

Table des matières

1	Introduction générale	2
2	Contexte Et Cadre Du Projet	4
2.1	Contexte	4
2.2	Problématique	5
2.3	Motivation Et Objectif	6
3	Etat De L'art	7
3.1	la Reconnaissance Optique Des caractères	7
3.1.1	Histoire	7
3.1.2	Fonctionnement	8
3.1.3	Avantage	9
3.2	Outils de reconnaissance optique de caractères	9
3.2.1	GOCR	10
3.2.2	Ocrad	11
3.2.3	Tesseract	12
3.3	Logiciel De Reconnaissance Optique De Caracteres	14
3.3.1	FreeOCR	14
3.3.2	ABBYY	14
3.3.3	Ocr Online	16
4	Analyse Et Conception De La Solution	18
4.1	Analyse	18
4.1.1	Besoin fonctionnel	19
4.1.2	Besoin Non Fonctionnel	20
4.2	Conception	20
4.2.1	Modélisation	20
4.2.2	Diagramme de séquence	27
4.2.3	Diagramme de séquence du moteur de recherche	28
5	Implémentation De La Solution Et Analyse Des Résultats	38
5.1	Environnement Matériel Et Logiciel	38
5.1.1	Platefome web	38
5.1.2	Le langage BATCH	39
5.1.3	Photoshop	39
5.2	API Et Drivers Utilisés	39
5.2.1	Jquery	39
5.2.2	ImageMagick	39

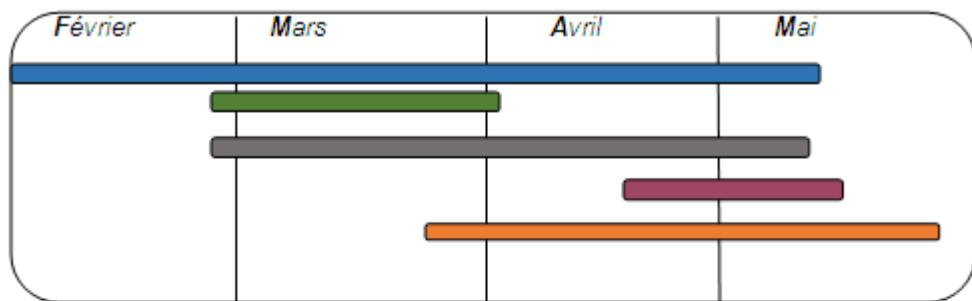
5.2.3	Tesseract	39
5.3	Analyse Et Test	39
5.3.1	Installation	39
5.3.2	Quelques Interfaces	40
5.4	Teste	48
5.4.1	Enregistrement du modèle :Image numérisée vide	48
5.4.2	Enregistrement du modèle :Texte Extrait par Tesseract	49
5.4.3	Table Champs :Marqueurs saisis Par l'utilisateur	50
5.4.4	Lancement du traitement : Image numérisée rempli	51
5.4.5	Texte Extrait par Tesseract après le traitement	52
5.4.6	Tableau contenant tous les mots du texte extrait après traitement	53
5.4.7	Table champ :Résutat du traitement	54
5.4.8	Affichage du Résultat	55
5.4.9	Nouveau Résultat après changemnt de marqueur Debut	56
6	CONCLUSION GENERALE	57
6.1	Synthèse	57
6.2	Difficultés rencontrées	57
6.3	Perspectives	58
7	Références et liens	59

Chapitre 1

Introduction générale

La mise en place d'une bonne infrastructure informatique est un excellent moyen pour une entreprise d'améliorer son organisation, son stockage de données et même sa productivité. L'informatique permet véritablement d'accroître l'efficacité opérationnelle d'une société en permettant d'améliorer sa réactivité tout en rendant certaines tâches faciles et automatiques. Le but de ce projet est de s'initier à l'élaboration d'un cadre de numérisation et de reconnaissance de documents papier imprimer, à partir d'un module OCR¹ de source libre, dans le but de concevoir une base de données qui sera constituée de toutes les informations provenant de ces documents. Ce qui permettra de gérer plus facilement les documents administratifs gérés par plusieurs Service public. Il a fallut nous organiser pour atteindre notre objectif, c'est la raison pour laquelle nous vous présentons ci-après une table qui montre l'évolution de notre travail .

1. OCR :Optical Character Recognition



Légende des couleurs :

- Recherche et Etude sur la reconnaissance optique de caractère
- Conception des différents diagrammes et cas d'utilisation
- Développement de l'application
- Tests
- Rédaction du rapport.

Figure 1 Durées approximatives des étapes de déroulement du projet

Fig 0-Evolution du projet

Chapitre 2

Contexte Et Cadre Du Projet

Ce chapitre décrit le contexte dans lequel nous avons effectué notre travail et formulé la problématique associée à ce dernier.

2.1 Contexte

Dans ce projet nous allons nous intéresser au Centre d'enregistrement et de révision des formulaires administratifs connus sous le nom de CERFA, géré par un organisme public qui centralise les documents officiels et fait le lien entre les usagers et les différentes administrations et pouvoirs publics français. Tous les formulaires administratifs possibles et imaginables sont gérés par cet organisme, ce qui nous laisse pensé à un champ d'action plutôt vaste.

Un formulaire Cerfa est un document officiel doté d'un identifiant unique et normé par le Ministère. Il offre la possibilité d'effectuer une demande claire pour une situation donnée. Il permet également à l'administration de disposer de tous les renseignements nécessaires à l'instruction ou à l'étude d'une demande tout en fournissant une réponse dans les meilleurs délais.

Ex : la demande d'allocation de parent isolé est le Cerfa n°10603*03, la déclaration trimestrielle pour l'allocation de parent isolé, étant le Cerfa n°10604*01, Les demandes d'entente préalable de l'assurance maladie sont par exemple à faire sur le Cerfa n°12040*01, la demande de certificat d'immatriculation d'une voiture correspond au Cerfa n°10672*03.

DEMANDE DE L'AIDE À LA CRÉATION ET À LA REPRISE D'UNE ENTREPRISE (ACCRE)



N° 13564*02

s'il y a plusieurs demandeurs, remplir autant de formulaires qu'il y a de demandeurs.

- Demande de l'ACCRE au moment de la déclaration d'entreprise : _____
 Demande de l'ACCRE postérieure au dépôt de déclaration d'entreprise (dans les 45 jours suivant la déclaration d'entreprise).
 Préciser le numéro SIRET de l'entreprise : _____

Création d'une entreprise individuelle : compléter les cadres 1, 2, 4 et 5.
 Création ou reprise d'une société : compléter tous les cadres de 1 à 5.

RÉSERVÉ AU CFE U E

Déclaration n° _____

Reçue le _____

Transmise le _____

DÉCLARATION RELATIVE AU DEMANDEUR

1 NOM DE NAISSANCE

Prénom : _____
 Numéro de Sécurité Sociale du demandeur : _____
 Domicile : rès., bâti., n°, voie, lieu-dit _____
 Code postal : _____ Commune / Pays : _____

Nom d'usage : _____

Nationalité : _____ Sexe : M F Né(e) le : _____

Numéro de téléphone personnel : _____

Perain : Commune de rattachement administratif : _____

Code postal : _____ Nom de la commune : _____

SITUATION DU DEMANDEUR

- Demandeur d'emploi indemnisé ou indemnisable
 Demandeur d'emploi non indemnisé inscrit à Pôle Emploi six mois au cours des dix-huit derniers mois
 Bénéficiaire : – du RSA – du RMI – de l'ASS – de l'ATA (1)
 Jeune de 18 à 25 ans révolus
 Personne de moins de 30 ans non indemnisée ou reconnue handicapée
 Salarié ou personne licenciée d'une entreprise en redressement, liquidation judiciaire ou sauvegarde qui reprend l'activité de l'entreprise
 Personne créant une entreprise implantée au sein d'une zone urbaine sensible
 Bénéficiaire du complément de libre choix d'activité

(1) Parmi les allocataires de l'allocation temporaire d'attente (ATA) sont éligibles à l'ACCRE : les bénéficiaires de la protection subsidiaire autorisés à exercer une activité, les ressortissants étrangers auxquels une carte de séjour temporaire a été délivrée, et les personnes en attente de réinsertion (anciens détenus et salariés expatriés non admis au régime d'assurance chômage).

Niveau de formation (cf. notice) : _____
 Motif d'inscription à Pôle Emploi (cf. notice) : _____
 Qualification du dernier Emploi occupé (cf. notice) : _____
 Date d'Inscription à Pôle Emploi : le _____

3 Pour une société

Dénomination sociale : _____

Le demandeur :

- détient avec sa famille plus de 50 % du capital dont 35 % au moins à titre personnel
 est dirigeant et détient directement ou avec sa famille au moins un tiers du capital du moins à titre personnel, aucun autre associé hors de sa famille ne détient plus de 5 %
 détient, avec les autres demandeurs d'ACCRE, plus de 50 % du capital de la société des demandeurs à la qualité de dirigeant, et chaque demandeur détient une part du moins égale à 10 % de la part détenue par le principal actionnaire ou porteur de parts

Nombre total d'associés (y compris le demandeur) : _____

Le demandeur est titulaire d'un contrat d'appui au projet d'entreprise (CAPE) :
 Le demandeur est en cours d'accompagnement dans le cadre du parcours NACRE :
 Nombre d'emplois (y compris le demandeur) : – créés _____ (en cas de création)
 – repris _____ (en cas de reprise)

5 J'atteste sur l'honneur que l'ACCRE ne m'a pas été accordée au cours des 3 dernières années et que les renseignements ci-dessus sont exacts, sous peine des sanctions prévues par la loi.

Date : _____ Signature du demandeur : _____

CADRE RÉSERVÉ À L'URSSAF

- Demande acceptée
 Demande refusée Motif : _____

N° d'enregistrement du dossier : _____ Date : _____

La loi n° 78-17 du 8 janvier 1978 modifiée, relative à l'informatique, aux fichiers et aux libertés s'applique aux réponses des personnes physiques à ce questionnaire. Elle leur garantit un droit d'accès et de rectification, pour les concernant, auprès des organismes destinataires de ce formulaire.

Fig 2-cerfa

2.2 Problématique

Les services publics constituent le coeur de l'activité de l'Etat et des collectivités territoriales. Ils ont connu une irrésistible progression et aujourd'hui, le concept de service public est confronté à l'évolution de la société française ainsi qu'aux attentes de l'Union Européenne. Il doit donc se montrer souple et faire preuve d'adaptabilité pour ne pas devenir complexe et inefficace, voire déçu. Ainsi, dans le soucis de satisfaire sa mission principale qui est de répondre dans les meilleurs délais, le service public doit disposer d'un système lui permettant de traiter, enregistrer et rechercher rapidement une demande (Cerfa).

2.3 Motivation Et Objectif

L'objectif de ce projet est de concevoir et d'implémenter un logiciel qui permettra d'automatiser la plus part des tâches liées aux formulaires Cerfa tout en minimisant au maximum l'effort humain, en réduisant le temps de traitement et en accélérant la recherche des documents par une structure bien organisée.

Conclusion :

En effet, la mise en place d'un système permettant de traiter et de sauvegarder d'une manière automatique le Cerfa permettra aux services publics d'accélérer le traitement de toutes les démarche administratives et de répondre dans un bref délai à toutes les requêtes de la population.

Chapitre 3

Etat De L'art

Dans ce chapitre, nous illustrerons le concept de la reconnaissance de caractères optique, en passant par son histoire, en décrivant son fonctionnement, et en donnant son utilité. Par la suite nous citerons quelques outils de reconnaissance de caractère optique et certains logiciels qui ce sont appuyé sur ces outils pour fournir des services garantissant la reconnaissance de texte.

3.1 la Reconnaissance Optique Des caractères

3.1.1 Histoire

En 1950, Frank Rowlett, qui avait cassé le code diplomatique japonais PURPLE, demanda à David Shepard, un cryptanalyste de l'AFSA (prédecesseur de la NSA américaine), de travailler avec Louis Tordella pour faire à l'agence des propositions de procédures d'automatisation des données. La question incluait le problème de la conversion de messages imprimés en langage machine pour le traitement informatique. Shepard décida qu'il devait être possible de construire une machine pour le faire, et, avec l'aide de Harvey Cook, un ami, construisit "Gismo" dans son grenier pendant ses soirées et ses week-ends.

Le fait fut rapporté dans le Washington Daily News du 27 avril 1951 et dans le New York Times du 26 décembre 1953 après le dépôt du brevet numéro 2 663 758. Shepard fonda alors Intelligent Machines Research Corporation (IMR), qui livra les premiers systèmes de ROC au monde exploités par des sociétés privées. Le premier système privé fut installé au Readers Digest en 1955, et, de nombreuses années plus tard, fut offert par le Readers Digest au Smithsonian, où il fut mis en exposition. Les autres systèmes vendus par IMR à la fin des années 1950 comprenaient un lecteur de bordereau de facturation à l'Ohio Bell Telephone Company et un numériseur (scanner de documents) à l'US Air Force pour la lecture et la transmission par télex de messages dactylographiés. IBM et d'autres utilisèrent plus tard les brevets de Shepard¹

1. brevets de Shepard :<http://www.plastifieuse.net/imprimerie8.php>

3.1.2 Fonctionnement

Un système ROC part de l'image numérique réalisée par un scanner optique d'une page (document imprimé, feuillet dactylographié, etc.), ou une caméra numérique, et produit en sortie un fichier texte en divers formats (texte simple, formats de traitements de texte, XML...).

Certains logiciels tentent de conserver l'enrichissement du texte (corps, graisse et police) ainsi que la mise en page, voire de rebâtir les tableaux et d'extraire les images. Certains logiciels comportent, en outre, une interface pour l'acquisition numérique de l'image.

Jusqu'à une date récente, le fonctionnement des systèmes ROC performants était peu connu car protégé par le secret industriel; les logiciels open-source disponibles (ex : GOcr) étant plutôt l'œuvre d'amateurs. La publication en open-source de systèmes performants (en particulier Tesseract en 2006) a quelque peu changé cette situation. Les étapes de traitement peuvent être schématisées ainsi :

A. Préanalyse de l'image

Le but est d'améliorer éventuellement la qualité de l'image. Ceci peut inclure le redressement d'images inclinées ou déformées, des corrections de contraste, le passage en mode bicolore (noir et blanc, ou plutôt papier et encre), la détection de contours.

B. Segmentation

Segmentation en lignes et en caractères (ou Analyse de page) : vise à isoler dans l'image les lignes de texte et les caractères à l'intérieur des lignes. Cette phase peut aussi détecter le texte souligné, les cadres, les images.

C. Reconnaissance proprement dite des caractères

Après normalisation (échelle, inclinaison), une instance à reconnaître est comparée à une bibliothèque de formes connues, et on retient pour l'étape suivante la forme la plus " proche " (ou les N formes les plus proches), selon une distance ou une vraisemblance (likelihood). Les techniques de reconnaissance se classent en quelques grands types¹ :

- **Classification par Caractéristiques (Features)** : Une forme à reconnaître est représentée par un vecteur de valeurs numériques - appelées features en anglais - calculées à partir de cette forme. Le nombre de features est de l'ordre de 100 à 300. Si les features sont bien choisies, une classe de caractères (par exemple l'ensemble des A majuscules) sera représentée par un " nuage " contigu de points dans l'espace vectoriel des features. Le rôle du classificateur est de déterminer à quel nuage (donc à quelle classe de caractères) la forme à reconnaître appartient le plus vraisemblablement. La classification fait généralement appel à divers types de réseaux de neurones artificiels entraînés sur de vastes bases de formes possibles.
- **Méthodes métriques** : Consistent à comparer directement la forme à reconnaître, au moyen d'algorithmes de distance, avec un ensemble de modèles appris. Ce type de méthode est peu utilisé et peu valorisé par les chercheurs, car souvent plus naïf et vraisemblablement moins efficace que les méthodes à base de features.

- **Méthodes statistiques** : Dans le domaine de la reconnaissance d'écriture manuscrite, il est fréquemment fait appel aux méthodes probabilistes/statistiques comme les chaînes de Markov.

D.Post-traitement

Utilisant des méthodes linguistiques et contextuelles pour réduire le nombre d'erreurs de reconnaissance : systèmes à base de règles, ou méthodes statistiques basées sur des dictionnaires de mots, de syllabes, de N-grammes (séquences de caractères ou de mots). Dans les systèmes industriels, des techniques spécialisées pour certaines zones de texte (noms, adresses postales) peuvent utiliser des bases de données pour éliminer les solutions incorrectes.

E.Génération

Génération du format de sortie, avec la mise en page pour les meilleurs systèmes.

3.1.3 Avantage

- **La recherche par Mot-clé(s)** : les utilisateurs peuvent facilement faire une recherche par mots-clés connexes parmi les documents stockés. Le contenu du texte d'une image est reconnu et indexé ; les utilisateurs peuvent rechercher un document par son titre, par le résumé, par le contenu ou d'autres mots clés pertinents présents dans ce document.
- **Gestion** : le contenu du texte d'un document devient éditabile comme un document ordinaire. Par conséquent, le contenu du document peut être facilement copié, collé et réutilisé.
- **Prise en charge multilingue** : la fonction ROC est disponible pour plusieurs langues.

3.2 Outils de reconnaissance optique de caractères

Nous allons comparer quelques outils de reconnaissance de caractère tout en effectuant un test de chacun d'eux sur la figure ci-dessous.

