être décrite précisément :

8.1. Prise en main du projet

8.1.1. Etude du contexte et des logiciels existants

Description de la tâche : cette tâche comporte :

- Découverte du sujet et du but du PFE
- Récupération des sources de RETRO2014 et lecture des rapports et documents correspondants
- Prise en main des logiciels RETRO2014

Livrables : Visual Studio 2012, rapport et document sur l'existant
Estimation de charge : 3 jours/homme (début le 18/09/2014 et fin le 02/10/2014).

8.1.2. Etude des méthodes de clustering

Description de la tâche : cette tâche comporte :

- Etude des méthodes de clustering
- Etude des méthodes de fouille visuelle de données
- Etude le libraire Accord.Net

Livrables : Rapport d'étude
Estimation de charge : 2 jours/homme (début le 25/09/2014 et fin le 02/10/2014).

8.2. Analyse le sujet

8.2.1. Conception des IHM et méthodes interactives

Description de la tâche : cette tâche comporte :

- Proposition et conception graphique de nouvelles IHM de « Analyse Cluster » et « Transcription Manual »
- Proposition la méthode d'analyse cluster et la méthode de fouille visuelle de données
- Proposition la méthode de transcription automatique

Livrables : Rapport d'étude et compte-rendu des IHM
Estimation de charge : 7 jours/homme (début le 09/10/2014 et fin le 20/11/2014).

8.2.2. Rédaction du cahier de spécification

Description de la tâche : cette tâche consiste à réaliser le cahier de spécification à remettre à la maitrise d'ouvrage pour valider le travail à effectuer durant ce projet. C'est une première étape importante du projet
Livrables : cahier de spécification
Estimation de charge : 4 jours/homme (début le 27/11/2014 et fin le 09/01/2015).

8.3. Développement du Retro2014

8.3.1. Rédaction du document de conception et spécification

Description de la tâche : cette tâche comporte de la rédaction du guide utilisateur et du guide de développements et la rédaction du document de conception et spécification.
Livrables : le guide utilisateur, le guide de développements et le document de conception et spécification.
Estimation de charge : 7 jours/homme (début le 08/01/2015 et fin le 26/02/2015).

8.3.2. Correction de l'existant

Description de la tâche : cette tâche consiste à corriger les sources de Retro2014 assurant le bon fonctionnement de logiciel actuel pour dérouler le développement du PFE
Livrables : sources Retro2014
Estimation de charge : 2 jours/homme (début le 08/01/2015 et fin le 15/01/2015).

8.3.3. Création et développement de IHM de Analyse Cluster + phase de test

Description de la tâche : cette tâche comporte :

- Développement IHM
- Développement du module « Analyse Cluster »

- ✓ Fonction « Analyse Cluster »

- Développements liés à la fonctionnalité :
  - ✓ Fonction « Average of Cluster »
  - ✓ Fonction « List Sous-Cluster Info »
  - ✓ Fonction « View One Sous-Cluster »
  - ✓ Fonction « Add To WorkSpace »
  - ✓ Fonction « View Hover Info »

Livrables : le code source de ces fonctions, des IHM, documentation (rapport technique), démonstration, résultat de test (cahier de test).

Le code source du module « Analyse Cluster » sera livré après avoir vérifié que les méthodes de clustering (à savoir ACP et méthode des k-médoïdes) ont été correctement implémentées (via les DLL). Il aura également été vérifié que la méthode de clustering implémentée dans RETRO2014 n'ait pas été modifiée.

Estimation de charge : le développement IHM estimé à 2 jours/homme (début le 20/01/2015 et fin le 21/01/2015). Le développement du module « Analyse Cluster » estimé à 5 jours/homme (début le 22/01/2015 et fin le 27/01/2015). Le développement liés à la fonctionnalité listées en dessus estimé à 6 jours/homme (début le 28/01/2015 et fin le 10/02/2015).

8.3.4. Implémentation de transcription automatiquement + phase de test

Description de la tâche : Cette tâche consiste à configurer la méthode de transcription et transcrire le Cluster automatiquement avec des modèles existant. Elle est liée à la librairie Accord.Net laquelle contient des méthodes d'apprentissage automatique.

Cette tâche comporte :

- Développement lHM
- Développement du module « Transcription Auto»
  - ✓ Fonction « Transcription Automatique »

Livrables : le code source, documentation (rapport technique), démonstration, résultat de test (cahier de test)

Estimation de charge : 8 jours/homme (début le 11/02/2015 et fin le 26/02/2015).

8.4. Phrase de test du système

8.4.1. Test complets

Description de la tâche : Cette tâche consiste à réaliser les différents tests permettant de vérifier que tous les éléments réalisés au cours de ce projet fonctionnent correctement. C'est une phase très importante du projet qui permettra de valider la fiabilité du résultat.

Livrables : le code source de Retro2014, l'exécutable de Retro2014, démonstration, résultat de test(rapport test)

Estimation de charge : 7 jours/homme (début le 03/03/2015 et fin le 17/03/2015).

8.4.2. Rédaction du cahier de recettes et validation

Description de la tâche : cette tâche consiste à réaliser le cahier de recettes et validation à assurer formellement que le projet est conforme au cahier de spécifications.

Livrables : le cahier de recettes et validation

Estimation de charge : 7 jours/homme (début 03/03/2015 et fin le 17/03/2015).

