République Algérienne Démocratique et Populaire
Ministère de L'Enseignement Supérieur et de la Recherche Scientifique
Université Mouloud Mammeri Tizi-Ouzou
Faculté de Génie Electrique & d'Informatique
Département d'Informatique

MEMOIRE

En vue de l'obtention de Diplôme de Master2 en informatique

Thème

Navigation dans une base d'images des manuscrits arabes anciens

Dirigé et proposé par :

Mr SOULAH Mohand Ou Rabah

Réalisé par :

Melle CHELLOUCHE Kaissa
Melle SAAD Samia

Promotion 2010/2011

Résumé :

La numérisation des manuscrits arabes anciens sous forme d'images donne une version très fidèle à l'original, mais le mode image ne permet pas l'accès direct à l'information ce qui rend ces images inaccessibles. Pour rendre le document numérisé plus utile, nous devons donc aller plus loin que le mode image et permettre d'accéder à ce que l'on appelle le mode texte ce qui nous oriente vers la recherche d'une méthode qui se base sur du texte pour indexer les images des manuscrits. Ainsi, le catalogage se propose comme une solution. En effet, constitué d'un ensemble de notices descriptives de manuscrits sur les quelle il est possible d'effectuer des recherches. Cette solution pose divers problèmes d'accès aux manuscrits arabes numérisés de fait qu'il donne une description très pauvre de contenu des manuscrits. Pour enrichir le catalogue nous proposons une solution qui consiste à annoter les images numérique des manuscrits pour associer à certaines de leurs parties des mots clés, des notes ou plus généralement, une représentation textuelle de leur contenu. L'indexation peut alors s'effectuer facilement sur ces représentations textuelles.

Introduction générale

Le document quelque soit son support, papier imprimé, ou document manuscrit est un moyen de communication et de transmission des connaissances humaines, siècle après siècle. Hélas, le temps, seul, est capable d'endommager de précieux trésors se trouvant dans de tels documents. La manipulation de ces œuvres constitue un danger supplémentaire de détérioration. La révolution électronique apparue ces dernières années et qui a bouleversé la vie quotidienne de l'humanité dans tous les domaines, apporte une solution concrète à ce problème.

Le passage du document de son aspect analogique vers une forme numérique se fait grâce à la technique de numérisation qui permet de préserver les manuscrits dans de meilleurs états et de donner la possibilité d'accès distant. L'Internet est un support concret qui permet d'atteindre cette finalité. La numérisation des manuscrits arabes anciens répond au double objectif de préservation et d'accessibilité à ces ressources.

Le mode image issu de la numérisation des manuscrits arabes anciens ne donne qu'une image du texte, c'est à dire qu'il ne permet pas d'effectuer l'accès par le contenu du manuscrit. Par contre, le mode texte permet des recherches plein texte sur le contenu. Par conséquent, la recherche d'une image d'un manuscrit particulier dans de une base d'images est d'une difficulté importante car leur contenu n'est pas reconnu comme les fichiers textuels. En effet il est impératif de mettre en place un système d'accès adéquat à ces ressources numérisées. Ainsi, le catalogage se propose comme une solution qui permettra d'effectuer des recherches sur les manuscrits. A ce niveau se pose le problème d'insuffisance de cet outil de fait qu'il décrit d'une manière très sommaire les manuscrits.

Notre travail, se veut d'apporter une solution pragmatique aux divers problèmes posés par l'accès aux manuscrits numérisés en utilisant le catalogue. Nous proposons ainsi les annotations comme outils d'enrichissement de ce catalogue.

L'objectif premier de ce projet est de développer un système permettant l'annotation des images des manuscrits arabes par des mots clés, des remarques ou plus généralement, une représentation textuelle de leur contenu. Le second objectif est de permettre à l'utilisateur de pouvoir réaliser des recherches sur les données saisies lors du processus d'annotation, et par conséquent, un outil d'indexation de données basé sur XML doit être créé, afin de pouvoir retrouver facilement une image à partir d'une annotation.

Pour mener à bien notre travail, nous avons organisé ce dernier en deux partie dont la première est l'état de l'art qui est structurée en trois chapitres comme suit :

Dans le premier chapitre, nous décrirons les divers aspects de la description des manuscrits arabes anciens et leurs préservations sous forme d'images numériques en utilisant la technique de numérisation.

Dans le deuxième chapitre nous aborderons les principes de recherche d'information et les outils d'indexations classiques utilisés dans ce domaine.

Le troisième chapitre traite les limites d'accès aux images numérique des manuscrit en utilisons le catalogue la description des annotations.

La deuxième partie nommée conception et implémentation est structurée en deux chapitre dont le premier mis en évidence la solution proposée et le modèle conceptuel de donnée.

Le Quatrième et dernier chapitre traite l'implémentation et la réalisation qui comporte la représentation de l'environnement de développement dont lequel notre application à été réalisée, les outils utilisées, quelques interfaces de notre application.

1. Introduction

De nombreuses bibliothèque possèdent des fonds patrimoniaux, parmi les quel on trouve les manuscrits arabes anciens qui constituent un patrimoine précieux pour l'humanité. L'accès à leur contenu devient un véritable problème. La numérisation est un outil de mise en valeur des ces documents manuscrits, elle permet de fournir des moyens nouveaux à la recherche d'information, de mieux diffuser les collections auprès du grand public en proposant en ligne des corpus d'ouvrages rares et disséminés. Parallèlement, la numérisation contribue à la sauvegarde et à la conservation des documents manuscrits.

Dans ce chapitre nous décrirons les divers aspects de la description des manuscrits arabes anciens et leurs numérisations sous forme d'image numérique.

2. Généralités sur les manuscrits

2.1 Définition d'un manuscrit

Un manuscrit est un document ou ouvrage écrit à la main, par opposition à ce qui est imprimé, ou tapé à la machine. Le manuscrit ancien a rapport à une époque lointaine durant laquelle l'imprimerie n'existait pas.

Les manuscrits arabes anciens traitent en général, les domaines suivants :

- L'histoire en particulier celle de l'époque coloniale.
- La théologie musulmane.
- L'astrologie.
- La littérature arabe.
- La dissertation en droit (tahrir fi l-fiqh).
- La médecine.
- La pharmacopée.
-……..

Par ailleurs, il existe certains manuscrits qui traitent simultanément des thèmes variés, ce qui rend particulièrement difficile leur classification [Soualah 2008].

2.2 Supports des manuscrits

Les anciens documents manuscrits existent, pour certains d'entre eux, depuis de plusieurs siècles. Leur survie et leur endurance face aux dégradations sont en grande partie dues à :

1. **la résistance des supports d'écritures**
   - **Le papyrus :** une plante aquatique à tige creuse qui est utilisé par les anciens.
   - **Le parchemin :** peau d'animal préparée sur laquelle on peut écrire.
   - **Le papier** : Le papier est apparu par la suite.

2. la qualité des anciennes recettes d'encres

Ces recettes sont faites à base de métaux lourds, de teintures végétales, de sèves animales, et de résines qui leur donnent une certaine indélébilité [Petra2010].

Il existe des règles à suivre dans l'usage des couleurs de l'encre ; le plus utilisé est le noir pour le représenter le texte. Le rouge est conseillé pour l'écriture des noms propres, des nombres, des citations, des termes techniques et pour le texte commenté [kaileh2004].

2.3 Valeur d'un manuscrit

Un manuscrit véhicule les connaissances d'une époque donnée il est donc témoin de cette ère. Notons par ailleurs, qu'un manuscrit est tout d'abord une œuvre produite d'une manière artisanale, utilisant des matériaux rares qui assurent une certaine durabilité. Il est de ce fait un document ancien qui peut être approché à travers ses caractéristiques matérielles et son histoire. Un manuscrit est un document unique, qui existe en un seul exemplaire, Par conséquent, il a une grande valeur par rapport à son contenu ou son histoire.

2.4 Présentation des manuscrits arabes anciens

Les manuscrits arabes existent dans beaucoup de bibliothèques à travers le monde et notamment dans les plus grandes : British Library et la Bibliothèque de France. On peut légitimement prétendre que dans ces deux bibliothèques on trouve les plus importantes collections en quantité et en qualité. Parmi les trésors s'y trouve, par exemple, par exemple un des premiers manuscrits du Coran, datant de la fin du huitième siècle, un Coran entièrement écrit en or pour le sultan Baybars II au Caire.

Par ailleurs, les bibliothèques privées, du monde arabe, sont héritées du père en fils. Souvent, leur contenu est mal organisé, ne présentant la moindre indication de classement sur contenu. La majorité des manuscrits, de ces bibliothèques, se trouvent en mauvaise état. Principalement, cela est du au temps et aux mauvaises conditions de conservation (poussière, termites, humidité et les variations de température).

2.5 Caractéristique d'un manuscrit arabe

Les manuscrits arabes anciens partagent un ensemble de caractéristiques qui peuvent être résumées comme suit [Soualah2008]:

- Début des manuscrits au verso du premier feuillet, alors que le recto est réservé à l'inscription du nom de l'auteur, au commanditaire de l'œuvre et parfois au cachet.
- Le début du texte peut être accompagné d'un décor particulier et représente souvent, le début de chaque chapitre, section ou sourate quand il s'agit d'une œuvre coranique.
- Le texte est écrit en longues lignes, à l'exception du texte poétique.
- Des règles d'usage de l'encre sont observées : la couleur rouge est souvent utilisée pour l'écriture des noms propres, des nombres et des citations.
- L'usage du texte encadré dans les manuscrits coraniques.
- Présence du texte dans les marges.

