République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahleb de Blida

Faculté des Sciences

Département d'informatique

Mémoire de fin d'études

Pour L'Obtention du diplôme de Master en *Informatique*

Option : *Ingénierie de Logiciel*

Théme :

Conception et déploiement d'une GED munie d'un outil d'extraction et de filtrage des données.

Réalisé par :

Nom : Bentrioua
Prénom : Naziha Khadoudja

Nom : Bouloufa
Prénom : Yassmine

Président : Mme. Chikhi I

Examinateur : Mme.Cheriguene

Promoteur : Mr.Kameche Abdellah

Encadreur : Mme Melaaz Magha

Soutenu le : **22 Octobre 2019**

ABSTRACT

This study is a part of a project to install an electronic document management system with an OCR for the archives center of Baba-Hassen in BNA, that in term will be able to be deployed at the national level. This work is only the first stage of the project.

The first part of our work is a selective study of the available GEDs, in order to choose the right GED for our needs, then treat and transform this GED to be able to do functional work and meets the requested requirements.

The second part is to answer the problems of digitization encountered by the archives center, which objective is to provide an OCR system that will be able to create a digital version of paper documents namely bank checks. The extraction of information sometimes can be complex, that's why we chose the convolutional neural networks specialized in form recognition that is done through a learning process. Despite the difficulties encountered, we reached an acceptable result.

Keywords: GED, OCR,the convolutional neural networks

RÉSUMÉ

La présente étude s'inscrit dans le cadre d'un projet visant à la mise en place d'une GED munie d'un OCR pour le centre des archives de la BNA de Baba-hassen, mais qui à terme pourra s'étendre au niveau nationale. Ce travail n'est que la première étape du projet .

La première partie de notre travail est une étude sélective parmi les GED disponibles, afin de choisir la GED adéquate à nos besoins, pour ensuite traiter et transformer cette GED pour qu'elle soit entièrement fonctionnelle et répond aux exigences demandées.

La deuxième partie est de répondre aux problèmes de numérisation rencontré par le centre des archives, dont l'objectif est de mettre à disposition un système OCR qui sera en mesure de créer une version numérique des documents papier à savoir les chèques bancaires et les bordereaux de versement AADL.
L'extraction d'information s'avère parfois être complexe, c'est pour cela qu'on a choisi les réseaux de neurones convolutionels spécialisés dans la reconnaissance de forme qui se fait à travers un processus d'apprentissage. Malgré les difficultés rencontrées on a abouti à un résultat acceptable.

Mots clés: GED, OCR,les réseaux de neurones convolutionels

REMERCIEMENT

En guise de reconnaissance, on tient à exprimer nos sincères remerciements à toutes les personnes qui ont contribué de prés ou de loin pour la réalisation de ce projet et la rédaction de ce mémoire.

Dans un premier temps on tient a remercier notre promoteur de projet Mr Kameche de l'université de Blida Saad Dahleb1,pour nous avoir encadré et orienté ainsi que sa patience et surtout ses judicieux conseils qui ont permis de donnée vie a ce projet.

Un grand merci a Mme Tafat chef du département des archives à la banque national d'Algérie à baba hassen et notre encadreuse Mme Melaaz ingénieur informatique à la BNA, pour nous avoir donné l'occasion extraordinaire de faire un stage pratique au sein de leur formidable équipe, on les remercie pour leur confiance et leur soutien inestimable.

On désire également remercier nos parents, nos proches, familles et amis qui ont fait en sorte de nous voir réussir dans notre parcours universitaire.

Enfin, on s'oserai oublier de remercier tout le corps professoral du département d'informatique de l'université Saad Dahleb de blida, pour leur travail énorme durant les années précédente, afin de nous offrir les conditions les plus favorable pour le déroulement de nos études.

SOMMAIRE

Page

Introduction générale. 1

1 Gestion Electronique de Documents **3**

1.1 Le fonctionnement de la GED : 3

1.1.1 L'acquisition des documents : 4

1.1.2 La gestion des documents : 5

1.1.3 Le stockage de données : 5

1.1.4 Diffusion des documents : 6

1.1.5 L'archivage : 6

1.2 Les composants d'une GED : 8

1.3 Les catégories de GED : 9

1.4 Les avantages de la GED : 9

1.5 Etude comparative des solutions GED proposées. : 11

1.6 Architecture d'Alfresco : 12

1.6.1 Couche de stockage et modélisation des données(Storage layer) 13

1.6.2 Couche de services (Repository) : 14

1.6.3 Couche APIs et protocoles : 15

1.6.4 Couche applications clientes : 16

1.7 Conclusion : 17

2 Reconnaissance optique des caractères **18**

2.1 Généralité sur les images : 18

2.1.1 L'image numérique : 18

2.2 Concept de l'OCR : 20

2.2.1 Chaîne de numérisation : 20

2.2.2 Reconnaissance du document : 23

2.3 Différents aspect de reconnaissance de l'écriture : 25

2.3.1 Reconnaissance en ligne : 26

2.3.2 Reconnaissance hors ligne . . . 26
2.4 Les différentes techniques de reconnaissance : . . . 26
2.4.1 Machine a vecteur de supports (SVM) . . . 26
2.4.2 Réseaux de neurones : . . . 27
2.5 Conclusion : . . . 33

3 Conception **34**
3.1 Diagramme UML . . . 34
3.1.1 Diagramme de cas d'utilisation : . . . 34
3.1.2 Diagramme de déploiement : . . . 42
3.1.3 Diagramme de séquence : . . . 43
3.1.4 Diagramme d'activité : . . . 47
3.1.5 Diagramme de classe : . . . 49
3.2 Conception du CNN . . . 50
3.3 Conclusion : . . . 55

4 Implémentation et Test **56**
4.1 Environnement de travail : . . . 56
4.1.1 Python : . . . 56
4.1.2 HTML et CSS . . . 57
4.1.3 PHP . . . 58
4.1.4 PostgreSQL . . . 59
4.1.5 Xampp . . . 60
4.1.6 Tomcat . . . 60
4.1.7 IntelliJ IDEA . . . 61
4.1.8 Keras . . . 61
4.2 Interfaces . . . 61
4.3 Implémentation de l'OCR . . . 80
4.3.1 Pré-traitement : . . . 83
4.3.2 Segmentation . . . 83
4.3.3 Post-traitement . . . 84
4.4 Test et résultat . . . 84
4.4.1 Cross validation : . . . 87
4.5 Conclusion : . . . 89

Bibliographie **96**

Liste des tableaux

Table		Page
1.1	Etude comparative des solutions GED.	11
3.1	La liste des acteurs.	35
4.1	Résultat sur MNIST.	87
4.2	Résultat sur Extended MNIST.	87
4.3	Résultat du Cross Validation.	88
4.4	Liste des abréviations..	95

Table des figures

Figure		Page
1.1	les étapes pour mettre en place une GED.	4
1.2	les étapes d'archivage d'un document.	6
1.3	L'architecture d'Alfrescot.	12
1.4	Couche de stockage.	13
1.5	Couche de services (Repository) .	15
1.6	Couche de protocoles et APIs.	16
1.7	Couche applications clientes.	16
2.1	Image matricielle Vs image vectorielle .	19
2.2	structure d'un système OCR .	20
2.3	Image en niveau de gris	21
2.4	Image binaire	22
2.5	fonction FindCountous.	23
2.6	Graphe SVM.	27
2.7	Architecture standard d'un réseau de neurone convolutionnel.	30
2.8	Convolution d'une image de 5 x 5 pixels avec un filtre de 3 x 3 pixels (foulée = 1 x 1 pixel)	31
2.9	Max Pooling by 2 x 2	32
2.10	Une couche entièrement connectée avec deux couches cachées.	32
3.1	diagramme de cas d'utilisation générale.	35
3.2	Diagramme de cas d'utilisation « Gestion des utilisateurs ».	36
3.3	Diagramme de cas d'utilisation « Gestion des sites ».	37
3.4	Diagramme de cas d'utilisation « Gestion des groupes ».	38
3.5	diagramme de cas d'utilisation « Gestion des workflow ».	39
3.6	diagramme de cas d'utilisation« Gestion des documents».	40
3.7	diagramme de cas d'utilisation« Gestion la numérisation».	41
3.8	diagramme de déploiement».	42
3.9	diagramme de séquence « Gestion des documents».	43
3.10	diagramme de séquence « Gestion des sites».	44
3.11	diagramme de séquence « Utilisation des sites ».	45

3.12 diagramme de séquence « Gestion des utilisateurs». 46
3.13 diagramme d'activité « gestion des documents ». 47
3.14 diagramme d'activité « Utilisation des documents ». 48
3.15 Diagramme de classe. 49
3.16 Architecture de notre réseau de neurones. 50
3.17 Flatten() 53
3.18 Graphe Relu. 54

4.1 Python. 56
4.2 HTML et CSS. 57
4.3 PHP. 58
4.4 PostgreSQL. 59
4.5 XAMPP. 60
4.6 Tomcat. 60
4.7 IntelliJ-IDEA. 61
4.8 Interface authentification GED. 62
4.9 Interface d'accueil GED. 63
4.10 Interface gérer user GED. 64
4.11 Ajouter user GED. 65
4.12 Modifier supprimer user GED. 66
4.13 Gérer groupe GED. 66
4.14 Section sites GED. 67
4.15 site GED. 67
4.16 Ajouter membres au site GED. 68
4.17 Options espace documentaire GED. 69
4.18 Démarrer Workflow GED. 70
4.19 Interface OCR. 71
4.20 Interface OCR. 72
4.21 Saisie manuelle. 73
4.22 OCR. 74
4.23 Recherche par date. 75
4.24 Interface bordereau AADL. 76
4.25 Saisie manuelle bordereau. 77
4.26 OCR bordereau. 78
4.27 Recherche bordereau. 79
4.28 Base de donnée. 80
4.29 Création d'un modèle. 80
4.30 Niveau de gris. 83
4.31 Conversion en binaire. 83

4.32 mesures de performance. 85

INTRODUCTION GÉNÉRALE

Actuellement, une quantité massive de documents et d'informations partagées qui circulent au sein des entreprises a été remarquée. La gestion de ces contenus est devenue de plus en plus pénible et fatigante vu le taux d'énergie et de temps consacrés à cette tâche, ainsi que pour la recherche d'informations qui demeure un problème au quotidien, de même pour la saisie manuelle de certaines informations qui représente un temps mort alors que ce temps peut être bénéfique pour la réalisation d'autres tâches jugées plus importantes.

Dans ce cadre, le centre des archives de la BNA (banque national d'Algérie) de Baba-Hassen , veut exploiter de nouvelles méthodes afin de faire face aux problèmes de la gestion documentaire ainsi que la saisie manuelle de ses chèques bancaires et celle des bordereaux.
Pour pallier à ces problèmes, une stratégie d'archivage doit être mise en place afin d'organiser et de classer les différents documents selon des critères spécifiques du centre des archives, et permettre de conserver et garder une trace numérique tout au long du cycle de vie du document.

Dans ce contexte, notre projet de fin d'étude consiste à déployer un système de GED open source, et la réalisation d'un OCR qui seront mis en place dans le but de répondre aux besoins du centre des archives de la BNA, afin de faciliter le travail quotidien des agents.

La structure de notre mémoire se présente comme ceci :

Chapitre 01 : Dans ce chapitre, on présentera le concept de GED , son architecture,les différents types de GED , et on comparera différentes solutions existantes sur le marché.

Chapitre 02 : par la suite, on parlera de la reconnaissance optique des caractères en présentant le processus de la chaîne de numérisation, les différents aspects de reconnaissance de l'écriture et les différentes techniques de la reconnaissance.

Chapitre 03 : sur ce chapitre on présentera la solution retenue et conception, qui repose sur la conception préliminaire en diagramme UML et la conception de notre OCR.

Chapitre 04 : Dans cette étape du travail ,on présentera une description des outils de travail utilisés,l'implémentation de notre OCR et l'intégration de la solution logicielle avec leurs différentes interfaces, on parlera aussi des résultats obtenus par notre OCR.

Enfin, une conclusion générale est présentée a la fin du mémoire qui repose sur les résultats obtenus, les problèmes rencontrés ainsi que les perspectives futures pour notre travail.

CHAPITRE 1

GESTION ELECTRONIQUE DE DOCUMENTS

La GED, ou en anglais EDM représente l'ensemble des moyens informatisés, utilisant d'une part des ressources matériels tel que les ordinateurs, les numériseurs optiques, disques magnétiques, serveurs etc. Et d'autre part des ressources logiciels tel qu'un logiciel documentaire ou un système de gestion de base de donnée. [Degeans, 1991].

Dans ce chapitre nous allons parler de l'intégration d'un système de gestion électronique de documents dans un réseau local ou externe d'une entreprise, qui permet principalement la dématérialisation de document papier en une version numérique, afin d'être en mesure de gérer ce document quel que soit son type. Cette intégration permet aussi d'exploiter et de partager l'information entre les différents services de l'entreprise, voir même sur des sites éloignés dans le but de simplifié le travail quotidien.

La finalité d'un système de GED est de permettre la conservation centralisée de la documentation d'une entreprise de telle manière que cette documentation puisse être retrouvée le plus sûrement et le plus rapidement possible pour être livrée dans les meilleurs délais possibles à son ou ses destinataires. C'est ainsi que l'on peut qualifier le système de gestion électronique de documents de véritable "mémoire de l'entreprise". Mais comme tout système informatique, un système de GED n'est qu'un outil sophistiqué mis au service de l'intelligence de l'être humain.

1.1 Le fonctionnement de la GED :

L'objectif principal de la GED est de faciliter et de réduire le temps de recherche d'un document pour une consultation sécurisée. Le fait de numériser tous les documents implique une réduction des

lots de papiers entre les structures de l'entreprise.

Le processus comprend plusieurs étapes :[FIGURE 1.1]

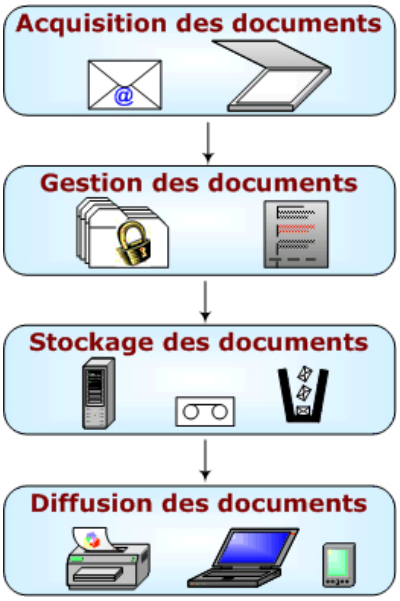

FIGURE 1.1. les étapes pour mettre en place une GED.

1.1.1 L'ACQUISITION DES DOCUMENTS :

Cette opération consiste à numériser un document papier avec des scanners. Différentes technologies sont disponibles pour extraire la partie textuelle de ces documents numérisés et ainsi enrichir ses métadonnées. Cette technologie se base sur des techniques d'OCR, de reconnaissance de codes à barres et d'ICR comportant par exemple des règles de vérifications linguistiques sur les mots reconnus. Les mots reconnus pourront ensuite être exploités par des fonctions de Text Mining qui permettront d'interpréter à des fins de classement thématique ou de pré-analyse les documents scannés.

1.1.2 LA GESTION DES DOCUMENTS :

Référencer :

Quelles que soient les sources venant alimenter le système de gestion des documents, l'outil de GED doit permettre d'aboutir à une version finale approuvée par les utilisateurs concernés. Le workflow lié à la validation d'un document est paramétrable et prendra en compte les droits d'accès et les profils des utilisateurs du système. il s'agit le plus souvent de la gestion du statut, de version et de la visibilité du document.

Classifier :

Cette opération consiste à ranger les documents dans un espace informatique accessible aux utilisateurs prévus. Le classement est réalisé automatiquement en s'appuyant sur les méta-données du document. La logique de classement déterminée (Alphabétique, numérique, chronologique…) sera traduite dans un plan de classement dans l'outil.

Indexer : Consiste à attribuer à un document une marque distinctive renseignant sur son contenu et permettant de le retrouver.
Elle se traduit par la recherche d'un symbole numérique ou nominal à partir de l'analyse du contenu du document. Ce symbole peut être :
– tiré d'une classification (indice). On parle alors d'indexation systématique.
– constitué d'un ou de plusieurs mots-clés. On parle alors d'indexation analytique ou d'indexation alphabétique matière." [Dewey, 2017]

1.1.3 LE STOCKAGE DE DONNÉES :

Plusieurs enjeux liés à cette étape : la notion de conservation visant à maintenir dans le temps la disponibilité d'un document. Cela induit une notion de durée indissociable du sort final du document (archivage prolongé, révision ou destruction). le stockage doit être adapté le mieux possible avec le volume des documents. Il doit aussi, en fonction de la fréquence de consultation et de l'importance des données, offrir un temps d'accès fiable, ainsi que des copies de sauvegarde sont nécessaire pour certain documents importants (contrat, facture, élément admis comme preuve) en cas de panne ou d'accident. [Cédric Ademain, 2017]

1.1.4 DIFFUSION DES DOCUMENTS :

Il faut garder a l'esprit que la finalité d'une GED est d'apporter une réelle ergonomie en termes de rapidité et de recherche, certains documents peuvent être désignés comme modifiables, d'autre doivent impérativement ne pas l'être, l'accès en lecture ou en écriture des informations, pour avoir un outil de GED sur mesure et vraiment efficace.

1.1.5 L'ARCHIVAGE :

Concerne particulièrement les documents qui doivent êtres légalement conservés, mais qui ne servent pas dans les opérations quotidiennes.

Dans la [FIGURE 1.2] on retrouve un schéma du processus d'archivage :

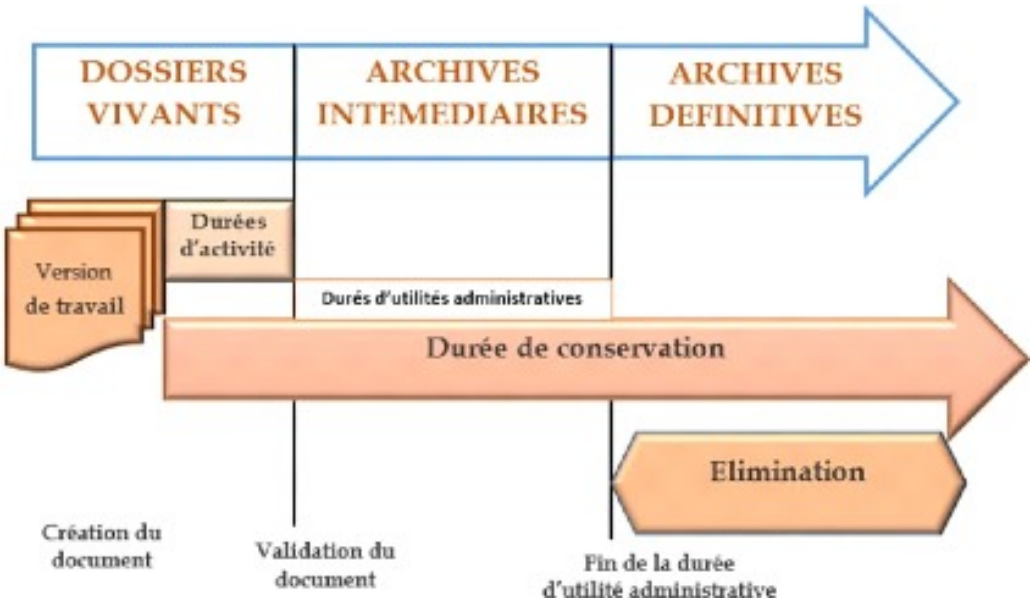

FIGURE 1.2. Les étapes d'archivage d'un document. [Freddy, 2015]

Le processus d'organisation et de gestion de l'ensemble des documents depuis leur création à leur sort final (Record management) identifie cinq étapes dans le cycle de vie d'un document qui sont résumées comme suit : [Chabin, 2007]

1. Création (ou capture) du document.
2. Validation du document.
3. Utilisation du document.

4. Fin de l'usage courant.
5. Échéance légale.

Selon les archivistes, le cycle de vie d'un document repose sur la théorie des trois âges , où chaque document passe par le processus suivant :

« Dossier vivant » on y retrouve les documents quotidiens conservés par la personne qui les a conçus.

« Archive intermédiaire » une fois la validation du document est faite, le document sera déplacé vers un autre endroit ou on ira le consulter en cas de besoin administrative, le document sera alors restituer à la demande.

« Archive définitive » elle présente la fin de la durée d'utilité administrative, ou le document sera archivé définitivement pour des raisons patrimoniales ou historiques.

Afin de relier la théorie des trois âges à celle du record management :

• Les dossiers vivants (ou archives actives) couvrant les étapes 1 à 3 du record management, prennent en compte les documents nécessaires au fonctionnement de l'entreprise.

• les archives intermédiaires (semi-statiques ou semi-actives) couvrant la période entre les étapes 4 et 5 du record management, ne sont plus utilisées couramment mais doivent être conservées de manière temporaire pour des raisons administratives ou juridiques.

• les archives définitives (statiques ou mortes) se positionnent après l'étape 5 du record management.

1.2 LES COMPOSANTS D'UNE GED :

Comme tout système informatique bien conçu, la présentation de ses éléments est nécessaire, voilà donc les quatre éléments constitutifs d'un système de gestion électronique de document. [Fallis, 2013].(c)

Le module de capture

Le module de saisie s'occupe de l'opération qui consiste a intégrer des données dans la mémoire d'un appareil électronique. Le module de saisie est constitué d'un numériseur et d'une carte de compression de données pour réduire le volume de stockage des documents numérisés.

Le module de pilotage

Le module de pilotage est constitué d'un ordinateur et d'un SGBD (classique ou relationnel). Sachant que le SGBD est un élément vital de la GED, car c'est lui qui assure la gestion et la recherche des documents numérisés indexés.

le module de pilotage comprend également un écran sur le quel tous les documents peuvent être visualisés, d'une part lors de leurs numérisation afin de vérifier la qualité des documents et d'autre part en phase de recherche afin de vérifier la pertinence du document.

Le module de stockage

Le module de stockage permet de mettre à disposition des utilisateurs des fichiers via un espace sécurisé. Ces documents stockés dans la GED sont ainsi diffusés de manière maîtrisée et contrôlée. Il s'agit des documents techniques ou commerciaux (les plans, les notices d'installation, les fiches techniques, les contrats…). L'accès a ces documents est assez rapide et intuitif.

Parmi les Supports physiques de stockage :
. DON, WORM, WMRM.
. Disques Durs tous types.
. Juke-Boxes.
. CD-RW, DVD-RW.

Le module d'impression

le module d'impression comprend une imprimante laser qui permet la restitution sous forme papier des documents recherchés et sélectionnés par l'utilisateur.

1.3 Les catégories de GED :

La GED aide à faire le passage des documents papiers en documents électroniques dont il existe plusieurs types, chacun est destiné à un usage bien précis : [Manuel, 2014]

La GED administrative : elle sert à numériser et classer les documents administratifs (factures, formulaires, réclamations…) souvent sous forme d'images, les stocker et conserver dans des supports numériques et enfin les diffuser via le réseau de l'entreprise (intranet par exemple) afin de permettre aux utilisateurs d'y accéder rapidement sans avoir à se déplacer ou à encombrer son bureau de dossiers physiques.

La GED Bureautique : elle sert à stocker les documents bureautiques (rapports, documentations…) dans leurs formats d'origines (WORD, EXCEL…) et elle regroupe un ensemble de progiciels qui permettent l'échange de ces documents, la consultation et parfois même leurs modifications est possible depuis n'importe quel poste de travail.

La GED documentaire : elle sert à faire la recherche documentaire qui est utilisée particulièrement dans des applications de bibliothèque, documentation scientifique…

Ce type de GED fait l'indexation des documents de divers types (texte, image…) pour faciliter la rechercher ainsi que l'accès à ces derniers.

1.4 Les avantages de la GED :

La GED offre une bonne organisation des documents numériques depuis leurs créations jusqu'a leurs archivages et parmi les fortes raisons de son utilisation dans les grandes entreprises :

- La sécurisation de stockage et d'accès aux documents.
- Le rassemblement des documents disséminés dans toute l'entreprise dans un unique dossier.
- Éviter les pertes de données en limitant au minimum la circulation des documents papiers.
- Recherche plus simple et plus rapide des documents.
- Faciliter l'accès instantané aux documents pour tous les utilisateurs quel que soit leurs lieux de travail.
- Un moteur de recherche fiable qui transmet les documents demandés, l'intérêt est de mettre tous les documents au même niveau d'un point de vue logique.

• En réduisant le temps de recherche de l'information, la GED permet d'améliorer la productivité en ne plus passant plus des heures a la recherche manuelle de documents, la productivité est ainsi performante et renforcée.

• La GED permet de réduire l'espace de stockage(plus besoins d'armoire ni de classeurs), cela implique un gain de coût, pas besoin de faire plusieurs documents pour les communiquer aux autres agents puisque désormais des copies seront accessibles à temps voulu.