8.5. Livraison finale

8.5.1. Rédaction du rapport et préparation de la soutenance

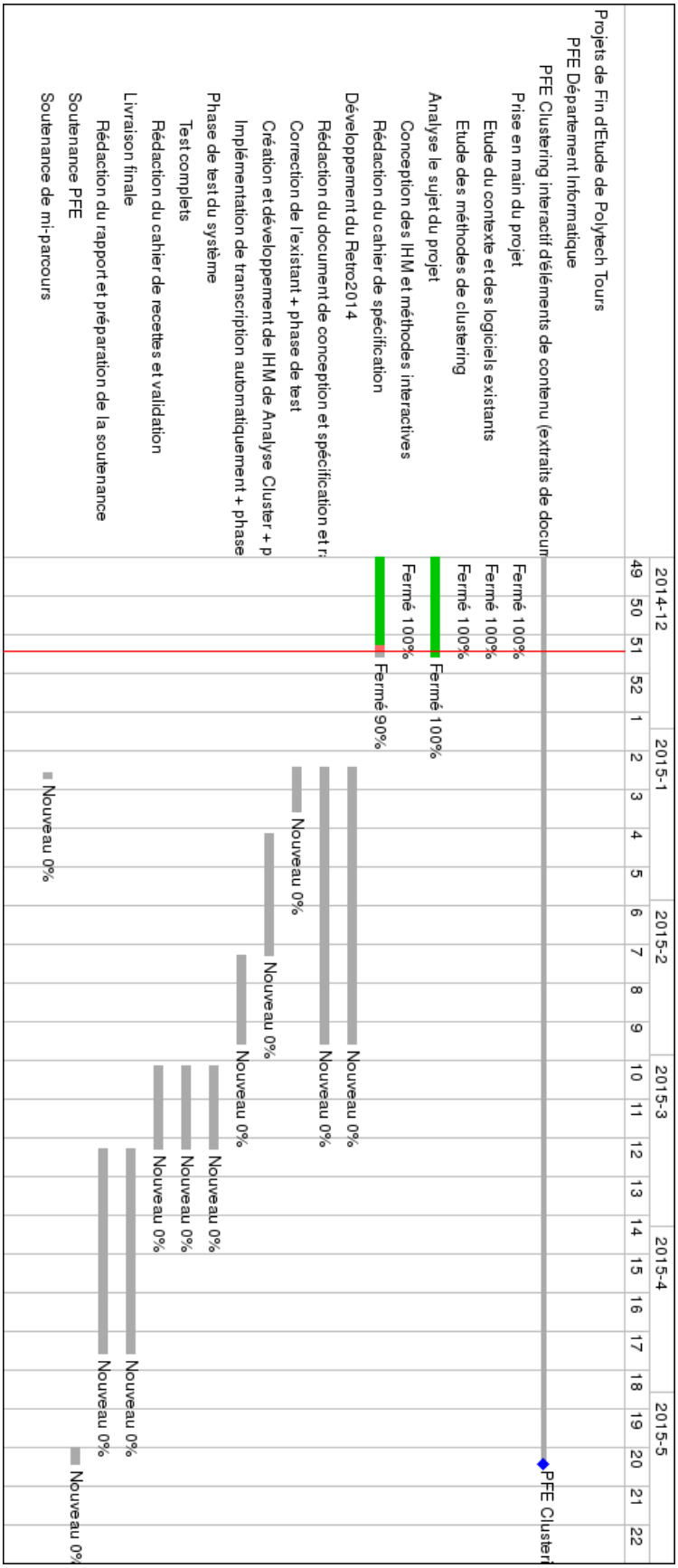
Description de la tâche : Cette tâche consiste à rédiger toute la partie rapport du projet ainsi que la préparation à la soutenance finale.

Livrables : Rapport e PFE accompagné des différents logiciels et codes sources réalisés pendant le projet.

Estimation de charge : 17 jours/homme (début le 18/03/2015 et fin le 23/04/2015).

9. Planning

Le planning (sous forme de diagramme de Gantt) synthétise l'ordonnancement de chacune des tâches en faisant apparaître leur éventuelle parallélisassions. Il indique également les dates clés de la réalisation du projet (soutenance, etc.) ainsi que les dates de remise des livrables.


2014-12
2015-1
2015-2
2015-3
2015-4
2015-5
49 50 51 52 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Projets de Fin d'Etude de Polytech Tours
PFE Département Informatique
PFE Clustering interactif d'éléments de contenu (extraits de docum
Prise en main du projet
Fermé 100%
Etude du contexte et des logiciels existants
Fermé 100%
Etude des méthodes de clustering
Fermé 100%
Analyse le sujet du projet
Fermé 100%
Conception des IHM et méthodes interactives
Fermé 100%
Rédaction du cahier de spécification
Fermé 90%
Développement du Retro2014
Nouveau 0%
Rédaction du document de conception et spécification et r
Nouveau 0%
Correction de l'existant + phase de test
Nouveau 0%
Création et développement de IHM de Analyse Cluster + p
Nouveau 0%
Implémentation de transcription automatiquement + phase
Nouveau 0%
Phase de test du système
Nouveau 0%
Test complets
Nouveau 0%
Rédaction du cahier de recettes et validation
Nouveau 0%
Livraison finale
Nouveau 0%
Rédaction du rapport et préparation de la soutenance
Nouveau 0%
Soutenance PFE
Nouveau 0%
Soutenance de mi-parcours
Nouveau 0%
PFE Clusteri

GLOSSAIRE

Dans cette partie on doit trouver, classés par ordre alphabétique, les définitions des termes courants utilisés, des termes techniques, abréviation, sigles et symboles employés dans l'ensemble du document.

PARADIIT (Pattern Redundancy Analysis for Document Image Indexation & Transcription) est une solution de transcription de documents anciens, formée de plusieurs applications développées par l'équipe RFAI.

BIBLIOGRAPHIE

(s.d.). Récupéré sur Site web Balsamiq Mockups: http://balsamiq.com/products/mockups/
(s.d.). Récupéré sur Site web Sandcastle: http://sandcastle.codeplex.com/
(s.d.). Récupéré sur Site web Accord.Net: http://accord-framework.net/