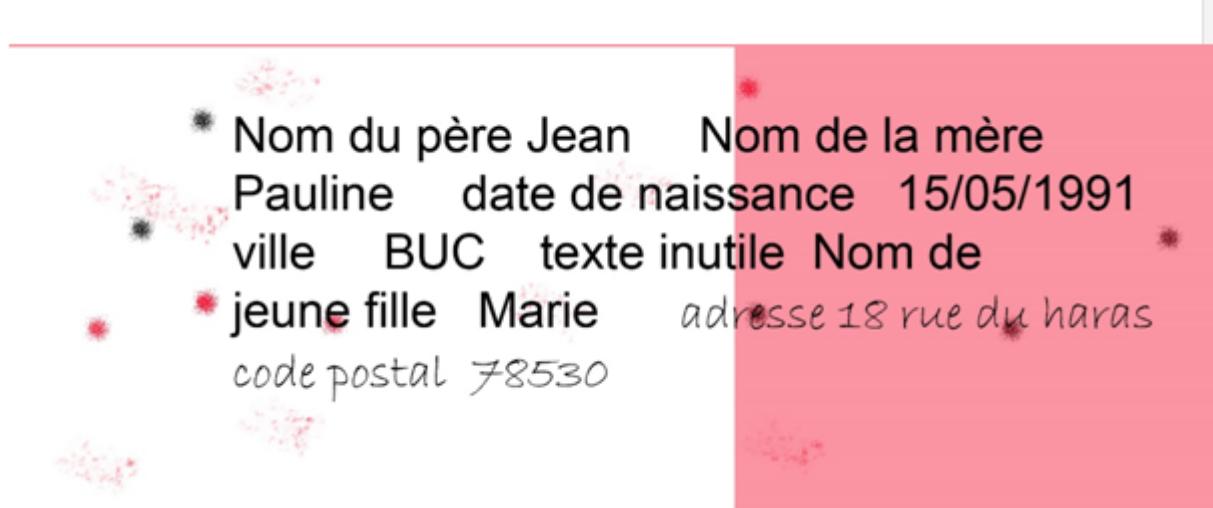
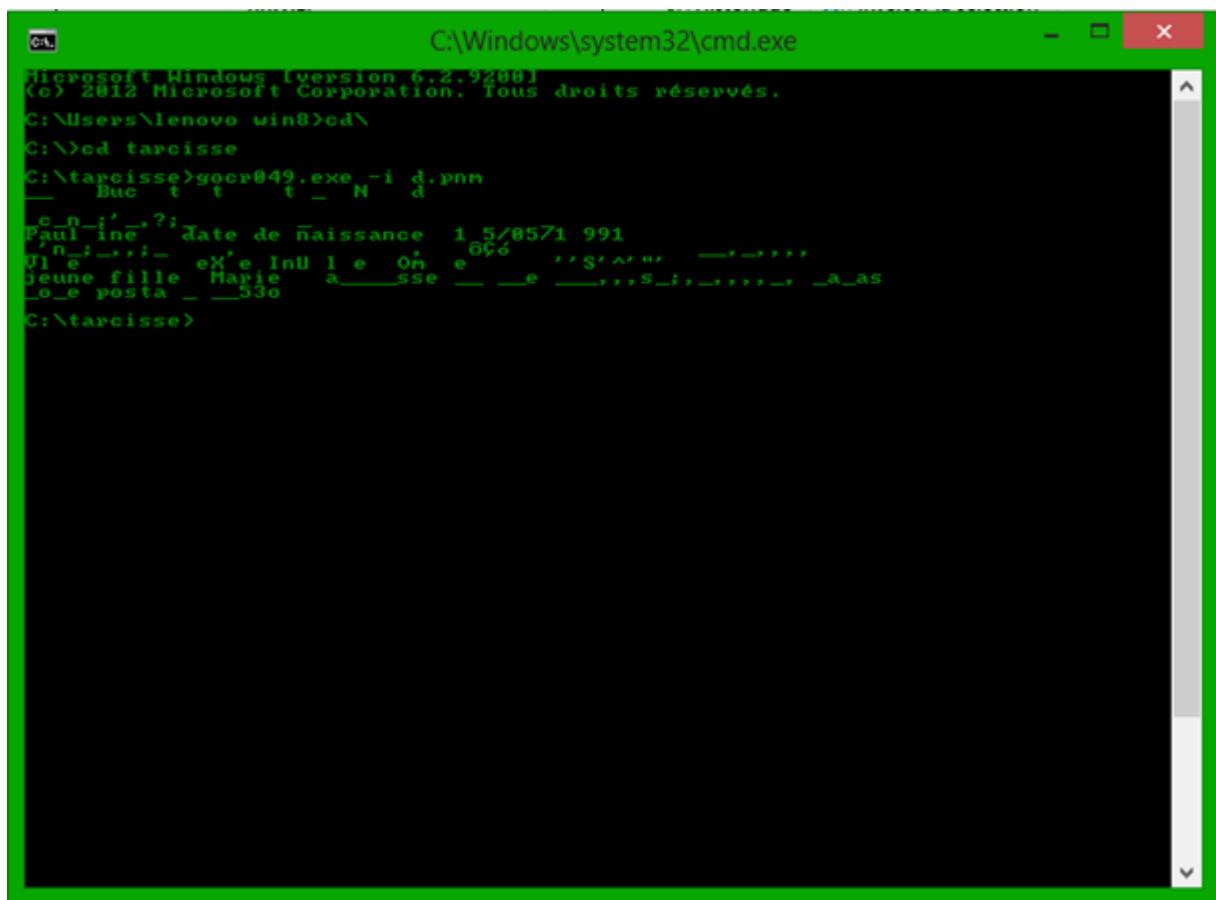


Fig 3-image de teste

3.2.1 GOCR

GOCR² est un outil de reconnaissance de caractères initialement mis en place par l'allemand **Jörg Schulenburg**. Il est capable de convertir une image numérisée (fichiers , pixmap portable ou PCX,PBM,BMP,GIF)) en un document pouvant être édité à l'aide d'un traitement de texte . Ses caractéristiques se reposent sur sa portabilité(OS), et détecte automatiquement les mots liées selon la diversité des langages. Il dispose également d'une version binaire sur Windows depuis octobre 2010 restreint au format ppm pour de raison de qualité. En exécutant l'image de la figure — sur GOCR en ligne de commande nous obtenons comme résultat :



The screenshot shows a Windows Command Prompt window titled 'C:\Windows\system32\cmd.exe'. The command entered is 'gocr849.exe -i d.pnm'. The output displays the text from the scanned document, which includes 'Pauline date de naissance 15/05/1991', 'Une jeune fille Marie assise', and 'lorsque posta 530'. The text is displayed in a monospaced font, with some characters appearing slightly distorted or illegible.

Fig 4-GOCR

Apres extraction, nous constatons qu'il y'a quelques restrictions lié à sa performance :

- Il est capable de gérer uniquement une colonne sans-serif de 20 à 60 pixel

2. GOCR :Optical Character Recognition GNU and GNU :GNU's NOT UNIX

- Il a une bonne extraction uniquement pour des documents présentés de manière simpliste ,de texte propre (avec une haute résolution,sans poussière).
- Il rencontre des difficultés d'extraction avec des polices sans empattement (Arial, Helvetica, Geneva et Verdana) qui se chevauchent ,ainsi qu'avec les textes manuscrits .Son extraction est plutôt mauvaise sur des polices hétérogènes que sur la police latin.

GOCR fournit aussi une bibliothèque permettant le développement de son propre OCR. " libgocr " codé entièrement en C.

3.2.2 Ocrad

Ocrad³ est un outil de reconnaissance optique de caractères, autorisé sous licence GNU GPL, développé par Antonio Diaz Diaz depuis 2003.Plusieurs version ont vu le jour notamment la version 0.7 publié en Février 2004, la version 0,14 publié en Février 2006 et la version 0,18 publié en mai 2009.C'est un outil principalement développé sous le langage C++ , basé sur une méthode d'extraction particulière.Le type d'image qu'il manipule est de type pixmap portable qui est par la suite convertit en un texte de format UTF-8. Il possède aussi un analyseur de mise en page, qui est en mesure de séparer des blocs de texte qui se trouvent sur les pages imprimées. En effectuant un teste sur l'image de la figure nous obtenons ceci comme résultat :

3. Ocrad :Optical Character Recognition

The screenshot shows a Microsoft Windows Command Prompt window titled 'C:\Windows\system32\cmd.exe'. The window contains the following text:

```
Microsoft Windows [version 6.2.9200]
(c) 2012 Microsoft Corporation. Tous droits réservés.
C:\Users\lenovo\win8>cd \
C:\>cd tarcisse
C:\tarcisse>ocrad.exe d.pbm o.txt
Nom du père Jean Nom de la mère
Pauline date de naissance 15/05/1991
ville Bug texte inutile Nom de
jeune fille Marie ad_5e_g vKé du hava5
Lode postal _g530
ocrad: bad magic number - not a pbm, pgm or ppm file.
C:\tarcisse>
```

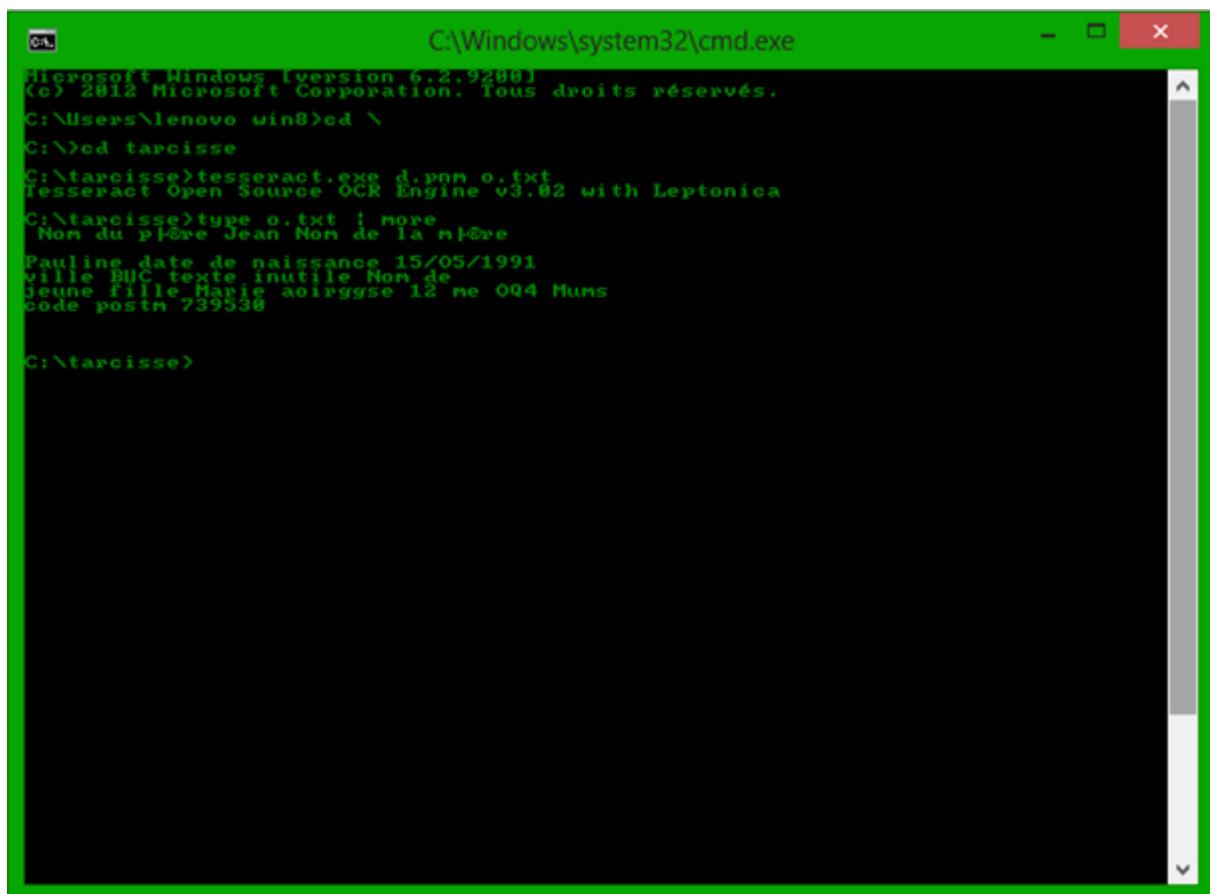
Fig 4-Ocrad

3.2.3 Tesseract

Tesseract est un outil initialement développé entre 1985 et 1994 par des ingénieurs dans les **laboratoires Hewlett Packard**(HP,société multitanionale de technologie de l'information) de Bristol ,en Angleterre, Greely Colorado, aux Etats-Uni comme logiciel privée, avec quelques modifications apportées en 1996 au port de Windows, et une certaine migration de C à C ++ en 1998.Etant donné que très peu de travaux ont été élaboré dans la décennie suivante, son code source a été rendu disponible en open source en 2005 par quelques ingénieurs de Hewlett Packard et de l' Université du Nevada, Las Vegas (UNLV).En 2006 son développement a été parrainé par Google. **Tesseract** est un moteur de reconnaissance simple (ecrit sous plusieurs langages de programmation,c++,php,java,c,pyton etc), dans le sens où il ne fournit pas d'interface utilisateur, n'effectue pas d'analyse de la mise en page et ne formate pas les résultats qu'il produit(version 2). Une autre de ses limitations est qu'il reconnaît uniquement les caractères US-ASCII et donc ne fonctionne correctement qu'avec des documents rédigés en langue anglaise.Sachant que l'acquisition de documents en niveau de gris ou en couleurs restait difficile, l'analyse d'un document produisait une sortie médiocre .La version 3.00 de **Tesseract** a apporté

une amélioration notamment, sur l'analyse des zones de la page avec une mise en forme structuré en sortie , et un certains nombre de nouveau format supporté, ajouté à l'aide de la bibliothèque Leptonica⁴. En plus de cela, **Tesseract** est capable de détecter si le texte est à espacement fixe ou proportionnel. Sa compilation et son exécution sont possibles aussi bien sous système GNU/Linux que système Microsoft. Les premières versions de **Tesseract** ne pouvaient reconnaître correctement que du texte en langue anglaise. Depuis la version 3, il est capable de reconnaître du texte écrit en français, anglais, arabe, bulgare, catalan, tchèque, chinois (simplifié et traditionnel), danois, allemand (standard et Fraktur script), grec, finnois, français, hébreu, croate, hongrois, indonésien, italien, japonais, coréen, letton etc.

Le test effectué sur l'image de la figure— donne le résultat suivant :



The screenshot shows a Windows Command Prompt window titled 'C:\Windows\system32\cmd.exe'. The window contains the following text:

```
Microsoft Windows [version 6.2.9200]
(c) 2012 Microsoft Corporation. Tous droits réservés.

C:\Users\lenovo\win8>cd \
C:\>cd tarcisse
C:\tarcisse>tesseract.exe d.pnm o.txt
Tesseract Open Source OCR Engine v3.02 with Leptonica
C:\tarcisse>type o.txt | more
Nom du père Jean Nom de la mère
Pauline date de naissance 15/05/1991
ville BUC texte inutile Nom de
Jeune fille Marie aolggse 12 ne 004 Mums
code postn 739538

C:\tarcisse>
```

Fig 5-Tesseract

4. Leptonica is a pedagogically-oriented open source site containing software that is broadly useful for image processing and image analysis applications

3.3 Logiciel De Reconnaissance Optique De Caractères

3.3.1 FreeOCR

FreeOCR est un logiciel de reconnaissance optique de caractères pour Windows. Il prend en charge la numérisation de la plupart des scanners Twain (protocole informatique standard destiné principalement à relier un scanner d'image à un ordinateur). FreeOcr comprend un programme d'installation de Windows, il est très simple à utiliser et prend en charge l'ouverture de documents multi-pages TIFF, les documents PDF et de fax Adobe ainsi que la plupart des types d'images TIFF compressé non pris en compte par le moteur Tesseract lui-même. La version V4 de **FreeOCR** incluant Tesseract V3 apporte plus de précision lors de l'analyse des pages. Cette précision est atteinte sans l'aide de l'outil de sélection de zone. Toutes les images sont extraites sous forme de texte brut ou exporter directement au format Microsoft Word.

Le résultat du test avec la même image de la figure — donne ceci :

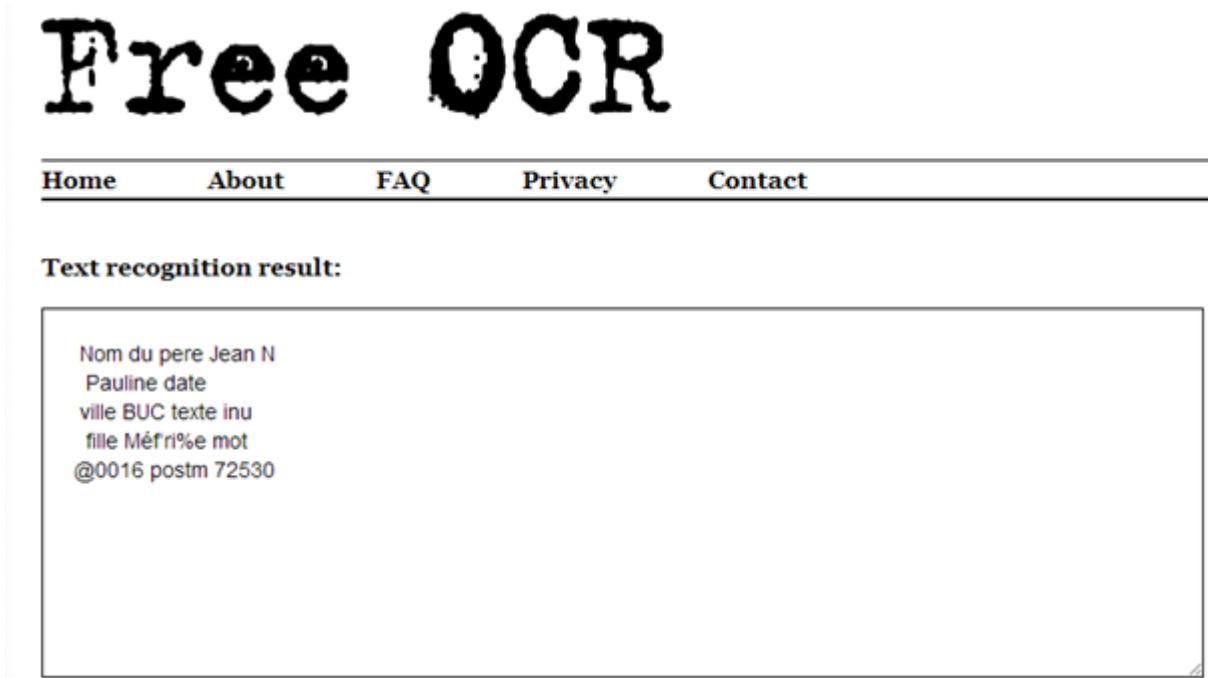


Fig 6-FreeCCR

3.3.2 ABBYY

ABBYY est une Entreprise fondée en 1989 à Moscou par **David Yang** fournissant des logiciels de reconnaissance optique de caractères (Fine Reader), de capture de documents et des logiciels d'enseignement assisté par ordinateur pour micro-ordinateurs et appareils mobiles. FineReader est un logiciel de reconnaissance de texte (OCR) qui convertit avec précision des documents numérisés,

des documents papier et des images en formats éditables y compris Microsoft Office et des PDF interrogeables - vous permettant de réutiliser leur contenu, de les archiver de manière plus efficace et de les récupérer plus rapidement. **FineReader** supprime le besoin de retaper des documents et garantit la disponibilité aisée d'informations importantes. Il fournit instantanément un accès au document entier et supporte 190 langues de reconnaissance (toutes stockés dans un dictionnaire) dans n'importe quelle combinaison. Par ailleurs le mécanisme de la reconnaissance optique de caractère de FineReader est le suivant :

- Premièrement, le programme analyse la structure de l'image du document
- Il divise la page en éléments tels que des blocs de textes , tableaux, images, etc Les lignes sont divisées en mots, puis - en caractères . Une fois que les personnages ont été choisies , le programme les compare à un ensemble d'images de motif .
- Il avance plusieurs hypothèses sur ce que ce personnage est .
- S'appuyant sur ces hypothèses le programme analyse les différentes variantes de la rupture de lignes en mots et les mots en caractères . Après le traitement de grand nombre de ces hypothèses probabilistes , le programme prend finalement la décision, vous présentant le texte reconnu

En outre, **ABBYY FineReader** fournit un dictionnaire qui permet d'effectuer une analyse secondaire des éléments de texte au niveau du mot. Avec le soutien du dictionnaire , le programme assure une analyse encore plus précise et la reconnaissance de documents et simplifie encore la vérification des résultats de la reconnaissance. Le programme assure une analyse encore plus précise et la reconnaissance de documents et simplifie encore la vérification des résultats de la reconnaissance.