- Les textes des manuscrits sont souvent accompagnés d'annotations sur les marges ou parfois dans le corps même du texte [kaileh2004], qui ont été ajoutées par les différents lecteurs ou par les auteurs eux-mêmes et qui ont, parfois, autant de valeur que le texte principal.
- Le texte d'une page d'un manuscrit est parfois accompagné d'un dessin particulier réalisé par des enluminures faisant un art artistique qui a comme rôle parfois, purement esthétique. Mais souvent il représente des symboles relatifs l'histoire intellectuelle de culture arabe.

2.6 La numérisation des manuscrits arabes anciens

La numérisation est un outil de mise en valeur des documents manuscrit, elle permet de fournir des moyens nouveaux à la recherche scientifique, de mieux diffuser les collections auprès du grand public en proposant en ligne des corpus d'ouvrages rares riche en information. Parallèlement, la numérisation contribue à la sauvegarde et à la conservation des documents patrimoniaux [Monique 2000]. C'est dans cette optique que se situe le projet de numérisation des manuscrits arabe anciens.

2.6.1 Définition de numérisation

la numérisation est une opération qui permet le passage d'un objet externe (manuscrit) à sa représentation interne c'est-à-dire stocké sur un support numérique (CD, disque dur…) après conversion de l'information physique en informations numérique grâce à des codes binaires (0 et 1) qui sont capables d'être traité par des systèmes numériques par la suite(visualisation, modification, stockage, transmission….).

2.6.2 Outils de numérisation

La numérisation est du aux systèmes dites otiques qui peuvent être classé en deux catégories principales : les caméras numériques et les scanners.

1. **scanner** : un scanner permet de numériser les manuscrits, le résultat de ce processus sont des fichiers images de type TIFF qui sont sauvegardé sur le disque dur avant de les transférer sur d'autres supports de stockage.

2. **caméra numérique :** la caméra numérique se caractérise par un espace mémoire pour stocker les images résultantes qui sont toujours de type TIFF (exemple : une carte mémoire de 64 Mo est une valeur suffisante pour stocker temporairement jusqu'à 80 pages d'un manuscrit).

2.6.3 Les modes de numérisation des manuscrits

1. Mode image

Le texte contenu dans une page d'un manuscrit est représenté sur un mode photographique. Ce type de document est obtenu par la numérisation directe du document. On obtient ainsi une copie électronique du document appelée image. Cette méthode est simple à réaliser et d'un coût relativement faible, mais elle génère cependant des fichiers encombrants. Ce mode

interdit toute recherche sur le texte du l'image. Par conséquent, on a besoin de les indexer pour faciliter l'accès au document.

2. Mode texte

Ce mode est obtenu soit par saisie directe du texte contenu dans un document manuscrit (transcription), soit par reconnaissance optique de caractères (OCR), à partir d'un document en mode image en lui appliquant des algorithmes qui analysent les dessins des caractères manuscrits: il les reconnaît et les traduit en données numériques. Ce mode ne permet pas de conserver la présentation initiale du document et permet la recherche sur le contenu du texte, et la navigation au sein du document.

2.6.4 Objectifs de la numérisation des manuscrits arabes anciens

- **Favoriser l'accès aux manuscrits :**

La numérisation permet de valoriser et de faciliter l'accès à l'information en offrant de nouveaux modes de consultation pour le public, car la majorité des manuscrits arabes originaux se trouve en unique exemplaire et ils sont fragiles.

- **Diffuser les manuscrits :**

La numérisation permet alors, par cette entremise, l'accès multiple et simultané aux documents numérisés

- **Facilité l'accès :**

La numérisation offre une grande souplesse de diffusion, les documents deviennent alors consultables facilement.

- **Sauvegarder les manuscrits :**

La numérisation n'est pas immuable à 100 %, toutefois, la conservation des manuscrits sous forme numérique assure une certaine longévité.

Grace à la numérisation, on évite la manipulation physique des manuscrits, leur conservation est ainsi améliorée. De plus, il est possible de les reproduire et de les diffuser sans dégradation. [Karinne2003]

- **Participation à la recherche d'information** :

La numérisation permet l'échange de connaissances et de compétences professionnelles.

Grâce à la liaison entre les fichiers images et les fichiers textes s'y rapportant, il est plus facile et plus rapide de choisir les fichiers nécessaires à une utilisation ultérieure. Il sera alors possible d'utiliser les images des manuscrit numérisés pour une multitude de projets, tels que: sites Internet, publications, création de base d'image …. etc

3 .Les images

Avant d'entrer dans le vif de sujet, il est important de comprendre la nature des objets que nous allons manipuler; nous nous intéressons à la notion d'image ? Qu'est-ce qu'une image ?

3.1 Notion d'image

Une image désigne un élément visuel. Il peut représenter des éléments concrets du monde extérieur ou des éléments abstraits. Elle peut être représentée par la peinture, la sculpture, le dessin, la photographie ... [Jerôme 2005].

L'image est un moyen de communication universel dont la richesse du contenu permet aux êtres humains de tout âge et de toute culture de se comprendre.

C'est aussi le moyen le plus efficace pour communiquer, chacun peut analyser l'image à sa manière, pour en dégager une impression et d'en extraire des informations précises.

Informatiquement, une image sera une représentation numérique en mémoire d'un sujet imprimé sur une rétine artificielle. On parle alors, d'image numérique.

3.2 Image numérique

Une image numérique est un ensemble structuré d'informations enregistrée sur un support numérique, qui, après affichage sur l'écran, ont une signification pour l'œil humain.

3.2.1 Catégories d'images numériques

Les images numériques appartiennent à deux grandes familles : bitmap et vectorielle.

- **Images vectorielles :**

Les images vectorielles sont des représentations d'entités géométriques telles qu'un cercle, un rectangle ou un segment. Ceux-ci sont représentés par des équations mathématiques.

Les formes plus complexes sont subdivisées en segments de droite ou de courbe, ce type d'images peut être manipulé avec beaucoup de facilité.

- **Images bitmap (matricielle) :**

Le non " bitmap" vient du forma BMP (bitmap) qui est le format standard de la prise numérique d'une image. Dans la suite de ce mémoire nous nous intéressons aux images matricielles.

Une image bitmap est constituée d'un ensemble de points appelé pixels qui se réduit à une matrice, Dans cette structure visuelle, chaque pixel est représenté par une couleur ou un niveau de gris.

Donc une image bitmaps est une matrice bidimensionnelle de valeurs numériques f (i,j) où :

i, j : coordonnées cartésiennes d'un pixel de l'image. f (i, j) : niveau de gris ou couleur d'un pixel.

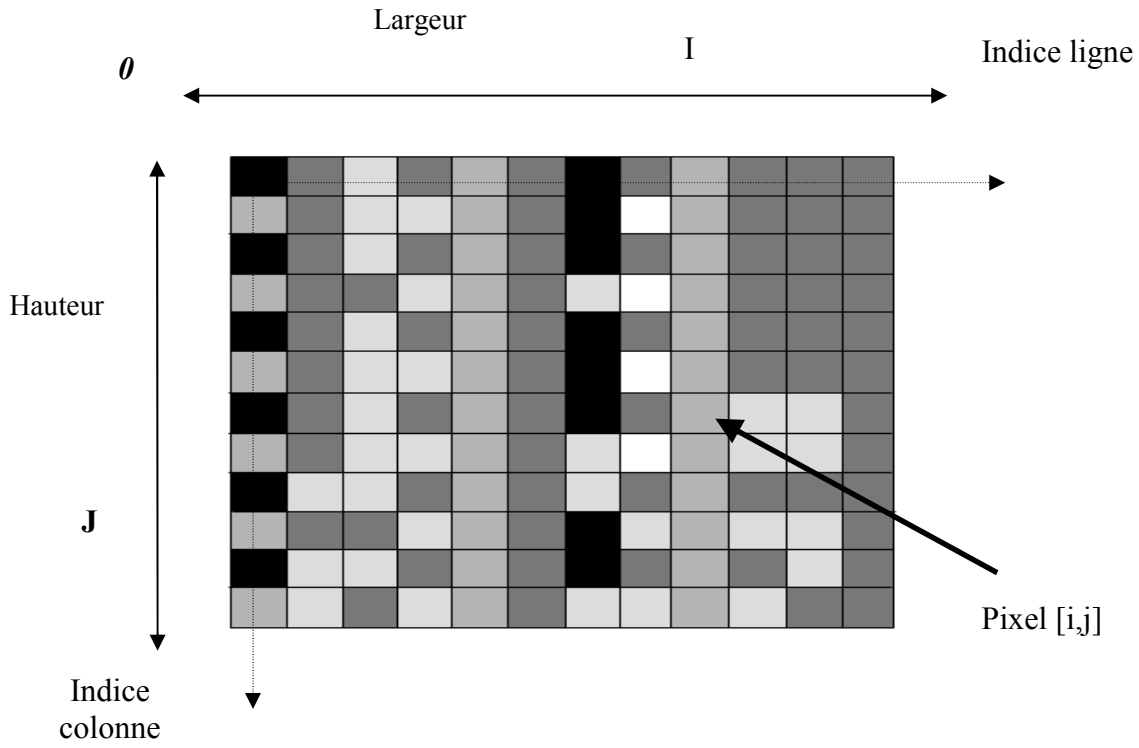


Fig.1.Présentation d'une image bitmap

3.2.2 Définition d'un pixel

Un pixel est une Contraction de l'expression anglaise " Picture éléments ": éléments d'image, le pixel est le plus petit point de l'image, c'est la plus petite unité d'information dans une image, Chaque pixel a une valeur. Cette valeur est stockée dans une case. Elle est codée sur un certain nombre de bits déterminant la couleur ou l'intensité du pixel (suite de 0 et de 1). Elle dépond du codage utilisé, c'est-à-dire le nombre de bits pour encoder chaque pixel.