1.5 ETUDE COMPARATIVE DES SOLUTIONS GED PROPOSÉES. :

Plusieurs solutions GED existent sur le marché informatique, afin de choisir celle qui convient à nos besoin on a effectué une étude comparative en se basant sur : le stockage, l'éditeur, l'environnement de travail et les fonctionnalités proposées par ces GED. Cette étude se résume dans le tableau suivant :

GED	Stockage	Editeur	Fonction	Environnement	autre caractéristiques
Report2web [Report2Web, 2017]	Aucune contrainte concernent le stockage , la data est stocké chez la société redwood-Software	Redwood-Software	Elle regroupe les principales fonctionnalités d'une Ged, visionneuse, revisionneuse et versions et recherche de document .	Peut importe votre environnement de travail, elle sera utilisable via votre navigateur.	Pas de problème de mise à jour à rechercher, tout est pris en charge par RedwofSoftware. Elle ne posséde pas d'OCR.
RecFind [RecFind, 2018]	Le Saas (Software as a service) permet de profiter d'un stockage de la data externalisé, un service pris en charge par Knowledgeone	Knowle-dgeone	Visionneuse, révisions, versions et recherche sont des fonctionnalités d'une Ged qui sont disponible	Windows, Mac, Linux	Pas besoin de gérer la mis à jour.Elle ne posséde pas d'OCR.
ImagingMadeSimple [ImagingMade-Simple, 2017]	La data de votre programme sera placées auprès de Imminet technologies.	Imminet Technologies	Elle regroupe toute les fonctionnalités d'une GED a savoir visionneuse, révisions, versions et recherche.	Accès facile via le web puisqu'il s'agit d'une application Saas.	Elle possède un système OCR qui est payant.
Tresorit [Tresorit, 2019]	C'est Tresorit qui gère le stockage avec une data externalisé.	Tresorit	Elle permet d'accéder a toute les fonctionnalités d'une GED a savoir visionneuse, revisions, versions et recherche.	Multiplatformes	Payant 12.5 dollar/mois par utilisateur.
Novaxel [NOVAXEL, 2019]	favorise les échanges via un mur d'activité, Chaque application métier dispose de son espace structuré dans l'outil.	Novaxel	Elle permet de numériser, classer, partager et archiver vos contenus. Son classement facilite la fonction recherche	elle s'implante simplement au sein de votre entreprise.	une communication sociale : chaque utilisateur, interne ou externe à l'organisation, peut interagir avec l'écosystème dès qu'on lui en donne l'accès et la permission.
Alfresco [Alfresco, 2019]	Les données de votre organisation seront stockés dans des serveurs à taille dynamique.	Alfresco Software	Création normalisée des documents, Sécurisation des originaux, indexation/ recherche optimisée, Classement des documents, Outils pour améliorer la diffusion.	Multiplateforme	En mode SaaS, elle reste personnalisable parce qu'elle est open source. Alfresco incarne une plateforme 2.0 qui inclut un système OCR payant.

TABLE 1.1 – Etude comparative des solutions GED.

A partir de ces résultats on a éliminé les GED qui ne répondent pas à nos besoins pour les raisons suivantes :

Report2Web : elle demande de la connexion pour son serveur de stockage, ce qui peut causer une perte de données, en plus les données sont stockées chez une autre société.

RecFind et ImagineMadeSimple : les données seront placées et pris en charge par leurs éditeurs.

Tresorit : elle demande un payement de 12.5 dollars/mois par utilisateur.

Novaxel : on ne peut pas adapter son interface selon les exigences du client.

Enfin de compte on a opté pour Alfresco, car en plus des fonctionnalités qu'elle propose (méta-données, types de documents, workflow documentaire, gestion de catégories, outils de collaboration, recherche, gestion de plusieurs bases indépendantes…), elle est gratuite et personnalisable selon les besoins de l'environnement de travail.

L'architecture d'Alfresco va être présentée dans la section suivante [Alfresco, 2019]

1.6 ARCHITECTURE D'ALFRESCO :

L'architecture d'Alfresco se constitue essentiellement des quatre couches suivantes :[FIGURE 1.3]

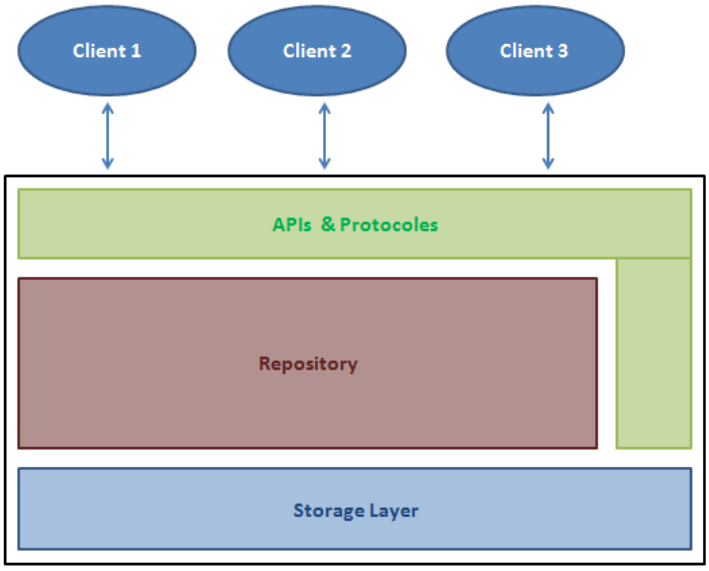

FIGURE 1.3. L'architecture d'Alfresco. [Alfresco, 2019]

1.6.1 COUCHE DE STOCKAGE ET MODÉLISATION DES DONNÉES(STORAGE LAYER)

Dans Alfresco un contenu est composé d'un document plus des méta-données, Alfresco utilise un système de fichier binaire (.bin) qui est formé d'une suite d'octets afin d'être indexé sur Lecene qui est une bibliothèque opensource qui permet d'indexer et de rechercher du texte, elle est utilisée dans certains moteurs de recherche. Cette couche est également composée d'un SGBD relationnel (par défaut PostgreSQL).[FIGURE 1.4]

Alfresco définit des modèles de données qui sont extensibles, personnalisables et souples, on peut ainsi créer :

• Des types de contenues, un contenu ne peut avoir plus qu'un type.
Exemple : type = contrat (méta-données : sujet, date d'envoi, valeur…).
• Aspect : qualifié un contenu, il peut avoir un ou plusieurs aspects exemple : aspect= client (méta-données : nom, référence, contrat…), donc la définition est réutilisable dans d'autre type.

• Les propriétés et les associations : peuvent être affectées pour définir un type ou un aspect.
Exemple : un client peut avoir plusieurs documents associés (Rapport, contrat, Émail…)

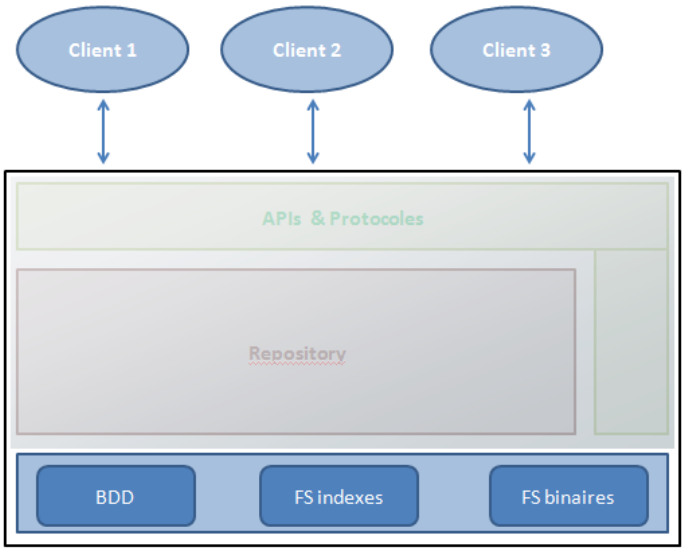

FIGURE 1.4. Couche de stockage. [Alfresco, 2019]

1.6.2 Couche de services (Repository) :

Les services permettent de faire le pilotage des contenus stockés sur l'entrepôt Alfresco.
Framework Spring permet de construire et définir l'infrastructure d'une application java, pour faciliter le développement et les tests, ainsi les services Alfresco reposent sur le Framework Spring qui inclut :[FIGURE 1.5]
• Services par des interfaces publiques (Repository Fondation Services).
• Composants qui implémentent ces services (Repository Implementation).
• Une configuration entièrement XMl.
• Transaction et sécurité (Permissions).

On cite principalement les services qui permettent le pilotage des contenues stockés dans l'entrepôt Alfresco :

1) Les actions et les règles :

Rechercher des actions sur les contenus (comme l'envoi de mail, déplacement de contenu…). Les règles s'appliquent à un espace car elles ajoutent de l'intelligence à cet espace et elles sont classées par :
• Un événement déclencheur (contenu entrant/sortant/modifié).
• Un ensemble de conditions (sur le contenu, le type …).
• Une action à appliqué.

2) Transformations et extractions des méta-données :

Transformation vers des formats différents Word=>PDF , Word=>Flash…
Avec une extraction automatique des documents (auteur, titre, description …)

3) Audit :

Permet le suivi et la traçabilité du contenu afin d'éviter un conflit d'édition sur le même contenu en même temps, l'audit du cycle de vie du contenu comprend la création la modification et la suppression.

4) WorkFlow :

Alfresco intègre un processus de Workflow qui permet d'assigner une tâche à une personne spécifique, on peut créer un workflow autonome, ou encore y joindre un fichier, l'utilité du workflow est de vous aider à suivre les tâches que vous et d'autre utilisateurs devez accomplir.

5) Sécurité :

Les permissions définissent les droits d'accès et d'opérations sur les contenus, les permissions sont agrégées pour définir des rôles par défaut : lecteur, éditeur, contributeur, collaborateur, coordinateur et les verrous pendant l'édition de contenu permettent de limiter les risques de conflit.
Plusieurs autres services pour la gestion de contenu sont disponible dont :
La recherche/l'authentification/Groupe et utilisateurs/Navigation/Cycle de vie...

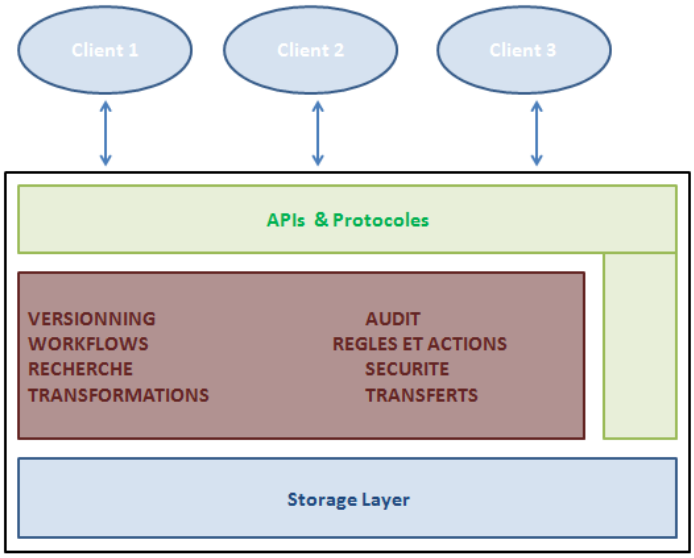

FIGURE 1.5. Couche de services (Repository). [Alfresco, 2019]

1.6.3 COUCHE APIS ET PROTOCOLES :

Les contenus d'alfresco peuvent être manipulés par des APIs afin d'en faire une plate-forme documentaire au sein du système d'information. Les applications clientes communiquent avec Alfresco via les APIs et les protocoles pour permettre l'interopérabilité d'Alfresco et les différentes applications du SI.[FIGURE 1.6]

L'entrepôt d'Alfresco est accessible par de différents protocoles, sans installation sur les postes des clients.
Les protocoles : FTP, webDAV, LDAP, NFS, CIFS, CMIS.
Les APIs : SOAP, web scripts, java API .

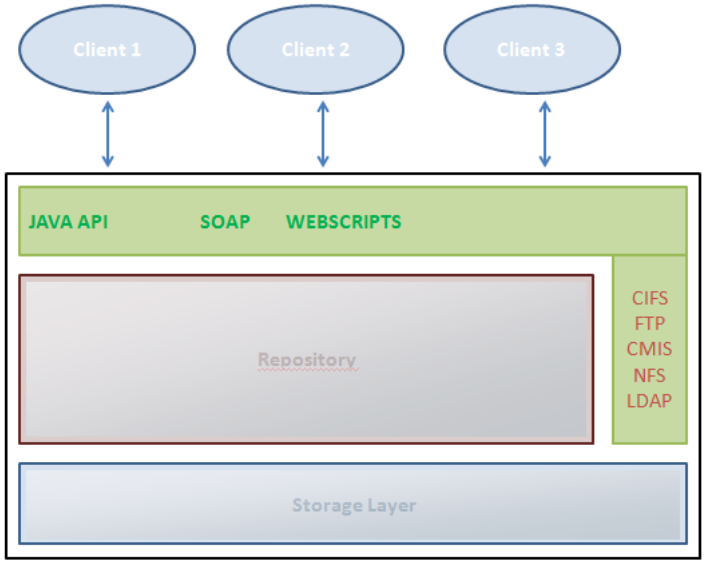

FIGURE 1.6. Couche de protocoles et APIs. [Alfresco, 2019]

1.6.4 COUCHE APPLICATIONS CLIENTES :

Différentes applications clientes pilotent les contenus de l'entrepôt Alfresco. Alfresco share est une interface web fournie par alfresco qui expose une partie des fonctionnalités du moteur, on peut créer nos propre interface web pour répondre a un besoin métier spécifique.[FIGURE 1.7]

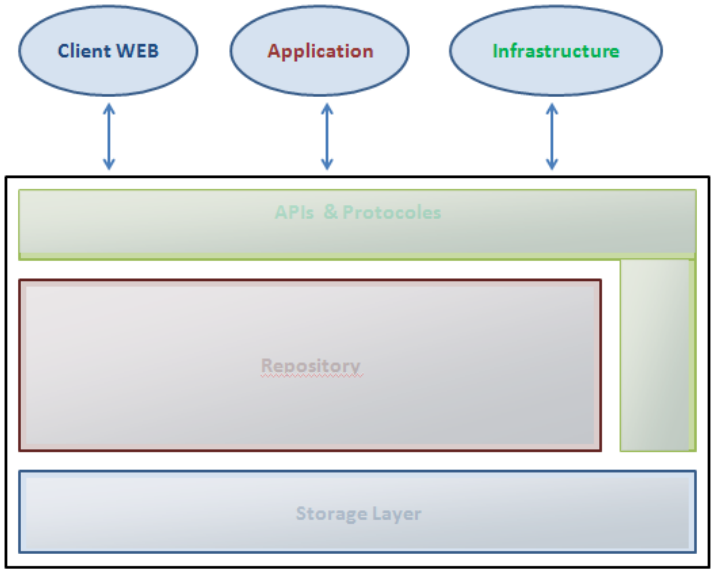

FIGURE 1.7. Couche applications clientes. [Alfresco, 2019]

1.7 Conclusion :

Les solutions GED proposent les outils adéquats pour la gestion, le stockage, la diffusion et l'archivage des documents ainsi que d'autres fonctionnalités qui facilitent pour les agents de l'entreprise l'exploitation quotidienne des documents. De même les solutions GED sont de plus en plus riches en nouvelles fonctionnalités tel que l'intégration de processus de reconnaissance optique des caractères. Nous avons aussi parlé des grandes fonctionnalités de notre solution GED Alfresco, ainsi que sa structure .

CHAPITRE 2

RECONNAISSANCE OPTIQUE DES CARACTÈRES

Dans ce chapitre, nous allons traiter la difficulté rencontrée par les employés du centre des archives, qui est la saisie manuelle des informations du chèque bancaire de chaque client tel que (N° de compte, nom prénom du client, date de comptabilisation, montant, N° de chèque) ainsi que les bordereaux de versement AADL (nom et prénom du client,N° de compte AADL,N° de bordereau,date de comptabilisation). Cette tâche demande énormément de temps et de concentration pour l'employé. pour cela une solution du domaine de l'intelligence artificielle est présentée : l'OCR OCR ou ROC en français qui a pour but de retranscrire le texte typographié dans une image. Il permet aussi de convertir différents types de documents tels que les papiers scannés, les fichiers PDF ou les photos numériques en fichiers modifiables et intérrogeables.

2.1 Généralité sur les images :

Le stockage d'information physique est devenu un problème de plus en plus répondu, car ce stockage exige certaines conditions tel que l'espace de conservation (local) qui doit être entretenu sous certains critères comme l'humidité et la température qui peuvent être très coûteux , une solution s'offre aux sociétés : l'acquisition numérique qui consiste à enregistrer les documents sous forme d'image numérique qui seront ensuite traités par des OCR pour extraire les informations qu'ils contiennent.

2.1.1 L'image numérique :

Une image numérique contient un nombre fini de pixels qui représente la dimension de l'image. Les pixels sont situés sur une grille régulière et à chaque pixel un niveau de gris ou de couleur lui est associé. Dans une image couleur (RVB) un pixel est codé sur trois octet, un octet pour chaque couleur

R : rouge, V : vert, B : bleu.

2.1.1.1 TYPES D'IMAGES NUMÉRIQUES :

On distingue 2 types d'images numériques, matricielles ou vectorielles qu'on définit de la manière suivante :[FIGURE 2.1]

Image matricielle (Bitmap) :

Est une image constituée d'une matrice de points (pixels) ou chaque point représente une couleur, donc la juxtaposition de points de couleurs nous donne notre image, un point important à souligné c'est le fait de perdre en qualité d'image si on l'agrandi beaucoup. Les formats les plus répondus à ce type d'image sont :BMP,PNG,GIF,JPEG...

Image vectorielle :

Utilise également la technique du pixel,sauf que leurs positions et leurs couleurs ne sont pas figées (mais plutôt calculées dynamiquement par le logiciel, on peut agrandir l'image autant de fois que l'on désire on ne perdera pas en qualité et les formats les plus répondu pour ce type d'image sont (odg,svg,ai)

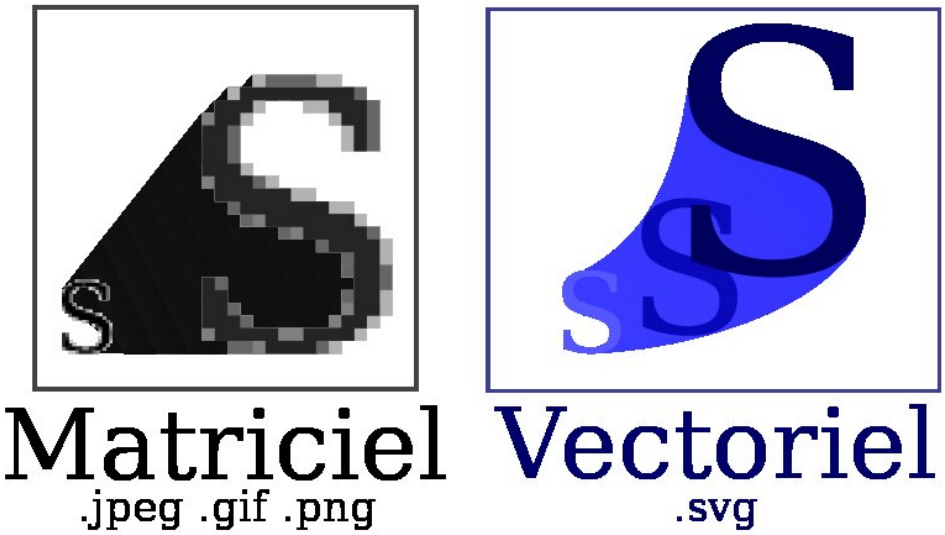

FIGURE 2.1. Image matricielle Vs image vectorielle .

2.2 CONCEPT DE L'OCR :

L'OCR est une technique qui permet à l'ordinateur la transformation d'un texte écrit sur papier en un texte éditable sous forme de fichier texte informatique, qui s'applique sur des documents de bonnes qualités tels que des livres, des rapports , des tickets ,des bordereaux, chèques etc

2.2.1 CHAÎNE DE NUMÉRISATION :

La reconnaissance de caractères est réalisée à l'aide de système appelé OCR, son but est d'être en mesure d'extraire du texte qui est une association de caractères appartenant à un alphabet formant des mots d'un vocabulaire donné, il est important de souligner qu'il doit être en plus capable d'extraire et d'identifier ces caractères sur plusieurs styles de polices , écriture manuscrite et différentes langues. La structure d'un système OCR comporte trois parties principales : l'acquisition et traitement d'image, reconnaissance du document et la vérification contextuelle (post-traitements)[FIGURE2.2]. [Baka and Fillali, 2016]

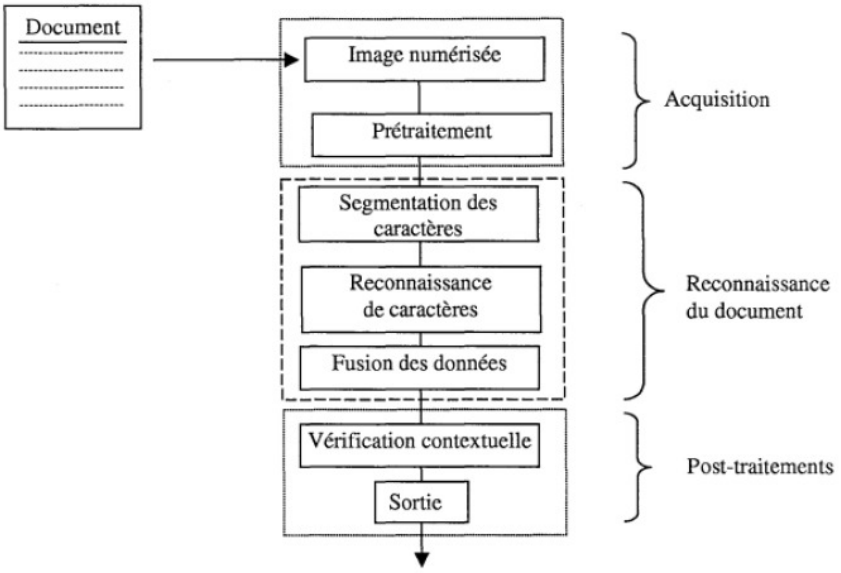

FIGURE 2.2. structure d'un système OCR .

2.2.1.1 ACQUISITION DES IMAGES DES DOCUMENTS :

Généralement les scanners sont utilisés pour sauvegarder le document en format numérique, les solutions utilisées pour l'OCR sont de 300dpi au minimum. Il est important de noter qu'à ce niveau

l'image n'est qu'une simple matrice de pixels qu'il faudra exploiter pour extraire les informations. Après la numérisation du document, on va lui effectuer un pré-traitement dont le but est d'avoir une image nettoyée sans bruit qui possède une lisibilité claire, afin de faciliter au système Ocr la reconnaissance et l'extraction des informations nécessaires, en s'appuyant sur les filtres nécessaires et en jouant avec les niveaux de gris on peut facilement favoriser la localisation ainsi que la reconnaissance des caractères. Le pré-traitement consiste également à réduire la masse d'information à traiter pour ne garder que l'essentiel. [VAN, 2001]

En image numérique toute information parasite qui pollue la transparence et la clarté d'une image représente un bruit numérique qui nécessitera un nettoyage. Il provient de l'éclairage des dispositifs optiques et électroniques du capteur. C'est un parasite qui représente certains défauts (poussière, petits nuages, baisse momentanée de l'intensité électrique sur les capteurs, ...etc.), il se traduit par des tâches de faibles dimensions et dont la distribution sur l'image est aléatoire. [Sarah, 2015]

Parmi les filtres nécessaires on cite :

Image au niveau de gris : Une image en couleur est représenté par une matrice de N lignes et P Colonnes (NxP) tel que chaque case représente un pixel, et pour chaque pixel un niveau de rouge, vert et bleu lui est attribué (R, V, B) cette valeur va de 0 à 255, et pour passer au niveau de gris chaque pixel est noir, blanc, ou à un niveau de gris entre les deux. Cela signifie que les trois composants ont la même valeur.
Pour obtenir ce résultat on applique généralement cette formule qui donne le niveau de gris [FIGURE2.3] en fonction des 3 composants :

Gris = (0.299*R+0.587*V+0.114*B) .

FIGURE 2.3. Image en niveau de gris

image binaire Une image binaire [FIGURE2.4] est une image numérique qui n'a que deux valeurs possibles pour chaque pixel.
Généralement, les deux couleurs utilisées pour une image binaire sont le noir et le blanc. La couleur utilisée pour les objets dans l'image est la couleur de premier plan tandis que le reste de l'image est la couleur de fond. Dans l'industrie de la numérisation de documents, on parle souvent de «bi-tonal».

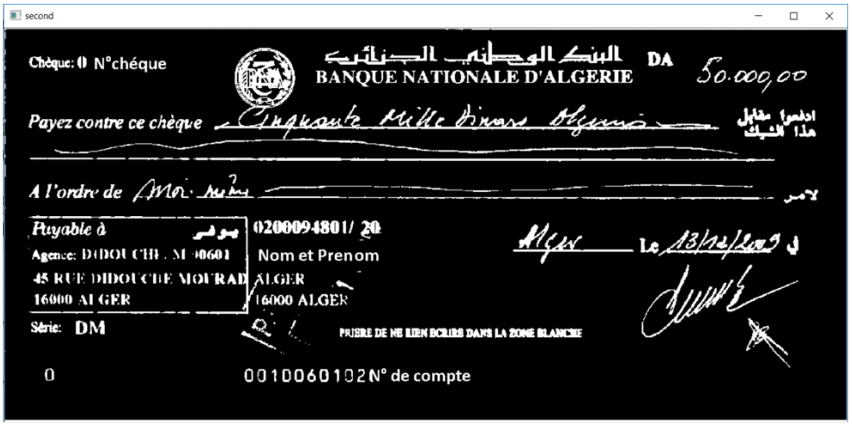

FIGURE 2.4. Image binaire.