En effectuant un test avec l'image de la figure - nous obtenons comme résultat :

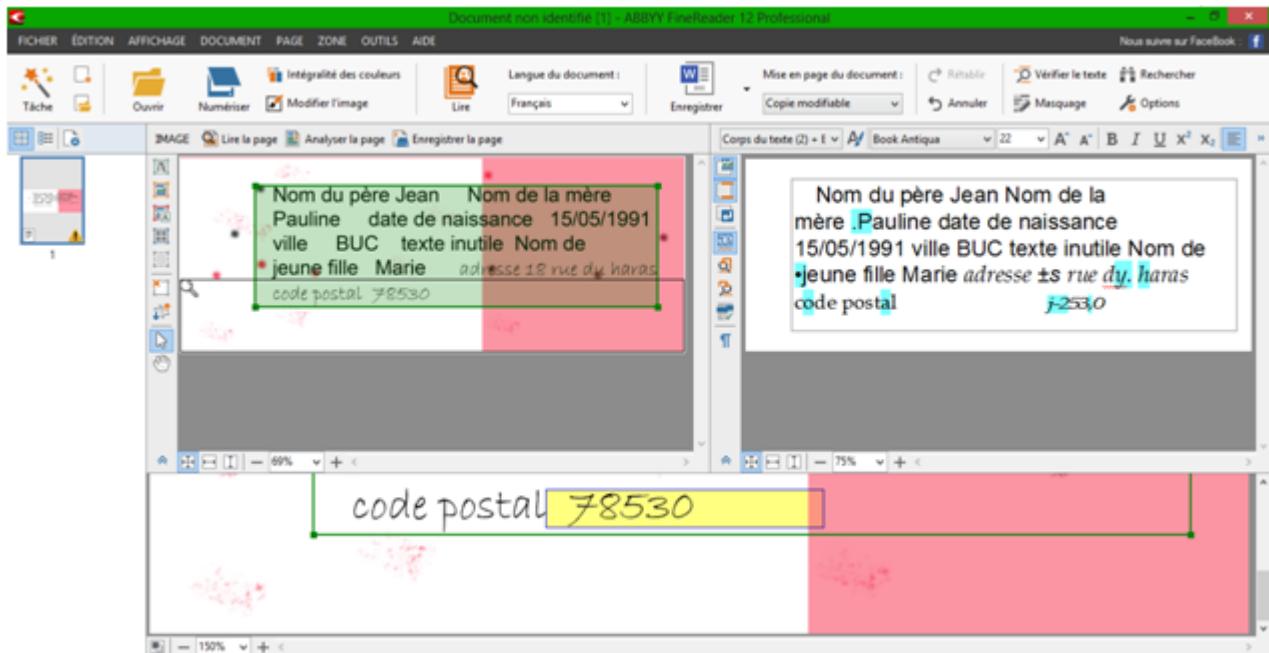


Fig 7-ABBY

3.3.3 Ocr Online

Il existe des outils gratuit en ligne qui permettent d'extraire du texte à partir d'une image numérisé. Parmi lequel :

- I2OCR : I2OCR est une version en ligne de reconnaissance optique des caractères libre (OCR) qui extrait le texte à partir d'images de sorte qu'il peut être édité, formaté, indexé, recherche, ou traduite. Avec plus de 60 langues de reconnaissance il prend en charge les formats d'image majeurs et Analyse multi colonne de document tout en étant 100 % GRATUITE illimités Mises. Le test effectué avec I2OCR a donné ceci comme résultat :

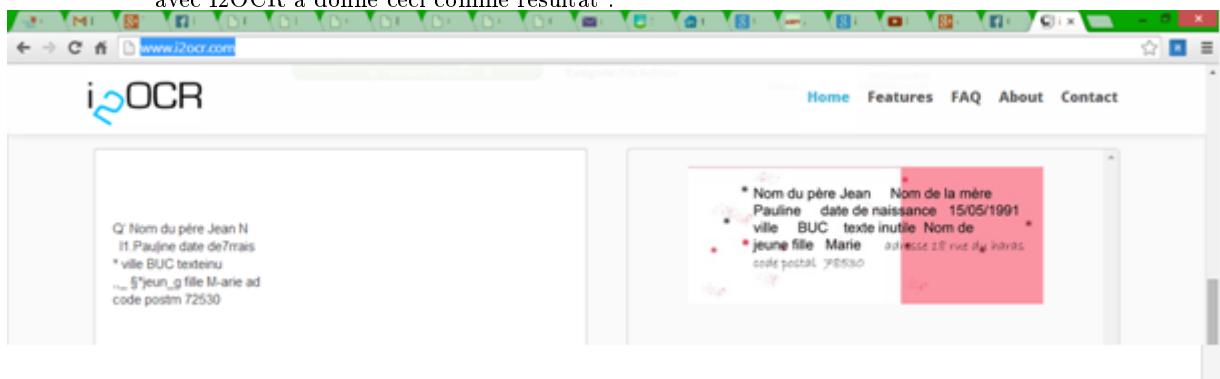


Fig 8-I2OCR

Outre que I2OCR voici la liste d'autres Ocr en ligne :

- FREE-ONLINE-OCR
- Newocr

- Onlineocr
- Abbyyonline (uniquement 20 documents par mois)
- Free-ocr
- Drive
- Finereaderonline
- Ocronline
- Les logiciel payants :

Conclusion :

La reconnaissance de caractère est un domaine actif de recherche pour la science informatique depuis la fin des années 1950. Au début, on pensait qu'il s'agissait d'un problème facile, mais il apparut qu'il s'agissait d'un sujet beaucoup plus intéressant. Il faudra encore de nombreuses décennies aux ordinateurs, s'ils y parviennent un jour, pour lire tous les documents avec la même précision que les êtres humains.

Parmi tous les outils et logiciel OCR existant TESSERACT paraît être l'outil qui donne le plus de précision sur l'extraction des données et répond le plus à nos besoins :

- Il offre une qualité d'extraction supérieure aux autres (GOCR, OCRAD) comme le montre différents testes,
- Il dispose d'une version exécutable en ligne de commande sur Windows et l'Unix en local ce qui le rend non dépendant des connexions extérieure et nous introduit au traitement automatique grâce aux fichiers batch tout en évitant le captcha proposé par les services web ocr en ligne,
- Il est également gratuit, open source et Disponible en plusieurs langue dont le français,

Chapitre 4

Analyse Et Conception De La Solution

Dans ce chapitre nous allons étudier les difficultés que rencontre les personnels administratif dans le traitement de Cerfa ,dans le but de concevoir un logiciel minimisant au maximum l'effort humain tout en offrant une rapidité de traitement et un mode de gestion de contenu visant à accélérer les recherches.

4.1 Analyse

Crée depuis le 18 juillet 1966, le formulaire Cerfa n'a pas été destiné à un traitement automatique. Donc, il fallait trouver un moyen de l'adapter à un traitement automatique sans toucher à l'existant.

4.1.1 Besoin fonctionnel

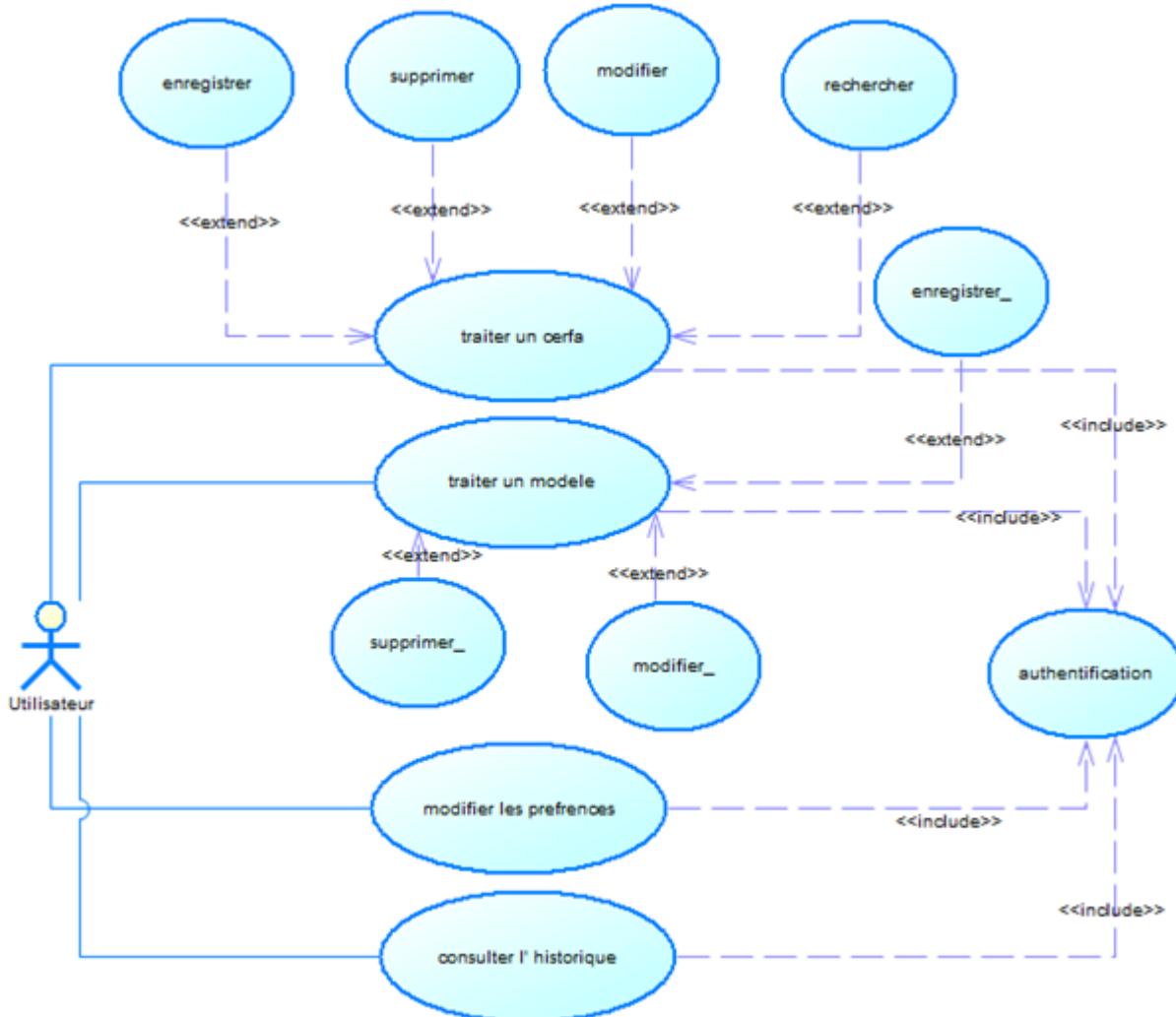


Fig 9-Différentes tâche de l'utilisateur

Ce diagramme illustre les différentes tâches de l'utilisateur. En effet le logiciel permettra à l'utilisateur de :

- **Traiter un Cerfa** : Le logiciel offre une interface permettant à l'utilisateur d'enregistrer le Cerfa dans la base de données, de le rechercher, de le modifier ou de le supprimer si l'il Y'a nécessité.
- **Traiter un modèle** : Comme nous l'avons dit plus haut, le Cerfa n'a pas été prévu pour un traitement automatique, pour s'y faire, un modèle doit être enrégistrer au préalable par le logiciel en vue de traiter tout les Cerfas correspondant à ce modèle. Ainsi le logiciel doit donner à l'utilisateur la possibilité d'enregistrer, modifier, supprimer, d'importer ou d'exporter un modèle.
- **Modifier Les préférences** : L'utilisateur aura la possibilité de modifier les paramètres d'extraction, le droit d'accès et les comptes, nous y reviendrons un peu plus en détail plus loin dans le chapitre suivant.

- **Consulter l'historique** : Chaque tâche effectuer par un utilisateur sera enregistré dans l'historique, et en fonction des paramètres définis, les utilisateurs pourront ou non le consulter .

4.1.2 Besoin Non Fonctionnel

- **La non redondance** : Pour chaque Cerfa enregistré il est indispensable d'éviter les doublons,
- **La réduction de l'effort humain** : Les différentes interfaces du logiciel doivent offrir non seulement la facilité d'utilisation, mais aussi la facilité d'exécuter automatiquement certaine tâche ne nécessitant aucun effort humain.
- **La performance** : Le temps de traitement et de recherche offert par le logiciel doit être largement inférieure à ce que l'utilisateur aurait mis, sans le logiciel. Il doit également donner la possibilité de faire plusieurs traitements en parallèle, sans pour autant compromettre, les tâches déjà en exécution.

4.2 Conception

Dans cette partie nous allons apprendre pas à pas le déroulement des séquences principales qui ont permis de développer ce logiciel ainsi que la manière dont nous avons apporté les modifications sur chacune des séquences pour nous permettre d'améliorer la manière dont nous avons structuré la base des données pour organiser les différents champs en vue d'accueillir des modèles multiples et complexes. Nous avons utilisé **merise** pour la modélisation de la base de données et UML pour le diagramme de séquence.

4.2.1 Modélisation

Première approche

La première approche consistait à enregistrer dans la base de données tout le texte extrait d'un Cerfa vide, et de le conserver comme modèle. Au moment de l'enregistrement d'un Cerfa rempli, extraire le texte de ce Cerfa, soustraire à ce texte , le texte du modèle pour ne garder que les champs entrés par l'utilisateur.

Nom du père.....Nom de la mère
.....date de naissance

ville.....texte inutile Nom de
jeune fille.....adresse.....
code postal.....

Nom du père Jean Nom de la mère
Pauline date de naissance 15/05/1991
ville BUC texte inutile Nom de
jeune fille Marie adresse 18 rue du haras
code postal 78530

Nom du père Jean Nom de la mère
Pauline date de naissance 15/05/1991
ville BUC texte inutile Nom de
jeune fille Marie adresse 18 rue du haras
code postal 78530

Jean
Pauline
15/05/1991
BUC
Marie
18 rue du haras
78530

Fig 10-Modélisation-première approche

Malheureusement cette approche a posé quelques problèmes majeurs. En effet :

- Un simple décalage d'un seul mot ou caractère entraînait des grandes erreurs
- L'enregistrement Dépend à grande partie de la qualité de l'extraction
- Certaines Données extraites sont muettes (sans libellé)
- Il y'avait une insuffisance d'informations sur le modèle

Deuxième approche

La deuxième approche consistait à retenir le plus d'informations possible sur les modèles avec l'introduction d'un nouveau concept (libellé, marqueur de début, marqueur de fin).

Ce concept vise à isoler les informations essentielles du Cerfa tout en nous indiquant les marqueurs qui nous permettront de les identifiées. Ainsi comme dans certain langage de programmation nous avons le ";" qui indique à l'analyseur lexicale la fin d'une instruction et le début d'une autre. Ces marqueurs nous aiderons à identifier les champs utiles.

En appliquant ce nouveau concept à l'exemple précédent, nous avons décidé de choisir comme marqueur de début **nom du père** et comme marqueur de fin **nom de la mère** en ayant respectivement comme libellés :**nom du père,nom de la mère** ou un synonyme selon votre vouloir pour indiquer de quoi il s'agit. Une fois le modèle enregistré, tout le Cerfa appartenant à ce modèle seront traité de la même manière.

Nom du père.....Nom de la mère
date de naissance

ville.....texte inutile Nom de
 jeune fille.....adresse.....
 code postal.....

LIBELLE

Nom du père
 Nom de la mère
 Date
 Ville
 Nom de jeune fille
 adresse

MDEBUT

Nom du père
 Nom de la mere
 Date de naissance
 Ville
 jeune fille
 adresse

MFIN

Nom de la mère
 Date de naissance
 Ville
 texte inutile Nom
 adresse
 code postal...

Nom du père Jean Nom de la mère
 Pauline date de naissance 15/05/1991
 ville BUC texte inutile Nom de
 jeune fille Marie adresse 18 rue du haras
 code postal 78530

Nom du père:**Jean**
 Nom de la mère:**Pauline**
 Date:**15/05/1991**
 Ville:**BUC**
 Nom de jeune fille: **Marie**
 adresse:**18 rue du haras**
 code postal:**78530**

Fig 11-Modélisation-deuxième approche

Cette approche a permis de résoudre la majorité des problèmes de la première, mais tout comme la première approche elle dépend de la qualité d'extraction. Ex : "nom du père " est totalement différent de "n'm du ère" pour une qualité d'extraction à une différence près.

Troisième Approche

Toujours dans les soucis de renvoyez le minimum d'erreurs à l'utilisateur, nous avons introduit un deuxième concept qui est la **tolérance**. Basée sur la distance entre deux mots, la tolérance nous permet de trouver le marqueur de début ou de fin quel que soit la qualité de l'extraction. En appliquant le concept de la tolérance à l'exemple précédent pour un enregistrement de modèle de bonne qualité ,tout en gardant les mêmes marqueurs de début et de fin,nous avons constaté des erreurs pendant l'extraction du texte lors du traitement du Cerfa :

Nom du père.....Nom de la mère
date de naissance

ville.....texte inutile Nom de
 jeune fille.....adresse.....
 code postal.....

LIBELLE	MDEBUT	MFIN
Nom du père	Nom du père	Nom de la mère
Nom de la mère	Nom de la mere	Date de naissance
Date	Date de naissance	Ville
Ville	Ville	texte inutile Nom
Nom de jeune fille	jeune fille	adresse
adresse	adresse	code postal...

N'm du &ère Jean Nom de la ùère
 Pauline date de naissance 15/05/1991
 ville BUC texte inutile Nom de
 jeune fille Marie adresse 18 rue du haras
 code postal 78530

Nom du père

N'm (1)	du (0)	&ère (1)	2
Nom (0)	de (1)	la (4)	5
Nom (0)	de (1)	jeune (3)	4

Fig 12-Modélisation-troisième approche

En partant de la troisième approche, une seule question se pose, mais à combien fixé cette **tolérance**?En effet :

- Plus la tolérance est petite, plus le moteur est rapide et nécessite une bonne qualité d'extraction.
- Plus la tolérance est grande, plus le moteur de recherche est lent et nécessite moins une extraction de bonne qualité.