3.2.3 Le codage de l'image

- **Codage d'une image en noir et blanc :**

Chaque pixel est soit noir, soit blanc. Il faut un bit pour coder un pixel (0 pour noir, 1 pour blanc). Exemple : Une image de 10000 pixels codée, occupe 10000 bits en mémoire.

- **Codage d'une image en niveaux de gris :**

Chaque pixel est codé sur 8 bits. On a alors 256 possibilités (256 niveaux de gris : noir, gris foncé,…, blanc). L'image de 10 000 pixels occupe alors 10 000 octets en mémoire.

- **Codage d'une image en couleurs :**

La gamme de couleurs possibles est très vaste, il est appelé «espace de couleurs». Il en existe plusieurs types d'encodage de la couleur, parmi lesquels nous citons :

1. Codage RGB (Reed, Green, Blue):

Elle consiste à utiliser 24 bits pour chaque point (pixel) de l'image. Huit bits sont employés pour décrire la composante rouge (R), huit pour le vert (G) et huit pour le bleu (B). Il est ainsi possible de représenter environ 16 millions possibilités théoriques de couleurs différentes. Par ailleurs, l'œil humain n'est pas capable de distinguer autant de couleurs.

On peut aussi coder RGV sur 8 bits, dans ce cas on attache une palette de 256 couleurs à l'image. Ces 256 couleurs sont choisies parmi les 16 millions de couleurs de la palette RVB, ainsi pour chaque image, le programme recherche les 256 couleurs les plus pertinentes.

Chaque code (de 0 à 255) désigne une couleur, l'image occupe donc moins de place en mémoire qu'avec un codage 24 bit.

3.2.4 Caractéristique d'une image

- **Définition :** On appelle définition le nombre de pixels constituant l'image, c'est-à-dire sa « dimension informatique » (le nombre de colonnes de l'image que multiplie son nombre de lignes). Une image possédant 640 pixels en largeur et 480 en hauteur aura une définition de 640 pixels par 480, notée *640x480*.
- **La résolution:** C'est la clarté ou la finesse de détails atteinte par un moniteur dans la production d'images. Elle détermine le nombre de pixels par unité de surface, exprimé en *points par pouce* (**PPP**, en anglais **DPI** pour *Dots Per Inch*). Plus la valeur de la résolution est élevée, plus les pixels sont petits.

  Exemple: 18 pixels par pouce soit environ 7 pixels par cm (1pouce =2,54 cm)
- **Poids d'une image :** Le poids d'une image est l'espace mémoire occupé par celle-ci qui égal à son nombre de pixels total que multiplie le poids de chacun de ces éléments (qui est le nombre de bits pour coder un pixel).

Pour connaître le total des pixels, cela revient à calculer le nombre de cases du tableau, soit la hauteur de celui-ci que multiplie sa largeur.

Exemple : Voici le calcul pour une image 640x480 en utilisant un codage RVB (24 bits):

- Nombre de pixels :

  640 x 480 = 307200
- Le poids de l'image est ainsi égal à :

  307200 x 24 = 7372800 bit= 7200 ko
- **Bruit :**

Un bruit (parasite) dans une image est considéré comme un phénomène de brusque variation de l'intensité d'un pixel par rapport à ses voisins.

3.2.5 Compression d'images numérique

Les documents numérisés en mode image occupent beaucoup de place. On diminue le volume de stockage initial en le compactant. Les images sont alors codées selon des procédés de compression.

La compression d'image consiste à réduire la taille physique de blocs d'informations constituant l'image en la codant dans un autre format utilisant des algorithmes de compression.

3.2.6 Format de fichier image numériques

Un Format d'image permet de décrire une image avec une structure de donnée et en lui appliquant un algorithme de compression. Suivant l'algorithme utilisé, on obtient des formats d'image différents. Parmi les plus utilisé on trouve :

1. Le format PNG (*Portable Network Graphics*,)

Il est un format compressé et très adapté au transfert sur un réseau. L'avantage du format PNG, est d'une part de pouvoir être rendu transparent et d'autre part, la compression proposée est sans perte d'information, autrement dit, il est possible de restaurer exactement l'image d'origine (avant compression).

- **Le format PNG existe en 2 versions :**

-PNG 8 bits : Elle est limitée à 256 couleurs.

-PNG 24 bits : Elle est plus évoluée. Elle supporte plusieurs millions de couleurs et permet de créer de beaux effets de transparence.

2. Le format GIF (Graphique Interchange Format)

Il est un format d'image très utilisé qui permet un bon affichage mais qui est limité à 256 couleurs. Cette limitation dans la gamme des couleurs permet d'obtenir une taille de fichier relativement petite. Il utilise une technique de compression sans perte comme le PNG et permet l'utilisation de couleurs transparentes. Il est couramment utilisé pour les barres d'outils et les icônes

3. Le format JPEG (Joint Photographic Experts Group)

Les images de type JPEG ont généralement l'extension ".jpg", mais aussi parfois ".jpeg". Ce format est très répandu et connu comme étant le format le plus puissant qui est utilisé pour diffuser des images de grande qualité sur le Web. Il permet aussi bien de traiter les images en couleur qu'en niveaux de gris. Il est aussi compressé mais sa compression est une compression avec pertes , notamment de qualité c'est-à-dire dire qui ne conserve pas la qualité.

4. Le format TIFF (Tagged Image File Format)

Le format TIF permet de stocker des images de taille importante, sans perte de qualité et sans aucune perte d'information.TIFF est le format le plus répandu dans le domaine de la numérisation et le stockage d'image (création et conservation) car il indépendant des plates formes ou des périphériques utilisés, c'est-à-dire qu'il peut être lu par la plupart des plates-formes. Le seul inconvénient du format TIFF réside dans sa taille d'image, qui est grande, donc ce format n'est pas adapter au transfert d'image sur le web.

4. Conclusion

Les manuscrits arabes anciens représentent des œuvres précieuses qui peuvent intéresser différents types de populations. La numérisation de ce type de documents est motivée par plusieurs objectifs dont les principaux est la préservation et l'accès distant.

1. Introduction

Tout processus de recherche de document se compose essentiellement d'un processus d'indexation et d'un processus d'interrogation. Un système de recherche du document est capable de donner des réponses pertinentes à une requête et pour se faire une phase d'indexation soit susceptible de fournir une représentation sémantique des documents.

La recherche documentaire classique se base essentiellement sur l'extraction des informations qui seront considérées comme des mots clés de la recherche.

Dans ce chapitre nous essayons de donner d'une part un aperçu sur la recherche d'information d'autre part sur l'indexation en recherche d'image. Partant des définitions qu'on retrouve dans la littérature, nous présenterons en détail les étapes de processus de l'indexation dans ce qui suit.

2. La recherche d'information

D'après [Chevalier2011] la **Recherche d'Information** peut se définir comme :

-Action, méthodes et procédures ayant pour objet d'extraire d'un ensemble de documents les informations voulues. Dans un sens plus large, toute opération (ou ensemble d'opérations) ayant pour objet la recherche, la collecte et l'exploitation d'informations en réponse à une question sur un sujet précis.

Une notion proche de la **Recherche d'Information (RI)** est la **Recherche Documentaire** étant définie dans [Chevalier2011] par :

- Action, méthodes et procédures ayant pour objet de retrouver dans des fonds documentaires les références des documents pertinents. Ensemble des techniques et modalités permettant de sélectionner l'information dans un fonds documentaire structuré en fonction de critères de recherches propres à l'utilisateur[Chevalier2011] .

2 .2 Définition d'un système de recherche d'information

Les systèmes de recherche d'information (appelés aussi moteurs de recherche dans le contexte du Web), notés par SRI, sont l'un des moyens les plus utilises par les utilisateurs pour trouver de l'information [Chevalier2011].

Un système de recherche d'information (SRI) est un ensemble de logiciel assurant l'ensemble des fonctions nécessaires à la recherche d'information qui, lorsqu'il est soumis à certains stimuli, fournit des réponses. La qualité d'un SRI se mesure à la qualité des réponses qu'il est capable de fournir.

Nous présentons. Dans cette section le processus sur lequel reposent les systèmes de recherche d'information[Chevalier2011] .

Ce processus est appelé processus en U et peut être schématise comme le présente la Figure.