2.2.2 RECONNAISSANCE DU DOCUMENT :

elle comporte 3 étapes essentielles :

2.2.2.1 LA SEGMENTATION :

Dans les systèmes de reconnaissances l'étape de segmentation est très importante car elle permet de diviser l'image en différentes imageâtes (mots, caractère ou graphèmes) qui font partie d'un alphabet prédéfini selon l'objectif visé. Le but est d'identifier les zones contenant l'information (texte typographié ou manuscrits, signature, logos, code barre…). [Susmith, 2019]

Région d'image (ROI) : Le terme région en traitement d'image constitue à identifier les pixels adjacents qui ont la même couleur (couleur du fond) ,ce qui représente la zone du caractère.

Contours : La détection de contours se fait en repérant les pixels d'une image Bitmap dont les niveaux de gris correspondent à un changement brutal de couleur . Dans une image numérique, les contours se situent entre les pixels appartenant à des régions ayant des intensités moyennes différentes ; il s'agit de contours de type « saut d'amplitude ». Un contour peut également correspondre à une variation locale d'intensité présentant un maximum ou un minimum ; il s'agit alors de contour « en toit ».

Parmi les algorithmes pour la détection des contours on a utilisé celui d'OpenCV findcountours [FIGURE 2.5]

```
#find contours
im2,ctrs, hier = cv2.findContours(gsblur.copy(), cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)
```

FIGURE 2.5. Fonction FindCountours.

Plusieurs techniques nous sont proposées on compte parmi elles :

RLSA (Run length smothing algorithme) :

Le principe de segmentation est de relier les pixels noirs entre eux si leurs distance est inférieures à un certain seuil (sachant que ce seuil est utilisé pour regroupé les pixels et peut être déterminé dans la phase de pré analyse du document). Cette technique nous permet de segmenter en lettre, en mots, en paragraphes selon le seuillage utilisé ,cet algorithme n'est pas robuste mais marche très bien pour

des rotations de quelque degré. [Augereau, 2013]

Espace blanc :

Cette technique est basée sur l'analyse du fond du documents et de retrouver les plus grands rectangles blancs et espaces blancs afin de mettre les zones particulières en évidence pour définir la structure du document . [Augereau, 2013]

Texture :

Plusieurs approches ont vu le jour, on utilise parmi elles le réseau de neurone pour extraire un ensemble de symbole textuelles, les caractéristique des textures sont obtenues en calculant le produit de convolution obtenu par l'image d'entrée.
On propose également une technique basée sur l'analyse multi résolution pour la classification des pixels du document à la fin on obtient la segmentation de chaque pixel de l'image.
Une autre technique considère que les zones de texte sont différentes des textures des zones de fond, de graphique…, le principe est d'appliquer des filtres de Gabor sur l'ensemble des pixels, cet méthode est robuste à la détection des textes typographiques comme le texte manuscrit. [Augereau, 2013]

2.2.2.2 CLASSIFICATION :

L'approche statistique : contient un ensemble de méthodes basées sur le modèle probabiliste qui nous permet de définir l'appartenance d'une forme à une classe avec un minimum de risque d'erreur.

La classification Bayésienne :
On va utilisé des mesures faites sur le caractère pour déduire les probabilités d'appartenance de la forme aux classes prédéfinies, afin de faire le meilleur choix qui maximise la probabilité d'appartenance à une classe.la classification Bayésienne est basée sur les probabilités conditionnelles (la règle de Bayes),sachant que les probabilités conditionnelles s'expriment par : la probabilités qu'un événement se produise sachant qu'un autre événement s'est déjà produit . [Abdelhak, 2011]

Théorème de Bayes P(A|B)=P(B|A).P(A)/P(B)

Le terme p(A|B) se lit : la probabilité que l'événement A se produise sachant que l'événement B s'est déjà réalisé.Autrement dit : la probabilité que le caractère A appartient a une classe, sachant que la classe du caractère B est déjà prédéfinie

L'approche stochastique : L'approche stochastique est essentiellement utilisé pour la reconnaissance manuscrite, elle utilise un modèle pour la reconnaissance qui considère la forme d'un caractère comme un signal continu observable dans le temps, la comparaison consiste à chercher une forte adéquation dans le graphe avec une suite d'éléments observés dans la chaîne d'entrée . Les méthodes les plus utilisées pour cette approche sont les modèles de Markov cachés(MMC) .
C'est un processus stochastique à temps discret dans le quel l'évolution future dépend de l'état présent et du hasard. L'image est segmentée en mots d'images tel que chaque segment est transmis à un module chargé d'estimer la probabilité pour que chaque segment apparaît quand l'état correspond de la chaîne de Markov et un certain état. [Chevalier, 2004]

L'approche structurelle : Elle repose sur la structure physique des caractères de manière générale elle permet de faire la description de forme complexe à partir de formes élémentaires, les caractéristiques sont directement extraites des données en entrée du système. [Soua, 2016]

2.2.2.3 Post-traitement :

Le post-traitement est le processus ultérieur de la classification, le but principal est d'améliorer le taux de reconnaissance en faisant des corrections pour lever l'ambiguïté dans la reconnaissance de certain caractères, ces corrections peuvent se présenter sous forme de correction morphologique, en jouant sur la largeur moyenne de chaque caractère, ou bien orthographique à l'aide de dictionnaire et thésaurus. On peut également avoir les connections sur les connaissances linguistiques au niveau :

Lexical :

On retient que les mots reconnus du dictionnaire et on rejette les lettres ambiguës

Syntaxique et sémantique :

afin de choisir le mot correspondant parmi ceux qui ont été retenu dans l'étape précédente.

2.3 Différents aspect de reconnaissance de l'écriture :

Il n'existe pas de système d'OCR universel qui permet la reconnaissance de n'importe quel caractère en différentes fontes, plusieurs critères rentrent en jeux, tout dépend de l'application finale visée et du type de données en entrée. Cependant on peut distinguer deux types de reconnaissance ayant chacun ses outils propres d'acquisition et ses algorithmes correspondants :

2.3.1 Reconnaissance en ligne :

La reconnaissance des caractères se fait en temps réel, il s'agit de reconnaître l'écriture au fur et a mesure de son tracé, la reconnaissance en ligne est généralement utilisée à l'écriture manuscrite, et elle représente un avantage majeur car la correction et la modification se fait de manière interactive. En revanche, elle nécessite un matériel beaucoup plus coûteux(tablette avec stylet).

2.3.2 Reconnaissance hors ligne

Il s'agit de reconnaître des textes à partir de documents écrits au préalable qui seront numérisés à l'aide d'un scanner, ce mode se rapproche du mode de reconnaissance visuelle, les informations recueillis se présentent sous une image de pixels. La reconnaissance hors ligne est plus difficile à cause de l'absence d'informations temporelles. Et c'est la technique utilisée dans notre projet.

2.4 Les différentes techniques de reconnaissance :

Dans un système OCR l'étape de reconnaissance vise à identifier les caractères non reconnus et les attribuer à leurs classes d'appartenances. Plusieurs on trouvé la mise en correspondance des pixels ou des caractéristiques, les réseaux de neurones, les SVM et les modèles de Markov cachés.

2.4.1 Machine a vecteur de supports (SVM)

Dans l'apprentissage machine, les SVMs sont des modèles d'apprentissage supervisés avec des algorithmes d'apprentissage associés qui analysent les données utilisées pour la classification et la régression.

Un modèle SVM est une représentation des exemples sous forme de points dans l'espace, mappés de telle sorte que les exemples de catégories distinctes soient divisés par un écart clair et aussi large que possible.
En plus d'effectuer une classification linéaire, les SVMs peuvent efficacement effectuer une classification non linéaire, en mappant implicitement leurs entrées dans des espaces de fonctions de grande dimension.

À partir d'un ensemble d'exemples d'apprentissage, chacun d'eux appartenant à l'une ou à l'autre des deux catégories, un algorithme d'apprentissage SVM crée un modèle attribuant de nouveaux exemples à une catégorie ou à l'autre, ce qui en fait un classifieur linéaire binaire non probabiliste. [Kushal, 2018]

Le classifieur SVM linéaire trace une ligne droite entre deux classes(Cette ligne sera sélectionnée par l'algorithme LSVM , qui non seulement sépare les deux classes mais reste aussi éloignée que possible

des échantillons les plus proches). Tous les points de données qui se trouvent d'un côté de la ligne seront étiquetés comme une classe et tous les points qui se trouveront de l'autre côté seront étiquetés comme la seconde.[FIGURE2.6] [Cory, 2019]

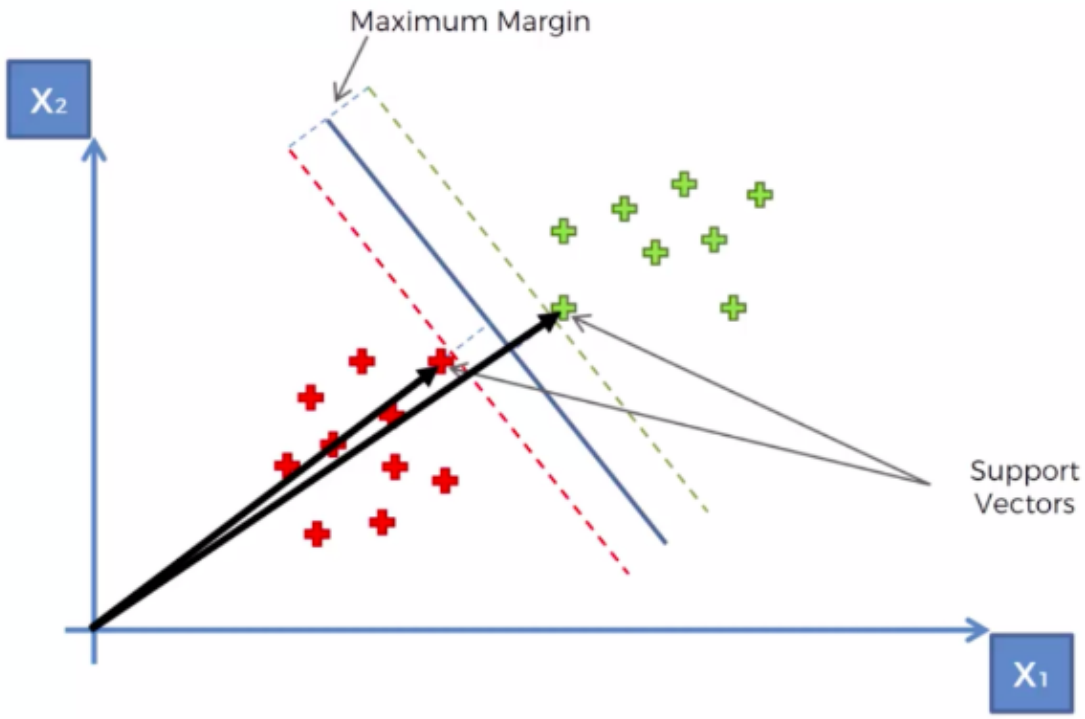

FIGURE 2.6. Graphe SVM.

2.4.2 RÉSEAUX DE NEURONES :

Les réseaux de neurones artificiels sont des algorithmes d'apprentissage et d'optimisation basés sur des concepts inspirés du fonctionnement d'un cerveau humain. [laurant Bastien, 2019]

Ils sont généralement constitués des cinq composants suivants :

1. Un graphe dirigé qui représente la topologie du réseau.
2. Une variable d'état associée à chaque nœud.
3. Un poids associé à chaque connexion.
4. Un biais associé à chaque nœud.
5. Une fonction de transfert pour chaque nœud qui détermine l'état d'un nœud en fonction des poids de ses liens entrants, son biais et les états des nœuds reliés à ce nœud. Cette fonction est habituellement

une fonction Relu.

2.4.2.1 CARACTÉRISTIQUE DE RÉSEAU DE NEURONE

La caractéristique des réseaux de neurones est leurs capacités à apprendre (par exemple à reconnaître une lettre, un son...), mais cette connaissance n'est pas acquise dès le départ. La plupart des réseaux de neurones apprennent par l'algorithme d'apprentissage. [laurant Bastien, 2019]

Il y a deux algorithmes principaux :

• **L'apprentissage supervisé :**
les résultats corrects (c'est-à-dire les valeurs que l'on désire que le réseau obtienne en sortie) sont fournis au réseau, si bien que celui-ci peut ajuster ses poids de connexions pour les obtenir. Après l'apprentissage, le réseau est testé en lui donnant seulement les valeurs d'entrée mais pas les sorties désirées, et en regardant si le résultat obtenu est proche du résultat désiré.

• **L'apprentissage non supervisé :**
on ne fournit pas au réseau les sorties que l'on désire obtenir. On le laisse évoluer librement jusqu'à ce qu'il se stabilise.

2.4.2.2 TOPOLOGIE DES RÉSEAUX DE NEURONES

On peut classer les réseaux de neurones en deux grandes catégories, selon la dépendance de l'évolution de ceux-ci en fonction explicite du temps.

•**Les réseaux statiques ou réseau à couche (FEED FORWARD) :**

C'est le cas de réseaux statiques, ou le temps n'est pas un paramètre significatif. En d'autres termes, la modification d'entrée n'entraîne que la modification stable de la sortie, mais elle n'entraîne pas le retour de l'information de cette entrée.
Les réseaux statiques (FEED FORWARD) sont des réseaux à couches, constitués d'une couche d'entrée, une couche de sortie et entre les deux au moins une couche composée de nombreux éléments de traitements non linéaires, appelée couches cachées.
Les signaux des entrées se propagent de la première couche à la couche de sortie en passant par les couches cachées, Il n'y a pas des communications entre les unités de la même couche, d'où le nom de feedforwarde. Les liens dirigés connectant les neurones sont appelés les inter-connections.
On distingue des réseaux à deux couches tel que le perceptron et l'adeline (adaptative linaire neurone)

qui sont caractérisés par :
•la simplicité de réglage d'apprentissage.
•la facilité de détermination de l'influence d'un neurone d'entrée sur l'erreur d'un neurone de sortie d'en déduire les modifications à apporter au lien qui les relie.
•La limitation au calcul de fonction très simple. Ces réseaux ne pouvaient résoudre que des problèmes simples de classification. Pour des problèmes complexes, une solution consiste à organiser le réseau en plusieurs couches. [laurant Bastien, 2019]

• **Les réseaux dynamiques (récurrents) :**

Comme leurs noms l'indique, contiennent des dé-bouclages partiels ou totaux entre neurones, ils représentent donc une évolution dépendante du temps.
Il faut bien distinguer la dépendance théorique, pour laquelle l'état du réseau à un certain instant dépend de son état à l'instant ou aux instants précédents, du temps nécessaire à obtenir une réponse, dans le cas d'une réalisation matérielle ou d'une simulation sur ordinateur.

Les critères motivant les choix d'un type de réseau sont la simplicité de mise en œuvre et l'efficacité des algorithmes d'adaptation appelés à répondre aux performances désirées du système, quelle que soit sa complexité.
L'opérateur non linéaire réalisé par un réseau, bouclé ou non, dépend des valeurs des coefficients de pondération du réseau.
Pour qu'un réseau effectue une tâche donnée, il faut donc ajuster la valeur de ses coefficients.
Une tâche est définie par un ensemble d'exemples, ou couples (valeurs des entrées et valeurs des sorties désirées correspondantes), tels les couples (forme classe) en classification, ou les couples (commande sortie mesurée du processus) en modélisation ces couples constituent l'ensemble d'apprentissage. [laurant Bastien, 2019]

2.4.2.3 Le CNN (réseau de neurone convolutionnel)

Est l'une des techniques d'apprentissages supervisées en profondeur les plus puissantes, sa structure finale est très similaire à celle des réseaux de neurones réguliers(RegularNets), nous utilisons une fonction de perte (rétentropie ou softmax) un optimiseur (adam optimizer), et un ensemble de couches. Les couches les plus importantes sont : les couches convolutives, les couches de regroupement et la couche de connexion.
Les CNNs sont principalement utilisés pour la classification des images bien qu'on peut trouver d'autre domaine d'application.Dans la figure[FIGURE2.7] on présente une architecture standard d'un CNN.

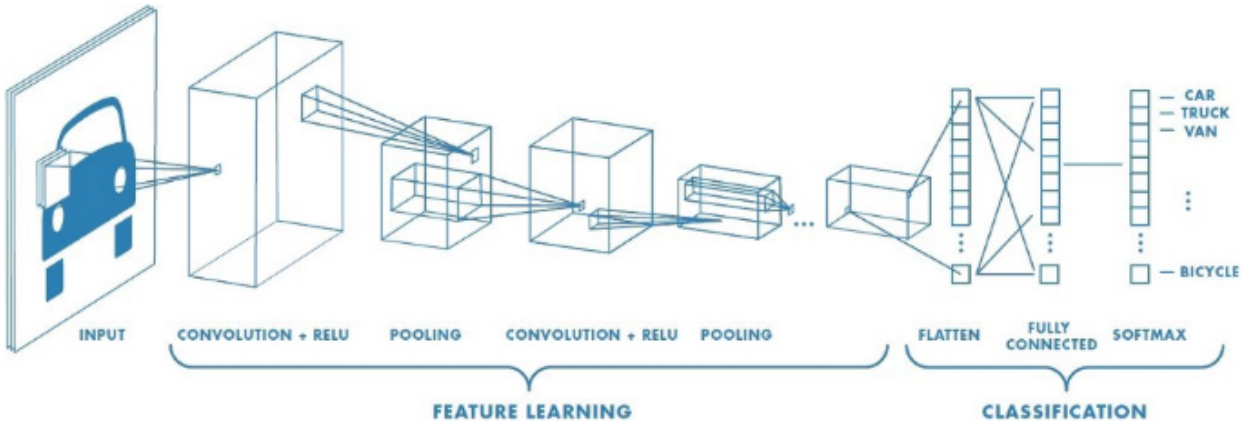

FIGURE 2.7. Architecture standard d'un réseau de neurone convolutionnel. [Orhan, 2018]

C'est un réseau feed-forward capable d'extraire les attributs topo-logiques d'une image. Sa première couche cachée est utilisée pour extraire les attributs caractéristiques de l'image d'entrée et sa dernière couche est utilisée pour la classification. Les deux premières couches du réseau peuvent être considérées comme des extracteurs d'attributs caractéristiques.

les principales couches du CNN sont définies comme ceci :

La couche convolutionnelle :

C'est la toute première couche du CNN, étant donné que les pixels ne sont associés qu'aux pixels adjacents et proches, cette couche consiste à réduire et filtrer l'image avec un filtre de pixels plus petit sans perdre la relation entre les pixels, lorsque nous appliquons une convolution a une image (5x5) en utilisant un filtre (3x3) avec une foulée de (1x1 décalage de 1 pixel a chaque étape), nous aurons une sortie de (3x3),voir [FIGURE2.8] la complexité est réduite a 64%.

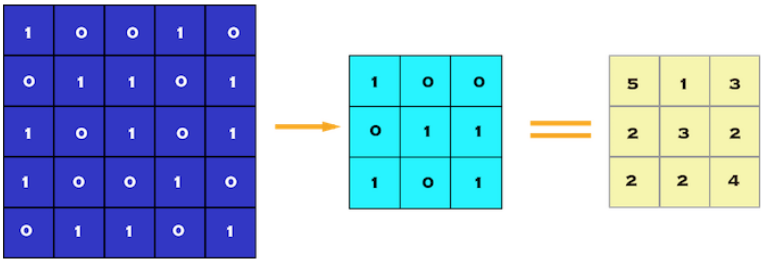

FIGURE 2.8. Convolution d'une image de 5 x 5 pixels avec un filtre de 3 x 3 pixels (foulée = 1 x 1 pixel). [Orhan, 2018]

La couche de pooling (regroupement) :

Un autre outil très puissant utilisé par le CNN s'appelle le pooling. Le Pooling est une méthode permettant de prendre une large image et d'en réduire la taille spatiale tout en préservant les informations les plus importantes qu'elle contient. Après chaque couche convolutionelle on utilise la couche de regroupement afin de réduire le nombre de paramètre en sélectionnant les valeurs maximales a l'intérieur de ces pixels, car c'est l'une des technique les plus courantes, elle permet de résoudre le problème de sur-ajustement.
Après avoir effectué le pooling l'image obtenu représente le quart du nombre de pixel de l'image de départ. Maxpooling est l'une des techniques de pooling les plus courantes,et peut être démontrée comme suit :[FIGURE2.9]

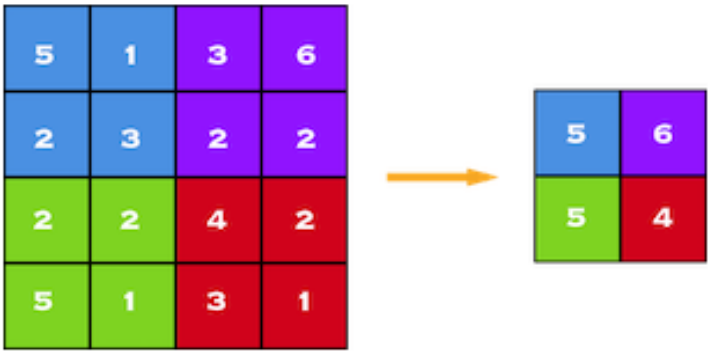

FIGURE 2.9. Max Pooling by 2 x 2. [Orhan, 2018]

Un ensemble de couches entièrement connectées :

Les couches entièrement connectées sont les principaux blocs de construction des réseaux de neurones traditionnels (réseau régulier) afin de classer nos images.
Trouvons ces dernières étapes après réduction de notre complexité spatio-temporelle grâce aux couches de convolution et pooling, ou chaque paramètre est lié pour déterminer la relation et l'effet réels de chaque paramètre sur les étiquettes.[FIGURE2.10]

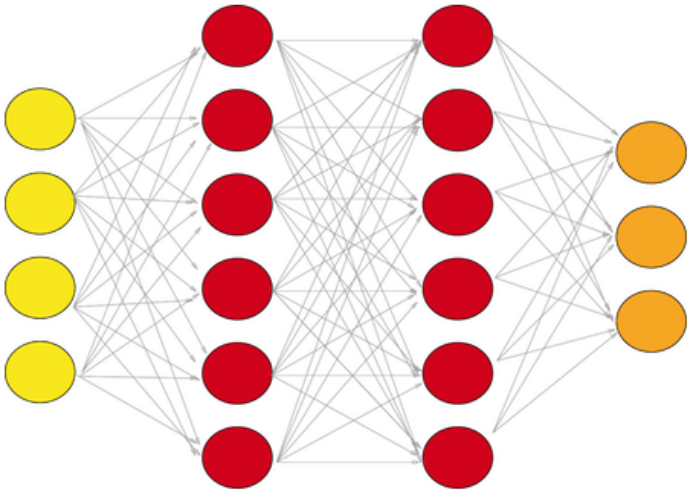

FIGURE 2.10. Une couche entièrement connectée avec deux couches cachées. [Orhan, 2018]

2.5 Conclusion :

Dans ce chapitre nous avons parlé de la reconnaissance de caractère optique en détaillant les différentes étapes de la chaîne de numérisation, les différentes techniques de reconnaissance ,ainsi que les différents types de réseaux de neurones, parmi eux la solution retenue qui est le (CNN) car il est le plus utilisé dans le traitement d'images ainsi que la multiclassification , en expliquant ses différentes couches.

Dans le prochain chapitre on parlera de la Conception de notre CNN suivie par celle de l'OCR et les différents diagrammes UML.

CHAPITRE 3

CONCEPTION

Dans ce chapitre, on va présenter en premier lieu la phase de modélisation, qui a sert à décrire de manière concrète et compréhensible l'état et le fonctionnement de notre système afin de le modéliser avec des diagrammes UML. En deuxième lieu, on va détailler l'implémentation de notre CNN ainsi que le processus de réalisation de notre OCR.

3.1 Diagramme UML

Afin de modélisé le fonctionnement de notre système on a utilisé le langage UML qui est un langage de modélisation graphique à base de pictogrammes conçu pour fournir une méthode normalisée pour visualiser la conception d'un système. Il est connu par sa simplicité ainsi que sa flexibilité marquante.

3.1.1 Diagramme de cas d'utilisation :

Il permet d'identifier les possibilités d'interaction entre le système et les acteurs (intervenants extérieurs au système), c'est-à-dire toutes les fonctionnalités fournies par le système. Il est l'un des diagrammes les plus structurants dans l'analyse d'un système.

Identification des acteurs :

L'objectif de l'ensemble des cas d'utilisation c'est de décrire les exigences du fonctionnement du système, chaque cas d'utilisation correspond a une fonction métier de notre système.[TABLE3.1]

Liste des acteurs	Rôle de l'acteur
Admin	Gère les utilisateurs ainsi que toutes les fonctionnalités de la GED.
User	S'occupe de la gestion des documents et des tâches.