Dans tout le deux cas nous avons constatez que la tolérance est fonction de la qualité de l'extraction, comme toutes les imprimantes, n'ont pas la même qualité d'extraction, la tolérance sera un paramètre à varier par l'utilisateur suivant les résultats fournis par le moteur. Sur la base de tout ce qui a été dit, voici la manière dont nous avons modélisé la base de données pour accueillir toutes ces informations :

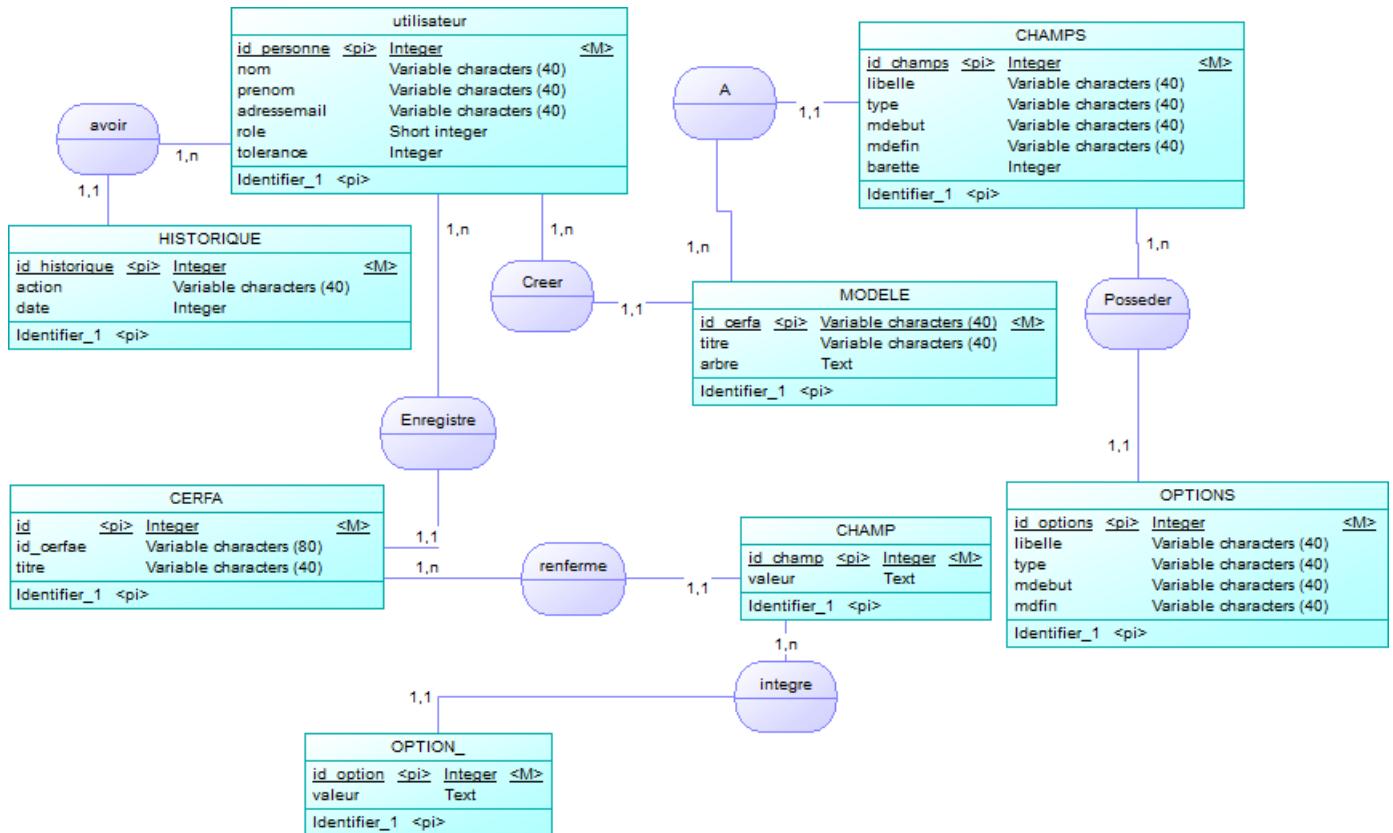


Fig 13-Modélisation des données

Chaque utilisateur disposera d'un compte(id,nom,role,parametre) qui lui permettra d'enregistrer un modèle de son choix.Selon l' étude que nous avons fait du Cerfa, nous avons distingué trois types de données :

- le champ de type texte,

- la case à cocher,
- et le tableau,

Outre que le champ de type texte, les autres champs ont des options. L'utilisateur pourra également enregistrer un Cerfa qui aura la même structure qu'un modèle au niveau des tables, mais au niveau de colonnes il sera composé d'un triplé (père, identifiant, valeur). Toute les tâches effectuées par l'utilisateur seront enregistré dans l'historique en vue d'avoir une trace sur l'utilisation du logiciel. En générant le modèle physique nous obtenons ceci :

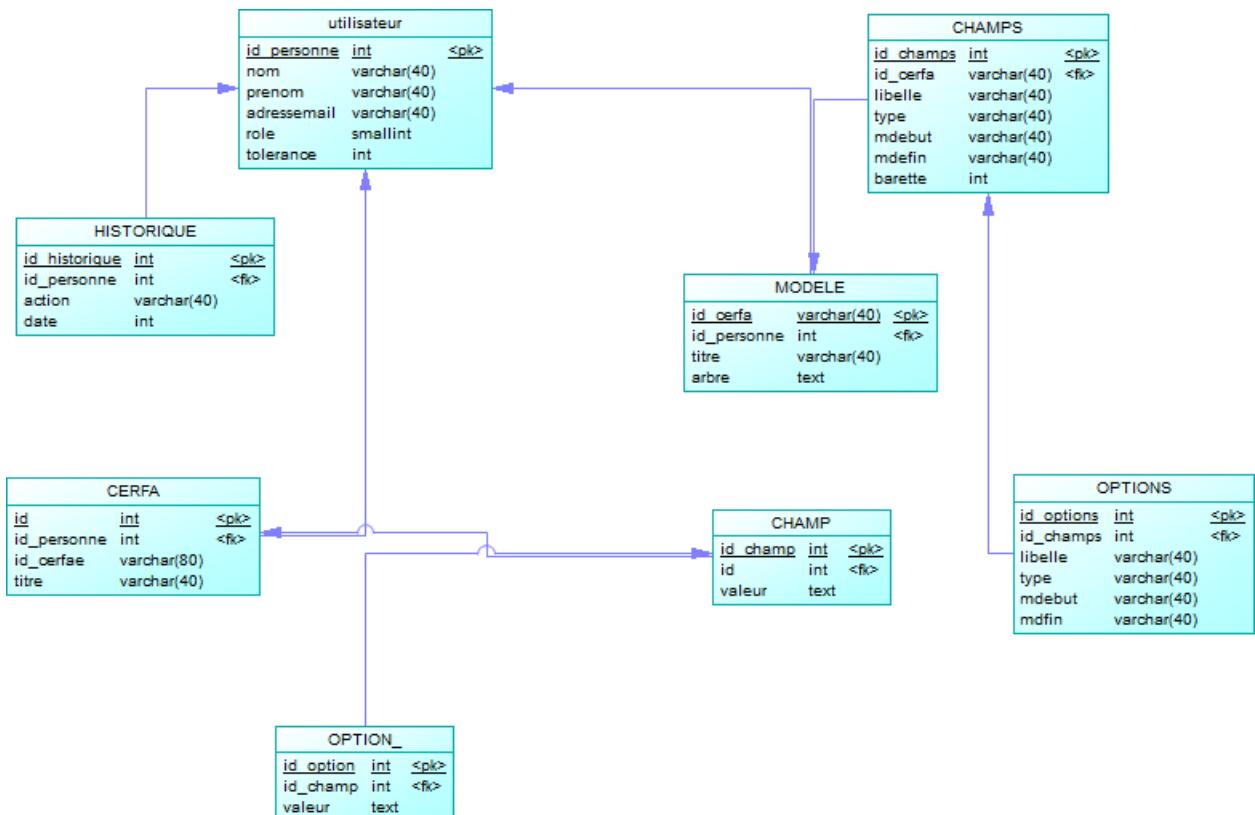


Fig 14-modèle physique de la base de données

4.2.2 Diagramme de séquence

Diagramme de séquence d'extraction des données

La premier étape consiste à extraire du texte provenant d'une image, et parmis tous les outils de reconnaissance optique nous avons retenu Tesseract, comme le processus d'extraction que nous avons employés pour Tesseract a déjà été cité si haut, nous allons simplement dresser le diagramme de séquence qui l'illustre :

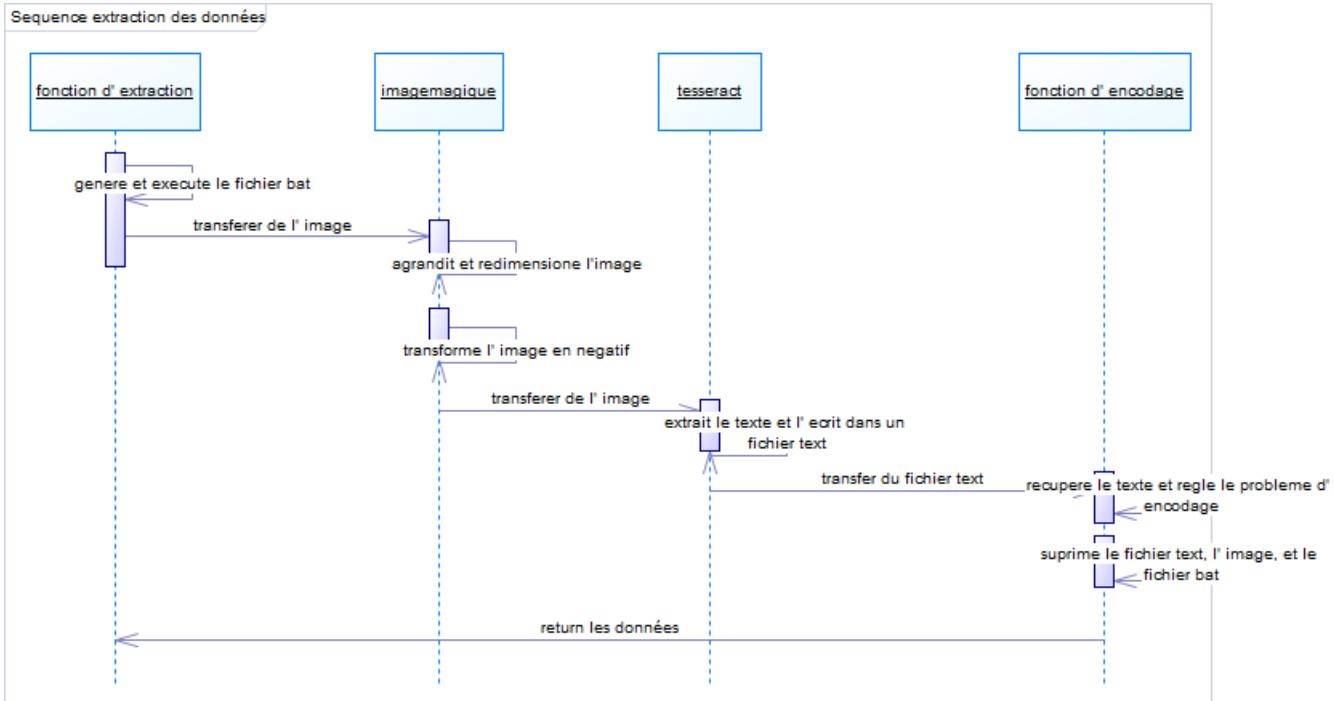


Fig 15-Tranitions l'ors de l'extraction de Texte

4.2.3 Diagramme de séquence du moteur de recherche

Après l'extraction des données, il faudra les organiser d'une manière structurée pour faciliter et accélérer la recherche. Ainsi nous avons mis au point un moteur de recherche basé sur le modèle enregistré au préalable pour pouvoir reconnaître d'une manière efficace les différents champs et les données utiles à extraire.

Le moteur de recherche est la partie phare de notre application. Elle est composé de plusieurs structures qui permettent d'obtenir des résultats qui dépendent de la qualité de l'extraction de données extraits à partir d'une image numérisée. Ces structures qui ne sont rien d'autres que des algorithmes s'exécutent de manière progressive et successive.

Dans un premier temps, le texte brut extrait par le moteur TESSERACT est converti en un tableau (**X**) de chaînes de caractères. A partir de ce tableau, le moteur récupère tous les numéros contenant des " ", des " * " ou sans ces caractères spécifiques, existant dans (**X**). Ensuite le moteur se connecte à la base de données et recherche l'existence de chaque numéro (stocké dans un tableau) trouvé dans la table "**modèle**" qui lui, contient tous les numéros de cerfa entrés par l'utilisateur lors de l'enregistrement d'un modèle.

Lorsque le numéro de cerfa a été identifié, le moteur va extraire les données à partir des marqueurs débuts et les marqueurs fins correspondant à ce numéro. Le cas contraire, l'utilisateur sera dans l'obligation de saisir le numéro de cerfa qui n'a malheureusement pas été retrouvé (enregistré dans la base de données) à cause de la qualité de l'extraction du texte provenant de l'image numérisée. Pour ce dernier cas, l'action du moteur sur ce numéro sera le même que les

précédentes actions décrites au dessus.

Pour mieux éclairer les explications ci-dessus nous vous proposons un schéma qui donne en détail les actions appliquées par le moteur de recherche à ce sujet :

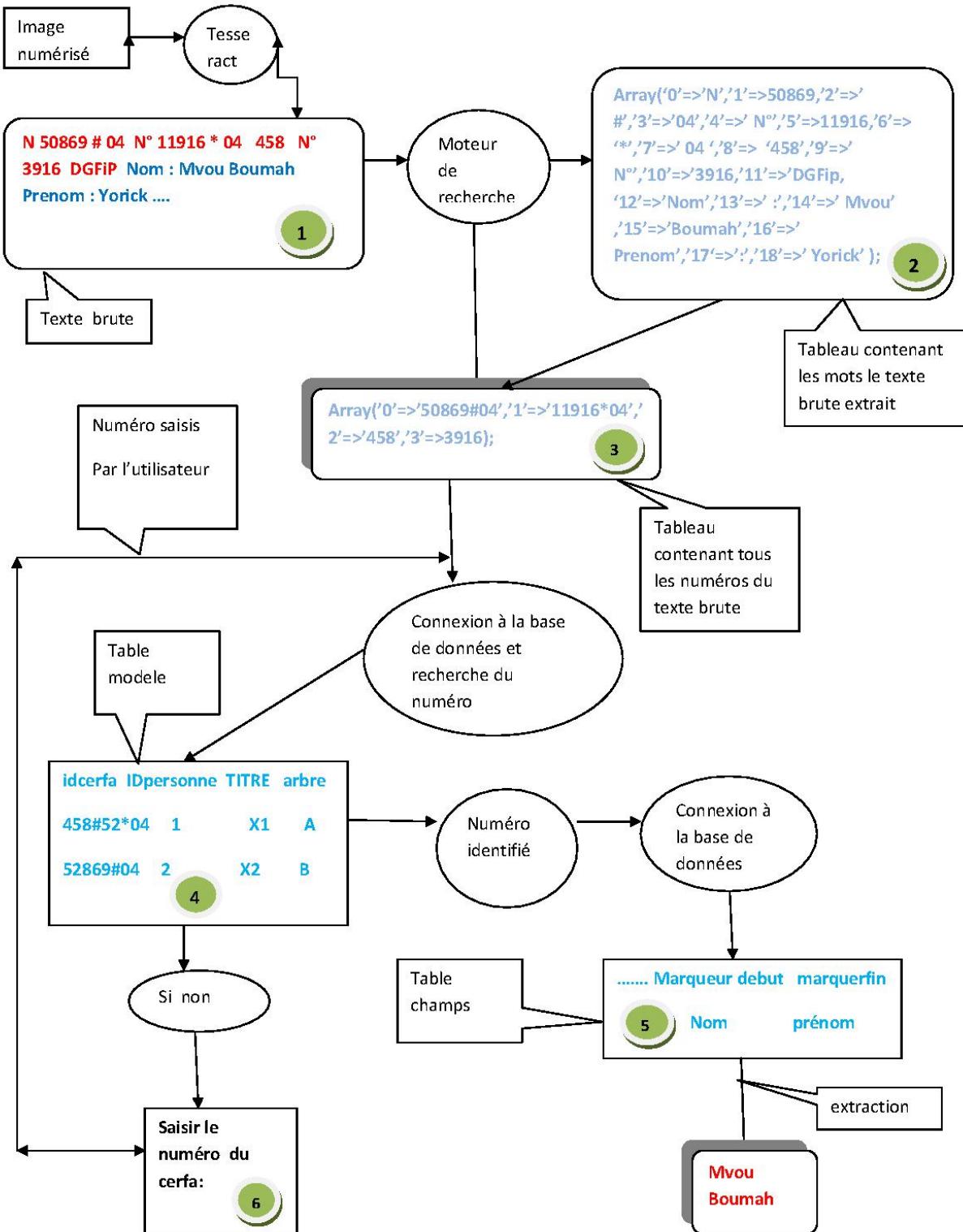


Fig 16-Moteur de recherche-Méthode d'extraction de données

Dans un premier temps l'application extrait le texte de l'image numérisée par le moteur Tesseract(1), ensuite le moteur de recherche découpera le texte en chaînes de caractère puis les stockera dans un tableau Y (2). A partir de ce tableau le moteur de recherche identifiera tous les numéros (avec ou sans les caractères " " et " * ") existant dans Y (3). Dans un troisième temps le moteur de recherche se connectera à la base de données en cherchant à identifier le modèle de cerfa existant (4) à partir de tous les numéros identifiés. Pour ce dernier cas :

- **Soit le modèle est identifié :** Dans ce cas le moteur de recherche se reconnecte à la base données et identifie les marqueurs début et fin de manière successive tout en récupérant le texte délimités par ces marqueurs(5).
- **Soit il ne l'a pas été :** Dans ce cas l'application invite l'utilisateur à saisir un numéro de modèle de cerfa existant dans la base de données et rattaché au modèle de cerfa qu'il souhaite enregistré(6).

Les marqueurs de début et de fin étant indispensable pour une extraction de donnée limpide et efficace, nous avons mis au point un algorithme qui permet de détecter ces marqueurs malgré la qualité de l'extraction du texte (de l'image numérisée). Ce algorithme consiste à déterminer les marqueurs de début et de fin dans le texte extrait en utilisant le facteur de "**tolérance**" qui vise à déterminer de manière approximatif ou exacte les marqueurs qui s'identifient le plus à ceux (celui) saisis par l'utilisateur lors de l'enregistrement du modèle. Tout d'abord , l'algorithme commence à mémoriser l'ensemble des mots que l'on nommera par "**A**" du texte extrait dont la distance avec la première chaîne de caractère du marqueur prédéfini est inférieur ou égale au facteur "**tolérance**". Par la suite , selon la taille du marqueur (prédéfini par l'utilisateur) l'algorithme déterminera l'ensemble des sous mots ayant la même taille que le marqueur prédéfini à partir de l'ensemble "**A**". L'ensemble des sous mots dont la somme des distances avec l'ensemble des mots qui définissent le marqueur prédéfini est la plus petite ,sera considéré comme étant le marqueur qui se rapproche le plus du marqueur enregistré dans la base de données.

Pour une meilleure perception des choses, nous illustrons nos explications par le schéma ci-dessous.