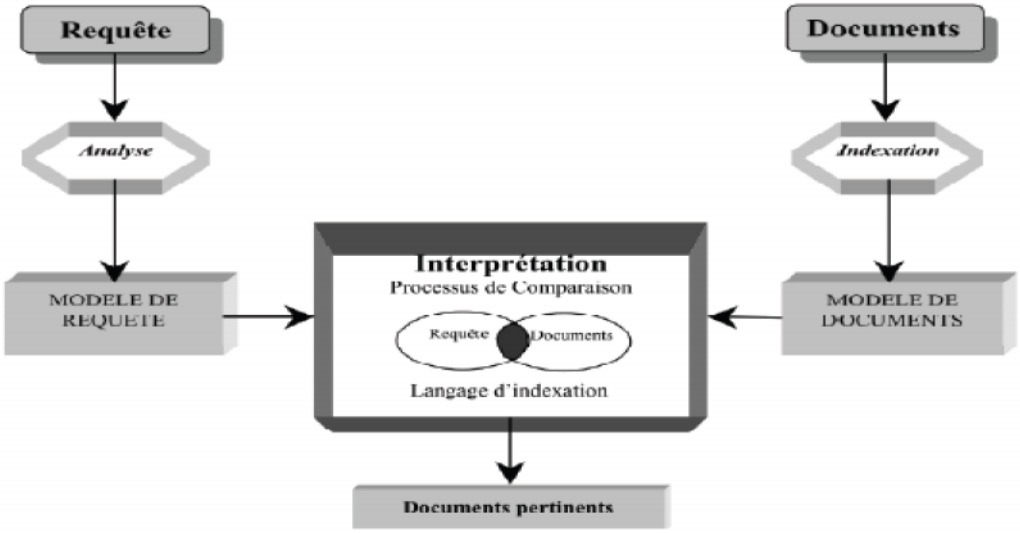


Fig.2. Le processus en U de la recherche d'information[Dahak2006].

Dans cette figure, cinq principales étapes se dégagent :

- La phase de traitement des documents : qui constituent le fonds documentaire du système.

Des qu'un document est ajoute au fonds documentaire, ce dernier est analysé afin de construire l'équivalent d'une notice bibliographique. Le système identifie donc et extrait les caractéristiques les plus importantes des documents permettant ainsi de répondre aux besoins d'un utilisateur

- **phase d'indexation**. Cette phase permet à l'utilisateur de retrouver des documents par rapport a leur contenu.

- La phase formulation des besoins d'un utilisateur : qui consiste à traduire le besoin mental en une requête qui sera soumise au système.

- La phase de recherche : qui consiste à identifier les documents répondants aux besoins des utilisateurs.

- La phase présentation des résultats : Le système retourne les documents jugés comme les plus pertinents à la demande de l'utilisateur [Chevalier2011].

2.3 Exemple d'un système de recherche d'information : moteur de recherche d'image.

Un moteur de recherche aide l'utilisateur à trouver ce qu'il cherche bien que ce dernier formule ses requêtes de façon pauvre. Un moteur de recherche ne peut travailler qu'à partir de l'information dont il a connaissance, c'est-à-dire l'ensemble des métadonnées qui auront été fournies sur les images. Les moteurs de recherche opèrent en deux étapes. Dans un premier temps, une phase d'enregistrement des images, des métas donnés et de structuration de la

base. Généralement l'utilisateur n'intervient pas à cette étape. La seconde phase, interactive, correspond au cycle requête / recherche des images / présentation des résultats.

Les premiers systèmes de recherche d'images utilisaient des mots-clés (descripteurs) associés aux images pour les caractériser. Grâce à cette association de mots-clés, il suffit d'utiliser les méthodes basées sur le texte pour retrouver les images contenant les mots-clés. Plusieurs moteurs de recherche proposent ces méthodes pour la recherche d'images [Jerome2005].

Les moteurs de recherche sont aujourd'hui perçus comme une application autonome, mais ils vont de plus en plus tendre à s'intégrer dans les différents environnements applicatifs. Typiquement, pour les bases de données d'images professionnelles telles qu'une base d'image des manuscrits, on se dirige vers une interconnexion beaucoup plus forte entre les moteurs de recherche et les applications d'enrichissement du contenu [Jerome2005].

2.4 Composant de système de recherche d'information

2.4.1. Le processus d'indexation

Le processus d'indexation est mis en œuvre afin de rendre la recherche acceptable, il convient d'effectuer une étape primordiale sur la base documentaire[Dahak2006]. Cette phase consiste à analyse chaque document de la collection, car dans un SRI un document est considéré comme un support qui véhicule de l'information ; La phase d'indexation permet donc, de capturer cette information afin de créer un ensemble de mots-clés qui correspond au contenu sémantique du document, la représentation de ce dernier est appelé un index de document *[Mehadi2010]*.(Ce point ça sera détaillé par la suite.) indexation manuelle, se différencient par l'agent mettant en œuvre le processus de l'indexation des documents :

- ✓ Dans le cas d'une indexation manuelle, c'est le documentaliste qui effectue l'analyse des documents, pour identifier le contenu d'un document.
- ✓ Dans le cas d'une indexation automatique, c'est le système de recherche d'information qui génère les index des documents. L'indexation assistée par fois l'utilisateur afin de valider ou corriger une représentation proposé par le système.

2.4.2 .Le processus d'interrogation

C'est la phase d'interaction entre le système et l'utilisateur. Ce dernier exprime son besoin d'information via un langage de requête que le système va se charger de traduire. Cette traduction se fait selon le modèle de requête et a pour but de comprendre les besoins de l'utilisateur et de les exprimer dans un formalisme similaire a celui mis en œuvre lors de l'indexation des documents *[Dahak2006][Mehadi2010]*.

IL faut, établir une comparaison sémantique entre les concepts figurants dans un document et ceux figurants dans la requête. La comparaison entre requête et document aboutit rarement à des équivalences strictes, mais plutôt à des équivalences partielles : le document correspond à une partie seulement de la requête*[Dahak2006][Mehadi2010]*. Le premier document de la liste renvoyée par le système est celui qui est considéré par le système comme le plus pertinent, c'est-à-dire celui qui répond le mieux a la requête.

2.5 Les modèles classiques de la RI

La figure suivante illustre les modèles de la RI

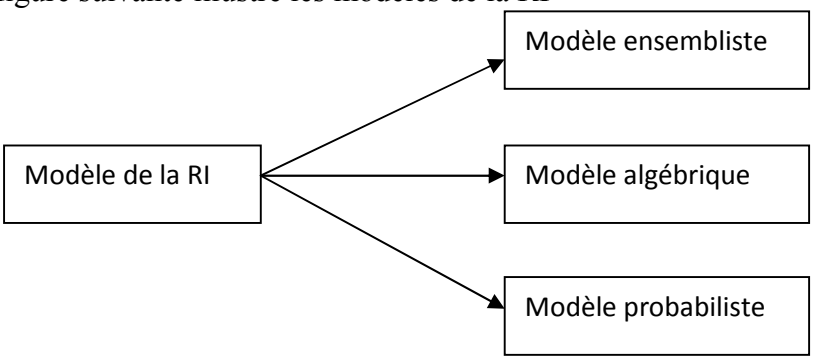


Fig.3.Modèles de la RI

2.5.1. Les modèles basés sur la théorie des ensembles

Ce modèles se base sur la théorie des ensembles, ainsi une requête est représentée par un ensemble de termes séparés par des opérateurs logiques (OR, AND, NOT). Le document est, quand à lui, représenté par une liste de mots-clé. Ces modèles permettent d'effectuer des opérations d'union, d'intersection et de différence lors de l'interrogation. Le modèle le plus connu et le plus simple de cette catégorie est le modèle booléen. On y retrouve également le modèle booléen étendu et le modèle flou [Dahak2006].

- Le modèle booléen

Le modèle booléen représente les documents et les requêtes par une liste de descripteurs reliés entre eux

par des opérateurs logiques (« et », « ou », « et non »)(« and », « or » et « not »). Par exemple, une image I peut être représentée par une liste de descripteurs reliés par l'opérateur « et » (□) :

$I = di_1 \square di_2 \square \ldots . \square di_n$

Plus généralement, une requête Q est représentée par une liste de descripteurs reliés par l'ensemble des opérateurs logiques :

$Q = dq_1 \square \ dq_2 \square \quad \neg dq_n$

La correspondance entre les images et la requête se traduit par une implication logique :

$I \rightarrow Q$

Ainsi, le système effectue une classification binaire en deux classes, positive et négative, correspondant respectivement aux images qui satisfont la requête et à celles qui ne la satisfont pas.

Le modèle est simple, implémenté efficacement dans le monde des documents textuels. Il présente des difficultés importantes d'adaptation au monde des documents visuels. En particulier, les descripteurs sont généralement des valeurs numériques et une simple comparaison par égalité n'a plus grand sens.

Par suite, la séparation rigide entre documents retrouvés et documents écartés est beaucoup trop stricte[Idrissi2008].

Le modèle booléen est le plus simple et le plus répandu des modèles de RI. C'est également le premier à s'imposer dans le domaine de la recherche d'information. Il s'appuie sur l'utilisation des opérateurs logiques manipulés grâce à l'algèbre de Boole. Il consiste à formuler une question avec une liste de termes séparés par des opérateurs logiques (ET, OU, NON), et à rechercher les documents correspondant à cette requête [Idrissi2008].

2.5.2.Les modèles algébriques

Les modèles algébriques regroupent tous les modèles de RI qui utilisent une représentation vectorielle des documents et des requêtes [Piwowarski2003] dans lesquels, la pertinence d'un document vis-à-vis d'une requête est définie par des mesures de distance dans un espace vectoriel. La similarité est calculée algébriquement en se basant sur la représentation du document et de la requête. Le représentant le plus connu de cette catégorie est le modèle vectoriel [Dahak2006][Idrissi2008].

- Modèle vectoriel

Dans le modèle vectoriel, les images I de la base et l'image requête Q sont représentées par un vecteur de descripteurs dans un espace d'attributs à n dimensions[Dahak2006]. Les éléments de ce vecteur représentent les pondérations des descripteurs utilisés [Idrissi2008] :

$$I = (wi_1, wi_2, \dots . . wi_n)$$

$$I = (wq_1, wq_2, \dots , wq_n)$$

La correspondance entre les images de la base et la requête s'effectue en terme d'une fonction de similarité (ou de dissimilarité) entre leurs vecteurs.