TABLE 3.1 – La liste des acteurs.

diagramme de cas d'utilisation général :

Le diagramme de cas d'utilisation est utilisé pour détecter et consigner les besoins des utilisateurs, de façon plus générale notre cas d'utilisation exprime sous format textuelle la manière dont un acteur utilise le système afin d'effectuer son travail. Voilà un scénario qui représente le fonctionnement globale de nôtre système.[FIGURE3.1]

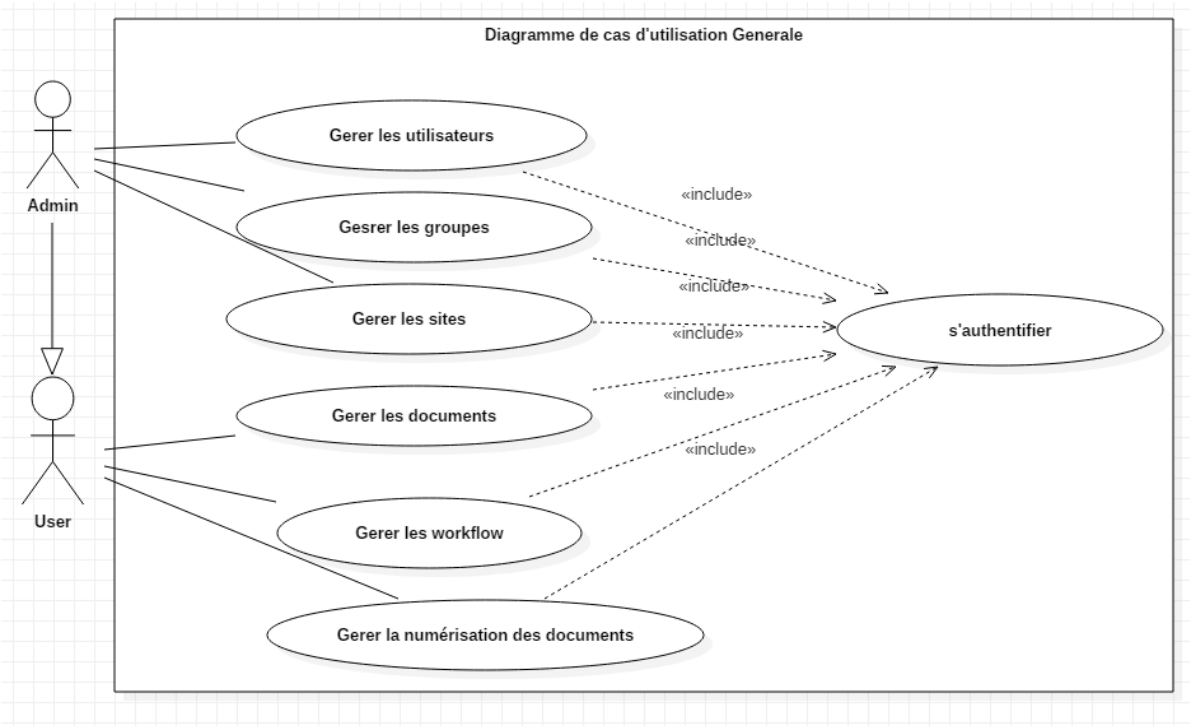

FIGURE 3.1. diagramme de cas d'utilisation générale.

Notre diagramme de cas d'utilisation général décrit les fonctionnalités globales de notre système. Il possède deux acteurs principaux (Admin, User), l'utilisateur s'occupe de la gestion des documents et celle des workflows ainsi que la gestion de la numérisation des documents. l'admin de sa part hérite des fonctionnalités de l'utilisateur et en plus il gère les utilisateurs, les sites ainsi que les groupes.
Les deux acteurs ne peuvent pas effectuer leurs tâches s'ils ne s'authentifient pas.

Diagramme de cas d'utilisation « gestion des utilisateurs »

Dans la [FIGURE3.2] on va présenter l'ensemble des fonctionnalités à effectuer lors de la gestion des utilisateurs.

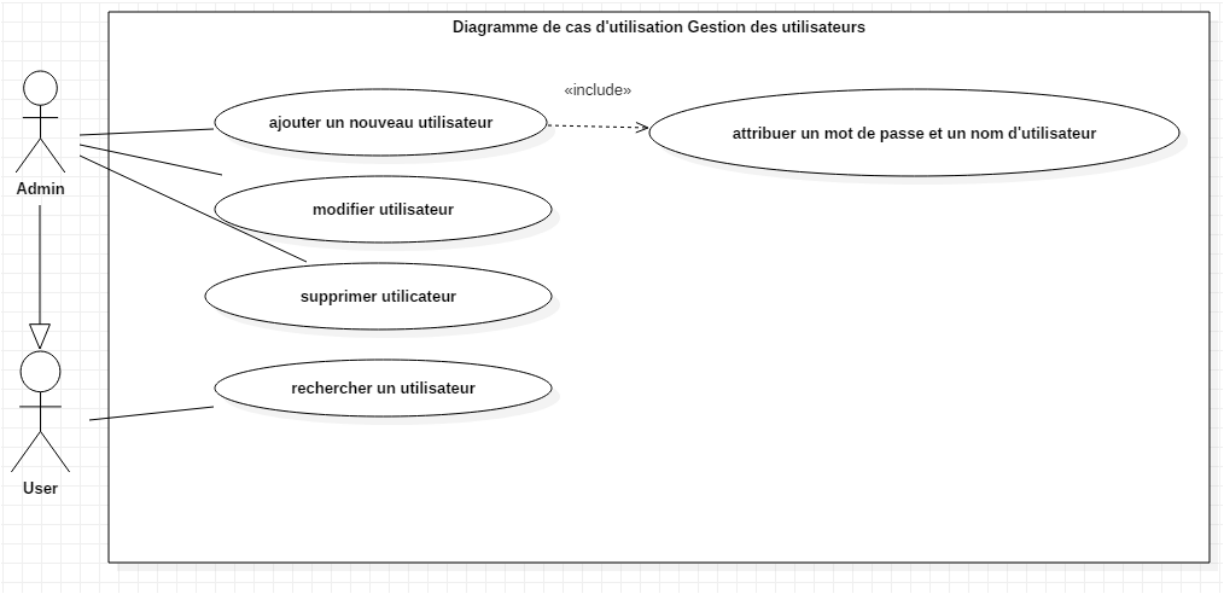

FIGURE 3.2. Diagramme de cas d'utilisation« Gestion des utilisateurs ».

Ce diagramme décrit clairement la gestion des utilisateurs, ou le USER peut seulement effectuer une recherche alors que l'admin peut soit ajouter, modifier ou supprimer un utilisateur en plus de la recherche.
Lors de l'ajout l'admin doit attribuer à l'utilisateur un mot de passe ainsi qu'un nom d'utilisateur unique.

Diagramme de cas d'utilisation « Gestion des sites » :

Dans la [FIGURE3.3] on va présenter l'ensemble des fonctionnalités à effectuer lors de la gestion des sites .

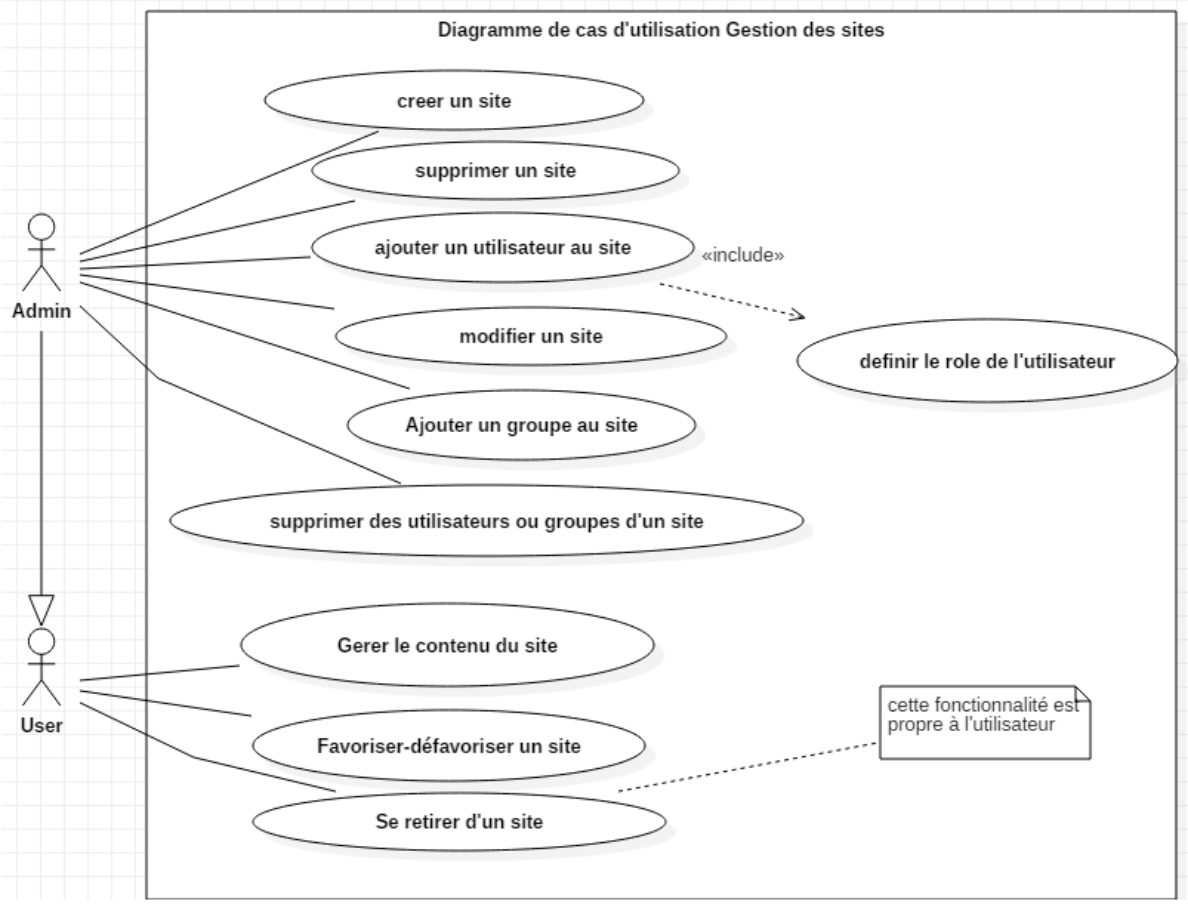

FIGURE 3.3. Diagramme de cas d'utilisation« Gestion des sites ».

Pour ce qui de la gestion des sites, l'utilisateur s'occupe de la gestion du contenu du site, favorise/défavorise un site ou se retirer du site (cette fonctionnalité est propre à l'utilisateur).
L'admin hérite de ces fonctionnalités en plus il peut créer, supprimer ou modifier un site, ajouter un utilisateur à un site, ajouter un groupe à un site ou supprimer des utilisateurs/groupes d'un site.

Diagramme de cas d'utilisation « Gestion des groupes » :

Dans la [FIGURE3.4] on va présenter l'ensemble des fonctionnalités à effectuer lors de la gestion des groupes.

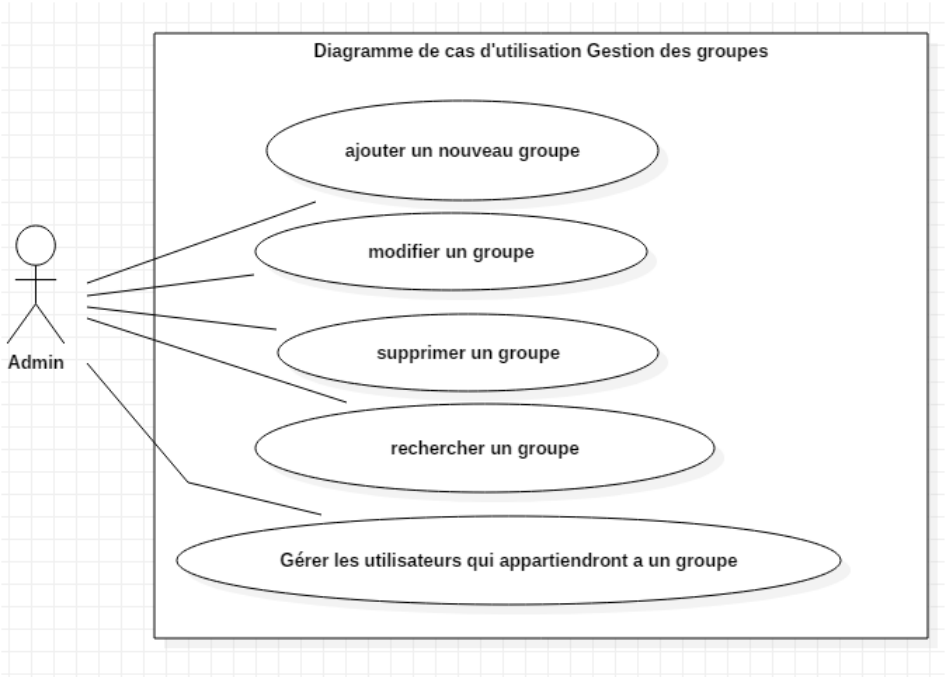

FIGURE 3.4. Diagramme de cas d'utilisation« Gestion des groupes ».

Dans ce diagramme l'admin est le seul qui peut effectuer des tâches, il peut ajouter, modifier, supprimer ou rechercher un groupe et aussi gérer les utilisateurs qui appartiendront à un groupe.

diagramme de cas d'utilisation « Gestion des workflow » :

Dans la [FIGURE3.5] on va présenter l'ensemble des fonctionnalités à effectuer lors de la gestion des tâches.

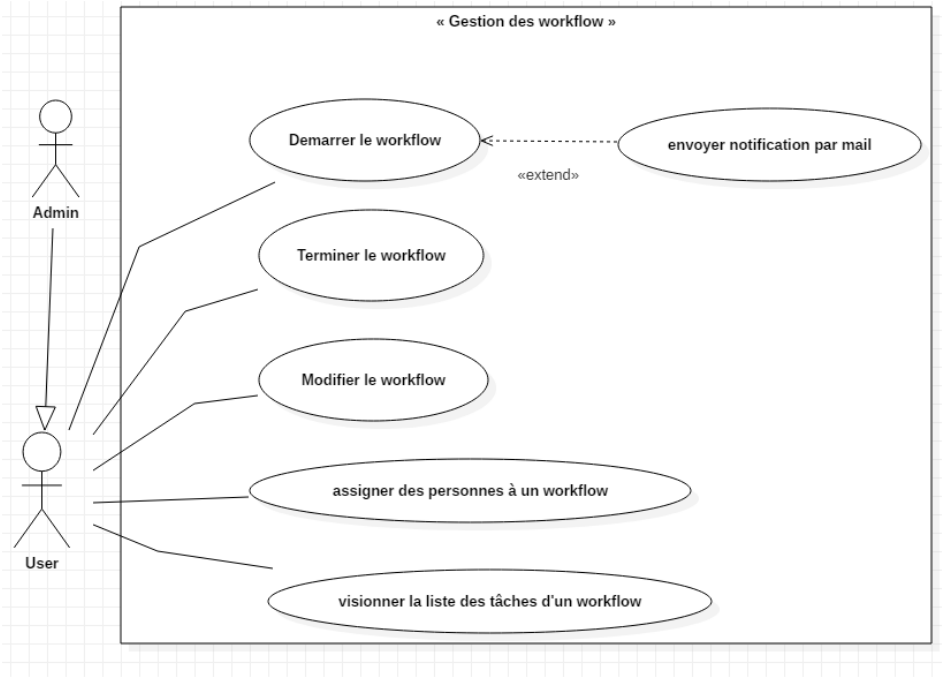

FIGURE 3.5. diagramme de cas d'utilisation « Gestion des workflow ».

Pour la gestion des workflows, l'utilisateur peut démarrer, terminer ou modifier un workflow, assigner des personnes à un workflow et visionner la liste des tâches d'un workflow.
Dans ce diagramme l'admin possède les même fonctionnalités que l'utilisateur.

diagramme de cas d'utilisation « Gestion des documents» :

Dans la [FIGURE3.6] on va présenter l'ensemble des fonctionnalités à effectuer lors de la gestion des documents.

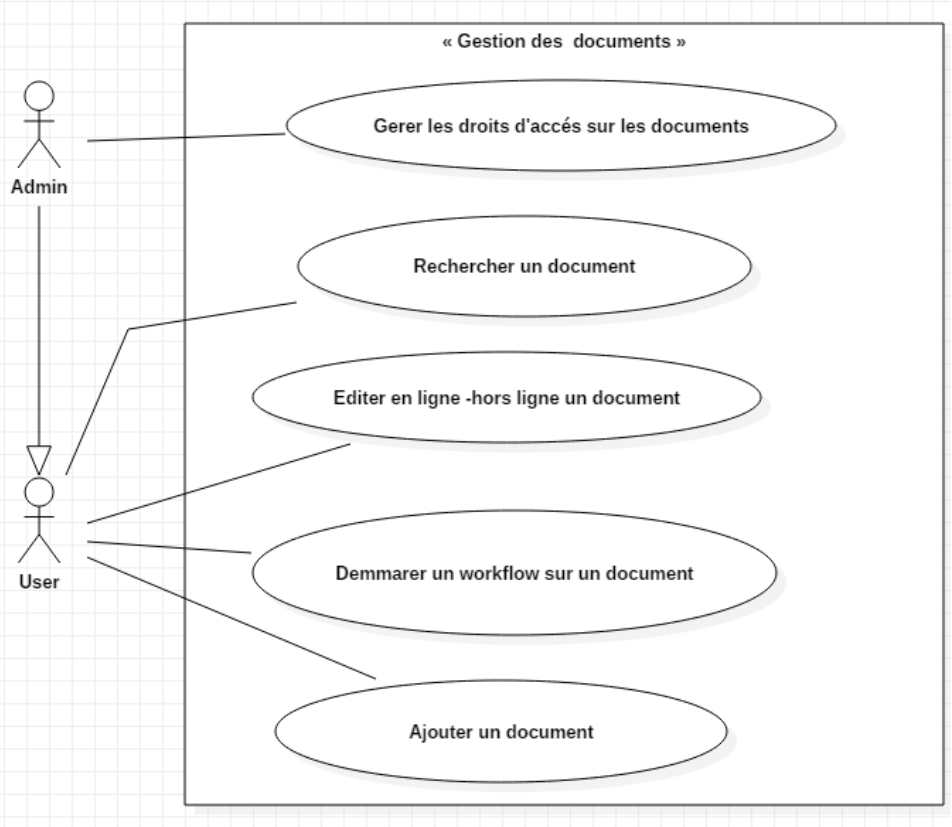

FIGURE 3.6. diagramme de cas d'utilisation « Gestion des documents».

Pour le diagramme de cas d'utilisation de la gestion des documents, l'utilisateur peut rechercher ou ajouter un document, éditer en ligne ou hors ligne un document, démarrer un workflow sur un document.
L'admin de sa part hérite des fonctionnalités de l'utilisateur et en plus il gère les droits d'accès sur les documents.

diagramme de cas d'utilisation « Gestion de la saisie» :

Dans la [FIGURE3.7] on va présenter l'ensemble des fonctionnalités à effectuer lors de la saisie des informations.

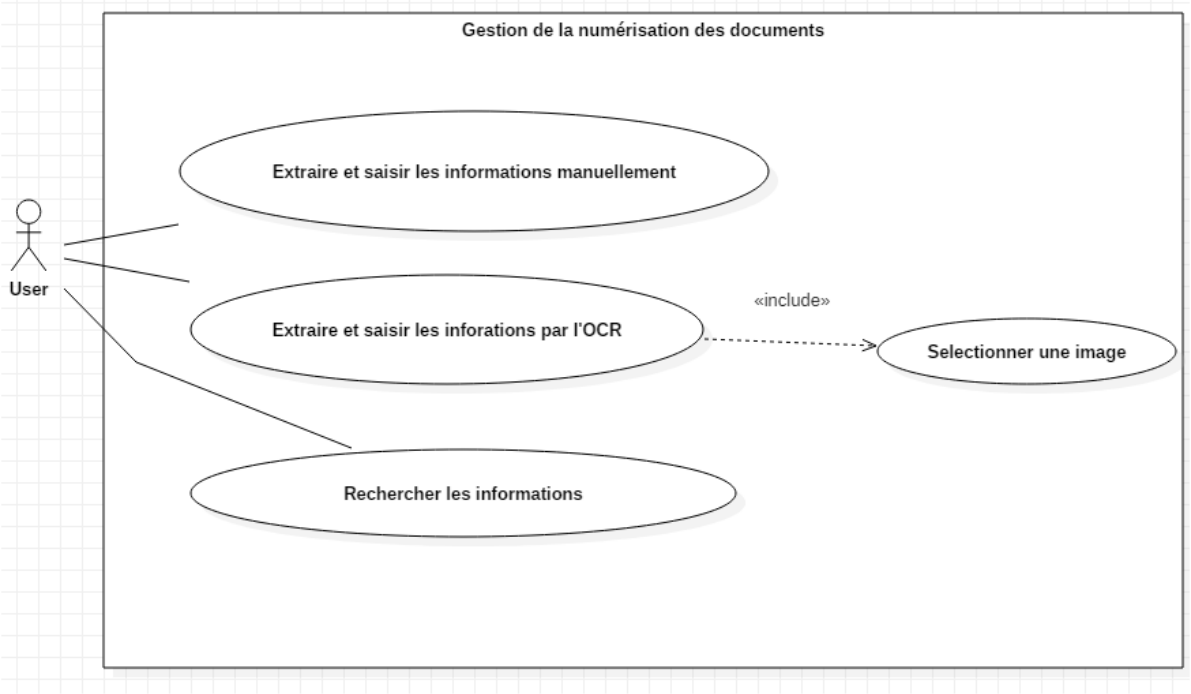

FIGURE 3.7. diagramme de cas d'utilisation « Gestion de la numérisation».

Pour le cas de la numérisation des documents, l'utilisateur peut extraire et saisir les informations manuellement ou par OCR (il doit sélectionner une image d'abord) et aussi rechercher les informations concernant un document stocké.

3.1.2 DIAGRAMME DE DÉPLOIEMENT :

Le diagramme de déploiement sert à représenter l'utilisation de l'infrastructure physique du système et la manière dont les composants du système sont répartis ainsi que leurs relations entre eux. Les caractéristiques des ressources matérielles physiques et des supports de communication peuvent être précisées sur ce diagramme.[FIGURE3.8]

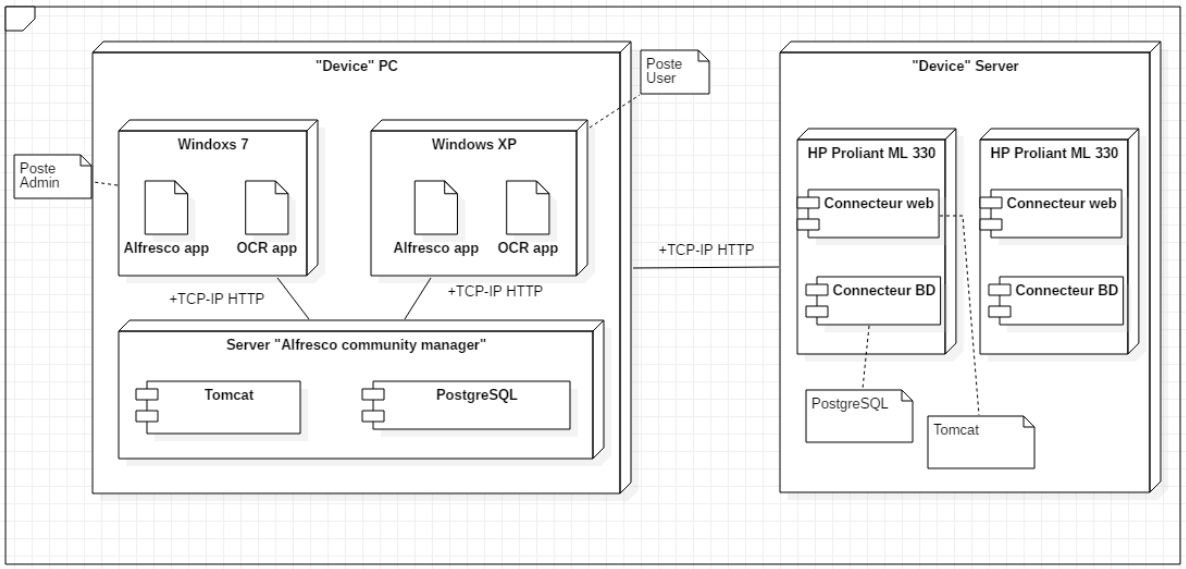

FIGURE 3.8. diagramme de déploiement.

3.1.3 DIAGRAMME DE SÉQUENCE :

Le diagramme de séquence permet de mieux expliquer les interactions entre acteur ou objet à travers certains scénarios du diagramme de cas d'utilisation, donc nous avons détaillé comment les éléments du système interagissent entre eux en s'échangeant des messages.

diagramme de séquence « Gestion des documents» :

Dans la [FIGURE3.9] on présente le diagramme de séquence pour la gestion des documents.