NB :X(mot1,mot2) retourne le nombre de caractère inexistant dans le mot1 et le mot2 dis(mots1,mots2) calcule la somme des nombres de caractères inexistant dans le mots1 et le mots2) Voir un exemple plus bas :

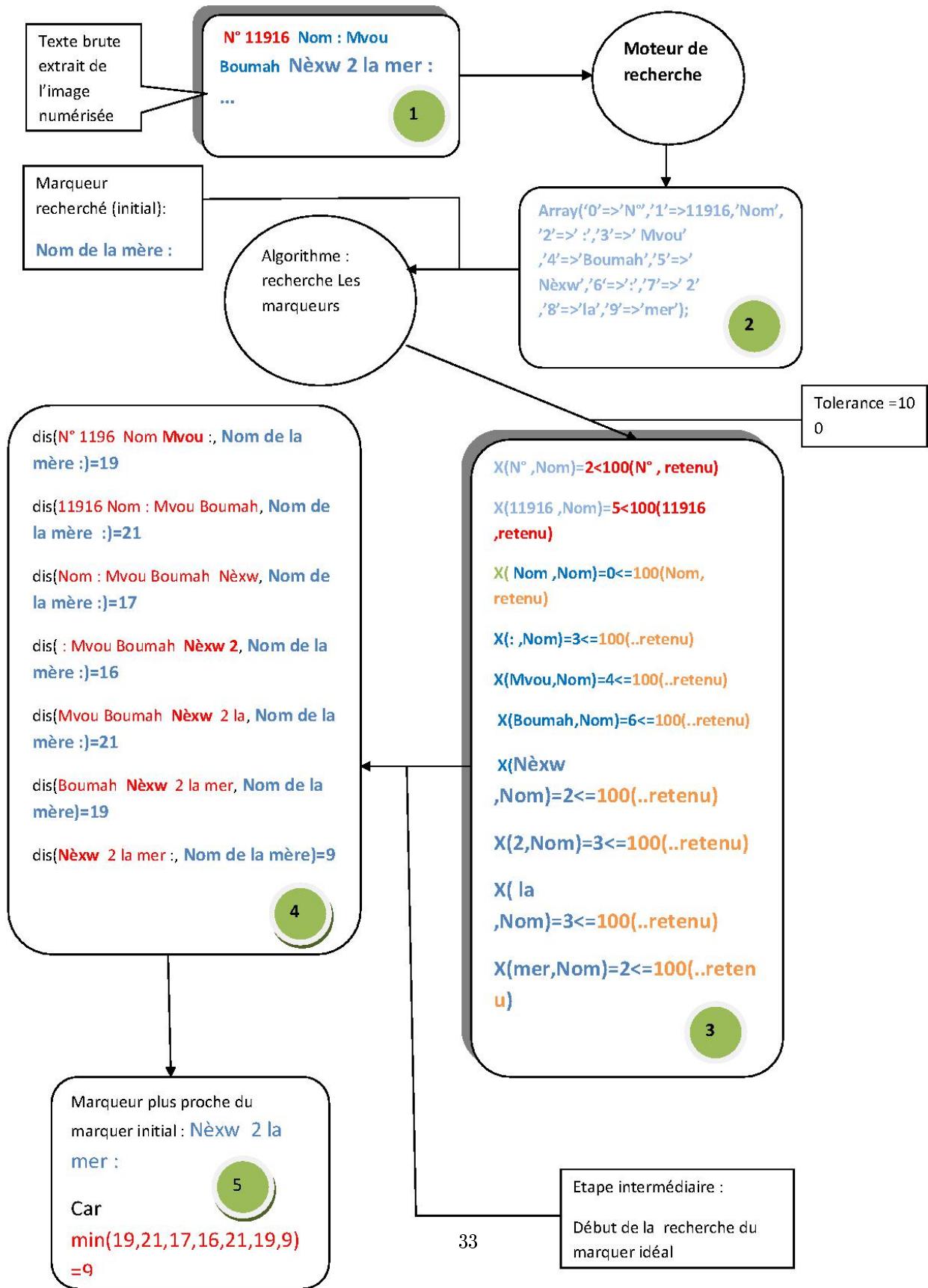


Fig 17-Moteur de recherche-Methode de reconnaissance des marqueurs 1

Les étapes (1) et (2) ont été interprétées dans le schéma précédent. Dans l'étape (3) nous remarquons que l'algorithme recherche l'ensemble des mots dont la distance avec la première chaîne de caractère du marqueur prédéfini ,est inférieur ou égale à la tolérance(=100).A partir de ce ensemble ,l'algorithme détermine l'ensemble des sous mots(Voir la partie 4 du schema) ayant la même taille que le marqueur prédéfini qui se rapproche le plus /ou exactement du marqueur initial (existant dans la base de données). La fonction " dis " de paramètre mots1 et mots2(par exemple) permet de calculer la somme des nombres de caractères manquant dans chaque sous mots des mots1 et mots2 comparer deux à deux.L'exemple ci-dessous nous aide à comprendre un peu mieux ce qui se passe :

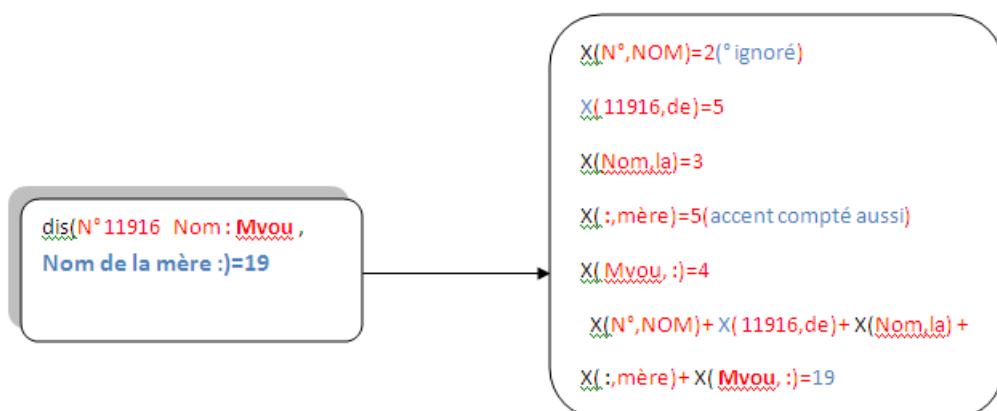


Fig 18-Moteur de recherche-calcule de la distance entre deux mots

X ,est la fonction php nommé **levenshtein** qui nous permet de calculer la distance entre deux mots .

Le moteur de recherche est aussi capable de gérer les cases à cocher. En effet à partir de nombreux tests d'extraction sous **tesseract** nous avons remarqué que les cases cochées(colorié complètement) sont traduit par un caractère particulier.Nous avons donc conclu que si ce caractère est lu entre le marqueur début et le marqueur fin qui le définit alors la case est cochée si non elle ne l'est pas.Voici ci après un schéma qui explique en détail les faits.

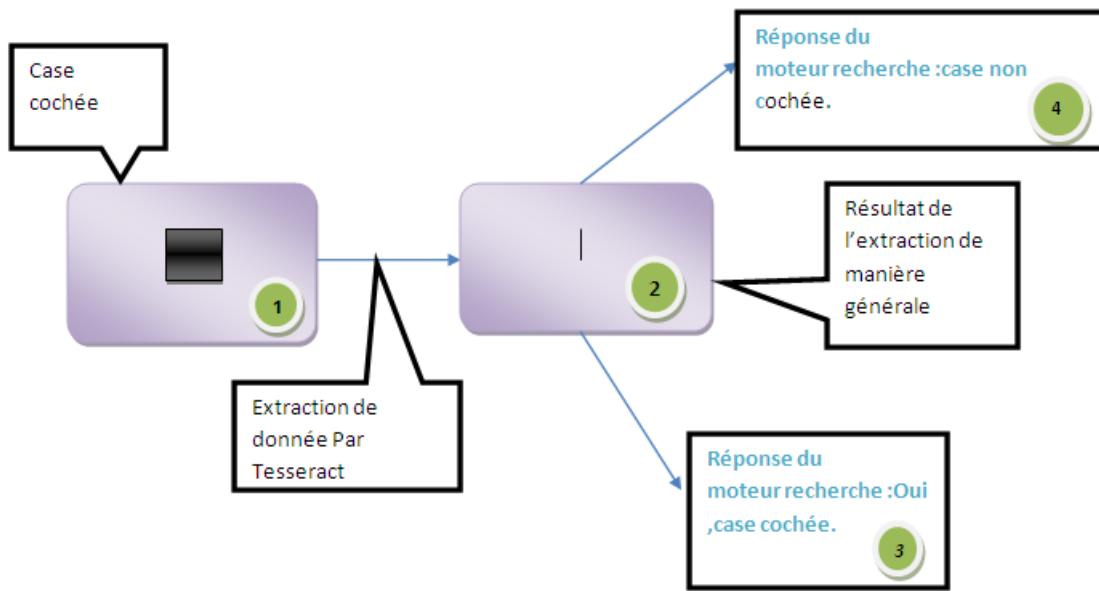
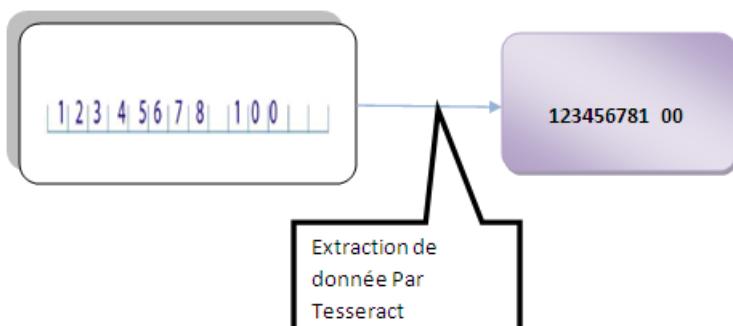


Fig 18.1 -Case à cocher

En (1) nous avons la case cochée. En (2) nous avons le caractère spécial qui correspond à la case cochée. Lorsque ce **caractère** est reconnu par l'application , alors elle considère qu'elle fait fasse à une case cochée. la case est belle et bien coché (3)si non elle est ne l'es pas(4). Pour permettre à l'application d'effectuer minitieusement la recherche sur ces cases nous avons ajouté un champ **barette** dans la table champs.Ce champ prend la valeur 1 ou 0.Si une ligne de la table **champs** à sa barette égale à 1 alors l'application saura qu'il est en face d'une case à cocher.

Deuxièmement, Le moteur de recherche est capable de lire des [numéros se trouvant dans les cases tout en supprimant les caractères superflus. Ci-dessous un



exemple :

Fig 18.2 -Case à numéro

Enfin,pour éviter la redondance des données dans la base de données, le moteur de recherche va exécuter un algorithme qui permettra de vérifier si le contenu du

modèle qui vient d'être extrait se rapproche partiellement ou complètement du contenu d'un des modèles existant dans la base. Il utilise une fois de plus le facteur **tolérance** qui est initialisé et fixé par la taille des données correspondant au modèle.

Dans un premier temps, le moteur de recherche va se connecter à la base de données en identifiant (à partir de la table " **cerfa**") tous les numéros de cerfa correspondant au modèle en plein traitement. Par la suite , il va rechercher toutes les données qui correspondent à ces numéros(à partir de la table " **champ** ") et va exécuter l'algorithme qui permet de vérifier si le modèle en plein traitement possède un contenu qui se rapproche des données déjà enregistré à partir du même modèle.

Cette reconnaissance consiste simplement à calculer la somme des distances que séparent l'ensemble des données qui viennent d'être extrait à ceux existant dans la table champs suivant le numéro de cerfa attaché au même modèle qui est en plein traitement.

Si cette somme est inférieur ou égale à la taille de données du modèle ,cela signifie qu'il existe un même modèle dont le contenu est identique ou se rapproche des données qui viennent d'être extrait du modèle . Si c'est le cas un message est adressé à l'utilisateur lui demandant la confirmation de l'enregistrement tout en lui signifiant l'existence proches de ces données dans la base de données, si non l'enregistrement se fait de manière automatique, aussi bien même lorsqu'il n'existe aucun numéro de cerfa rattaché au modèle. Ci-dessous un schéma qui vous permettra de mieux saisir nos explications :

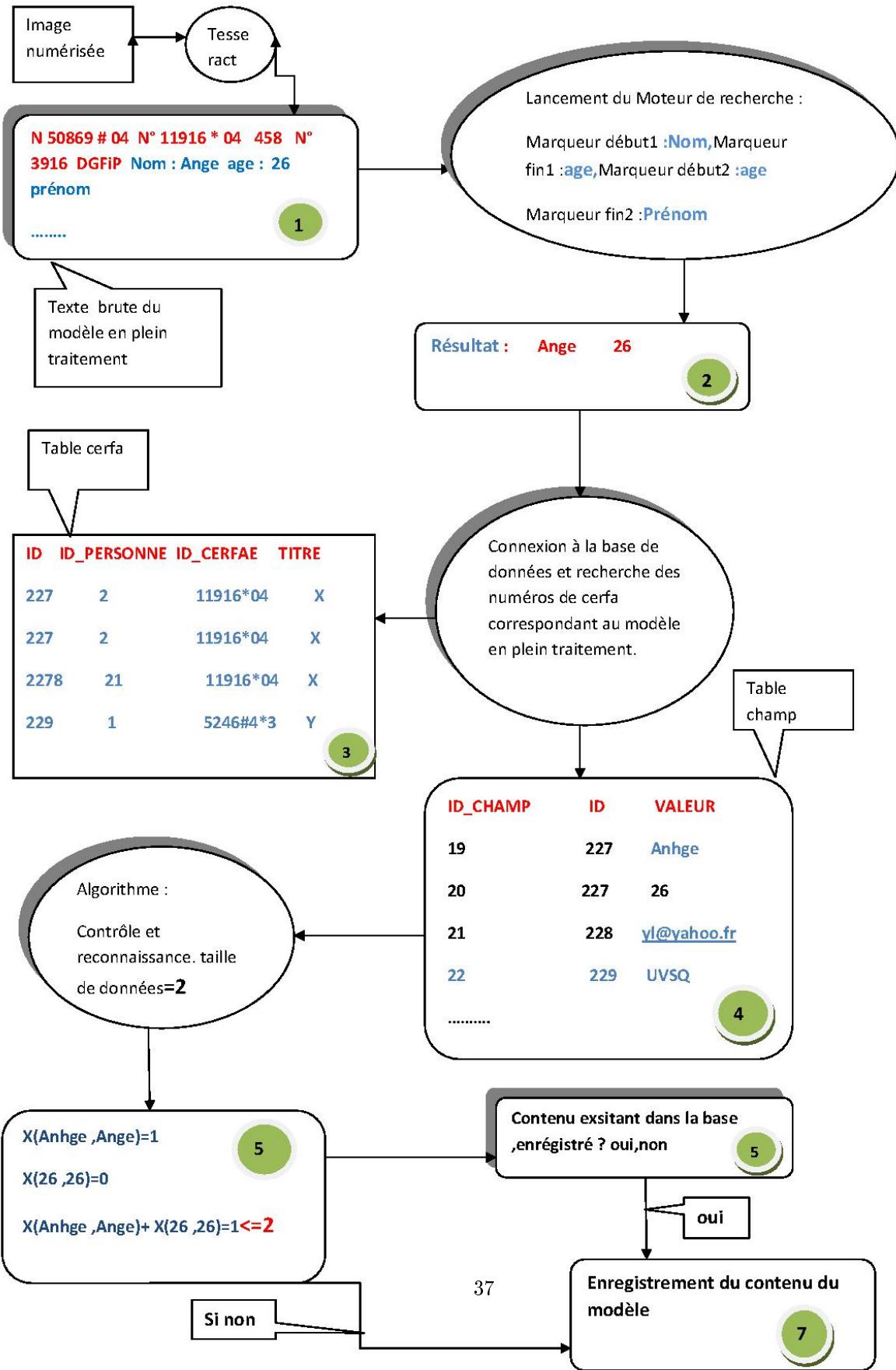


Fig 19-Moteur de recherche-Redondance de données

(1) et (2) permettent d'obtenir le résultat de l'extraction(3) à partir des marqueurs débuts et fin préenregistré par l'utilisateur. Par la suite le moteur de recherche va se connecter à la base de données en identifiant d'abord tous les numéros de cerfa enregistrés à partir du même modèle(3).Une fois les numéros identifiés,le moteur de recherche va lancer l'algorithme qui prendra compte des valeurs déjà insérées dans la table champ(4) afin de calculer la somme des distance séparant chaque valeur existant dans la table champ à celles qui viennent d'être extraits(5) .La somme étant inférieure ou égale à la taille des données(ici=2) l'utilisateur devra choisir d'accepter d'enregistrer ou non le contenu du modèle reconnu par l'application(6).Si il répond par non l'enregistrement du contenu du modèle n'aura pas lieu.Dans le cas où le modèle serait inexistant dans la table " modèle " ou Le cas où l'utilisateur acceptera L'enregistrement ,L'application enregistrera de manière automatique le contenu du modèle (7).

Conclusion Il était indispensable de construire quelques algorithmes permettant de gérer une recherche efficace sur les marqueurs,et capable d'identifier le numéro de cerfa à partir du modèle numérisée.

Chapitre 5

Implémentation De La Solution Et Analyse Des Résultats

Ce chapitre met la lumière sur la plateforme utilisée et les outils adoptés afin de mettre en uvre la solution. Nous y décrivons la démarche suivie pendant la réalisation et nous illustrons certaines fonctionnalités assurées à travers quelques interfaces.

5.1 Environnement Matériel Et Logiciel

5.1.1 Platefome web

Nous nous sommes dirigés vers la plateforme web notamment : HTML, PHP, JQUERY par ce qu'elle a répondu à la plus part de nos besoins et constitue la plateforme idéale pour un travail à accès multiple et offre plusieurs avantages :

Au niveau du client :

- Les même ressources sont partager et accéder par plusieurs utilisateurs en même temps,
- Pas besoin d'installer un logiciel spécifique au niveau du poste du client, il suffit d'utiliser le navigateur présent par défaut dans tout le système d'exploitation,

Au niveau de la programmation :

- Le HTML (HyperText Markup Language) qui est un format de données conçu pour représenter les pages web nous a permis de développer plus rapidement les interfaces et de faire une bonne mis en page grâce au CSS et une gestion dynamique des évènements coté client grâce à JavaScript.
- Le PHP (HyperText Preprocessor) est un langage, spécialement conçu pour le développement d'applications web. A la fois langage de programmation et de Template, le PHP nous a donner la possibilité de générer

dynamiquement du HTML ainsi nous avons préparé notre logiciel à accueillir facilement n'importe quel modèle quel que soit la complexité de sa structure.

5.1.2 Le langage BATCH

Le Batch est un langage de programmation très ancien, il permet de réaliser des tâches programmer à l'aide de simple lignes de commandes. Nécessitant une connaissance minimum de l'environnement DOS, il permet d'exécuter plusieurs lignes de commandes enregistré au préalable. Grâce au PHP nous avons générer dynamiquement de fichier batch pour rendre automatique toute les tâches exécutables sur le cmd pour éviter l'intervention humain.