2.5.3. Les modèles probabilistes

Dans le modèle probabiliste, une probabilité de pertinence de l'image, en réponse à la requête, est attribuée à chacun des descripteurs[Idrissi2008]. Cela suppose qu'il existe un sous-ensemble d'images R pertinentes que l'utilisateur veut retrouver parmi celles disponibles, les autres images NR étant considérées comme non-pertinentes.

Si $P(R|\vec{Q})$ est la probabilité que l'image I soit pertinente pour la requête Q et si $P(NR|\vec{Q})$ est la probabilité que l'image I ne soit pas pertinente pour la requête Q, alors la similarité entre l'image I et la requête Q est exprimée par[Idrissi2008]:

$$\sin(I, Q) = \frac{P(R|\vec{Q})}{P(NR|\vec{Q})}$$

La figure suivante représente les principaux modèle de la RI

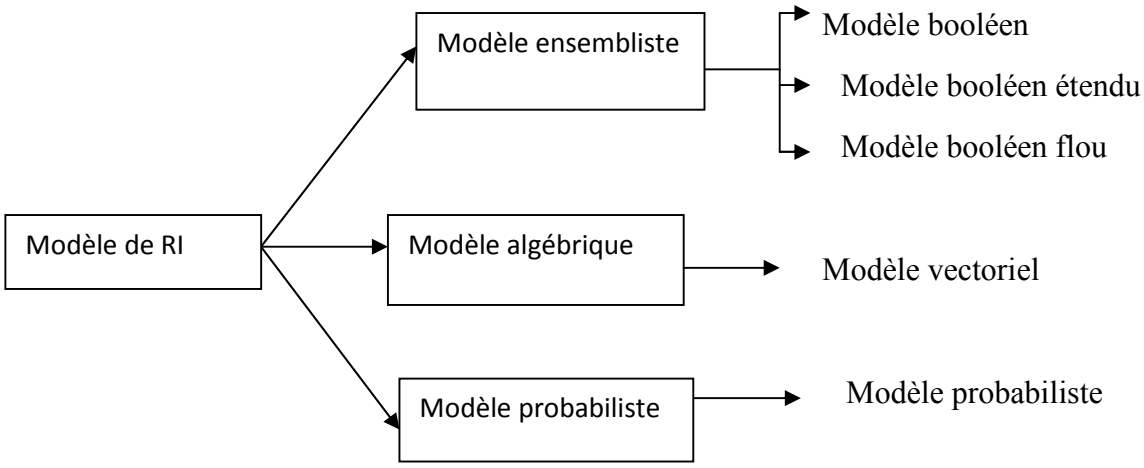


Fig.4. Principaux modèles de RI

2.6.Recherche d'images

Une métadonnée est une donnée à propos d'une donnée traitant ainsi des informations sur le contenu du document pour faciliter sa recherche, sa localisation et son identification. Les métadonnées d'une image est un ensemble structuré d'informations la concernant qui peuvent être son nom, sa date, sa qualité et la description de ce qu'elle contient… [Ouadah2008][Mokdem2010].

Il existe deux principes de recherche d'images :

-La recherche d'images par le contenu : L'architecture générale des systèmes de recherche d'images par le contenu est basée sur un calcul de similarité. Plus précisément, on classe les images résultat présentées à l'utilisateur selon la distance entre le vecteur descripteur de l'image exemple et les vecteurs descripteurs des images de la base[Jerôme2005].

- La recherche d'images par métadonnées :

Les premiers systèmes de recherche d'images utilisaient des mots-clés associés aux images pour les caractériser. Grâce à cette association de mots-clés, il suffit d'utiliser les méthodes basées sur le texte pour retrouver les images contenant les mots-clés. Plusieurs moteurs de recherche proposent ces méthodes de recherches d'images [Jerôme2005].

Notre objectif est l'accès aux images de manuscrits numérisés en utilisant les annotations des utilisateurs. Ainsi, notre intérêt se porte sur cette méthode qui fera office d'une large description dans ce qui suit.

2.6.1 Intérêt de la recherche d'image

Le but de la recherche d'images est de retrouver une (ou plusieurs) image(s) parmi une base d'images pour répondre à une requête d'un utilisateur. La recherche sur les bases d'images nécessite de posséder une base d'images.

3. Aperçus de l'indexation

3.1. Définition de l'indexe

Etant donnés une base de documents volumineuse à traiter, il est généralement admis que la recherche d'information doit s'appuyer sur une description de ces derniers qui résume les informations susceptibles d'être référencées par un utilisateur. Cette représentation est dénommée un index.

Une définition plus pratique est donnée dans [Dahak2006]qui définit un index comme une liste de mots retenus avec pour chacun d'eux les documents dans lesquels ils apparaissent.

Définit dans [Abbaci03] par : « Un index est une structure qui permet d'associer à chaque terme d'indexation, la liste des documents qui contiennent ce terme ».

3.2. Définition de l'indexation

Le but de l'indexation est de créer une représentation permettant de repérer et retrouver facilement l'information dans un ensemble de documents.

Plusieurs définitions de l'indexation sont présentes dans la littérature. On en retient quelques unes: On retrouve dans [Lancaster1998] cette définition : « Le but principal de l'indexation est de construire des représentations d'éléments publiés sous une forme adaptée pour le stockage dans tout type de base de données ».

[Paradis1996] précise qu'un index doit être une représentation synthétique de l'information et doit mettre en évidence la sémantique de cette dernière en vue d'une requête.

[Zargayouna2005] « L'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document"

La phase d'indexation est extrêmement importante dans la recherche d'image. Elle a un impact direct sur la pertinence des images recherchées. En revanche, si une image est mal indexée ;il risque de devenir inaccessible pour l'utilisateur.

3.3. Les formes d'un index

Les index peuvent prendre différentes formes allant de mots simples à des structures sémantiques plus complexes impliquant plusieurs concepts et relations. Les descripteurs représentent l'information atomique d'un index. Ils sont censés indiquer de quoi parle le document. On parle aussi d'unités élémentaires (en anglais "***Tokens***")[Dahak2006].

Le but étant de les choisir de manière à ce que l'index (qui réduit la représentation) perde le moins d'informations sémantiques possible. Habituellement les descripteurs sont des mots du document, des n-grammes ou des concepts.

Les mots du document : toute chaîne de caractères compris entre deux séparateurs (espace, virgule…), au niveau de l'indexation, on peut extraire les mots tels qu'ils sont présentés dans

le document. De même pour des fins de normalisation, on peut effectuer certaines transformations sur ce mot.

Les concepts : termes ou mots-clés: il s'agit d'expressions (pouvant contenir un ou plusieurs mots). Ces concepts sont le plus souvent entrés manuellement (cas de l'indexation manuelle, ou semi-automatique) et peuvent être écrits de manière libre par un utilisateur, ou, ce qui est souvent le cas, doivent être choisis parmi une liste de concepts (on parle alors de vocabulaire contrôlé).

Les N-grammes : Il s'agit d'une représentation originale d'un texte en séquences de N caractères consécutifs. On trouve des utilisations de bigrammes et trigrammes dans la recherche documentaire (ils permettent de reconnaître des mots de manière approximative et ainsi de corriger des flexions de mots ou même des fautes de frappe ou d'orthographe).

Le tableau ci-dessous, représente un exemple de ces différentes formes d'index :

Table		**Modèles de la recherche d'information structurée**
Mot	**Origine**	Modèles, de, la, recherche, d, information, structurée
	Lemme	Modèle, de, la, recherche, information, structure
	Racine	Modèl, d, l, recherch, inform, structur
Concept		R.I
Bigramme		Mo, od, de, el, le, es, s_, _d, de, e_, _l, la, …., ée

Fig.5.Différentes formes d'index

Dans la pratique, la forme la plus utilisée est la représentation par mots-clés. L'extraction automatique des concepts d'une collection de documents est souvent une entreprise très délicate, nécessite l'utilisation des techniques du traitement automatique du langage naturel ; vu que ces derniers sont directement liés à la langue utilisée.

L'indexation à base de concepts est souvent manuelle ou semi-automatique. Elle est inadaptée à de larges collections de documents. Les n-grammes, quant à eux, sont indépendants de la langue utilisée. Mais nécessitent, par contre, un espace mémoire assez important et plusieurs traitements doivent être effectués sur la requête dans un processus de recherche d'information. Ils sont plus utilisés pour la classification des documents que pour la recherche d'information [Jalam2002].

Dans ce qui suit nous considérons uniquement les descripteurs sous forme de termes (mots-clés).

3.4 Processus d'indexation

Le processus de l'indexation effectue le transfert de l'information contenue dans la description d'un document vers une représentation traitable par système informatique [Calabretto2003].