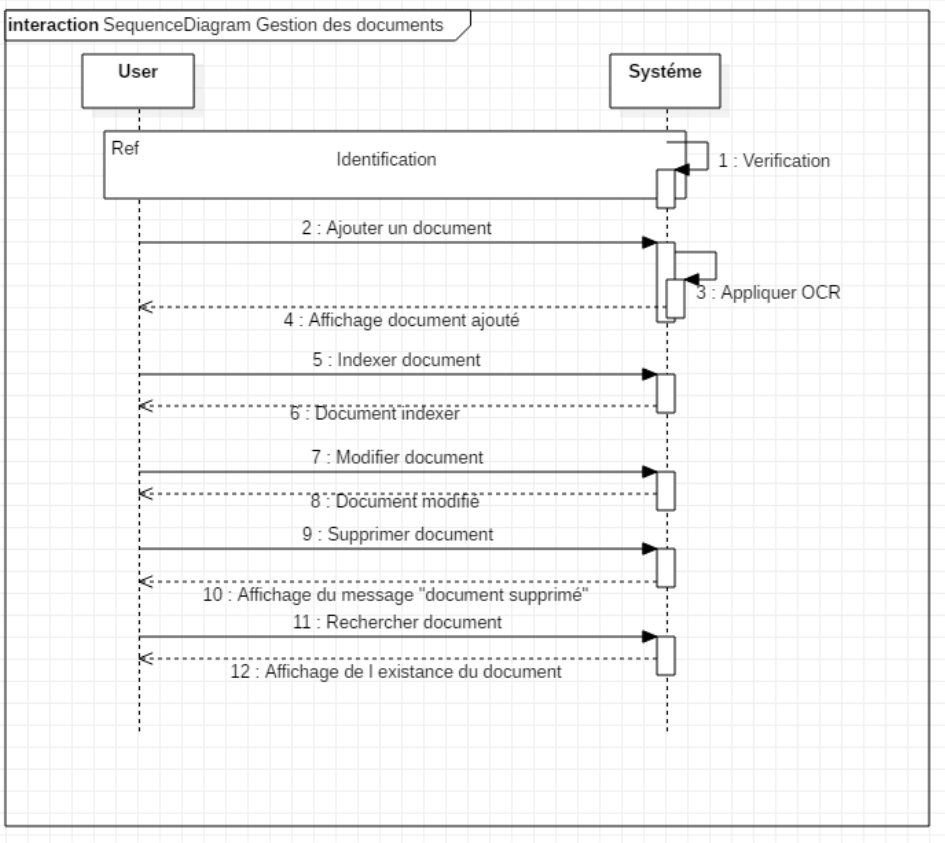

FIGURE 3.9. diagramme de séquence « Gestion des documents».

diagramme de séquence « Gestion des sites» :

Dans la [FIGURE3.10] on présente le diagramme de séquence pour la gestion des sites.

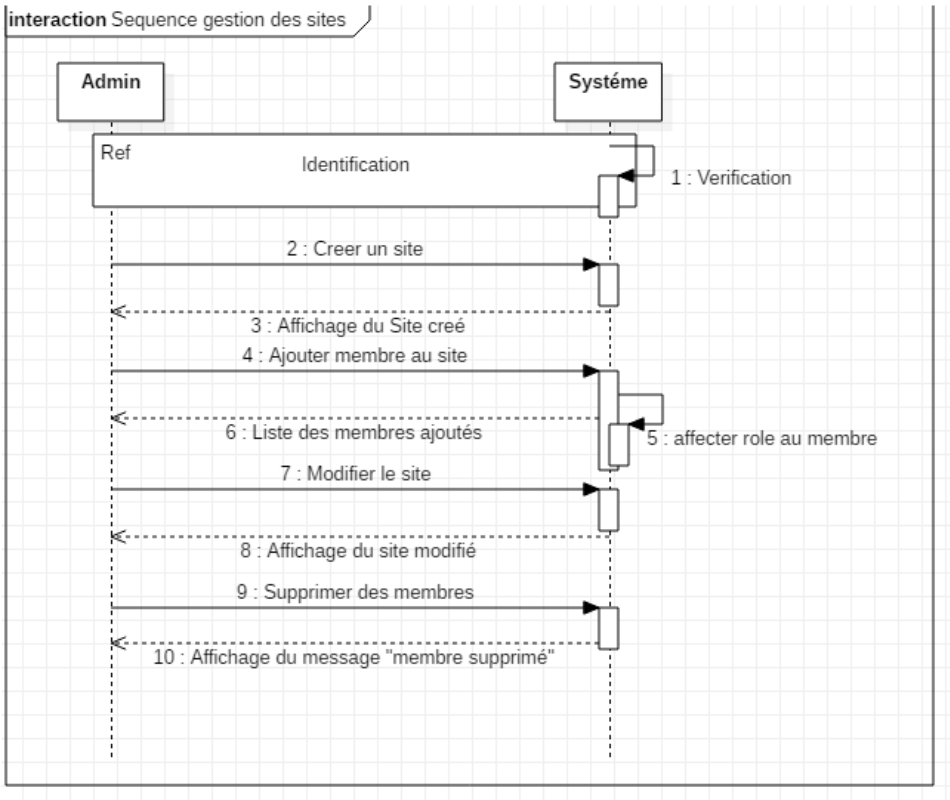

FIGURE 3.10. diagramme de séquence « Gestion des sites».

diagramme de séquence «Utilisation des sites» :

Dans la [FIGURE3.11] on représente le diagramme de séquence pour l'utilisation des sites.

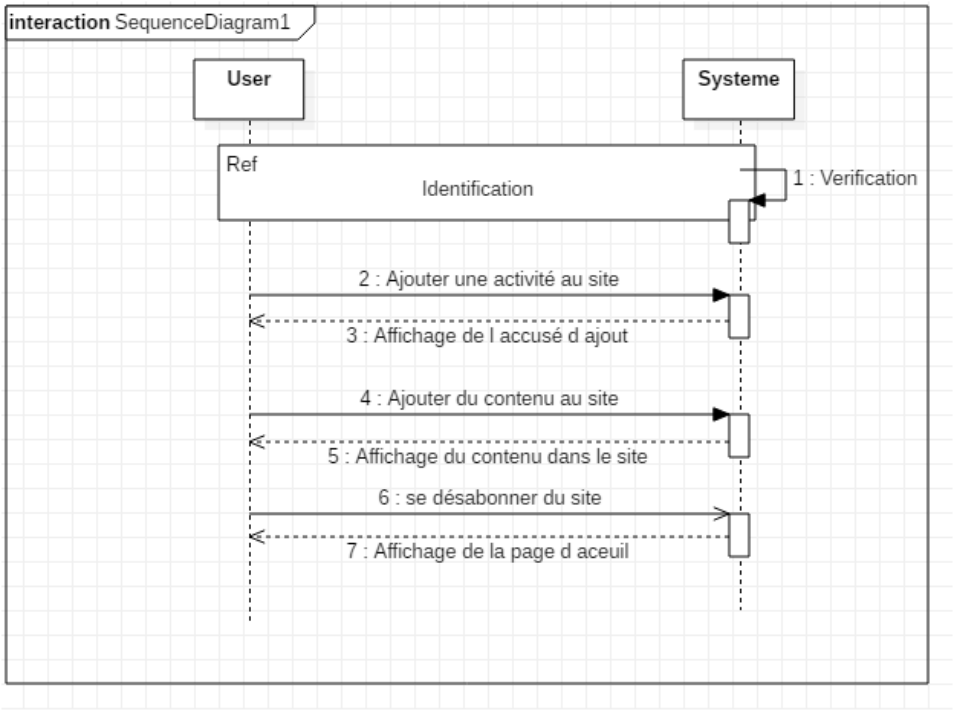

FIGURE 3.11. diagramme de séquence « Utilisation des sites ».

diagramme de séquence« Gestion des utilisateurs» :

Dans la [FIGURE3.12] on représente le diagramme de séquence pour la gestion des utilisateurs.

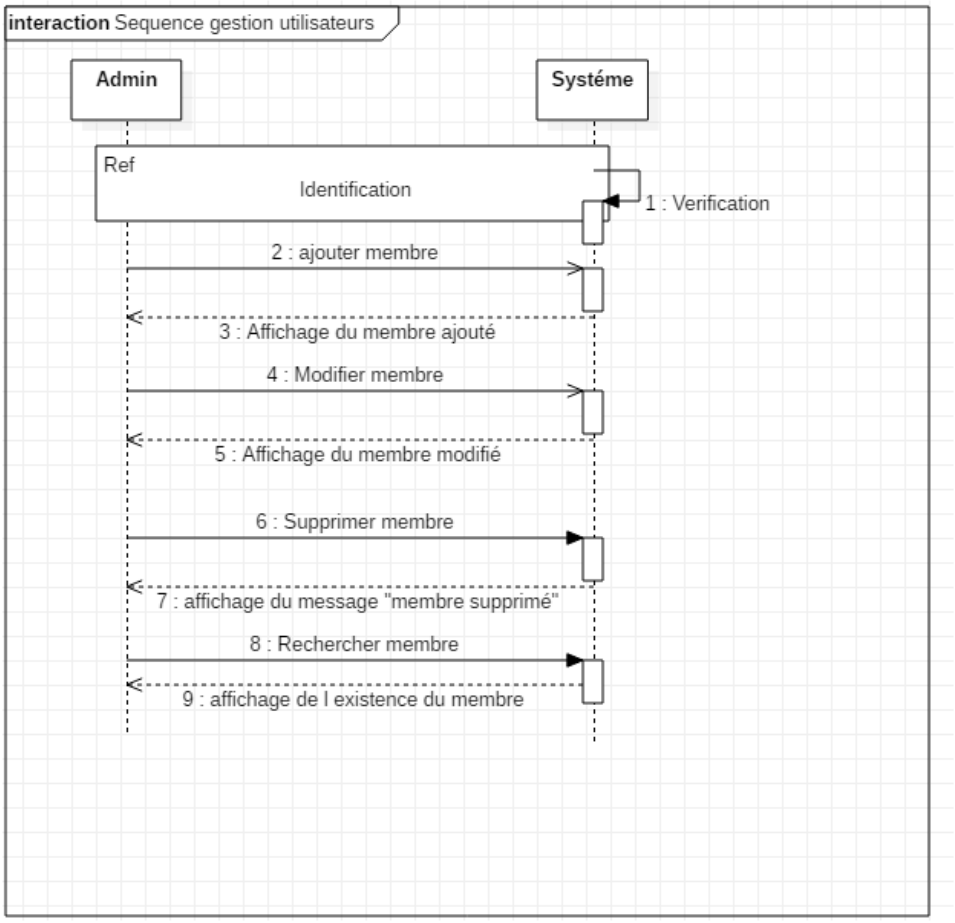

FIGURE 3.12. diagramme de séquence « Gestion des utilisateurs».

3.1.4 DIAGRAMME D'ACTIVITÉ :

Le diagramme d'activité permet de fournir une vue du comportement d'un système en décrivant les actions séquentielles d'un processus, le diagramme d'activité est un moyen de décrire graphiquement les étapes effectives dans un cas d'utilisation complexe, il permet de modéliser des éléments de l'architecture de notre système tels que le fonctionnement et l'utilisation.

diagramme d'activité« Gestion des documents» :

Dans la [FIGURE3.13] on représente le diagramme d'activité pour la gestion des documents.

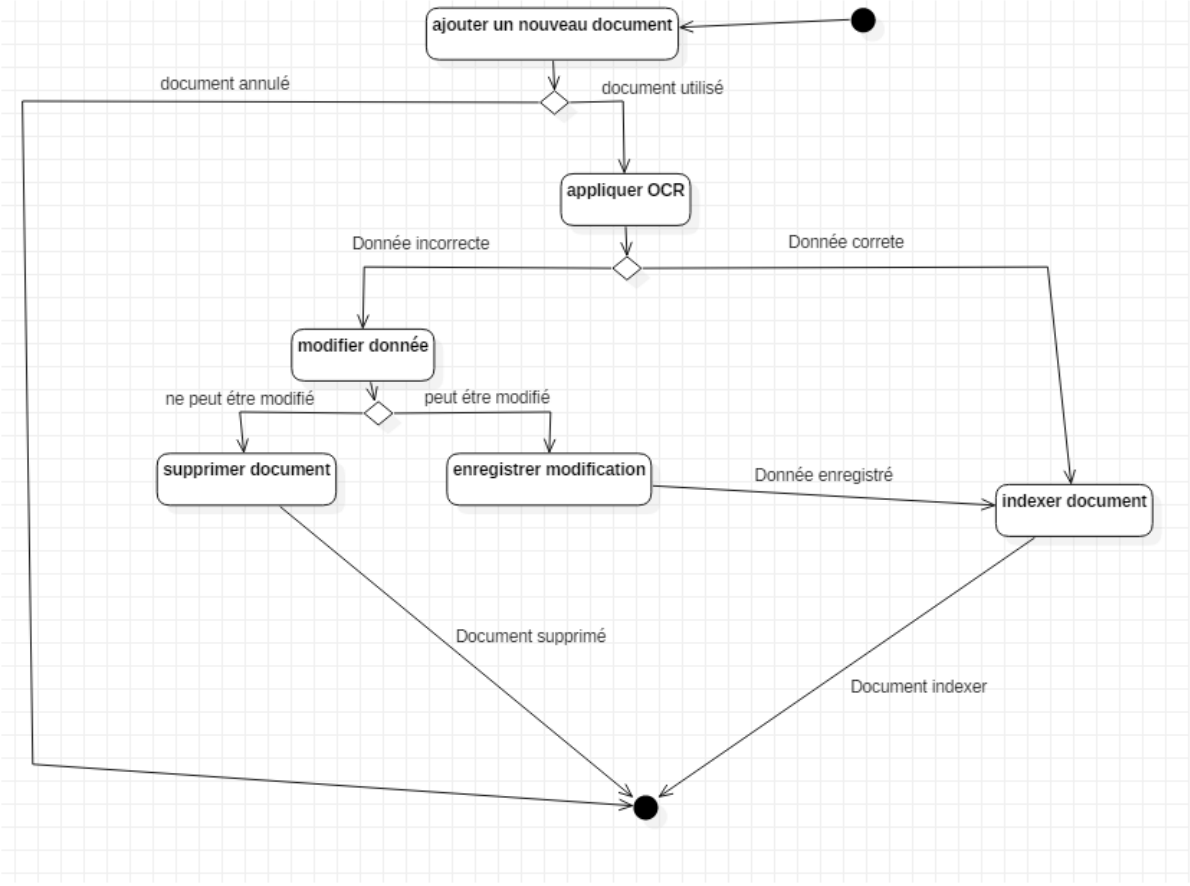

FIGURE 3.13. diagramme d'activité « gestion des documents ».

Pour le diagramme d'activité de la gestion des documents on doit d'abord ajouter un nouveau document qui peut être par la suite soit utilisé ou annulé (dans ce cas la l'opération se termine), s'il est utilisé on lui applique l'OCR. Si on aura des données correctes le document sera indexer et l'opération se termine, sinon si les données sont incorrectes on leurs effectue une correction. Si elles sont non

modifiables on supprime le document et l'opération se termine, sinon on enregistre les modifications et on indexe le document puis l'opération sera terminée.

diagramme d'activité« Utilisation des documents» :

Dans la [FIGURE3.14] on représente le diagramme d'activité pour l'utilisation des documents.

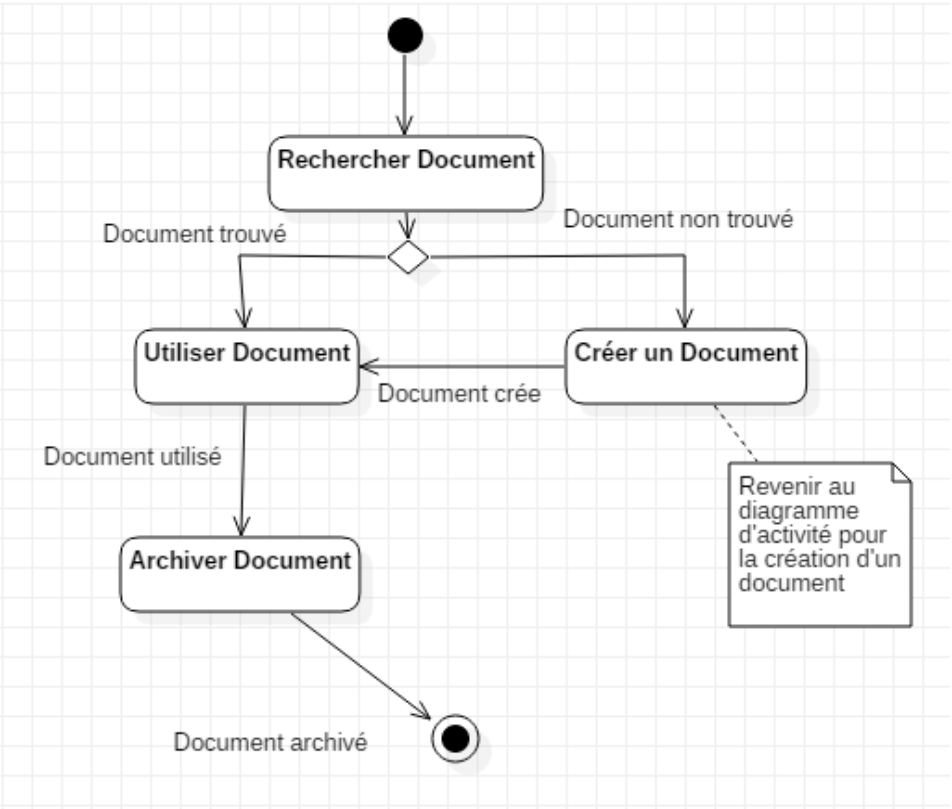

FIGURE 3.14. diagramme d'activité « Utilisation des documents ».

Pour l'utilisation des documents on doit d'abord rechercher un document. S'il existe on l'utilise puis on l'archive et l'opération sera terminée, sinon on doit d'abord créer le document puis l'utiliser et l'archiver.

3.1.5 DIAGRAMME DE CLASSE :

Le diagramme de classe décrit clairement la structure de notre système en modélisant ses classes, il permet de comprendre l'aperçus général des schémas de notre application, il regroupe un ensemble d'objets de notre base de données qui partagent les mêmes propriétés, comportement et actions.

Dans la [FIGURE3.15] on expose seulement les classes qui conviennent à nos besoins d'utilisation.

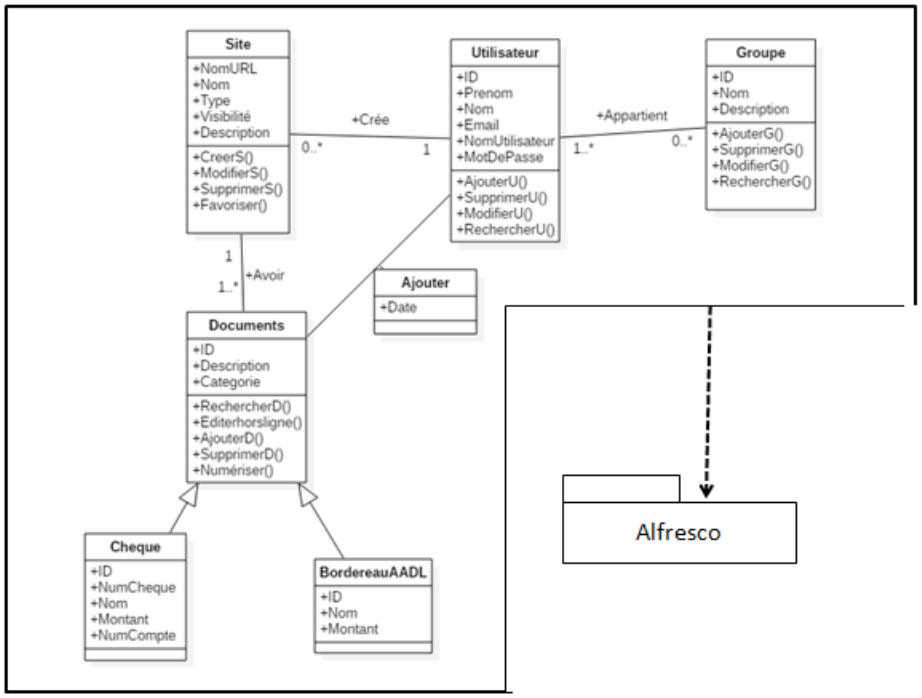

FIGURE 3.15. Diagramme de classe.

Dans notre diagramme de classe on a le package Alfresco lié à nos classes. On a 6 classes (site, utilisateur, groupe, document, chèque, bordereauaadl) où chaque classe possède des attributs, des méthodes et des liens.
L'utilisateur par exemple peut appartenir à 0 ou plusieurs groupes, créer 0 ou plusieurs sites et ajouter des documents à une date précise.
Le groupe peut contenir au moins un utilisateur qui est son créateur.
Le site peut être crée par un seul et unique utilisateur et peut avoir 1 ou plusieurs documents.
Le document peut être ajouté par un seul utilisateur et appartenir à un seul site, il peut être soit un chèque ou un bordereauaadl.

à ce niveau, on a bien décrit le fonctionnement de notre système avec les différents types de diagrammes(diagramme de classe, diagramme de cas d'utilisation, diagramme de séquence, diagramme d'activité,diagramme de déploiement).
Maintenant on passe à la conception de notre réseau de neurones convolutionnel pour entamer l'implémentation dans le chapitre suivant.

3.2 CONCEPTION DU CNN

Pour la conception de notre CNN on a utilisé un certain nombre de couches qui sont présentées comme ceci : [FIGURE3.16]

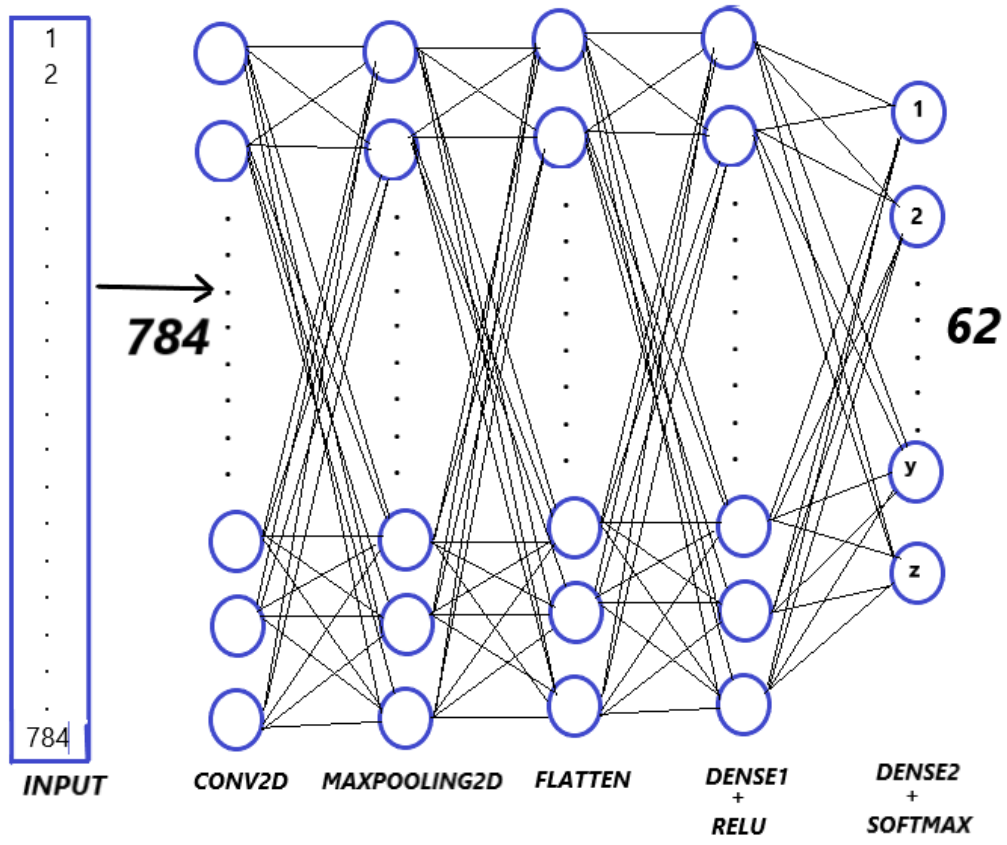

FIGURE 3.16. Architecture de notre réseau de neurones.

Les couches utilisées dans le modèle sont définit comme ceci :

Input :

En réalité c'est pas une couche, mais c'est la représentation des pixels de l'image d'entrée qui seront traités partie par partie (chaque neurone de la 1ère couche traite 9 pixels à la fois). Le nombres des pixels de l'image représente le nombre de neurones de la 1ère couche.

Conv2D :

La couche de convolution donne son nom au réseau de neurones convolutif car elle représente l'élément principal constitutif de ce réseau.