5.1.3 Photoshop

Photoshop est l'application de retouche d'images la plus performante au monde nous a permis de simuler différentes situations (formulaire mal scanné, formulaire datant, formulaire flou, etc) en vue de préparer notre logiciel à s'adapter à différent type de fichier en bon ou mauvais état. Il nous a également permis de concevoir la plus part des figures présentes dans ce rapport, certaines icônes de l'interface du logiciel et les différentes animations présentes dans les diapos.

5.2 API Et Drivers Utilisés

5.2.1 Jquery

JQuery est une bibliothèque JavaScript gratuite et très pratique, ayant une syntaxe courte et logique, compatible avec tous les navigateurs courant. Son utilisation nous a permis de rendre Ajax et la syntaxe de JavaScript compatible à la plus part de navigateur.

5.2.2 ImageMagick

ImageMagick est un logiciel en ligne de commande très puissant de manipulation d'images dans pratiquement tous les formats existant. Son insertion dans l'application permet d'améliorer la qualité de l'image en vue d'avoir une extraction de bonne qualité.

5.2.3 Tesseract

C'est l'Outil de reconnaissance optique de caractère que nous avons retenu pour l'extraction en vertu des raisons citées si haut.

5.3 Analyse Et Test

5.3.1 Installation

Pour déployer ce logiciel la première phase consiste à mettre en place un serveur web, dans notre cas WampServer qui est une plate-forme de développement Web sous Windows pour des applications Web dynamiques à l'aide du

serveur Apache2, du langage de scripts PHP et d'une base de données MySQL. Il possède également PHPMyAdmin pour gérer plus facilement vos bases des données. Une fois le serveur web prêt à être utilisé, il suffit de double cliquer sur l'exécutable du logiciel Cerfa et de spécifier le dossier www du serveur comme l'indique la figure suivante :

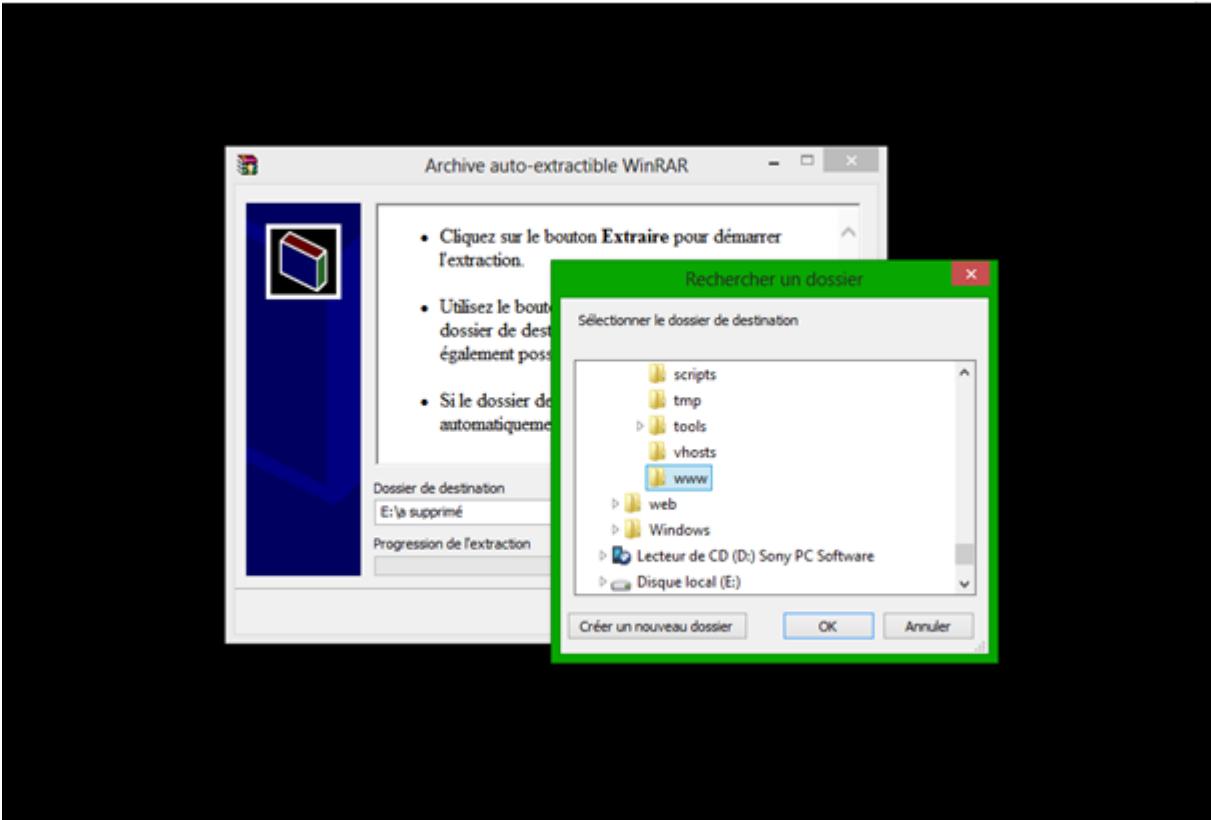


Fig 20-Moteur de recherche-Redondance de données

5.3.2 Quelques Interfaces

Une fois l'installation du logiciel terminé et le serveur activé, la première interface rencontrée par l'utilisateur sera l'interface d'authentification où il pourra se loger avec un login et un mot de passe (admin pour une première connexion et pourra le changer par la suite pour des raisons de sécurité).



Fig 21-Login

Une fois l'utilisateur logué il est redirigé vers la page d'accueil où il trouve l'ensemble de Cerfa déjà traité classé sous trois catégories. **-Les Cerfa récemment traités sont classé (au maximum 12 classé) de manière chronologique.**

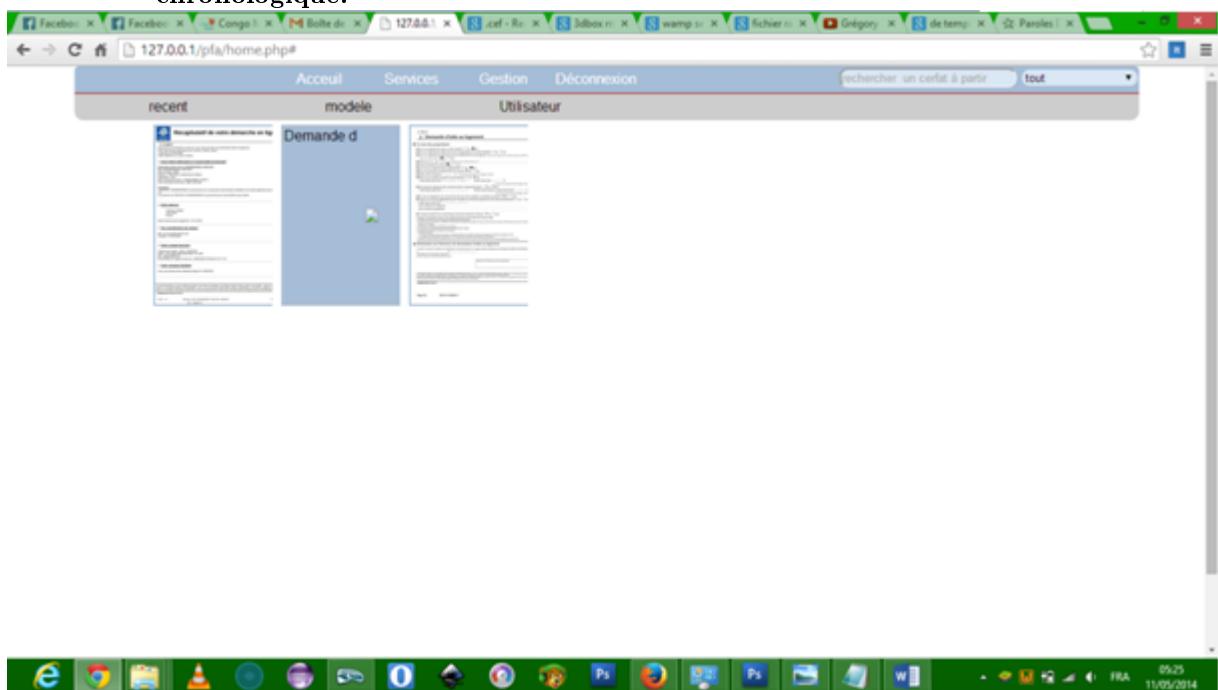


Fig 22-Affichage-Accueil

-La deuxième catégorie crée pour chaque modèle enregistré une armoire virtuelle, représenté par le numéro du modèle, le titre du modèle ainsi que le nombre de Cerfa traité avec ce modèle. Cela permet à l' utilisateur de naviguer plus rapidement pour rechercher un Cerfa dont il connaît le modèle.

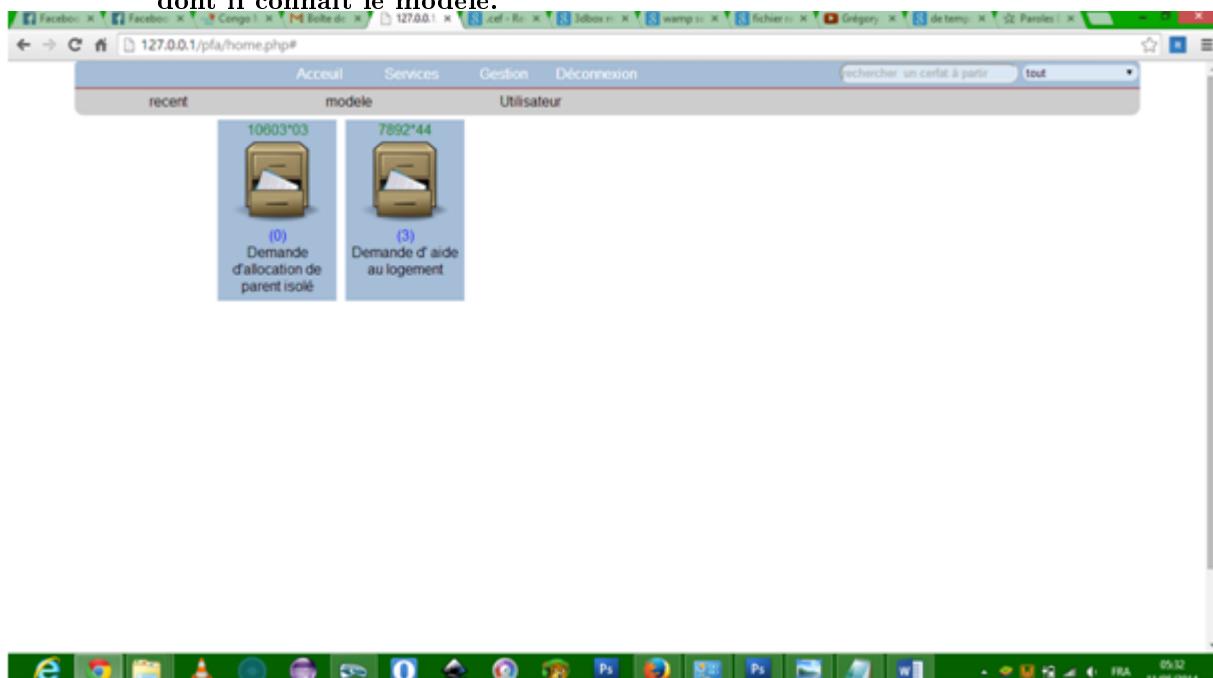


Fig 23-Armoire virtuelle

La troisième catégorie classe les documents traités selon l'utilisateur, ainsi un utilisateur pourra accéder plus rapidement à toutes les demandes dont il est à l' origine de traitement.

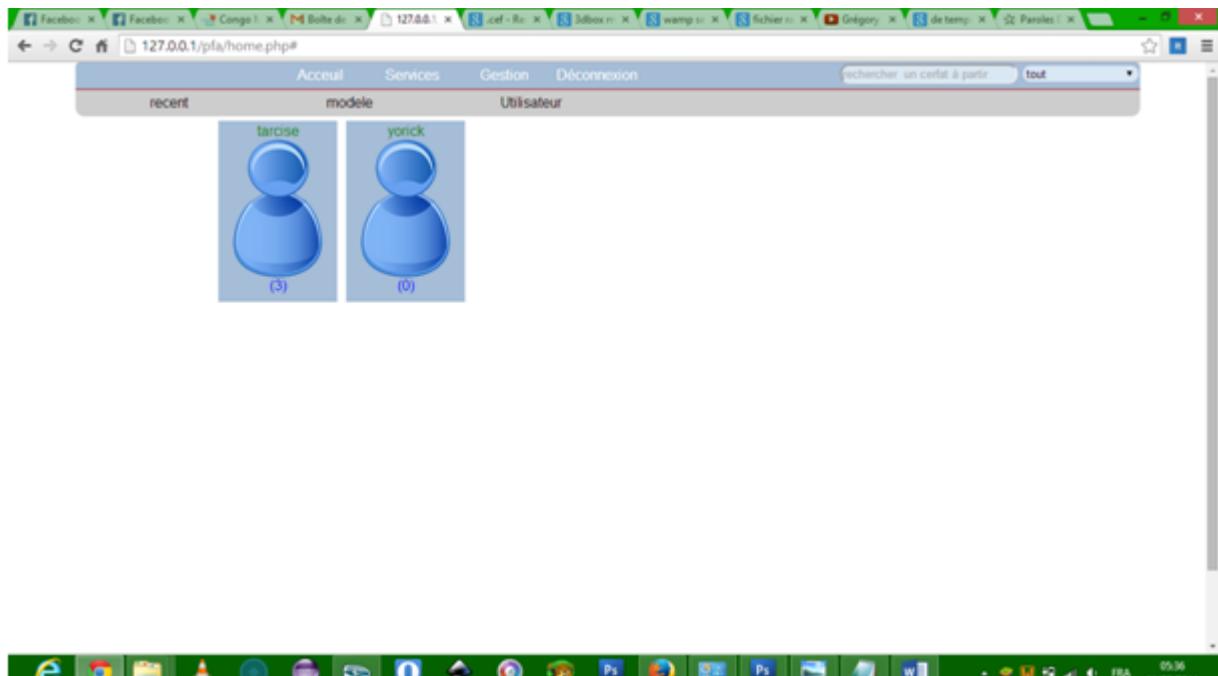


Fig 24-Profil des utilisateurs

Ainsi en cliquant sur l'un de Cerfa un onglet s'ouvre à droite de l'utilisateur, lui permettant de voir le Cerfa avec la possibilité de modifier les différents champs , de le supprimer après confirmation, ou de naviguer entre les différentes page pour le Cerfa à plusieurs pages.

Récapitulatif de votre démarche en ligne

N° 4749678
Récapitulatif du 21/01/2014 à 19:53:27, vous venez de faire une demande d'aide au logement.
Vous êtes alors devenu allocataire de la Caf des Yvelines située :
1 Rue DE LA FONTAINE
78261 MANTES LA JOLIE CEDEX

> **Vous-même (allocataire et responsable du dossier)**

Déclaration d'état civil de TSHOMOKONGO TARCISSE
Mme TSHOMOKONGO TARCISSE
Nom d'usage : EDDY
Nom(s) le : 15/01/1991 à KINSHASA CONGO
Nationalité : Autre
Nom et prénom du père : EDMAHAMBA TSHEFU
Nom et prénom de la mère : MBU PAULINE

Prestations
TARCISSE TSHOMOKONGO ne perçoit pas ou n'a pas perçu de prestations familiales d'un autre organisme que la Caf
Les parents de TARCISSE TSHOMOKONGO ne perçoivent pas de prestations pour laquelle.

> **Votre adresse**
18 B Rue HARAS
78530 BUC
France
Date d'entrée dans le logement : 01/11/2013

> **Vos coordonnées de contact**
Mail : sancisseandy@gmail.com
Portable : 0752524824

> **Votre compte bancaire**

Fig 25-Onglet cerfa

Juste à côté de l'accueil nous avons le menu service qui possède trois sous menu :

- **Enregistrer modèle :** Ce le sous menu qui nous permet d'enregistrer un modèle , il a été conçus de manière à faciliter l' ajout d' un modèle, il dispose d' une barre d' outils redimensionnable et mobile selon le bon vouloir de l' utilisateur, cette barre d' outils permet de spécifier les différent champs à ajouter au modèle ainsi que le type et voir mémé la position a laquelle ils doivent être ajouter. Juste en bas de cette barre nous avons une zone permettant d'uploader un Cerfa non rempli de préférence, le texte extrait de ce Cerfa permettra de former automatiquement un auto complet au différent champ du modèle et accélérer l'enregistrement du modèle. A gauche de cette zone l'utilisateur dispose d'un endroit où il doit spécifier les champs obligatoires du modèle, il s'agit bien de l'ID et du titre du Cerfa. Par contre les différents champs associés au modèle peuvent être ajouté, modifier ou supprimer à tout moment.

Lors de la soumission du modèle un contrôle minutieux est effectué sur les différents ajoutés, rappelant à l'utilisateur avec une grande précision le libellé ou le numéro du champ qui contient l'erreur si le libellé n'a pas été spécifié.

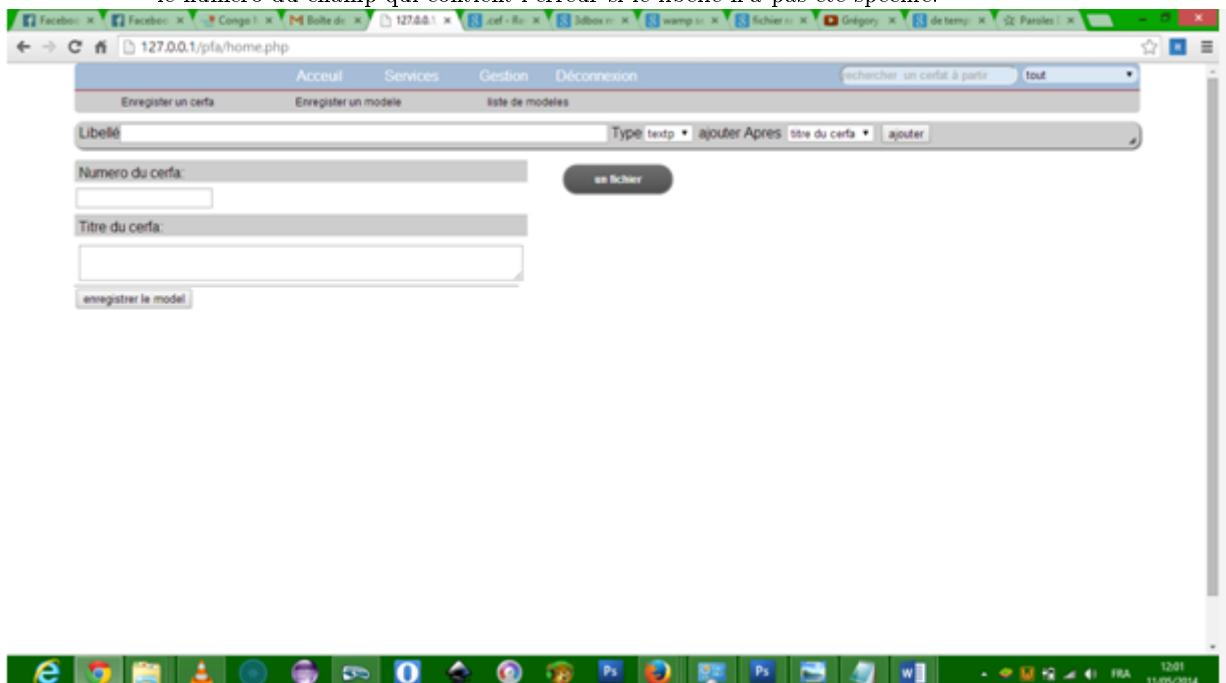


Fig 26-Enregistrement du modèle

- **Liste de modèles :** Apres l'enregistrement des modèles, cet onglet vous offre la possibilité de le voir, le supprimer, ou les modifier en cas de nécessiter.

id cerfa	Titre	enregistrer par	Action
10803'03	Demande d'allocation de parent isolé	tarcise eddy	voir supp
7892*44	Demande d'aide au logement.	tarcise eddy	voir supp

Fig 27-Liste des modèles

- **Enregistrer un Cerfa :** C'est la partie où l'utilisateur pourra uploader les différents fichiers pour un traitement automatique. Il est composé de plusieurs bouton permettant de sélectionner, désélectionner effectuer plusieurs traitement en parallèle, ou de sélectionner plusieurs Cerfa pour le traiter comme un seul, l'utilisateur dispose également d'une zone d'aperçu à droite où il pourra voir en grand le différent document uploadé.