A partir d'une collection de documents, le processus d'indexation nous renvoie une liste d'index structurée. On utilise ce résultat, le plus souvent, pour effectuer des recherches d'informations. Mais, il peut également servir à comparer et classifier des documents, proposer des mots-clés, faire une synthèse automatique de documents, calculer des co-occurrences de termes…

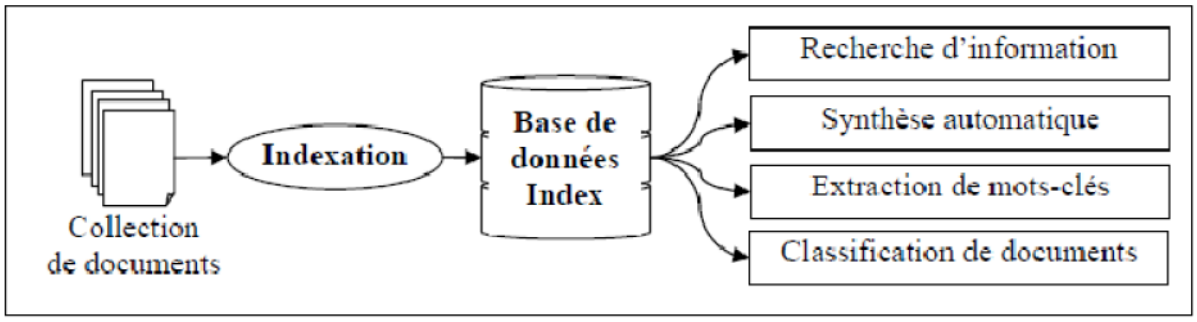
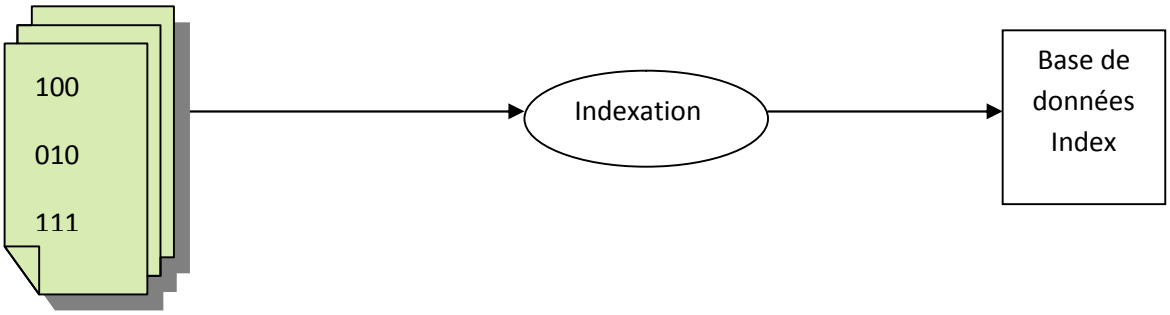


Fig.6.Processus d'indexation



Collection de document

Fig.7. processus d'indexation

3.4.1. Etape du processus d'indexation

Le processus d'indexation se compose de plusieurs étapes que nous avons schématisé ci-dessous

La figue suivante illustre les étapes du processus de l'indexation :

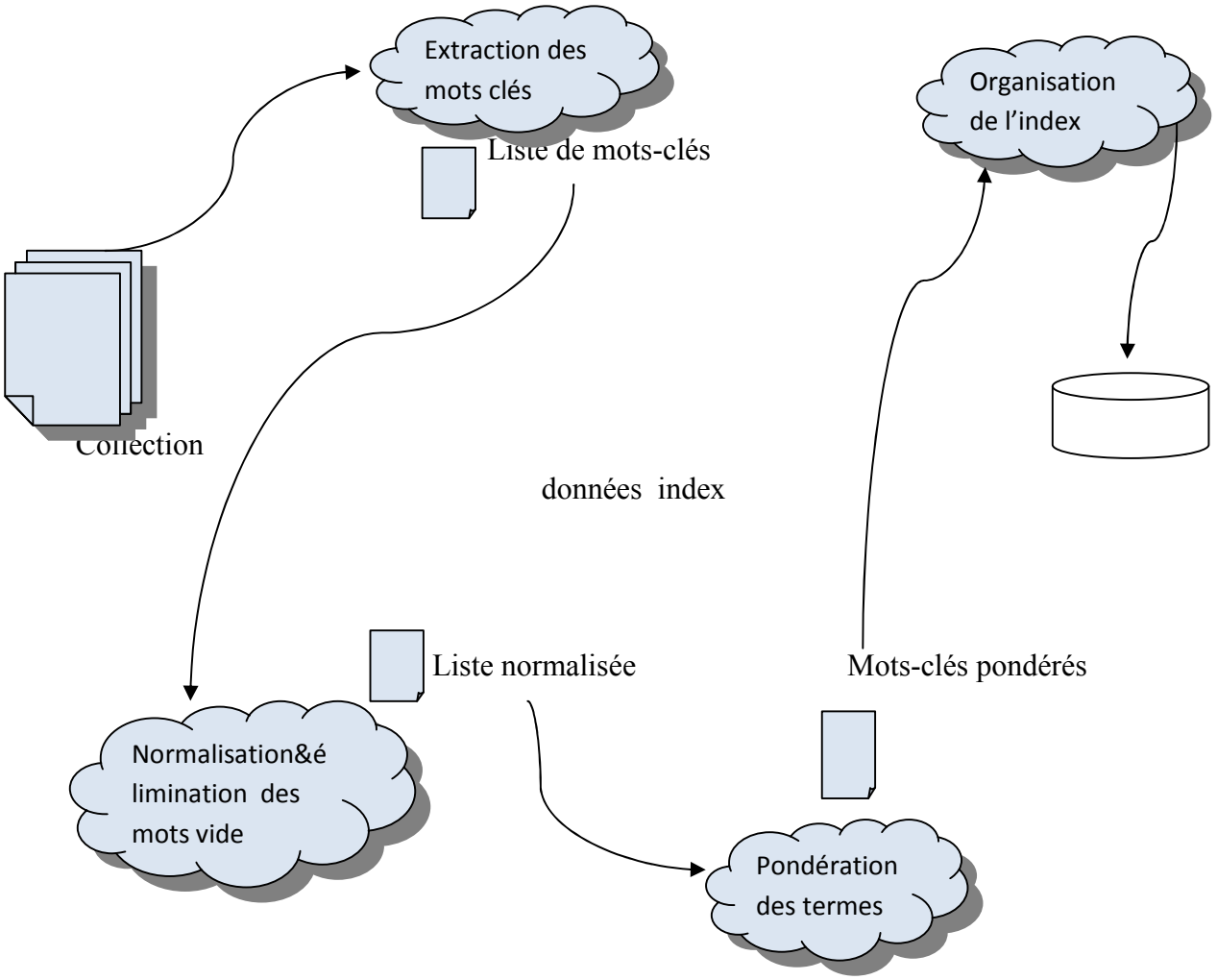


Fig.8. Processus d'indexation des documents XML

En général, L'indexation consiste d'abord à extraire les termes du document dans une certaine limite de fréquences. Les termes sélectionnés sont ensuite lemmatisés, ce qui permet une certaine normalisation des termes dans l'index.

3.4.1.1. L'extraction des mots-clés du document

Appelée ***Tokenization*** en anglais, l'extraction des mots-clés est une étape importante et constituera la base de tout le reste du processus d'indexation. Il faut que cette phase soit d'une qualité maximale[Dahak2006].

L'utilisation d'une extraction automatique des mots-clés ou l'utilisation d'une liste de mots-clés prédéfinie, détermine le type d'indexation. Orientée document dans le premier cas et orientée requête dans le second.

3.4.1.2.La normalisation des mots-clés du document :

Ce traitement consiste à retrouver pour un mot sa forme normalisée (masculin pour les noms, l'infinitif pour les verbes,..).Ainsi dans l'indexe ne sont conservées que les formes normalisées, ce qui offre un gain de place appréciable, mais surtout, si le même traitement est effectué sur la requête, cela permet d'être souple et rapide dans la recherche. Par exemple si l'utilisateur effectue une recherche à l'aide d'un descripteur (mot-clé) d, toutes les images de la base qui ont ce descripteur seront listées à l'utilisateur[Dahak2006].

3.4.1.3. L'élimination des mots vides

Cette étape revêt une importance certaine dans la mesure où elle constitue un facteur d'une grande influence dans la précision de la recherche. Le fait de ne pas éliminer les mots vides provoque inévitablement du bruit. L'élimination des mots vides qui sont des mots du langage courant et qui ne contiennent pas beaucoup d'information sémantique doit se faire aussi bien à l'indexation qu'à l'interrogation (élimination des mots vides(pronoms personnels, prépositions,...) de la requête)[Dahak2006].

3.4.1.4.Pondération des mots-clés

Après avoir choisi les termes (descripteur) d'indexation, il est possible de leur attribuer un poids permettant de préciser l'importance d'un terme par rapport à l'autre dans la description d'un document [Dahak2006]. Cette pondération dépend fortement du modèle de recherche utilisé pour l'appariement image-requête, la plupart des méthodes utilisées repose sur la combinaison de ces trois facteurs[Mehadi2010] :

- Un facteur de pondération locale qui mesure l'importance d'un terme dans une image
- Un facteur de pondération globale qui quantifie l'importance d'un terme dans le corpus (l'importance d'un terme au sien d'une base d'image)
- Un facteur de normalisation qui prend en considération la taille d'une image

Il existe plusieurs techniques de pondération des termes dont les trois les plus importantes sont décrites ci_dessous[Dahak2006]:

a) La fréquence d'occurrences

Cette fréquence prend en compte le nombre d'apparitions d'un mot dans un document (fréquent, rare). Les mots les moins fréquents du corpus qui quant à eux peuvent par exemple être issus de fautes d'orthographe ou de l'utilisation d'un vocabulaire trop spécifique à quelques documents du corpus. Par contre, un mot qui apparaît beaucoup dans un document possède certainement une information forte sur la sémantique du document[Dahak2006].