Une convolution est la simple application d'un filtre à une entrée qui entraîne une activation. L'application répétée du même filtre à une entrée donne une carte d'activations appelée carte de caractéristiques, indiquant l'emplacement et la force d'une caractéristique détectée dans une entrée, telle qu'une image. [Jason, 2019]

MaxPooling2D :

Une couche de regroupement est une nouvelle couche ajoutée après la couche de convolution. L'ajout de cette couche après la couche de convolution est un motif commun utilisé pour ordonner les couches dans un réseau de neurones de convolution qui peut être répété une ou plusieurs fois dans un modèle donné.
Elle agit séparément sur chaque carte de caractéristiques pour créer un nouvel ensemble du même nombre de cartes de caractéristiques regroupées.
Le regroupement implique la sélection d'une opération de regroupement, un peu comme un filtre à appliquer aux cartes de caractéristiques. La taille de l'opération de regroupement ou du filtre est inférieure à la taille de la carte de caractéristiques. concrètement, ce sont presque toujours 2×2 pixels appliqués avec une foulée de 2 pixels. [Jason, 2019]

Flatten :

Après avoir terminé les deux étapes précédentes, nous sommes maintenant supposés avoir une carte de fonctionnalités regroupée. Comme le nom de cette étape l'indique, nous allons littéralement aplatir notre carte de fonctionnalités regroupée dans une colonne.[FIGURE3.17]

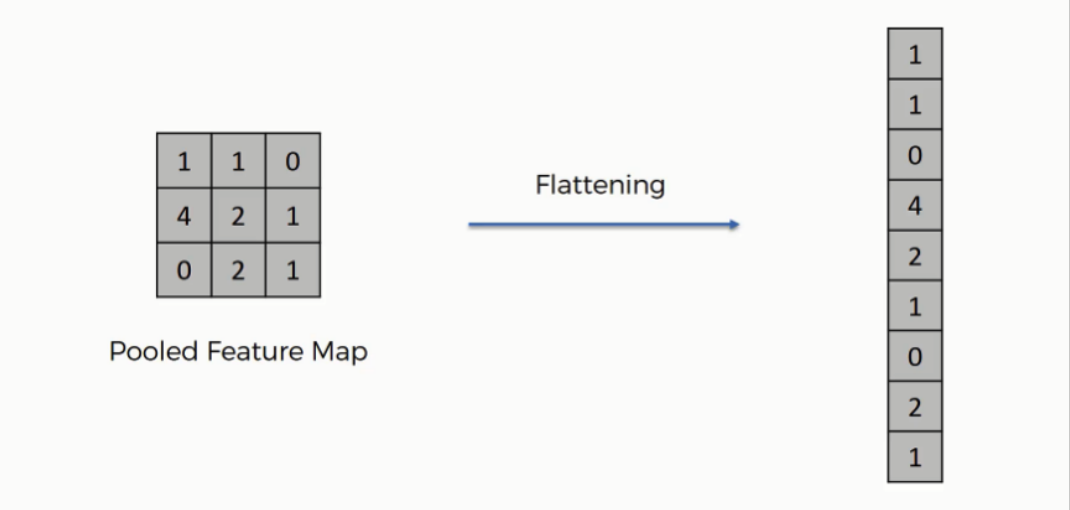

FIGURE 3.17. Flatten()

Dense :

Dense implémente l'opération, 128 représente l'unité de dimension pour l'espace de sortie.
L'activation est une fonction qui est appliquées a la sortie d'une couche de réseau neuronal, qui sera transmise comme entrée à la couche suivante, les fonctions d'activation sont une partie essentielles des réseaux de neurone car sans elles le réseau neuronal se réduit à un simple modèle de régression logistique.

Relu :

Il est possible d'améliorer l'efficacité du traitement en intercalant entre les couches de traitement une couche qui va opérer une fonction mathématique (fonction d'activation) sur les signaux de sortie.
La fonction RELU (abréviation de Unités Rectifié linéaires) **R(z)=max(0,z)** force les neurones à retourner des valeurs positives.

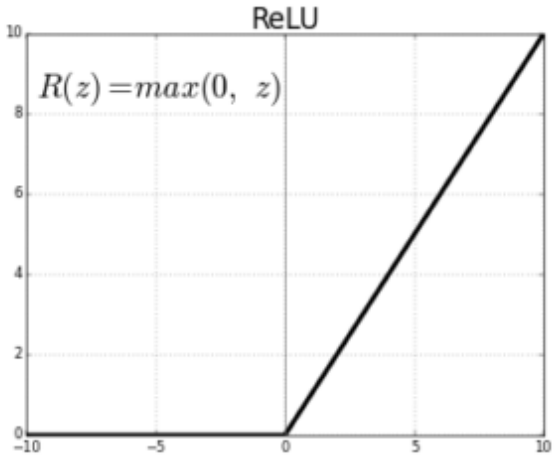

FIGURE 3.18. Graphe Relu.

Softmax :

La fonction d'activation de softmax est utilisée dans les réseaux de neurones lorsque nous voulons construire un classifieur à plusieurs classes qui résout le problème de l'affectation d'une instance à une classe lorsque le nombre de classes possibles est supérieur à deux.

La fonction softmax est en fait une fonction arg max. Cela signifie qu'il ne renvoie pas la plus grande valeur de l'entrée, mais la position des plus grandes valeurs.

Nous interprétons le résultat de la fonction softmax comme la probabilité de la classe. La fonction est définie par :

$$\sigma(\mathbf{z})_j = \frac{\mathrm{e}^{z_j}}{\sum_{k=1}^{K} \mathrm{e}^{z_k}} pour tout j \in \{1,\ldots,K\} j \in \{1,\ldots,K\}$$

C'est-à-dire que la composante **j** de notre vecteur est égale à l'exponentielle de la composante **j** du vecteur **Z** divisée par la somme des exponentielles de toutes les composantes de **Z**.

3.3 Conclusion :

Dans ce chapitre nous avons parlé de la conception du système avec les diagrammes UML, présenté les CNNs. Ces réseaux sont capables d'extraire des caractéristiques d'images présentées en entrée et de classifier ces caractéristiques. Ils implémentent aussi l'idée qui permet de réduire beaucoup de nombre de paramètres libres de l'architecture. Dans le but de réduire les temps de calcul, l'espace mémoire nécessaire, et également d'améliorer les capacités de généralisation du réseau.

Dans le chapitre suivant nous allons parler de l'implémentation de notre CNN ainsi que celle de notre OCR. Nous allons aussi présenter les différentes interfaces d'Alfresco et pour finir nos tests et résultats obtenus.

CHAPITRE 4

IMPLÉMENTATION ET TEST

Dans ce chapitre nous allons présenter notre environnement de travail (Python,HTML et CSS ,PHP,PostgreSQL,XAMPP,TOmcat), ainsi que la bibliothèque Keras pour l'apprentissage et la classification. on va aussi présenter et expliqué brièvement les différentes interfaces GED et OCR ,et la fin une discussion des résultats de traitement de notre OCR.

4.1 Environnement de travail :

4.1.1 Python :

Figure 4.1. Python.

Python est un langage de programmation interprété, multiparadigme est multiplateforme compatible avec de nombreux systèmes d'exploitation, il favorise la programmation impérative structurée, fonctionnelle et orientée objet.
Python est placé sous une licence libre (Python software fondation license).l'utilisation de python est adaptée dans de nombreux contexte grâce a ses nombreuse bibliothèques optimisées, il est cependant particulièrement utilisé comme langage de script pour automatiser des taches simples mais fastidieuses.
La version utilisée pour notre projet est python 3.6. [Lutz, 2009]

4.1.2 HTML ET CSS

FIGURE 4.2. HTML et CSS.

HTML : est un langage de balisage conçu pour représenter les pages web, il permet d'écrire de l'hypertexte ainsi que la structuration sémantique et logique pour mettre en forme le contenu des pages. [Gillies and Cailliau, 2000]

CSS : les feuilles en cascade forment un langage informatique qui décrit la présentation des documents HTML et HML, il est utilisé dans la conception des sites web. [Lie, 2005]

La combinaison Html et CSS est un véritable standard en informatique, les deux langages se trouvent à la base de tout projet web car ils ont un rôle qui les rend incontournables, le HTML va donc créer la structure des pages tandis que le CSS va nous permettre de modifier l'apparence des contenus de la page. Concernant l'environnement de travail, on a choisi Brackets comme éditeur de texte car il dispose d'une excellente ergonomie. [Gir, 2004]

4.1.3 PHP

FIGURE 4.3. PHP.

PHP : est un langage de programmation libre impératif orienté objet.

Il est considéré comme une des bases de la création de sites web dits dynamiques mais également des applications web,c'est un langage de script qui est particulièrement adapté au développement web. Rapide et flexible il intègre tous les outils nécessaires à la création de sites dynamiques.
Lorsqu'une page PHP est exécuté par le serveur, alors celui-ci renvois généralement au client (aux visiteurs du site) une page web qui peut contenir du HTML, XHTML, CSS, JavaScript ... [w3s, 2003]

4.1.4 PostgreSQL

FIGURE 4.4. PostgreSQL.

PostgreSQl : est un système de gestion de base de donnée relationnelle et objet (SGBDRO). c'est un outil libre créer par Micheal Stonebraker.

Ce système multiplateformes est largement connu et réputé à travers le monde, surtout pour être respectueux des normes ANSI SQL, ce projet est géré par une communauté de développeurs.

PostgreSQL fonctionne sous plusieurs systèmes d'exploitation dont Linux, MacOS, Windows(depuis la version 8.0)etc. plusieurs interfaces utilisateurs existent : psql, pgAdmin et PHPAdmin, pour notre projet on a utilisé PgAdmin. [Pos, 2005]

4.1.5 XAMPP

FIGURE 4.5. XAMPP.

Xampp : Est un ensemble de logiciels qui permettent de mettre facilement en œuvre un serveur web local, il offre une bonne souplesse d'utilisation car il est simple et rapide à installer, d'ailleurs il est a la porté d'un grand public car il n'nécessite pas de connaissance particulière, en plus il fonctionne sur plusieurs systèmes d'exploitation.

En d'autre terme XAMPP est un programme simple et gratuit qui permet aux utilisateurs d'héberger des sites web sur leur pc. [des, 2009]

4.1.6 TOMCAT

FIGURE 4.6. Tomcat.

Tomcat : Est un conteneur web libre de servlets il implémente les spécifications des servlets, utilisé en association avec un autre serveur web en général Apache.

Principe de fonctionnement :

• Le serveur web s'occupe des pages web traditionnelles (.html, .php par exemple).

• Il délègue à Tomcat les pages relevant spécifiquement d'une application web Java (Servlet, JSP...). [WIki, 2019]

4.1.7 IntelliJ IDEA

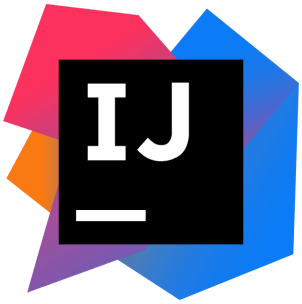

FIGURE 4.7. IntelliJ-IDEA.

Intellij IDEA : est un EDI disponible en deux versions, l'une communautaire, open source, sous licence Apache 2 et l'autre propriétaire, protégée par une licence commerciale. En plus de java Intellij permet également de supporter plusieurs langages comme HTML, CSS, PHP, Python, JS etc...
Il est disponible sous plusieurs Systèmes d'exploitations (Windows, Linux et MacOS) . . [will, 2015]

4.1.8 Keras

Keras est une interface de programmation d'application de réseaux de neurones de haut niveau, écrite en Python et capable de s'exécuter sur TensorFlow, CNTK ou Theano. Elle a été développé pour permettre une expérimentation rapide. Pouvoir faire de la recherche de qualité est essentiel pour pouvoir passer de l'idée au résultat le plus rapidement possible. [Git, 2008]

4.2 Interfaces

Nous allons présenter les différentes interfaces et fonctionnalités d'Alfresco qui a été personnalisé et modifié afin de correspondre à l'organisme qui va l'utilisé (le centre des archives), ainsi que les

différentes interfaces de notre OCR.

Interface d'authentification :

Au lancement de la GED on aura une interface d'authentification qui exige à l'utilisateur de saisir les informations demandées à savoir le nom d'utilisateur et mot de passe afin de se connecter à la plateforme.[FIGURE 4.8]

FIGURE 4.8. Interface authentification GED.

Interface d'accueil : :

Après avoir saisie les bonnes informations, vous aurez accès a la plateforme d'accueil qui se compose de :[FIGURE 4.9]

FIGURE 4.9. Interface d'accueil GED.

(1) tableau de bord utilisateur : quelque soit votre emplacement dans la GED vous aurez toujours ce tableau à votre porté pour vous faciliter la navigation.

(2) personnaliser le tableau de bord : vous pouvez modifier votre tableau de bord en changeant la disposition de vos dashlets ou encore les supprimés et ajouter d'autre celons vos besoins.

(3) Mon Profil : permet d'afficher l'identité de l'utilisateur en cours, une possibilité de mettre a jour le profil est mis en place afin d'ajuster des détails ou changer le mot de passe utilisateur.

(4) Mes activités : cette section permet d'afficher les dernières activités (celons une certaine échéance de temps) que l'utilisateur en cours à effectuer.

(5) Mes documents : ici une liste de documents récemment modifiés ou ajoutés par l'utilisateur sera affichée.

(6) Mon calendrier : permet de vous rappeler les prochains événements à venir.

(7) Mes tâches : cette section affichera une liste des tâches qui vous ont été assigné afin de les accomplir, une fois la tâche terminée elle disparaîtra de votre liste.

(8) Documents édités par moi : comme son nom l'indique, cette partie permet d'affiché les documents édités par l'utilisateur à savoir des documents, article de blog etc.

Interface gérer les utilisateurs : :

Afin de gérer les utilisateurs qui auront accès à la GED, seul l'administrateur aura la possibilité d'ajouter ou faire des modifications ou suppressions sur les utilisateurs.

Pour accéder à l'interface de la gestion des utilisateurs, on clique sur (1) « outils admin » dans le tableau de bord ensuite sur (2) « utilisateurs », une liste des utilisateurs existent s'affichera [FIGURE 4.10].

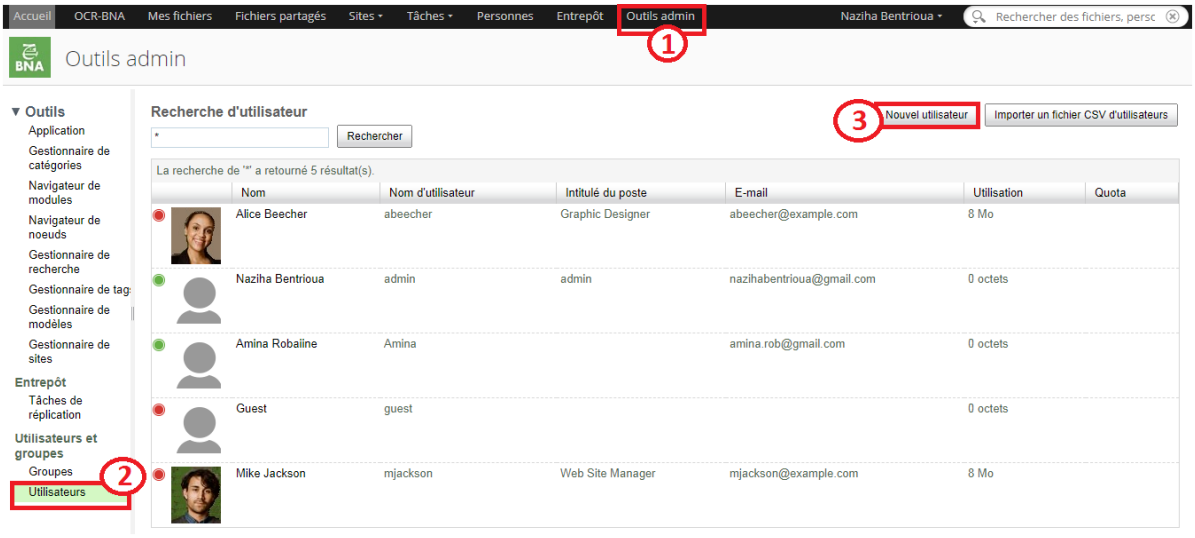

FIGURE 4.10. Interface gérer user GED.

Pour ajouter un nouvel utilisateur on clique sur (3) et une interface de ce genre [FIGURE 4.11] s'affichera pour remplir les informations nécessaire à la création de l'utilisateur.

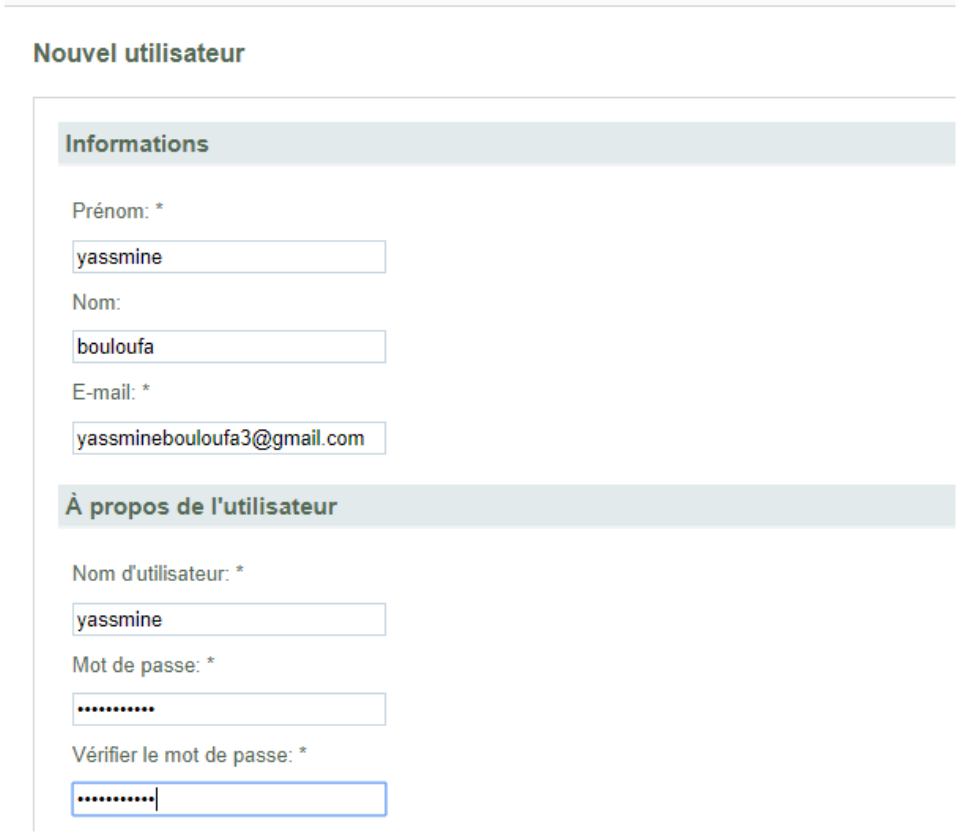

FIGURE 4.11. Ajouter user GED.

Pour faire des modifications et suppressions sur un utilisateur, on clique sur le nom d'utilisateur affiché ou recherché, et on aura accès à la modification ou suppression de la personne, dans la partie modification on peut assigné l'utilisateur à des groupes de travail.[FIGURE 4.12]

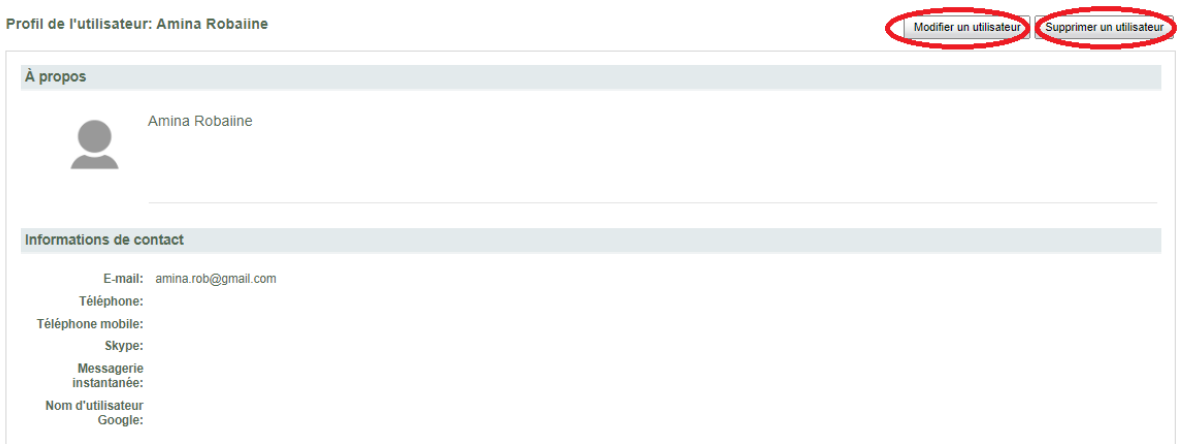

FIGURE 4.12. Modifier supprimer user GED.

Interface gérer les groupes :

Afin de gérer les équipes de travail de notre département (centre des archives), des groupes peuvent être crées afin d'organiser et faciliter la coordination entre groupes.[FIGURE 4.13]

FIGURE 4.13. Gérer groupe GED.

Interface gérer les sites :

Un site est une zone de travail dans laquelle les membres peuvent partager du contenu et travailler ensemble. Le créateur du site devient son gestionnaire et peut confier la gestion du site à d'autres membres.

Créer un site est simple et rapide, dans le tableau de bord on trouve la section « sites » dans laquelle on a la possibilité de créer un site en cliquant sur « créer site ».[FIGURE 4.14]

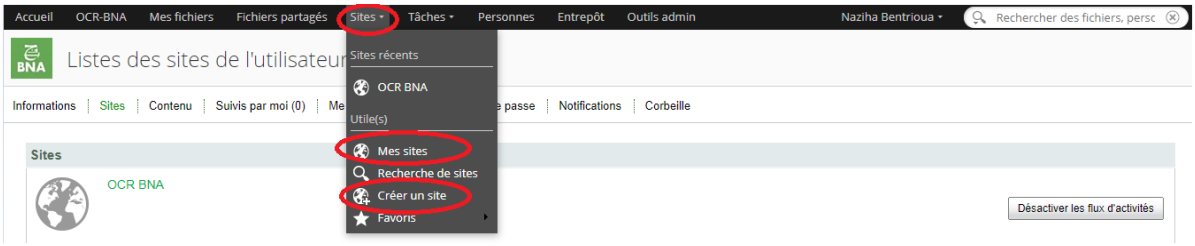

FIGURE 4.14. Section sites GED.

Dans la même section « site », on peut accéder à un site déjà créé on obtient une interface comme ceci[FIGURE 4.15] qui se compose de :

(1) Membres du site. (2) Contenus du site (espace documentaire). (3) Activité du site. (4) Mes discussions : (pour faciliter la communication entre les membres du site).

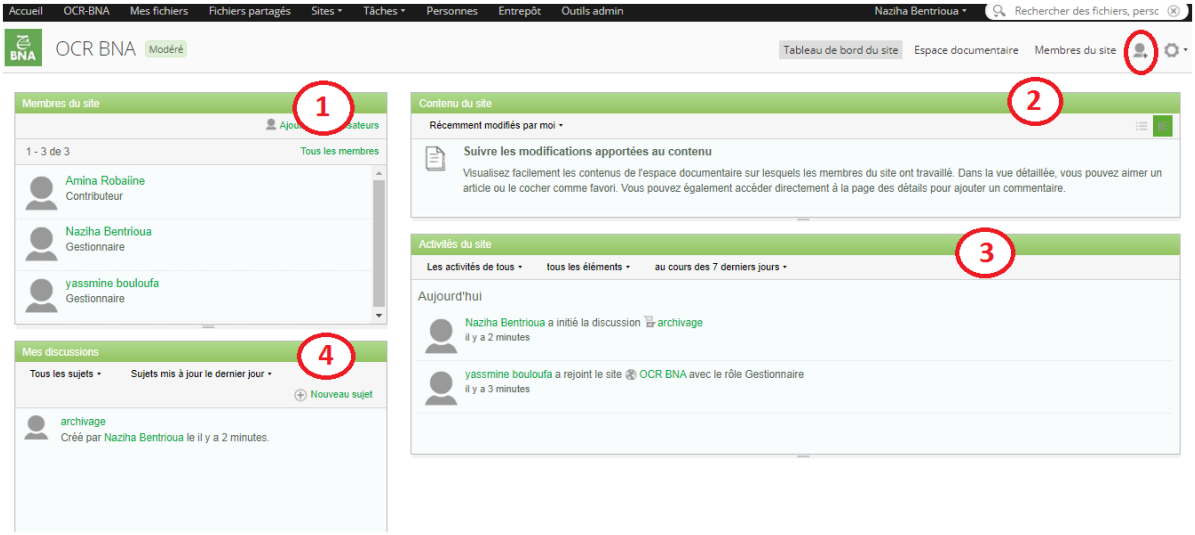

FIGURE 4.15. site GED.

Pour ajouter un utilisateur au site on clique sur l'icône « Ajouter membre », une interface se présentera [FIGURE 4.16] dans la quelle on trouvera 3 étapes à suivre :

(1) sélectionner l'utilisateur à ajouté.

(2) définir le rôle de l'utilisateur sur :
Gestionnaire : dispose des droits d'accès à l'ensemble du contenu du site.
Collaborateur : peut modifier mais ne pas supprimer le contenu du site.

Contributeur : dispose des droits d'accès du contenu dont il est le propriétaire uniquement.
Lecteur : dispose des droits d'accès au site comme lecteur uniquement.

(3) Ajouter l'utilisateur au site.