Fig 28-Traitement du cerfa

A la fin du traitement d'un Cerfa, un lien est automatiquement créé. Ce lien permet à l'utilisateur de voir à l'immédiat quel sont les différents extraits, et lesquelles doivent être modifier.

N° 789244

2 Demande d'aide au logement

► Si vous êtes propriétaire

- Vous avez emprunté pour acheter ou faire construire : oui non
- Vous avez emprunté pour réaliser des travaux d'augmentation de la surface habitable : oui non
- Vous avez emprunté pour réaliser des travaux d'amélioration de votre logement (peinture de papier peint, moquette, peinture, mobilier de cuisine ou de salle de bains) : oui non
- AIDE CAR (précisez : viager, location-vendeur, location-acquéreur, rachat de souche, etc.) : _____
- S'agit-il d'un PAP, PC ou PAS ? oui non
- Avez-vous un ou plusieurs prêts complémentaires ? oui non
- Etes-vous à jour dans le remboursement de votre prêt ? oui non
- Surface totale du logement : _____ m² Ne pas tenir compte des balcons, loggias, terrasses.
- Menez-vous en location une partie de votre logement ? oui non
- Si oui, depuis quelle date ? Surface totale louée : _____ m² Ne pas tenir compte des balcons loggias, terrasses.
- Votre logement comprend-il une ou plusieurs pièces à usage professionnel ? oui non
- Si oui, depuis quelle date ? Surface totale des pièces à usage professionnel : _____ m² Ne pas tenir compte des balcons loggias, terrasses.
- Etes-vous co-entrepreneur avec une personne autre que votre conjoint(e), concubin(e) ou pacs(e) ? oui non
- Assurez-vous des frais supplémentaires pour l'occupation d'un deuxième logement pour des raisons professionnelles ? oui non
- Si oui, depuis quelle date ?
- Non et adresse de l'employeur _____
- Non et adresse du propriétaire _____
- Le logement répond-il aux caractéristiques de décence énumérées ci-dessous : oui non

Principales caractéristiques de décence que le logement doit respecter (Arrêté 2002-129 du 20 janvier 2002)

- Le logement ne doit pas avoir fait l'objet d'un arrêt d'insalubrité ou de péril ;
- la toiture, les murs, les portes, les plafonds, les planchers, les installations électriques et de gaz ne présentent pas de risques manifestes pour la santé et la sécurité physique des occupants ;
- l'éclairage et la ventilation sont suffisants et sans danger ;
- il n'y a pas moins un accès extérieur avec un point d'eau potable froide et chaude ;
- l'assainissement et de chauffage est suffisante et sans danger ;
- l'isolation thermique _____

Fig 29-Lien du résultat de la recherche

Et juste après le service nous avons l'interface de gestion, qui nous permet de paramétriser le logiciel, les comptes et préférences et l'historique de l'utilisateur.

paramètre Compte Historique

Historique

Fig 29-Interface de gestion

5.4 Teste

5.4.1 Enregistrement du modèle :Image numérisée vide



N° 11916 * 04
N° 50869 # 04

Réinitialiser le formulaire

Valider et imprimer



N° 3916
DGFiP

DÉCLARATION PAR UN RÉSIDENT D'UN COMPTE OUVERT HORS DE FRANCE

(CODE GÉNÉRAL DES IMPÔTS : ART. 1649 A, 2^e ET 3^e AL. ; ART. 1758 ET 1736 IV)

1. IDENTITÉ DU (OU DES) DÉCLARANT(S)
<ul style="list-style-type: none">• NOM PATRONYMIQUE (ET NOM D'USAGE, S'IL Y A LIEU), PRÉNOMS, DATE ET LIEU DE NAISSANCE DU (OU DES) DÉCLARANT(S) : <input type="text"/>• DOMICILE : <input type="text"/>• QUALITÉ : <input type="text"/>
2. VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES TITULAIRE D'UN COMPTE OUVERT OU UTILISÉ À L'ÉTRANGER
2.1. ET VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES UN PARTICULIER N'AGISSANT PAS EN QUALITÉ D'EXPLOITANT D'UNE ACTIVITÉ DONNANT LIEU À DÉCLARATION SPÉCIFIQUE DE RÉSULTATS
<ul style="list-style-type: none">• NOM PATRONYMIQUE, PRÉNOMS, DATE ET LIEU DE NAISSANCE, DOMICILE DU (OU DES) TITULAIRE(S) DU COMPTE : <input type="text"/> <input type="text"/>
2.2. ET VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES UN PARTICULIER AGISSANT EN QUALITÉ D'EXPLOITANT D'UNE ACTIVITÉ DONNANT LIEU À DÉCLARATION SPÉCIFIQUE DE RÉSULTATS OU UNE PERSONNE MORALE (2)
<ul style="list-style-type: none">• FORME JURIDIQUE DE VOTRE ENTREPRISE : <input type="text"/> (3)• NOM PATRONYMIQUE, PRÉNOMS, DATE ET LIEU DE NAISSANCE, DOMICILE DU (OU DES) TITULAIRE(S) DU COMPTE : <input type="text"/> <input type="text"/>• DÉSIGNATION OU RAISON SOCIALE DU TITULAIRE DU COMPTE (2) : <input type="text"/>• NUMÉRO SIRET : <input type="text"/>• ADRESSE DU LIEU D'ACTIVITÉ, DU SIÈGE SOCIAL OU DU PRINCIPAL ÉTABLISSEMENT (2) : <input type="text"/> <input type="text"/>
3. VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES BÉNÉFICIAIRE D'UNE PROCURATION SUR UN COMPTE OUVERT OU UTILISÉ À L'ÉTRANGER*
3.1. ET VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES UN PARTICULIER N'AGISSANT PAS EN QUALITÉ D'EXPLOITANT D'UNE ACTIVITÉ DONNANT LIEU À DÉCLARATION SPÉCIFIQUE DE RÉSULTATS
<ul style="list-style-type: none">• NOM PATRONYMIQUE, PRÉNOMS, DATE ET LIEU DE NAISSANCE, DOMICILE DU (OU DES) TITULAIRE(S) DE LA PROCURATION : 49

* Sauf si cette procuration est utilisée au profit exclusif d'un non-résident.

Fig 30-cerfa vide

Avant l'enregistrement du modèle ,l'utilisateur doit saisir les marqueurs qui permettront d'extraire les données du document numérisée une fois rempli.Le numéro du cerfa est le **50869#04**.

5.4.2 Enrégistrement du modèle :Texte Extrait par Tes-seract

N° 3916 — I I-‘I I I R I M If R I I: N/I 'I' I () N/I 1.1: 2009 01 25290 PO — Février 2009 — 8 006859 Réinitialiser le formulaire I Valider et imprimer n E i .1 N° 11916 -X- 04 N° 50869 # 04 N° 3916 DGFIP Liberté ° Égalité ° Fraternité RÉPUBLIQUE FRANÇAISE DÉCLARATION PAR UN RÉSIDENT D'UN COMPTE OUVERT HORS DE FRANCE (CODE GÉNÉRAL DES IMPÔTS : ART. 1649 A, 2e ET 3e AL. ; ART. 1758 ET 1736 IV) 1. IDENTITÉ DU (OU DES) DÉCLARANT(SI - NOM PATRONYMIQUE (ET NOM D'USAGE, S'IL Y A LIEU), PRÉNOMS, DATE ET LIEU DE NAISSANCE DU (OU DES) DÉCLARANT(S) : | | - DOMICILE : | | - QUALITÉ : 2. VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES TITULAIRE D'UN COMPTE OUVERT OU UTILISÉ À L'ÉTRANGER 2.1. ET VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES UN PARTICULIER N'AGISSANT PAS EN QUALITÉ D'EXPLOITANT D'UNE ACTIVITÉ DONNANT LIEU A DÉCLARATION SPÉCIFIQUE DE RÉSULTATS - NOM PATRONYMIQUE, PRÉNOMS, DATE ET LIEU DE NAISSANCE, DOMICILE DU (OU DES) TITULAIRE(S) DU COMPTE : 2.2. ET VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES UN PARTICULIER ACISSANT EN QUALITÉ DEXPLOITANT D'UNE ACTIVITÉ DONNANT LIEU A DÉCLARATION SPÉCIFIQUE DE RÉSULTATS OU UNE PERSONNE MORALE (2) ' FORME JURIDIQUE DE VOTRE ENTREPRISE : (3) - NOM PATRONYMIQUE, PRÉNOMS, DATE ET LIEU DE NAISSANCE, DOMICILE DU (OU DES) TITULAIRE(S) DU COMPTE : ' DÉSIGNATION OU RAISON SOCIALE DU TITULAIRE DU COMPTE (2) : I °NUMÉROSIRET:||||||| - ADRESSE DU LIEU D'ACTIVITÉ, DU SIÈGE SOCIAL OU DU PRINCIPAL ÉTABLISSEMENT (2) : 3. VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES BÉNÉFICIAIRE D'UNE PROCURATION SUR UN COMPTE OUVERT OU UTILISÉ À L'ÉTRANGER* 3. I. ET VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES UN PARTICULIER N'AGISSANT PAS EN QUALITÉ D'EXPLOITANT D'UNE ACTIVITÉ DONNANT LIEU A DÉCLARATION SPÉCIFIQUE DE RÉSULTATS - NOM PATRONYMIQUE, PRÉNOMS, DATE ET LIEU DE NAISSANCE, DOMICILE DU (OU DES) TITULAIRE(S) DE LA PROCURATION : * Saufsi cette procuration est utilisée au profit exclusif d'un non-résident. m MINISTÈRE DU BUDGET DES COMPTES PUBLICS ET DE LA FONCTION PUBLIQUE

Fig 31-Extraction de texte

Avant que l'utilisateur saisisse les marqueurs il lui ai indispensable d'avoir à ses yeux le texte extrait.Ainsi le texte extrait et l'autocomplete (de saisis) lui permettra aisement d'introduire les marqueurs correspondant au document.

5.4.3 Table Champs :Marqueurs saisis Par l'utilisateur

Base de données ocr

Structure de la table champs

Colonne	Type	Null	Défaut
ID_CHAMPS	int(11)	Non	
ID_CERFA	varchar(40)	Non	
LIBELLE	text	Oui	NULL
TYPE	text	Oui	NULL
MDEBUT	text	Oui	NULL
MDEFIN	text	Oui	NULL
barette	int(11)	Non	

Contenu de la table champs

122 50869#04 NOM PATRYMONIQUE_PRENOMS ,DATE ET LIEU DE NAISSANCE	textp	DÉCLARANT(S) :	DOMICILE	0
123 50869#04 Domicile	textp	DOMICILE	QUALITÉ	0
124 50869#04 NOM PATRYMONIQUE_PRENOMS ,DATE ET LIEU DE NAISSANCE	textp	du COMPTE :	2.2. ET VOUS	0
125 50869#04 Forme juridique	textp	VOTRE ENTREPRISE : (3)		0
126 50869#04 NOM PATRYMONIQUE_PRENOMS ,DATE ET LIEU DE NAISSANCE	textp	du COMPTE :	DÉSIGNATION	0
127 50869#04 Raison sociale	textp	DU COMPTE (2) :	NUMÉROSIRET:IIIIIIIIIIII 1	
128 50869#04 Adresse du lieu d'activité	textp	ÉTABLISSEMENT (2) : 3. VOUS		0

Fig 32-Table champs

Une fois que l'utilisateur valide le modèle, les marqueurs associés au document sont insérés dans la table "champs"

5.4.4 Lancement du traitement : Image numérisée rempli

 N° 11916 * 04 N° 50869 # 04	<input type="button" value="Réinitialiser le formulaire"/> <input type="button" value="Valider et imprimer"/>	 N° 3916 DGFiP
---	--	---

DÉCLARATION PAR UN RÉSIDENT D'UN COMPTE OUVERT HORS DE FRANCE

(CODE GÉNÉRAL DES IMPÔTS : ART. 1649 A, 2^e ET 3^e AL. ; ART. 1758 ET 1736 IV)

1. IDENTITÉ DU (OU DES) DÉCLARANT(S)	
<ul style="list-style-type: none"> • NOM PATRONYMIQUE (ET NOM D'USAGE, S'IL Y A LIEU), PRÉNOMS, DATE ET LIEU DE NAISSANCE DU (OU DES) DÉCLARANT(S) : <input type="text" value="TSHOMOKONGO TARCISSE"/> • DOMICILE : <input type="text" value="15 RUE 18 B RUE DU HARAS"/> • QUALITÉ : <input type="text" value="MVOULAKOINS"/> 	
2. VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES TITULAIRE D'UN COMPTE OUVERT OU UTILISÉ À L'ÉTRANGER	
<p>2.1. ET VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES UN PARTICULIER N'AGISSANT PAS EN QUALITÉ D'EXPLOITANT D'UNE ACTIVITÉ DONNANT LIEU À DÉCLARATION SPÉCIFIQUE DE RÉSULTATS</p> <ul style="list-style-type: none"> • NOM PATRONYMIQUE, PRÉNOMS, DATE ET LIEU DE NAISSANCE, DOMICILE DU (OU DES) TITULAIRE(S) DU COMPTE : <input type="text" value="M'VOU YORRIK"/> <input type="text" value="petit texte en minuscule"/> 	
<p>2.2. ET VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES UN PARTICULIER AGISSANT EN QUALITÉ D'EXPLOITANT D'UNE ACTIVITÉ DONNANT LIEU À DÉCLARATION SPÉCIFIQUE DE RÉSULTATS OU UNE PERSONNE MORALE (2)</p> <ul style="list-style-type: none"> • FORME JURIDIQUE DE VOTRE ENTREPRISE : <input type="text" value="FK"/> (3) • NOM PATRONYMIQUE, PRÉNOMS, DATE ET LIEU DE NAISSANCE, DOMICILE DU (OU DES) TITULAIRE(S) DU COMPTE : <input type="text" value="ABCDEFGHIJK9MNOPQRSTUVWXYZ"/> <input type="text"/> • DÉSIGNATION OU RAISON SOCIALE DU TITULAIRE DU COMPTE (2) : <input type="text"/> • NUMÉRO SIRET : <input type="text" value="1 2 3 4 5 6 7 8 _ 1 0 0 _ _"/> • ADRESSE DU LIEU D'ACTIVITÉ, DU SIÈGE SOCIAL OU DU PRINCIPAL ÉTABLISSEMENT (2) : <input type="text" value="15 RUE ALI BELAHOUINE 18852"/> <input type="text"/> 	
3. VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES BÉNÉFICIAIRE D'UNE PROCURATION SUR UN COMPTE OUVERT OU UTILISÉ À L'ÉTRANGER*	
<p>3.1. ET VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES UN PARTICULIER N'AGISSANT PAS EN QUALITÉ D'EXPLOITANT D'UNE ACTIVITÉ DONNANT LIEU À DÉCLARATION SPÉCIFIQUE DE RÉSULTATS</p> <ul style="list-style-type: none"> • NOM PATRONYMIQUE, PRÉNOMS, DATE ET LIEU DE NAISSANCE, DOMICILE DU (OU DES) TITULAIRE(S) DE LA PROCURATION : 	

Fig 32.1-Cerfa rempli

Ceci n'est que l'image qui va être traiter par l'application

5.4.5 Texte Extrait par Tesseract après le traitement

N° 3916 — I M I R IM I: R I I N 11T UNI LI 2009 0125290 PO — Février
2009 — 8 006859 Réinitialiser le formulaire | Valider et imprimer | E | _1
N° 11916 * 04 N° 50869 # 04 N° 3916 DGFiP Liberté ° Égalité ° Fraternité
RÉPUBLIQUE FRANÇAISE DÉCLARATION PAR UN RÉSIDENT D'UN
COMPTE OUVERT HORS DE FRANCE (CODE GÉNÉRAL DES IMPÔTS :
ART. 1649 A, 2e ET 3e AL. ; ART. 1758 ET 1736 IV) 1. IDENTITÉ DU (OU
DES) DÉCLARANT(S) - NOM PATRONYMIQUE (ET NOM D'USAGE, S'IL Y
A LIEU), PRÉNOMS, DATE ET LIEU DE NAISSANCE DU (OU DES)
DÉCLARANTS : | TSHOMOKONGOTARCISSE ° DOMICILE : || 15 RUE 18
B RUE DU HARAS MVOULAKOINS - QUALITÉ : | 2. VOUS [OU L'UN DES
MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES TITULAIRE D'UN
COMPTE OUVERT OU UTILISÉ À L'ÉTRANGER 2. 1. ET VOUS [OU L'UN
DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES UN PARTICULIER
N'AGISSANT PAS EN QUALITÉ DEXPLOITANT D'UNE ACTIVITÉ
DONNANT LIEU A DÉCLARATION SPÉCIFIQUE DE RÉSULTATS ' NOM
PATRONYMIQUE, PRÉNOMS, DATE ET LIEU DE NAISSANCE, DOMICILE
DU (OU DES) TITULAIRE(S) DU COMPTE : M'VOU YORRIK petit texte en
minuscule 2.2. ET VOUS [OU L'UN DES MEMBRES DE VOTRE FOYER
FISCAL] (1) ÊTES UN PARTICULIER AGISSANT EN QUALITÉ
D'EXPLOITANT D'UNE ACTIVITÉ DONNANT LIEU A DÉCLARATION
SPÉCIFIQUE DE RÉSULTATS OU UNE PERSONNE MORALE (2) ' FORME
JURIDIQUE DE VOTRE ENTREPRISE : FK (3) - NOM PATRONYMIQUE,
PRÉNOMS, DATE ET LIEU DE NAISSANCE, DOMICILE DU (OU DES)
TITULAIRE(S) DU COMPTE : ABCDEFGHJK9MNOPQRSTUVWXYZ °
DÉSIGNATION OU RAISON SOCIALE DU TITULAIRE DU COMPTE (2)t | -
NuMERoSIRET:II2I3I4I5I6I7I8I III0IOI | | - ADRESSE DU LIEU D'ACTIVITÉ,
DU SIÈGE SOCIAL OU DU PRINCIPAL ÉTABLISSEMENT (2) : 15 RUE ALI
BELAHOUINE 18852 . VOUS [OU L'UN DES MEMBRES DE VOTRE
FOYER FISCAL] (1) ÊTES BÉNÉFICIAIRE D'UNE PROCURATION SUR UN
COMPTE OUVERT OU UTILISÉ À L'ÉTRANGER* 3.1 . ET VOUS [OU L'UN
DES MEMBRES DE VOTRE FOYER FISCAL] (1) ÊTES UN PARTICULIER
N'AGISSANT PAS EN QUALITÉ DEXPLOITANT D'UNE ACTIVITÉ
DONNANT LIEU A DÉCLARATION SPÉCIFIQUE DE RÉSULTATS - NOM
PATRONYMIQUE, PRÉNOMS, DATE ET LIEU DE NAISSANCE, DOMICILE
DU (OU DES) TITULAIRE(S) DE LA PROCURATION : * Saufsi cette
procuration est utilisée au profit CXCIUSIIidILIn non-résident. A
MINISTÈRE DU BUDGET DES COMPTES PUBLICS ET DE LA FONCTION
PUBLIQUE

Cette partie reste cachée à l'utilisateur, nous vous l'avons montré pour faire la comparaison avec le texte extrait (du document vide) et afin d'expliquer le résultat ci-après.