Un mot qui est fréquent c'est un mot informatif dans un document s'il y est présent souvent mais qu'il n'est pas présent trop souvent dans les autres documents du corpus [Dahak2006]. Cette loi illustre donc, l'importance d'un terme en fonction de sa fréquence dans un corpus

b) La valeur de discrimination

Par "discrimination", on réfère au fait qu'un terme distingue bien un document des autres documents. C'est-à-dire, un terme qui a une valeur de discrimination élevée doit apparaître seulement pour un petit nombre de documents. Un terme qui apparaît dans tous les documents n'est pas discriminant[Dahak2006].

c) Tf*Idf

Le terme tf*idf est très connu dans le milieu de la recherche d'information. Cela désigne un ensemble de schémas de pondération de termes. ***tf*** signifie « *term frequency* » et ***idf*** « *inverted document frequency* ». Par ***tf***, on désigne une mesure qui a un rapport avec l'importance d'un terme pour un document[Dahak2006]. En général, cette valeur est déterminée par la fréquence du terme dans le document.

3.5 Structure d'index

Pour répondre rapidement aux requêtes, des structures de stockage sont nécessaire pour mémoriser les informations sélectionnées lors de processus d'indexation [Sauvagnat2005].

Les structures d'index sont :

3.5.1 Les fichiers Séquentiels (*Sequential files*)

Un fichier séquentiel est le moyen le plus simple de stocker un fichier de données puisque l'on stocke les enregistrements les uns à la suite des autres dans leur ordre d'insertion [Dahak2006].

Jusqu'au milieu des années 70, tous les systèmes textuels utilisaient les fichiers séquentiels du fait de l'utilisation de bandes magnétiques. Une requête sur un document consistait alors à parcourir toute la bande jusqu'à ce que l'on trouve le bon document, d'où une lenteur certaine du système malgré l'optimisation des algorithmes de recherche. L'amélioration du processus était bloqué par le matériel (bandes magnétiques) ce qui changea radicalement avec l'arrivée d'un nouveau système de stockage plus rapide : le disque dur [Dahak2006].

3.5.2 Fichiers inversés (Inverted files)

Cette structure est à la base de tous les systèmes de recherche. Un système à base de fichiers inversés contient trois composants principaux :

Un dictionnaire **:** Le fichier dictionnaire contient tous les mots ou groupes nominaux spécifiques pouvant servir de mots-clés pour l'indexation et la recherche dans l'ensemble des fichiers à traiter. A chaque entrée du dictionnaire est associé le nombre de fois où l'entrée apparaît dans l'ensemble documentaire.

Un fichier de hachage : Ce fichier contient pour chaque entrée du dictionnaire une liste décrivant dans quel fichier apparaît cette entrée. Cette méthode permet de restreindre l'étude sur les fichiers qui nous intéressent et pas les autres. A signaler que dans certains cas, la position dans le fichier est aussi stockée.

Les fichiers de données : Qui représentent les documents du corpus à indexer. La figure suivante illustre la représentation d'un système à base de fichiers inversés.

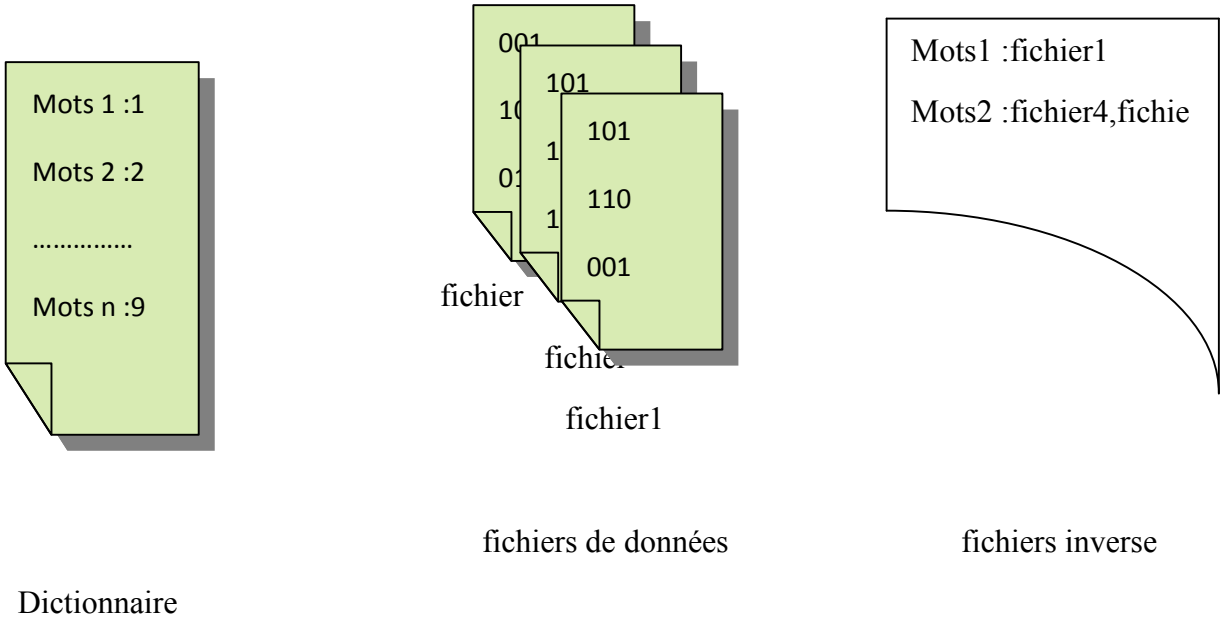


Fig.9. Système à base de fichiers inversés

Ce système de fichiers, s'il est rapide pour trouver un résultat, consomme énormément de place de stockage, les fichiers index étant parfois aussi gros que les fichiers de données, surtout dans le cas où les positions où apparaissent les mots clés dans les fichiers sont stockées. Les mises à jour sont aussi coûteuses puisqu'il faut refaire l'index à chaque ajout.

3.5.3 Matrice document/terme

[weigel2002] présente la matrice document/terme comme une matrice à deux dimensins(m,n).

pour chaque terme d'indexation ti ($1 \leq i \leq m$), chaque document dj ($1 \leq j \leq n$) prend la valeur 1 si une occurrence de ti apparait dans dj et 0 sinon*[Mehadi2010]*.

	&1	&2	&3	&4
XML	1	0	1	0
JAVA	0	1	1	1
SAX	0	0	0	1
DOM	0	0	1	0

Fig.10. la matrice document/terme

4. Indexation de documents semi-structurés

Indexation de document semi-structuré doit impérativement passer par deux étapes :

4.1. Indexation du contenu

Dans le cas de la recherche d'information, le processus d'indexation consiste à extraire de descripteurs (information, mots-clés) des documents**:** [Dahak2006]. L'indexation de document ne doit pas se faire indépendamment de la structure du document car un terme figure à un emplacement précis dans le document.

4.2. Indexation de la structure

La structure d'un document peut être indexée selon des granularités variées, à vrai dire toutes les informations utilisées dans le processus d'indexation ne sont pas forcement structurées [Fellag2006]. En effet nous avons trois approches pour l'indexation d'information structurelle :

4.2.1. Indexation basée sur les champs

Dans cette méthode d'indexation, le document est représenté comme un ensemble de champs et du contenu associé à chaque champ [Dahak2006]. Pour permettre une recherche restreinte à certains champs, les termes de l'index sont construits en combinant le nom du champ avec les termes du contenu.

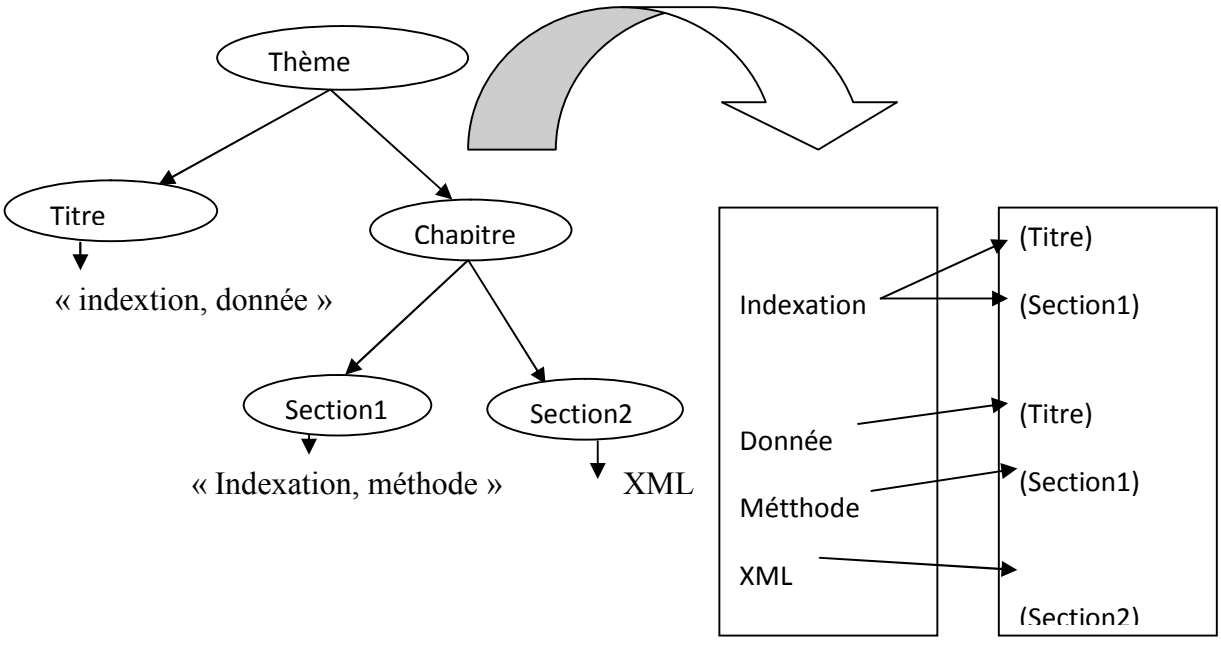


Fig.11. indexation basée sur les champs

4.2.2. Indexation basés sur le chemin

Les index basés sur les chemins utilisent les chemins du document comme unité structurelle de base, ils stockent les chemins menant aux contenus en commençant par la racine du document[Dahak2006].