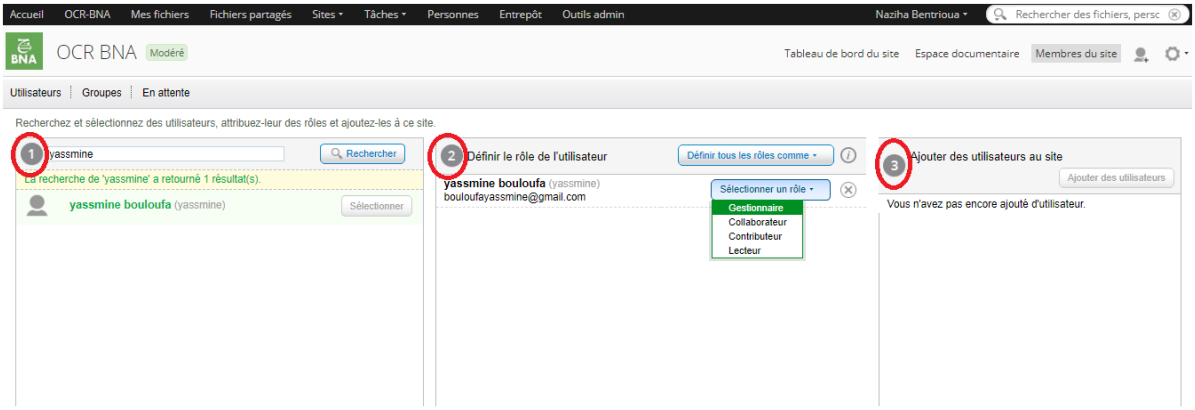

FIGURE 4.16. Ajouter membres au site GED.

Interface gérer les documents :

L'espace documentaire donne une vue global des différents documents ajoutés au site, plusieurs options sont mis à disposition y compris la modification et suppression tel que [FIGURE 4.17] :

FIGURE 4.17. Option espace documentaire GED.

(1) Gérer les droits d'accès : cette fonction prévaut sur les rôles assignés par défaut, un membre peut avoir plus d'accès ou moins sur un document par rapport aux autres contenus documentaires.

(2) Éditer en hors-ligne : Elle permet de télécharger un document sur votre ordinateur et le verrouiller dans l'espace documentaire afin d'empêcher d'autre utilisateur de le modifier en parallèle.

(3) Recherche par filtrage : on peut filtrer notre recherche et cela dépend des besoins recherché ex. (tous les documents, modifié par moi, modifié par d'autre, récemment modifié, récemment ajouté, mes favoris).

(4) démarrer un workflow : aide à organiser les tâches que les utilisateurs doivent accomplir, on peut assigner chaque tâche à une personne avec un document précis sur une échéance de temps à respecter.[FIGURE 4.18]

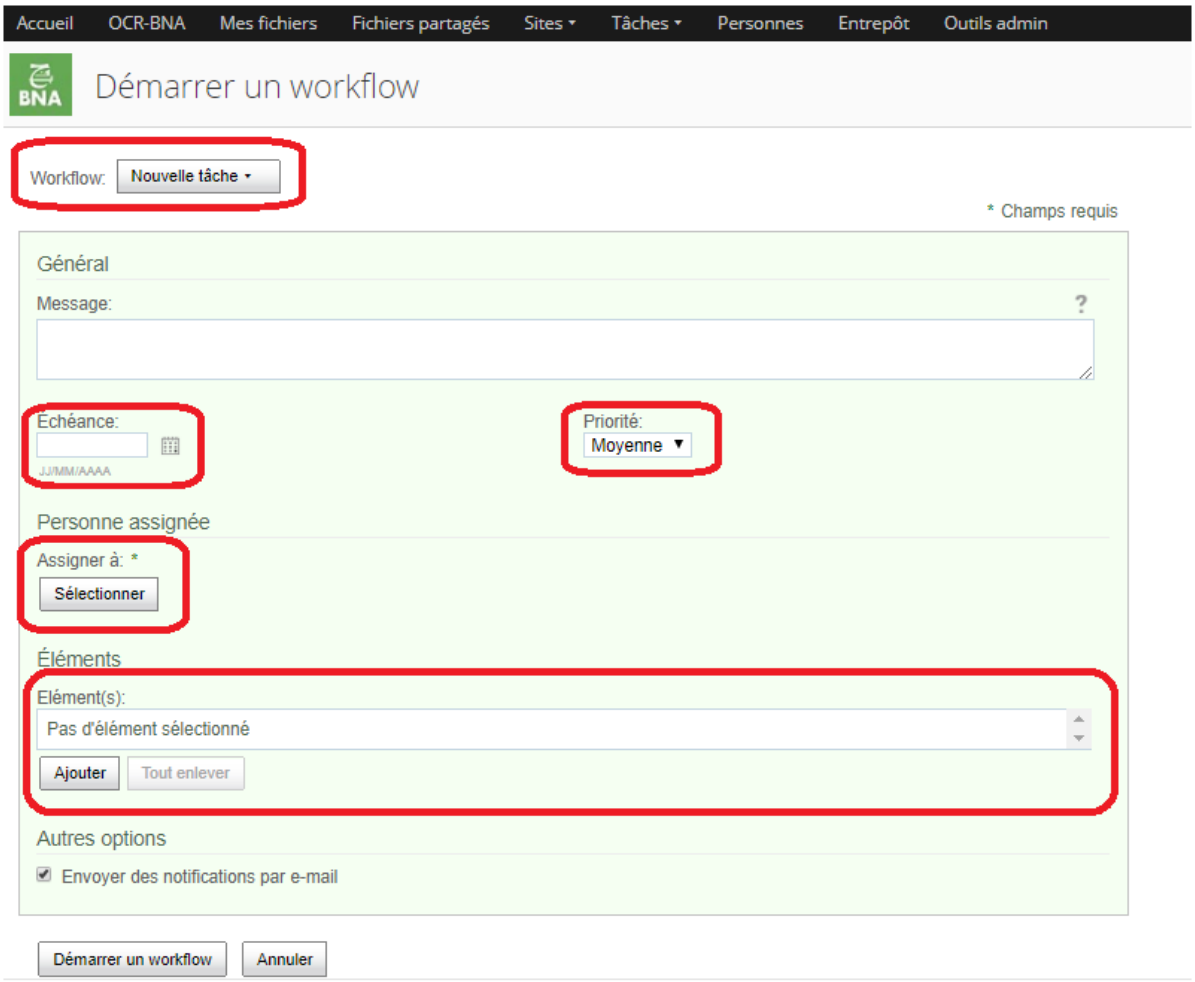

FIGURE 4.18. Démarrer Workflow GED.

Interface OCR :

Concernant notre OCR la 1ère interface [FIGURE 4.19] propose de choisir un type de document sachant que notre OCR s'applique sur les chèques bancaires et les bordereaux de versements AADL de la BNA :

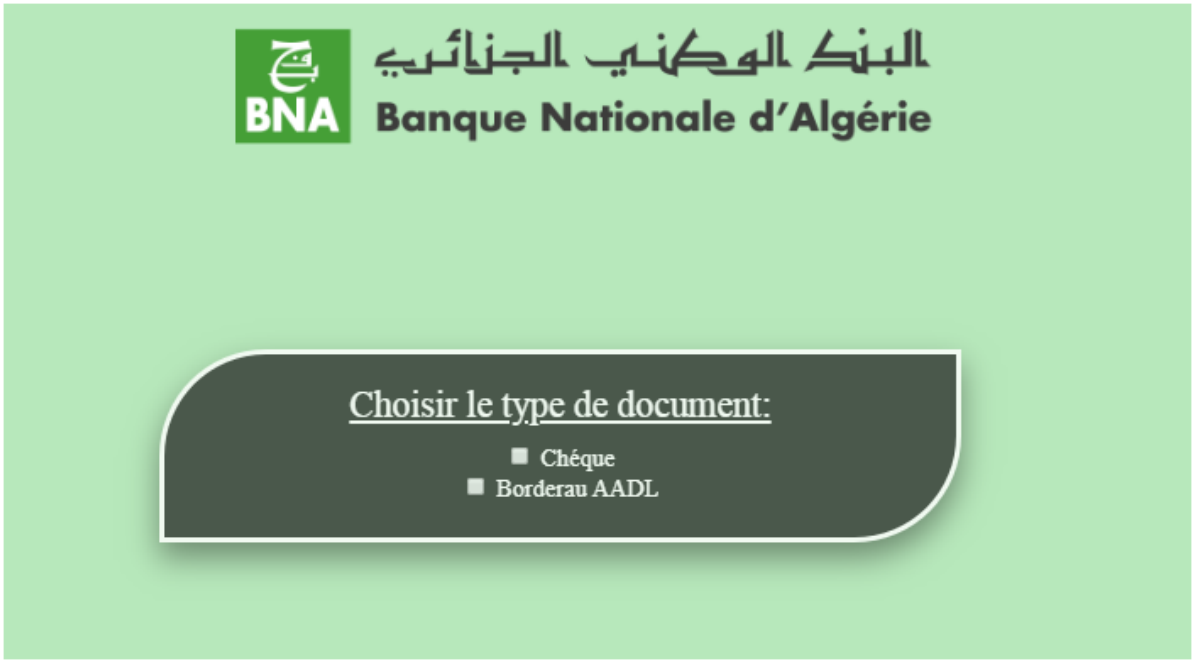

FIGURE 4.19. Interface accueil OCR.

une fois le type "chèque" sélectionné on obtient 3 catégories : [FIGURE 4.20]

(1) Insertion manuelle.

(2) Appliquer OCR.

(3) Recherche.

FIGURE 4.20. Interface chèque.

En cliquant sur L'insertion manuelle on aura une interface comme ceci, elle se compose d'un formulaire dont chaque case correspond à une donnée essentielle à l'archivage du chèque bancaire du client . Cette catégorie nécessite une insertion manuelle des données afin de les ajoutées à la base de données en cliquant sur le bouton « valider ».[FIGURE 4.21]

FIGURE 4.21. Saisie manuelle.

La 2ème catégorie représente l'application de l'OCR, son utilisation est simple il suffit de choisir l'image du chèque scanné, un traitement de quelque seconde sera effectuer le temps d'appliquer des filtres, de segmenter et reconnaître les caractères des zones qui nous intéresses afin d'envoyer chaque information dans sa case adéquate du formulaire, la possibilité de modifier le résultat obtenue est possible pour être ajouté à la base de données.[FIGURE 4.22]

FIGURE 4.22. OCR.

Et finalement la 3ème catégorie représente la recherche, [FIGURE 4.23], l'utilisateur a la possibilité de faire une recherche bien précise ,il suffit d'introduire les informations qu'on veut rechercher (par exemple une recherche par N° de compte ou la date). Et là on aura le résultat de la recherche sous forme de tableau.

FIGURE 4.23. Recherche par date.

Concernant le 2ème type de document "Bordereau AADL",il suffit de cliquer sur ce type pour avoir l'interface suivante [FIGURE 4.24], une fois le type "Bordereau AADL" sélectionné on obtient 3 catégories :

(1) Insertion manuelle.

(2) Appliquer OCR.

(3) Recherche.

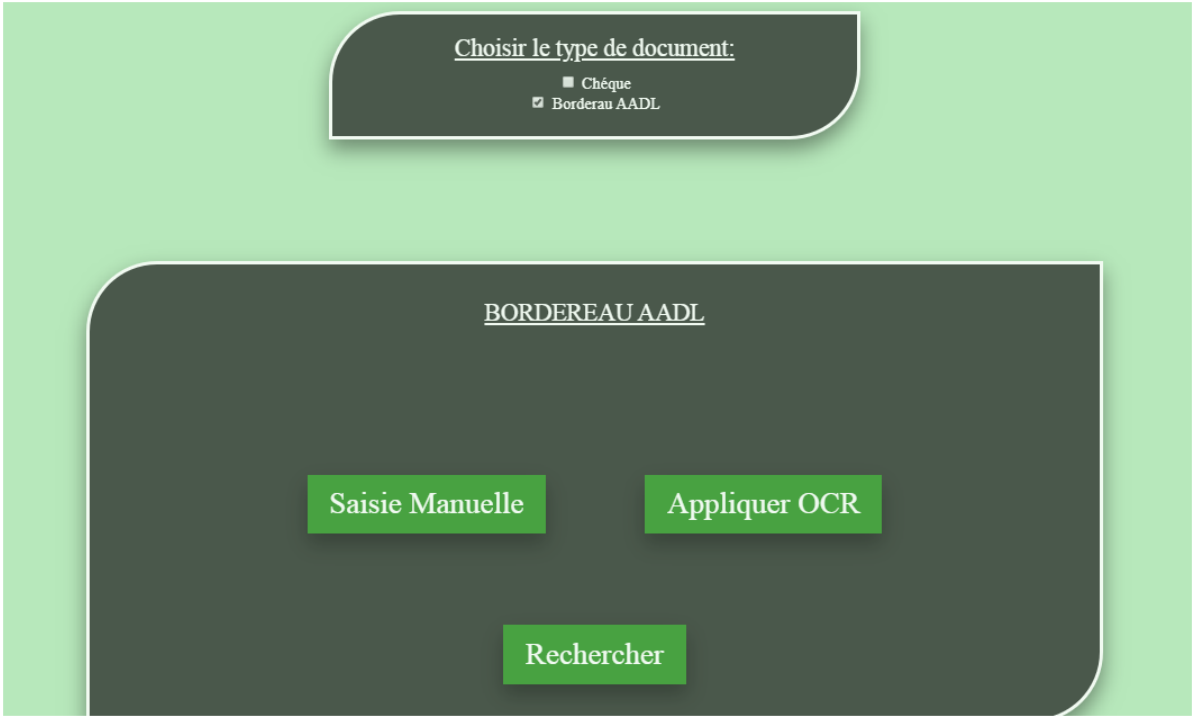

FIGURE 4.24. Interface bordereau AADL.

En cliquant sur L'insertion manuelle on aura une interface comme ceci, elle se compose d'un formulaire dont chaque case correspond à une donnée essentielle à l'archivage du bordereau d'AADL du client . Cette catégorie nécessite une insertion manuelle des données afin de les ajoutées à la base de données en cliquant sur le bouton « valider ».[FIGURE 4.25]

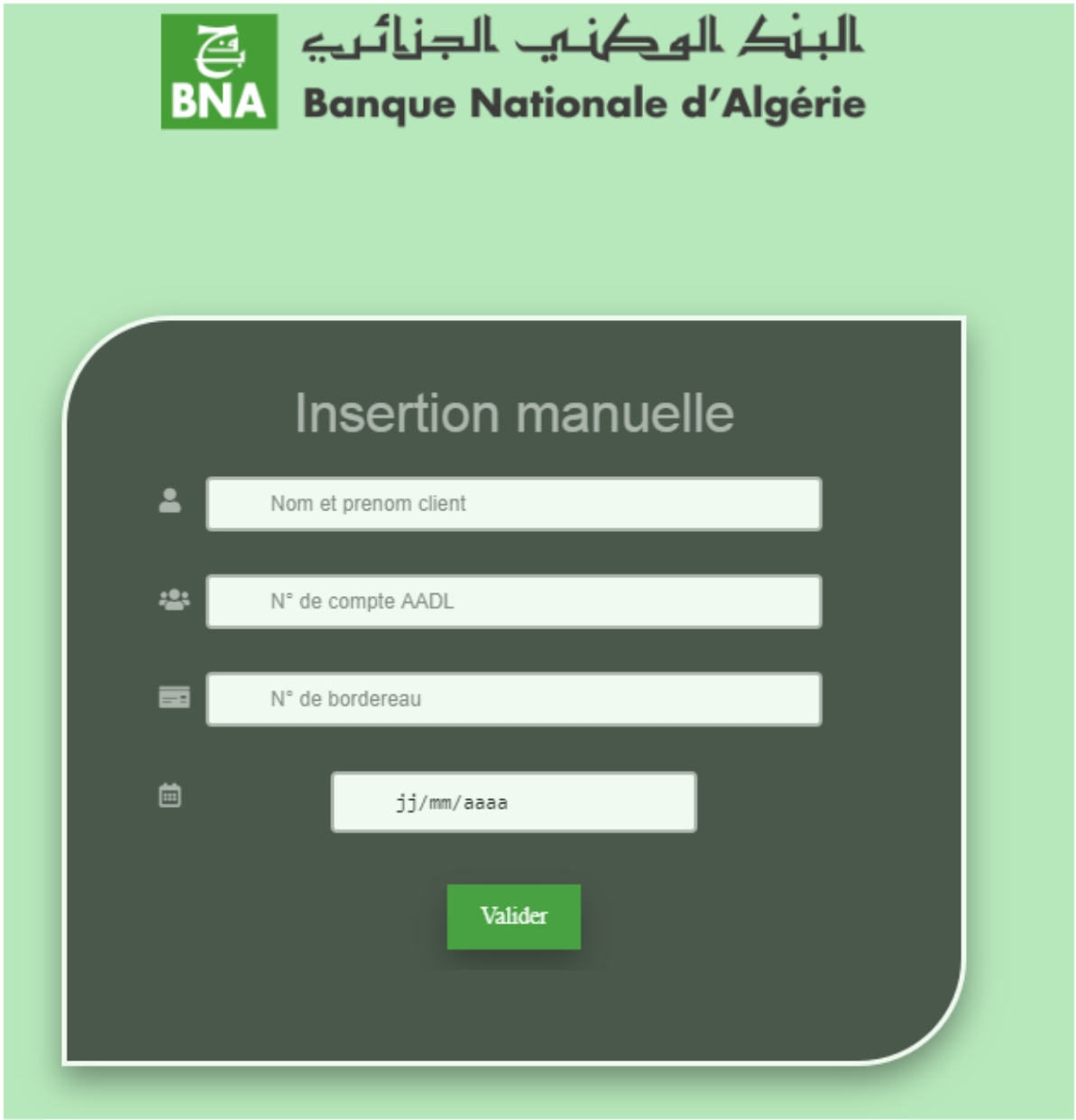

FIGURE 4.25. Saisie manuelle bordereau.

La 2ème catégorie représente l'application de l'OCR, son utilisation est simple il suffit de choisir l'image du bordereau scanné, un traitement de quelque seconde sera effectuer le temps d'appliquer des filtres ,de segmenter et reconnaître les caractères des zones qui nous intéresses afin d'envoyer chaque information dans sa case adéquate du formulaire), la possibilité de modifier le résultat obtenue est possible pour être ajouté à la base de données.[FIGURE 4.26]

FIGURE 4.26. OCR bordereau.

Et finalement la 3ème catégorie représente la recherche, [FIGURE 4.27], l'utilisateur a la possibilité de faire une recherche bien précise il suffit d'introduire les informations qu'on veut rechercher (par exemple une recherche par date). Et là on aura le résultat de la recherche sous forme de tableau.

FIGURE 4.27. Recherche borderau.

L'ensemble des informations traitées précédemment sont stockées dans une base de données afin de garantir la sauvegarde des données pour pouvoir les restaurées dans le besoin d'une recherche par exemple.[FIGURE 4.28] Ils sont stockés sous PostgreSQL.

Data Output | Explain | Messages | Notifications

	id integer	nom character varying (300)	montant integer	ordre character varying (300)	cheque character varying (300)	date date	numcompte double precision
7	7	GUETTAF KHEIR	23000	Moi méme	9681286	2019-09-11	1020001291574
8	8	I HAMAD ATEF	156200	Moi méme	372990	2019-09-03	1020001985415
9	9	MERAH AHME	6000	Moi méme	9535263	2019-09-11	1020001969313
10	10	ROBAINE AMINA	8546000	Moi méme	9535263	2019-09-06	1020001969313
11	11	BOULANDJAS LIAS	50000	moi meme	718515	2019-09-10	20001989295
12	12	BOULANDJAS LIAS	5000	moi meme	718515	2019-09-10	20001989295
13	13	GUETTAF KHEIR	23000	moi meme	9681286	2019-09-10	1020001291574
14	14	OULD HAMOUDA WASSILA	15400	moi meme	1432538	2019-09-10	1020002001323

FIGURE 4.28. Base de donnée.

4.3 IMPLÉMENTATION DE L'OCR

Avant de procéder à l'implémentation de notre OCR on devait implémenter d'abord CNN, et pour cela on a utilisé Keras, qui est une bibliothèque pour l'apprentissage en profondeur, elle prend en charge les réseaux convolutionnels et les réseaux récurrents, ainsi que la combinaison des deux.

Le bout de code suivant représente le type du modèle ainsi que les couches qu'on a utilisé pour la construction de notre CNN :[FIGURE4.29]

```
def define_model():
    model = Sequential()
    model.add(Conv2D(28, (3, 3), activation='relu', input_shape=(28, 28, 1)))
    model.add(MaxPooling2D((2, 2)))
    model.add(Conv2D(56, (3, 3), activation='relu'))
    model.add(MaxPooling2D((2, 2)))
    model.add(Flatten())
    model.add(Dense(56, activation='relu', kernel_initializer='he_uniform'))
    model.add(Dense(123, activation='softmax'))
    # compile model
    model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['acc',f1_m,precision_m, recall_m])
    return model
```

FIGURE 4.29. Création d'un modèle.

Dans Keras on assemble des couches pour construire un modèle séquentiel afin de construire un réseau simple et entièrement connecté. [FIGURE3.16] Tel que 784 représente le nombre de pixels d'une

image en entrée (28*28) et 62 représente nos classes de sortie (1,2,3…a,A,b,B…).

Les couches utilisées dans le modèle sont définit comme ceci :

keras.layers.Conv2D :

cette couche crée un noyau de convolution qui est convoluté avec l'entré de la couche pour produire un tenseur de stories cette couche consiste à filtrer l'image avec un filtre de pixel plus fin et réduire la taille sans perdre la relation entre les pixels , Kernalsize() permet de paramétrer la hauteur et la longueur de la fenêtre de convolution 2D . Lorsqu'on l'utilise en 1ère couche dans un modèle on doit fournir le inputshape qui est paramétré à (28,28,1).

keras.layers.MaxPooling2D :

c'est la 2ème couche de notre modèle elle permet de réduire la taille spatiale et le nombre de paramètre. nous avons sélectionné une taille de regroupement poolsize() qui permet de réduire verticalement et horizontalement les dimension spatial en sélectionnant les valeurs maximales.

Keras.Layers.Flatten :

Les couches de convolution et de regroupement ont des tenseurs multidimensionnels comme sortie. Afin d'utiliser la couche Dense, on doit transformer ces tenseurs en un tenseur 1D en utilisant Flatten().

keras.layers.Dense :

Dense implémente l'opération, 128 représente l'unité de dimension pour l'espace de sortie.
L'activation est une fonction qui est appliquées a la sortie d'une couche de réseau neuronal, qui sera transmise comme entrée à la couche suivante, les fonctions d'activation sont une partie essentielles des réseaux de neurone car sans elles le réseau neuronal se réduit à un simple modèle de régression logistique.

Relu :

Un élément important dans l'ensemble du processus est l'Unité linéaire rectifiée ou ReLU. Les mathématiques derrière ce concept sont assez simples, chaque fois qu'il y a une valeur négative dans un pixel, on la remplace par un 0. Ainsi, on permet au CNN de rester en bonne santé (mathématiquement parlant) en empêchant les valeurs apprises de rester coincer autour de 0 ou d'exploser vers l'infinie. [Charle, 2017]

Le résultat d'une couche ReLU est de la même taille que ce qui lui est passé en entrée, avec simplement toutes les valeurs négatives éliminées.

Softmax :

Softmax est bien pour la classification, elle calcule la distribution des probabilités de chaque classe cible parmi toutes les classes possibles.

Par exemple, si on donne en entrée la couleur des pixels d'une image de chat, on la multiplie par une matrice de poids qui lui correspond, afin de la transformer en un vecteur de T élément(logits) et ainsi chaque logit sera le score d'un animal. Si le score du chat est le plus important, alors la probabilité donnée par la fonction softmax que l'image est un chat sera la plus importante, d'après l'étude de la couleur des pixels. Mais on peut travailler sur d'autres caractéristiques, et ainsi obtenir d'autres probabilités, afin de déterminer l'animal sur la photo. Au fur et à mesure que l'intelligence artificielle aura d'exemples, plus la matrice de poids s'affinera, et plus le système sera performant : on parle d'apprentissage automatique. [SAIMADHU, 2017]

Après avoir ajouté les couches précédentes à notre modèle séquentiel, vient la phase de compilation ou on a utilisé :

optimizer=tf.keras.optimizers.Adam() :

l'optimisation Adam est une méthode de descente de gradient stochastique basée sur une estimation adaptative des moments du 1er et second ordre.
La méthode est efficace en termes de calcul, requit peu de mémoire et convient bien aux problèmes de donnée/paramètres.

loss=tf.keras.losses.categoricalcrossentropy() :

cette fonction est utilisée lorsqu'il existe au moins 2 classes d'étiquettes, les étiquettes sont fournis sous formes d'entrée.

metrics=['accuracy'] :

précision de classe elle calcule la fréquence a la quelles les prédictions correspondent aux étiquettes.