5.4.6 Tableau contenant tous les mots du texte extrait après traitement

```
Array ([0] => N° [1] => 3916 [2] => — [3] => I [4] => M [5] => l' [6] => R [7] => IM [8] =>
I: [9] => R [10] => I [11] => l' [12] => N [13] => 11T' [14] => UN/i [15] => LI' [16] =>
2009 [17] => 0125290 [18] => PO [19] => — [20] => Février [21] => 2009 [22] => — [23]
=> 8 [24] => 006859 [25] => Réinitialiser [26] => le [27] => formulaire [28] => I [29] =>
Valider [30] => et [31] => imprimer [32] => I [33] => E [34] => j [35] => _1 [36] => N°
[37] => 11916 [38] => * [39] => 04 [40] => N° [41] => 50869 [42] => # [43] => 04 [44] =>
N° [45] => 3916 [46] => DGFiP [47] => Liberté [48] => ° [49] => Égalité [50] => ° [51]
=> Fraternité [52] => RÉPUBLIQUE [53] => FRANÇAISE [54] => DÉCLARATION [55]
=> PAR [56] => UN [57] => RÉSIDENT [58] => D'UN [59] => COMPTE [60] => OUVERT
[61] => HORS [62] => DE [63] => FRANCE [64] => (CODE [65] => GÉNÉRAL [66] =>
DES [67] => IMPÔTS [68] => : [69] => ART. [70] => 1649 [71] => A, [72] => 2e [73] =>
ET [74] => 3e [75] => AL. [76] => ; [77] => ART. [78] => 1758 [79] => ET [80] => 1736
[81] => IV [82] => 1. [83] => IDENTITÉ [84] => DU [85] => (OU [86] => DES) [87] =>
DÉCLARANT(S) [88] => - [89] => NOM [90] => PATRONYMIQUE [91] => (ET [92] =>
NOM [93] => D'USAGE, [94] => S'il [95] => Y [96] => A [97] => LIEU), [98] =>
PRÉNOMS, [99] => DATE [100] => ET [101] => LIEU [102] => DE [103] =>
NAISSANCE [104] => DU [105] => (OU [106] => DES) [107] => DÉCLARANRS) [108]
=> : [109] => | [110] => TSHOMOKONGOTARCISSE [111] => ° [112] => DOMICILE
[113] => : [114] => | [115] => | [116] => 15 [117] => RUE [118] => 18 [119] => B [120]
=> RUE [121] => DU [122] => HARAS [123] => MVOULAKOINS [124] => - [125] =>
QUALITÉ [126] => : [127] => | [128] => 2. [129] => VOUS [130] => (OU [131] => L'UN
[132] => DES [133] => MEMBRES [134] => DE [135] => VOTRE [136] => FOYER [137]
=> FISCAL] [138] => (1) [139] => ÉTES [140] => TITULAIRE [141] => D'UN [142] =>
COMPTE [143] => OUVERT [144] => OU [145] => UTILISÉ [146] => À [147] =>
L'ÉTRANGER [148] => 2. [149] => 1 [150] => . [151] => ET [152] => VOUS [153] =>
[OU [154] => L'UN [155] => DES [156] => MEMBRES [157] => DE [158] => VOTRE
[159] => FOYER [160] => FISCAL] [161] => (1) [162] => ÉTES [163] => UN [164] =>
PARTICULIER [165] => N'AGISSANT [166] => PAS [167] => EN [168] => QUALITÉ
[169] => DEXPLOITANT [170] => D'UNE [171] => ACTIVITÉ [172] => DONNANT [173]
=> LIEU [174] => A [175] => DÉCLARATION [176] => SPÉCIFIQUE [177] => DE [178]
=> RÉSULTATS [179] => ' [180] => NOM [181] => PATRONYMIQUE, [182] =>
PRÉNOMS, [183] => DATE [184] => ET [185] => LIEU [186] => DE [187] =>
NAISSANCE, [188] => DOMICILE [189] => DU [190] => (OU [191] => DES) [192] =>
TITULAIRE(S) [193] => DU [194] => COMPTE [195] => : [196] => M'VOU [197] =>
YORRIK [198] => petit [199] => texte [200] => en [201] => minuscule [202] => 2.2.
[203] => ET [204] => VOUS [205] => [OU [206] => L'UN [207] => DES [208] =>
MEMBRES [209] => DE [210] => VOTRE [211] => FOYER [212] => FISCAL] [213] =>
(1) [214] => ÉTES [215] => UN [216] => PARTICULIER [217] => AGISSANT [218] =>
EN [219] => QUALITÉ [220] => D'EXPLOITANT [221] => D'UNE [222] => ACTIVITÉ
```

[223] => DONNANT [224] => LIEU [225] => A [226] => DÉCLARATION [227] => SPÉCIFIQUE [228] => DE [229] => RÉSULTATS [230] => OU [231] => UNE [232] => PERSONNE [233] => MORALE [234] => (2) [235] => ' [236] => FORME [237] => JURIDIQUE [238] => DE [239] => VOTRE [240] => ENTREPRISE [241] => : [242] => FK [243] => (3) [244] => - [245] => NOM [246] => PATRONYMIQUE, [247] => PRÉNOMS, [248] => DATE [249] => ET [250] => LIEU [251] => DE [252] => NAISSANCE, [253] => DOMICILE [254] => DU [255] => (OU [256] => DES) [257] => TITULAIRE(S) [258] => DU [259] => COMPTE [260] => : [261] => ABCDEFGHJK9MNOPQRSTUVWXYZ [262] => ° [263] => DÉSIGNATION [264] => OU [265] => RAISON [266] => SOCIALE [267] => DU [268] => TITULAIRE [269] => DU [270] => COMPTE [271] => (2) [272] => t [273] => | [274] => - NuMERoSIRET;1121314151617181 [275] => IIIIOIOI [276] => | [277] => | [278] => - [279] => ADRESSE [280] => DU [281] => LIEU [282] => D'ACTIVITÉ, [283] => DU [284] => SIÈGE [285] => SOCIAL [286] => OU [287] => DU [288] => PRINCIPAL [289] => ÉTABLISSEMENT [290] => (2) [291] => : [292] => 15 [293] => RUE [294] => ALI [295] => BELAHOUINE [296] => 18852 [297] => . [298] => VOUS [299] => [OU [300] => L'UN [301] => DES [302] => MEMBRES [303] => DE [304] => VOTRE [305] => FOYER [306] => FISCAL] [307] => (1) [308] => ÉTES [309] => BÉNÉFICIAIRE [310] => D'UNE [311] => PROCURATION [312] => SUR [313] => UN [314] => COMPTE [315] => OUVERT [316] => OU [317] => UTILISÉ [318] => À [319] => L'ÉTRANGER* [320] => 3.1 [321] => . [322] => ET [323] => VOUS [324] => [OU [325] => L'UN [326] => DES [327] => MEMBRES [328] => DE [329] => VOTRE [330] => FOYER [331] => FISCAL] [332] => (1) [333] => ÉTES [334] => UN [335] => PARTICULIER [336] => N'AGISSANT [337] => PAS [338] => EN [339] => QUALITÉ [340] => D'EXPLOITANT [341] => D'UNE [342] => ACTIVITÉ [343] => DONNANT [344] => LIEU [345] => A [346] => DÉCLARATION [347] => SPÉCIFIQUE [348] => DE [349] => RÉSULTATS [350] => - [351] => NOM [352] => PATRONYMIQUE, [353] => PRÉNOMS, [354] => DATE [355] => ET [356] => LIEU [357] => DE [358] => NAISSANCE, [359] => DOMICILE [360] => DU [361] => (OU [362] => DES) [363] => TITULAIRE(S) [364] => DE [365] => LA [366] => PROCURATION [367] => : [368] => * [369] => Saufsi [370] => cette [371] => procuration [372] => est [373] => utilisée [374] => au [375] => profit [376] => CXCLUSIflDILn [377] => non-résident. [378] => A [379] => MINISTÈRE [380] => DU [381] => BUDGET [382] => DES [383] => COMPTES [384] => PUBLICS [385] => ET [386] => DE [387] => LA [388] => FONCTION [389] => PUBLIQUE)

Fig 32.2-Tableau contenant les mots du document

Cette partie n'est pas visible par l'utilisateur, nous vous le montrons juste pour la clarification des différents résultats de l'extraction.

5.4.7 Table champ :Résutat du traitement

← T →	ID_CHAMP	ID	VALEUR
<input type="checkbox"/> Copier <input type="checkbox"/> Effacer	175	263	NOM PATRONYMIQUE (ET NOM D'USAGE,
<input type="checkbox"/> Copier <input type="checkbox"/> Effacer	176	263	15 RUE 18 B RUE DU HARAS MVOULAKOINS
<input type="checkbox"/> Copier <input type="checkbox"/> Effacer	177	263	MVOU YORRIK petit texte en minuscule
<input type="checkbox"/> Copier <input type="checkbox"/> Effacer	178	263	FK
<input type="checkbox"/> Copier <input type="checkbox"/> Effacer	179	263	ABCDEFGHIJKLMNPQRSTUVWXYZ °
<input type="checkbox"/> Copier <input type="checkbox"/> Effacer	180	263	12345678
<input type="checkbox"/> Copier <input type="checkbox"/> Effacer	181	263	15 RUE ALI BELAHOUINE 18852

hss Tout cocher / Tout décocher Pour la sélection : Modifier Effacer Exporter

Fig 32.3-Table champ

Ceci n'est que le contenu de la table **champ** qui stocke le contenu de tous les modèles traités par l'application.

5.4.8 Affichage du Résultat

un teste
NOM PATRYMONIQUE,PRENOMS ,DATE ET LIEU DE NAISSANCE
NOM PATRONYMIQUE (ET NOM D
Domicile
15 RUE 18 B RUE DU HARAS MVOULAKOINS
NOM PATRYMONIQUE,PRENOMS ,DATE ET LIEU DE NAISSANCE
M
Forme juridique
FK
NOM PATRYMONIQUE,PRENOMS ,DATE ET LIEU DE NAISSANCE
ABCDEFGHIJKLMNPQRSTUVWXYZ ↵
Raison sociale
12345678
Adresse du lieu d' activit ↵
15 RUE ALI BELAHOUINE 18852

Fig 32.4-Resultat-affichage

Nous constatons que la première valeur du champs n'est pas celle attendu. En effet Le moteur de recherche a simplement exécuté toutes les démarches que nous vous avons expliqué plus haut. Il détecte premièrement le marqueur Debut (le premier marqueur de la table champs) qui est "DECLARANT(S)" à la ligne (7) du texte ensuite il extrait les informations jusqu'à rencontré le marqueur de fin qui est ici "DOMICILE" à la ligne (9) du coup tout le texte compris entre ces marqueurs sera extrait(de la ligne 7 à 9) et sera ensuite modifiable par l'utilisateur. L'utilisateur est capable de modifier les résultats pour introduire les bonnes informations si le rendu n'est pas celui attendu.

5.4.9 Nouveau Résultat après changement de marqueur Debut

En effet lorsque l'on change le marqueurs debut du premier champ de la table **champ** par **DECLARANRS** :) ligne(9) du deuxième texte extrait ,on obtient le résultat suivant :

+ Options			
	ID_CHAMP	ID	VALEUR
<input type="checkbox"/> Modifier Copier Effacer	1	266	TSHOMOKONGOTARCISSE °
<input type="checkbox"/> Modifier Copier Effacer	2	266	15 RUE 18 B RUE DU HARAS MVOULAKOINS
<input type="checkbox"/> Modifier Copier Effacer	3	266	M'YOU YORRIK petit texte en minuscule
<input type="checkbox"/> Modifier Copier Effacer	4	266	FK
<input type="checkbox"/> Modifier Copier Effacer	5	266	ABCDEFGHIJKMNOPQRSTUVWXYZ °
<input type="checkbox"/> Modifier Copier Effacer	6	266	12345678
<input type="checkbox"/> Modifier Copier Effacer	7	266	15 RUE ALI BELAHOUINE 18852

Tout cocher / Tout décocher Pour la sélection : Modifier Effacer Exporter

Fig 32.5-Nouveau résultat

CONCLUSION Nous avons pu voir les textes extraits par Tesseract avant et après l'insertion des données.Nous pouvons déduire que la qualité de l'extraction de données de l'image numérisée est un facteur important à la réussite du traitement.

Chapitre 6

CONCLUSION GENERALE

6.1 Synthèse

L'objet de ce mémoire portait sur la conception et la réalisation d'un outil de supervision à distance des plateformes hétérogènes, c'est-à-dire supervise les services à partir de n'importe quelle situation physique et logique répondant aux besoins des entreprises. Le chapitre deux nous a permis de mieux appréhender la notion de supervision et de faire un tour d'horizon sur les difficultés que rencontrent quelques solutions existantes afin de les surmonter pour concevoir un logiciel qui réduit au maximum l'effort humain.

6.2 Difficultés rencontrées

Tout au départ nous avons voulu concevoir l'application sous un langage autre que php, par exemple c++. Mais faute de certains outils d'amélioration sur l'image numérisée permettant de supprimer les bruits, de transformer l'image en gris, de faire des zooms nettes afin d'avoir une meilleure qualité d'extraction, nous n'avons pas pu arriver au bout de notre objectif. Il était donc assez difficile de résoudre le problème de la reconnaissance de caractère avec des API libres et disponibles autre que le PHP qui offrait un large champ de possibilité. Travailler sur un projet qui fait intervenir plusieurs entités n'est pas toujours une tâche aisée. A côté de cette difficulté, nous pouvons noter celle liée à l'encodage des caractères lorsque l'on mélange plusieurs API et langage de programmation.

La conception d'un ensemble d'algorithme permettant de traiter au mieux le problème d'extraction n'a pas du tout été facile au départ. Le facteur de tolérance n'a pas été refléchi de manière instantanée, nous étions confrontés à des problèmes d'ordonnancement dans l'extraction de données à cause des marqueurs des champs qui changeaient parfois d'une extraction à une autre. La recherche du numéro de cerfa qui était pourtant assez simple nous posait problème à cause du fait que nous ne savions pas au départ quelle était la forme du format à choisir. Nous avons alors fait un raisonnement générale qui nous a permis de traiter l'ensemble des numéros de cerfa existants. Nous avons toutefois fait face à toutes ces difficultés en faisant appel à notre sens de la détermination. Ce projet nous aura ainsi permis entre autres d'affirmer nos connaissances en matière de gestion de projets.

6.3 Perspectives

- Toujours dans les soucis de réduire l'effort humain nous travaillons sur un nouveau format de fichier le 'cef' qui permettra à l'utilisateur d'importer, d'exporter et de cloner le modèle de Cerfa d'une entité à une autre.
- En vue d'élargir le filtre de recherche et de permettre à un utilisateur de rechercher un document avec tous les critères possible nous travaillons également sur un modèle de requête 'csql' (exécutable en ligne de commande à travers l'interface du logiciel et plus proche du langage humain pour faciliter et rendre son utilisation accessible même aux utilisateurs occasionnels).
- Passe du serveur apache au serveur node.js
- Gestion des droits d'accès avancés et un contrôle d'historique poussé selon le bon vouloir et les exigences de l'entreprise.
- L'ouverture du logiciel aux personnes handicapées en y ajoutant la reconnaissance vocale.
- Et enfin l'amélioration de la qualité d'extraction et du moteur de recherche.

Apres plusieurs testes, nous avons constaté que le traitement d'un Cerfa d'une page prenait en moyenne 1 minute pour un être humain et 15 secondes pour un traitement automatique, ce qui nous donne un gain de 45 secondes (75%) pour une opération de traitement. En considérant qu'une journée est constituée de 8 heures de travail (6 heures de gain) et qu'une année est composé de 365 jours (273.75 jour de gain) nous voyons toute l'utilité et la nécessité de passé par un traitement automatique en vue de répondre dans les plus brefs délais à toute les demandes possibles et imaginables.

Chapitre 7

Références et liens

- Tesseract :
<http://linuxfr.org/news/tesseract-ocr>
- [http://en.wikipedia.org/wiki/Tesseract_\(software\)Ocrad](http://en.wikipedia.org/wiki/Tesseract_(software)Ocrad) :
<http://en.wikipedia.org/wiki/Ocrad>
- GOCR :
<http://en.wikipedia.org/wiki/GOCR>
- Quelques informations sur les outils de reconnaissance :
<http://romain.riedinger.free.fr/cdc.pdf>
- FRee ocr :
<http://www.paperfile.net/index.html>
- Ocr Abbyy :
http://finereader.abbyy.com/about_ocr/whatisocr/Ocrgratuits :
<http://roget.biz/4-ocr-gratuits-en-ligne-et-sans-logiciel>
- Imageagick :
<http://www.imagemagick.org/>
- Lecture sur php :
<http://www.php.net/>
- Cerfa :
<http://www.service-public.fr/>
- Html,jquery,javascript :
<http://www.w3schools.com/>
- Tesseract :
<https://code.google.com/p/tesseract-ocr/w/list>