Lors de l'implémentation de notre OCR on a passé par les phases suivantes :

4.3.1 PRÉ-TRAITEMENT :

Cette étape est principalement effectuée pour garantir à notre système la faciliter à identifier les caractères d'une image, une large gamme de fonctions de pré-traitement est mis en œuvre en fonction de nos besoin. Parmi ces fonctions on a utilisé :

```
ref = cv2.cvtColor(ref, cv2.COLOR_BGR2GRAY)
```

FIGURE 4.30. Conversion en niveau de gris.

Cv2.cvtcolor [FIGURE4.30] cette fonction permet de convertir l'image en couleur à une image au niveau de gris.

```
ref = cv2.threshold(ref, 0, 255, cv2.THRESH_BINARY_INV |cv2.THRESH_OTSU)[1]
```

FIGURE 4.31. Conversion en binaire.

Cv2.threshold [FIGURE4.31]cette fonction permet de convertir notre image en niveau de gris en une image binaire(les valeurs des pixels prennent seulement la valeur noir ou blanc).

4.3.2 SEGMENTATION

Concernant la segmentation on a travaillé avec une méthode basée sur les ROI.

Dans un premier temps on a effectué un découpage en zone contenant les informations à extraire de notre document, car il pourra y avoir des régions dans l'image qui ne contiennent aucune information.

Pour extraire les contours de la zone nous utilisant Cv2.findcontours qui produira une liste de contours désordonnée, avant de bouclé sur cette liste de contours on doit d'abord les trié de gauche à droite. Dans cette boucle on va calculer le rectangle de délimitation de chaque contour de cette liste avec la fonction Cv2.boundingrect , si la hauteur et la largeur du caractère sont supérieures ou égales à la hauteur et à la largeur minimales, on va effectuer une extraction du ROI de l'image à l'aide des coordonnée (x,y) et de la largeur/hauteur de notre rectangle englobant, puis on le redimensionne selon la taille des images de notre dataset pour le faire passer par notre modèle entraîné afin prédire sa classe

d'appartenance.

A la fin on effectue une concaténation entre les caractères prédis pour reconstituer le mot.

4.3.3 Post-traitement

Cette phase sert à évoluer les taux de reconnaissance des caractères en effectuant une correction manuelle sur les caractères mal prédis, puis ajouter l'image de ces caractères à notre dataset pour l'entraîner sur ces images pour une meilleurs reconnaissance.

4.4 Test et résultat

MNIST est une base de données qui contient des chiffres manuscrits (0 à 9), elle est considérée comme un sous-ensemble d'un ensemble de données plus vaste (EMNIST). [Orhan, 2018]

Elle est le "Hello World" de l'apprentissage automatique sur qui on peut former un algorithme pour tester une nouvelle technologie ou un nouveau modèle afin de s'assurer de leurs bons fonctionnements.

MNIST est divisée en deux jeux de données :

Le jeu d'apprentissage (trainX) qui contient 60000 exemples de chiffres écrits à la main.

Le jeu de test (testX) qui contient 10000 images de chiffres sur lesquelles on effectue les tests de bon fonctionnement de nos modèles.

Ces images ont la même taille et à l'intérieure les chiffres sont centrés et leur taille est normalisée.

Elle peut être utilisée dans l'apprentissage supervisé pour la classification car elle associe à chaque image de chiffre la classe qui lui appartient.

Pour notre cas, on a opté pour MNIST munie d'une extension pour les caractères alphabétiques afin d'enrichir nos données.

Interprétation des mesures de performance :

Une fois que nous avons construit notre modèle, la question la plus importante qui se pose est de savoir quelle est la qualité de ce modèle. L'évaluation de notre modèle est donc la tâche la plus

importante du projet, qui définit la qualité de nos prédictions.[FIGURE 4.32]

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

FIGURE 4.32. Mesures de performance. [Renuka, 2017]

Les vrais positifs et les vrais négatifs sont les observations correctement prédites et donc affichées en vert. Nous voulons minimiser les faux positifs et les faux négatifs afin qu'ils apparaissent en rouge. Ces termes sont un peu déroutants. Alors prenons chaque terme un par un et comprenons-le bien. [Renuka, 2017]

True Positives (TP) -

Ce sont les valeurs positives correctement prédites, ce qui signifie que la valeur de la classe réelle est égale à la valeur de la classe prédite .

True Negatives (TN) -

Ce sont les valeurs négatives correctement prédites, ce qui signifie que la valeur de la classe réelle est égale à la valeur de la classe prédite.

Faux positifs et faux négatifs, ces valeurs se produisent lorsque votre classe actuelle est en contradiction avec la classe prédite.

Faux positifs (FP) -

Quand la classe réelle est non et la classe prédite est oui.

Faux Négatifs (FN) -

Lorsque la classe réelle est oui mais que la classe prédite est non.

Une fois ces quatre paramètres effectués, on peut calculer les scores de précision, Accuracy, Recall et F1 score.

Accuracy-

est la mesure de performance la plus intuitive. On peut penser que, si nous avons une grande précision, notre modèle est meilleur. Oui, la précision est une excellente mesure, mais uniquement lorsque nous disposons de jeux de données symétriques dans lesquels les valeurs de faux positif et de faux négatif sont presque identiques. Pour notre modèle, nous avons 0,752 , ce qui signifie que notre modèle est d'environ . 75% précis.

Accuracy = TP + TN / TP + FP + FN + TN

Recall -

est le rapport entre les observations positives correctement prédites et toutes les observations de la classe réelle. Nous avons un Recall de 0,728, ce qui est bon pour ce modèle car il est supérieur à 0,5.

Recall = TP / TP + FN

Score F1 -

Le score F1 est la moyenne pondérée de la précision et du recall. Par conséquent, ce score prend en compte à la fois les faux positifs et les faux négatifs. Dans notre cas, le score F1 est 0.7303.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

1- Entraînement du modèle sur MNIST : Dans un premier temps, nous avons entraîné notre modèle sur la base de données MNIST simple (caractères numériques) pour voir les performances de notre architecture neuronale. On a effectué une étude paramétrique par rapport au nombre d'époques afin de voir au bout de quelle époque notre modèle entraîné sera stable et ainsi pour suivre l'évolution des valeurs des mesures par rapport aux nombres d'époques.

On a évalué notre modèle en utilisant l'algorithme de Adam comme optimiseur.
Les résultats obtenus sont résumées dans le tableau [TABLE 4.1]

Epoques	Accuarcy	Recall	F1-score
1000	75.6	1.035	0.9575
2000	80.54	1.030	0.9586
4000	89.681	1.015	0.96043
6000	98.69	1.006	0.96152
8000	98.698	1.006	0.96159
10000	98.7	1.006	0.9616

TABLE 4.1 – Résultat sur MNIST.

2- Entraînement du modèle sur Extended MNIST :

De la même manière ,on a entraîné notre modéle sur MNIST+caractères alphabétiques, on a effectué une étude paramétrique par rapport au nombre d'époques et on a obtenu les résultats présentés dans le tableau [TABLE 4.2]

Epoques	Accuarcy	Recall	F1-score
1000	69.95	0.7503	0.7445
2000	70.23	0.7452	0.7401
4000	71.98	0.7412	0.7390
6000	73.63	0.7308	0.7310
8000	75.21	0.7289	0.7304
10000	75.22	0.7288	0.7303

TABLE 4.2 – Résultat sur Extended MNIST.

On constate ici que l'accuracy se stagne aux alentours de 75 % Par contre , on voit ici qu'il y a une légère baisse des performances du modèle. Ceci est dû à la taille du nouveau dataset . La proportion des caractères alphabétique est inférieure à celle des caractères numériques. Un dataset plus complet pourra remédier à cela.

4.4.1 CROSS VALIDATION :

La validation croisée est une méthode statistique utilisée pour estimer l'habileté des modèles d'apprentissage automatique.

La procédure a un paramètre unique appelé k qui fait référence au nombre de groupes dans lesquels un échantillon de données donné doit être divisé. En tant que telle, la procédure est souvent appelée validation croisée des k-fold. Lorsqu'une valeur spécifique pour k est choisie, elle peut être utilisée à la place de k dans la référence au modèle, telle que k = 10 devenant une validation croisée de 10 fois. [Jason, 2018]

La validation croisée est principalement utilisée dans l'apprentissage automatique appliqué pour estimer les compétences d'un modèle d'apprentissage automatique sur des données invisibles. C'est-à-

dire d'utiliser un échantillon limité afin d'estimer comment le modèle devrait fonctionner en général lorsqu'il est utilisé pour faire des prédictions sur des données non utilisées pendant la formation du modèle.

La procédure générale est la suivante :
1-Mélangez le jeu de données de manière aléatoire.
2-Diviser le jeu de données en k groupes.
3-Pour chaque groupe unique :

- •Prendre le groupe comme un ensemble de données à retenir ou à tester.
- •Prendre les groupes restants comme un ensemble de données d'entraînement.
- •Ajuster un modèle sur le kit d'apprentissage et l'évaluer sur le kit de test.
- •Conserver le score d'évaluation et jeter le modèle.

4-Résumer les compétences du modèle à l'aide de l'échantillon de scores d'évaluation du modèle.

Il est important de noter que chaque observation de l'échantillon de données est affectée à un groupe individuel et reste dans ce groupe pendant la durée de la procédure. Cela signifie que chaque échantillon a la possibilité d'être utilisé dans le délai de conservation 1 fois et d'entraîner le modèle k-1 fois. [Jason, 2018]

En fixant la valeur de l'époque à 8000 (c'est l'époque où l'accuracy s'est stabilisée dans l'entraînement) et en variant la valeur de k de 5 à 10, on a obtenu les résultats suivants : [TABLE 4.3]

Valeur de K	**Précision**
5	74.521
6	75.21
7	74.845
8	75.112
9	75.157
10	75.013

TABLE 4.3 – Résultat du Cross Validation.

Discussions du résultat :

Après une évaluation répétée de notre modèle sur des valeurs différentes de K, on a aboutit à la valeur K=6 dans la [TABLE 4.3] qui était la plus optimal et la plus efficace pour notre cas, car elle nous a offert une meilleure précision. Cette valeur n'est jamais fixe, car elle peut changé selon le modèle de réseau de neurones utilisé.

4.5 Conclusion :

Nous avons présenté dans ce chapitre les outils de travail qui nous ont permis de réaliser notre application.

En premier nous avons présentés les outils et les langage utilisés, ensuite nous avons fait la présentation et explication de nos interfaces de travail concernant la GED et l'OCR, et pour finir on a parler des test, et résultats obtenus par notre OCR.

CONCLUSION GÉNÉRALE

De nos jours la croissance des documents est assez importante et devrai être gérée avec la méthode la plus rigoureuse qui soit, de même pour le problème de la saisie qui prend un temps considérable.

Le travail présenté dans ce mémoire aborde les différentes étapes nécessaires au déploiement de la GED Alfresco. Pour arriver au résultat attendu, nous avons commencé par comprendre le fonctionnement général de la GED, on a fait une étude comparative afin de choisir la GED qui correspond à nos attentes.

Par la suite on a étudié le concept de l'OCR et suivie les étapes de la chaîne de numérisation, afin de réaliser notre OCR avec la technique la plus adéquate, c'est-à-dire les CNNs, en présentant les différentes couches utilisées dans la classification tel que : la couche convolutionelle, la couche de pooling et la couche fully connected.

Nous avons rencontré quelques problèmes dans la phase de traitement de l'OCR. Qui a fait que le temps d'exécution était trop coûteux. Malgré les problèmes liés aux différents aspects techniques et architecturaux d'Alfresco, nous avons réussi à réaliser une application fiable et fonctionnelle. Ce travail pourra être amélioré par :

- L'extension de la reconnaissance optique de caractères sur d'autres types de documents de la

banque telle que les bordereaux, les rapports, les factures etc.

• Étendre le déploiement de cette application du réseau local du centre des archives a baba-hassen vers le territoire national.

• Prévoir un système de stockage externe vers des armoires de stockages ou bien le cloud.

ANNEXE

Indexation :

correspond à la représentation d'un texte ou d'un document par un indice ou un mot clé, avec l'aide ou non d'un langage documentaire, en vue d'en faciliter le repérage et la consultation.On distingue à ce titre : [Ged, 2017]

L'indexation par type : elle offre une description formelle du document en utilisant ses méta-données (type, auteur, titre, source, date, etc.) dont le vocabulaire est standardisé afin de permettre l'utilisation de ces méta-données par le plus grand nombre d'outils de recherche.

L'indexation par concepts ou mots-clés : elle vise plutôt le contenu du document pour faciliter les opérations de recherche. Il peut s'agir ici, pour le concepteur du système ou le créateur du document, de recenser les termes qui apparaissent le plus souvent; on parle alors d'indexation statistique. Il peut aussi s'agir d'un système plus évolué où le concepteur sélectionne les termes dans un thésaurus (liste de mots liés par des relations de hiérarchie ou d'équivalence) en rapport avec le document pour vous familiariser avec l'édition.

Archivage :

L'archivage électronique désigne le stockage à long terme de documents et données numériques.

L'archivage de contenus électroniques est l'ensemble des actions, outils et méthodes mis en œuvre pour réunir, identifier, sélectionner, classer, et conserver des contenus électroniques, sur un support sécurisé, dans le but de les exploiter et de les rendre accessibles dans le temps, que ce soit à titre de preuve (en cas d'obligations légales) ou à titre informatif.
Le contenu archivé est considéré comme figé et ne peut donc être modifié. Ceci est notamment possible

en garantissant l'authenticité via l'empreinte électronique, la signature électronique, la traçabilité des accès et bien d'autres moyens. La durée de l'archivage va en fonction de la valeur du contenu et porte le plus souvent sur du moyen ou long terme. La conservation est l'ensemble des moyens mis en œuvre pour stocker, sécuriser, pérenniser, restituer, tracer, transférer , les contenus électroniques archivés. [Universalis, 2019]

LISTE ABRÉVIATION

Abréviation	Désignation
GED	Gestion électronique des documents.
EDM	Electronic document management.
ICR	Intelligent Character Recognition.
OCR	Optical character recognition.
SGBD	Système de gestion de base de données.
API	Application Programming Interface.
SI	Système d'information.
ROC	Reconnaissance optique de caractères.
SVM	Support vector machine.
LSVM	Linear Support Vector Machine.
CNN	Convolutional Neural Networks.
UML	Unified Modeling Language.
RELU	Rectified linear units.
HTML	HyperText Markup Language.
CSS	Cascading Style Sheets.
PHP	Hypertext Preprocessor.
XAMPP	X (cross) Apache MariaDB Perl PHP.
XHTML	Extensible HyperText Markup Language.
RELU	Rectified linear units.
SGBDRO	Système de gestion de base de données relationnel et objet.
SQL	Structured Query Language.
EDI	Environnement de développement intégré.
ROI	Region of interest.

MNIST	Mixed National Institute of Standards and Technology.
EMNIST	Extended Mixed National Institute of Standards and Technology.
TP	True positive.
TN	True negative.
FP	False positive.
FN	False negative.
BNA	Banque national d'Algerie.

TABLE 4.4 – Liste des abréviations..

BIBLIOGRAPHIE

[Abdelhak, 2011] BouKHAROUBA Abdelhak. Contribution à la segmentation et à la reconnaissance de l'écriture arabe manuscrite. 2011.

[Alfresco, 2019] Alfresco. Alfresco / GED ECM / Guide Open Source, 2019. http://www.open-source-guide.com/Solutions/Applications/Ged-ecm/Alfresco.

[AlfresDoc, 2007] AlfresDoc. Démarrer avec Alfresco Share | Alfresco Documentation, 2007. https://docs.alfresco.com/using-fr/concepts/gs-intro.html.

[Augereau, 2013] Olivier Augereau. Reconnaissance et classification d'images de documents. feb 2013.

[Baka and Fillali, 2016] Abdeladim Baka and Hicham Fillali. *Traitement et reconnaissance des caracteres*. PhD thesis, 2016.

[Bruno, 2005] Bruno. Descente de gradient, 2005. https://www.math-info.univ-paris5.fr.

[Cédric Ademain, 2017] BPMS Cédric Ademain. 4 étapes à maitriser pour mettre en place un outil de GED - BPMS.info, 2017. https://www.bpms.info/4-etapes-a-maitriser-pour-mettre-en-place-un-outil-de-ged/.

[Chabin, 2007] Marie-Anne. Chabin. *Archiver, et apres ?* Djakarta, 2007.

[Charle, 2017] Crouspeye Charle. Comment les Réseaux de neurones à convolution fonctionnent, 2017. https://medium.com/@CharlesCrouspeyre/comment-les-r{é}seaux-de-neurones-{à}-convolution-fonctionnent-b288519dbcf8.

[Chevalier, 2004] Sylvain Chevalier. Reconnaissance d'écriture manuscrite par des techniques markoviennes : une approche bidimensionnelle et générique. dec 2004.

[Cory, 2019] Maklin Cory. vecteur de support - Vers la science des données, 2019. https://towardsdatascience.com/support-vector-machine-python-example-d67d9b63f1c8.

[Degeans, 1991] Isabelle Degeans. Étude sur la gestion électronique de documents FIFTI SA/rapport de stage. page 68 p, 1991.

[des, 2009] *Tutoriel XAMPP : Comment utiliser XAMPP pour exécuter votre propre serveur Web – Des Geeks et des Lettres*, 2009. https://desgeeksetdeslettres.com/web/xampp-plateforme-pour-heberger-son-propre-site-web.

[Dewey, 2017] Dewey. Indexation et classification | Enssib, 2017. https://www.enssib.fr/services-et-ressources/questions-reponses/indexation-et-classification.

[Fallis, 2013] A.G Fallis. No Title No Title, 2013.

[Fethallah *et al.*, 2017] Hadjila Fethallah, Merzoug Mohammed, Bekaddour Akkacha, and Smahi Mohammed Ismail. Classification des images par les réseaux de neurones. 2017.

[Freddy, 2015] ILUNGA KADIATA Freddy. Développement d'une application web pour l'optimisation du processus d'archivage et d'accès aux données d'une entreprise. Cas de Bell Equipement. - Freddy ILUNGA KADIATA, 2015. https://www.memoireonline.com/09/18/10304/m{_}Developpement-d-une-application-web-pour-l-optimisation-du-processus-d-archivage-et-da
html.

[Ged, 2017] Ged. Indexation, Classification et Recherche | Gestion Documentaire - GED.fr, 2017. https://www.ged.fr/indexation/.

[Gillies and Cailliau, 2000] James. Gillies and R. Cailliau. *How the Web was born : the story of the World Wide Web*. Oxford University Press, 2000.

[Gir, 2004] *Introduction au cours HTML et CSS - Pierre Giraud*, 2004. https://www.pierre-giraud.com/html-css-apprendre-coder-cours/introduction/.

[Git, 2008] tensorflow/metrics.py at r1.14 · tensorflow/tensorflow · GitHub, 2008. https://github.com/tensorflow/tensorflow/blob/r1.14/tensorflow/python/keras/metrics.py{#}L562-L594.

[Grierson and Schiefelbein, 2002] C. Grierson and J. Schiefelbein. *The Arabidopsis Book*. American Society of Plant Biologist, 2002.

[Hund, 2001] Thierry Hund. Etude de faisabilité, étude de marché et sélection d'un progiciel de GED pour le groupe SAUR. 2001.

[ImagingMadeSimple, 2017] ImagingMadeSimple. ImagingMadeSimple /Guide Open Source, 2017. http://www.open-source-guide.com/Solutions/Applications/Ged-ecm/ImagingMadeSimple.

[Jason, 2018] Browlee Jason. Une introduction en douceur à la validation croisée des k-fold, 2018. https://machinelearningmastery.com/k-fold-cross-validation/?fbclid=IwAR3dUhxFiRNQeXIh8n-FFLf13bwXWjUTOychk67nIyShT9dCpwTaKHrtj5s.

[Jason, 2019] Brownlee Jason. Comment les couches convolutives fonctionnent-elles dans les réseaux de neurones à apprentissage en profondeur ?, 2019. https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/.

[Jones and Smirnoff, 2006] M. Jones and N. Smirnoff. Nuclear dynamics during the simultaneous and sustained tip growth of multiple root hairs arising from a single root epidermal cell. *J. of Exp. Bot.*, 57(15) :4269–4275, 2006.

[Kushal, 2018] Kushal. Classification des données à l'aide de machines à vecteurs de support (SVM) en Python - GeeksforGeeks, 2018. https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-python/?fbclid=IwAR2j-l6AKUnrro9xooMjM-r{_}9qoJho{_}a8g-Fs5FKeg-ezS5BmKcejJy6a-E.

[laurant Bastien, 2019] laurant Bastien. Réseau de neurones artificiels : qu'est-ce que c'est et à quoi ça sert ?, 2019. https://www.lebigdata.fr/reseau-de-neurones-artificiels-definition.

[Lie, 2005] Håkon Wium Lie. *Cascading Style Sheets*. université d'Oslo, 2005.

[Lutz, 2009] Mark. Lutz. *Learning Python*. O'Reilly Media, oct 2009.

[Manuel, 2014] Manuel. Fiche ressource 4 - La gestion électronique des documents (GED), 2014. https://www.i-manuel.fr/SP{_}AD/SP{_}ADdocfic4.htm.

[Masucci and Schiefelbein, 1994] J. D. Masucci and J. W. Schiefelbein. The rhd6 mutation of arabidopsis thaliana alters root-hair initiation trhough an auxin- and ethylene-associated process. *Plant. Physiol.*, 106 :1335–1346, 1994.

[NOVAXEL, 2019] NOVAXEL. NOVAXEL ged - Trimarg, 2019. http://www.trimarg.fr/produit/novaxel-ged/.

[Orhan, 2018] Gazi Yalcin Orhan. Image Classification, 2018. https://towardsdatascience.com/image-classification-in-10-minutes-with-mnist-dataset-54c35b77a38d.

[Payne and Grierson, 2009] R.J.H. Payne and C.S. Grierson. A theoretical model for rop localisation by auxin in arabidopsis root hair cells. *PLoS ONE*, 4(12) :e8337. doi :10.1371/journal.pone.0008337, 2009.

[Pos, 2005] *PostgreSQL*, 2005. https://sql.sh/sgbd/postgresql.

[RecFind, 2018] RecFind. RecFind : Logiciel de Gestion documentaire (GED) - Avis et prix, 2018. https://www.appvizer.fr/collaboration/gestion-documentaire-ged/recfind.

[Renuka, 2017] Joshi Renuka. Accuracy, Precision, Recall F1 Score : Interpretation of Performance Measures, 2017. https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

[Report2Web, 2017] Report2Web. Report2Web : Logiciel de Gestion documentaire (GED) - Avis et prix, 2017. https://www.appvizer.fr/collaboration/gestion-documentaire-ged/report2web.

[Rigas *et al.*, 2001] S. Rigas, G. Debrosses, K. Haralampidis, F. Vicente-Angulo, K. A. Feldman, A. Grabov, L. Dolan, and P. Hatzpoulos. Trh1 encondes a potassium transporter required for tip growth in arabidopsis root hairs. *The Plant Cell*, 13 :139–151, 2001.

[SAIMADHU, 2017] POLAMURI SAIMADHU. Difference Between Softmax Function and Sigmoid Function, 2017. http://dataaspirant.com/2017/03/07/difference-between-softmax-function-and-sigmoid-function/.

[Sarah, 2015] Ponchin Sarah. Bruit numérique : définition, comment l'éviter ou le corriger, 2015. https://www.linternaute.com/photo_numerique/retouche-photo/1261891-bruit-numerique-definition-comment-eviter-corriger/.

[Senn, 2009 accessed February 3 2014] Mark Senn. *Using LaTeX for Your Thesis*, 2009 (accessed February 3, 2014). http://engineering.purdue.edu/~mark/puthesis.

[Soua, 2016] Mahmoud Soua. Extraction hybride et description structurelle de caractères pour une reconnaissance efficace de texte dans les documents hétérogènes scannés : Méthodes et Algorithmes parallèles. nov 2016.

[Susmith, 2019] Reddy Susmith. Segmentation in OCR, 2019. https://medium.com/@susmithreddyvedere/segmentation-in-ocr-10de176cf373.

[Tresorit, 2019] Tresorit. Tresorit : Logiciel de Gestion documentaire (GED) - Avis et prix, 2019. https://www.appvizer.fr/collaboration/gestion-documentaire-ged/tresorit.

[Uniqtech, 2018] Uniqtech. Understand the Softmax Function in Minutes - Data Science Bootcamp - Medium, 2018. https://medium.com/data-science-bootcamp/understand-the-softmax-function-in-minutes-f3a59641e86d.

[Universalis, 2019] Universalis. ARCHIVAGE NUMÉRIQUE, La mise en pratique de l'archivage numérique - Encyclopædia Universalis, 2019. https://www.universalis.fr/encyclopedie/archivage-numerique/6-la-mise-en-pratique-de-l-archivage-numerique/.

[VAN, 2001] DROOGENBROECK VAN. Acquisition et traitement de l'image, 2001. https://orbi.uliege.be/bitstream/2268/1768/1/totaliBIO.pdf.

[w3s, 2003] *PHP Tutorial*, 2003. https://www..com/php/.

[WIki, 2019] WIki. Apache Tomcat — Wikipédia, 2019. https://fr.wikipedia.org/wiki/Apache{_}Tomcat.

[will, 2015] will. IntelliJ IDEA — Wikipédia, 2015. https://fr.wikipedia.org/wiki/IntelliJ{_}IDEA.