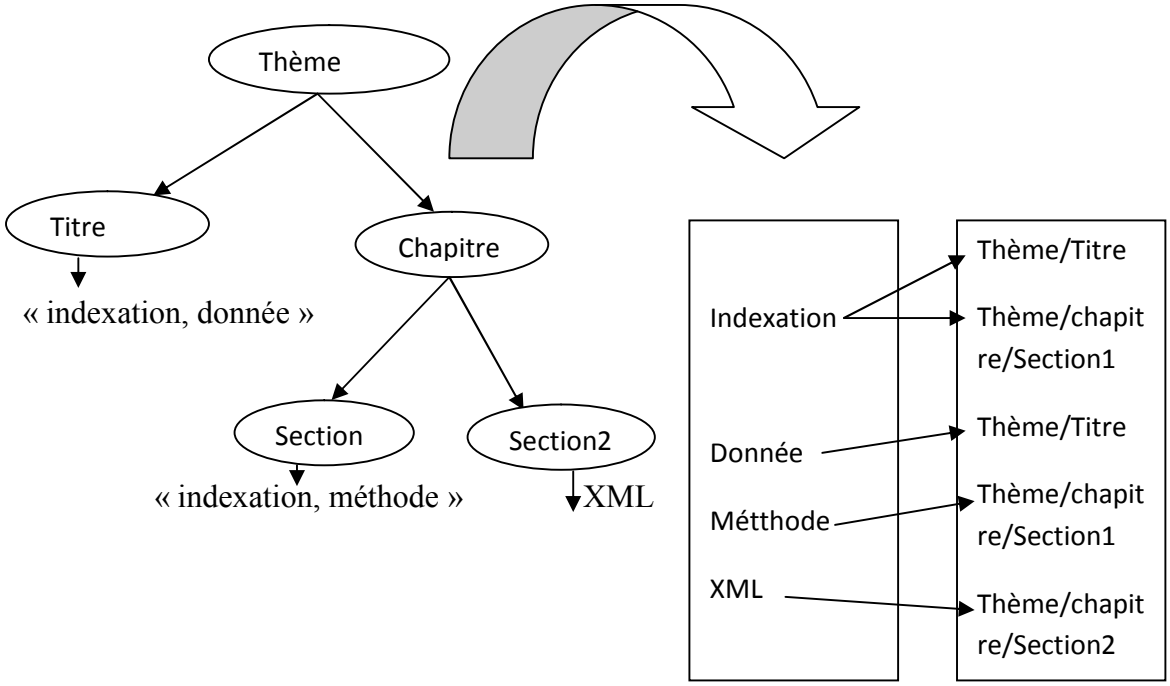


Fig.12. exemple d'indexation basée sur les chemins

Dans cette approche cependant, il devient difficile de retrouver les relations ancêtres-descendants entre les différents nœuds des documents. Les approches d'indexations basées sur des arbres le permettent quant à elles.

4.2.3. Indexation basés sur l'arbre

Dans cette approche, les nœuds de l'arbre du document sont numérotés dans l'index de sorte de pouvoir reconstruire la structure arborescente du document, plusieurs méthodes d'indexation d'identification structurelle des nœuds ont été proposées dans la littérature parmi elle la numérotation de Dietz[Mehadi2010].

La technique de la numérotation de Deitz est la première méthode qui a employé l'ordre de parcours d'arbre pour déterminer le rapport de descendance ou d'ascendance (ancêtre) entre n'importe quelle paire de nœuds d'arbre. La proposition de Deitz était[Mehadi2010]: pour deux donnés x et y d'un arbre T, x est un ancêtre de y si et seulement si x apparait avant u dans le parcours pré ordre de T et après y dans le parcours post ordre de T.

Dans la figure suivante, un arbre XML dont les nœuds sont annotés par la numérotation de Deitz est monté. Chaque nœud est marqué avec une paire de pré ordre et de post ordre. Dans l'arbre, nous pouvons dire que le nœud <4,3>est un ancêtre du nœud<5,2>, parce que le nœud <4,3> vient avant <5,2>, dans le pré ordre et après le nœud <5,2> dans le post ordre.

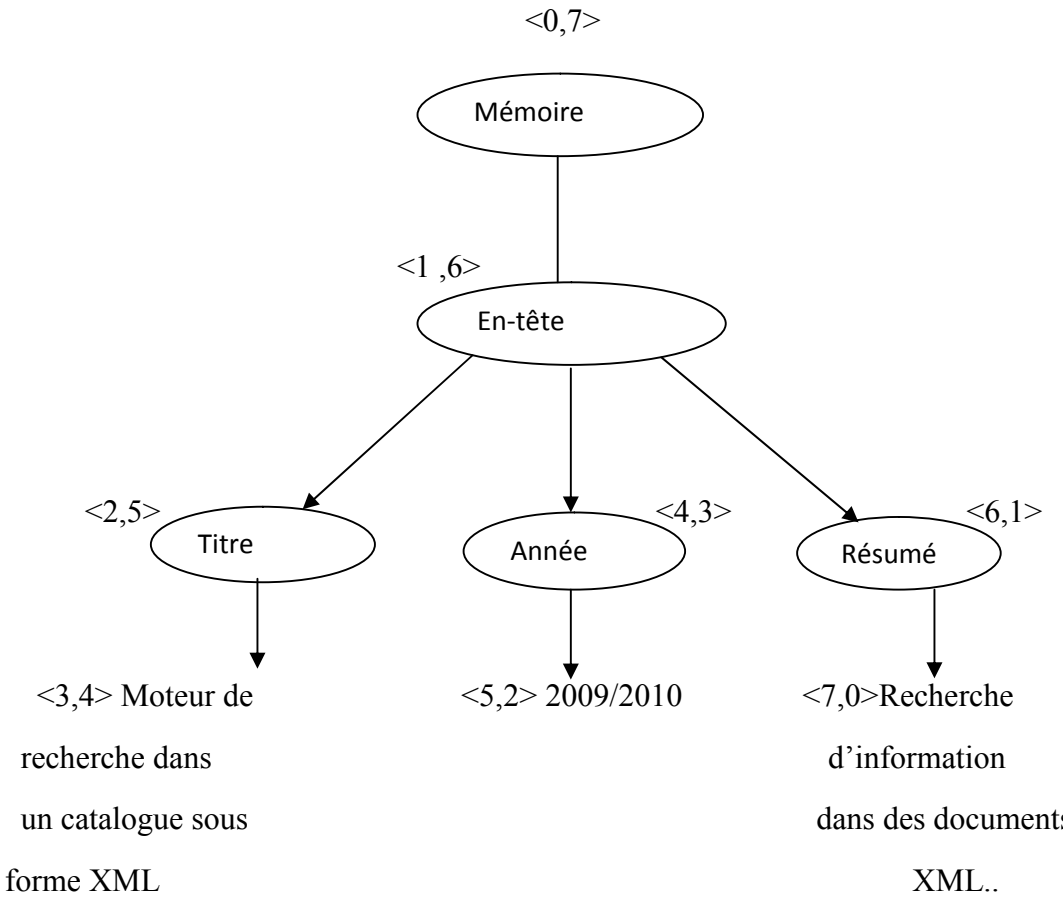


Fig.13.Représentation d'indexation basée sur les arbres

5.Conclusion

Un processus d'indexation de la recherche d'information structurée doit concilier le contenu et la structure d'un document pour pouvoir répondre efficacement aux besoins en information des utilisateurs.

Dans ce chapitre nous avons étudié le processus de l'indexation pour la recherche documentaire qui doit concilier le contenu et la structure d'un document pour pouvoir répondre efficacement aux besoins des utilisateurs. Nous avons présenté les techniques de création d'index.

Une bonne organisation de l'index permet une accessibilité plus ponctuelle dans la structure de document, et donc une réponse pertinente exhaustive à une demande utilisateur.

L'indexation peut ce faire sur des termes extrait d'un document, ces derniers sont des composants d'une notice écrite par l'utilisateur ; comment peut en définir l'annotation des manuscrits arabe (notice) ? Quelles est ces caractéristiques ?c'est ce que nous allons traiter dans le chapitre suivant.

1. Introduction

La numérisation des manuscrits arabes anciens sous forme d'images donne une version très fidèle à l'original, mais le mode image ne permet pas l'accès direct à l'information ce qui rend ces images inaccessibles. Pour rendre le document numérisé plus utile, nous devons donc aller plus loin que le mode image et permettre d'accéder à ce que l'on appelle le mode texte ce qui nous oriente vers la recherche d'une méthode qui se base sur du texte pour indexer les images des manuscrits. Ainsi, le catalogage se propose comme une solution. En effet, constitué d'un ensemble de notices descriptives de manuscrits sur les quelle il est possible d'effectuer des recherches. Cette solution pose divers problèmes d'accès aux manuscrits arabes numérisés de fait qu'il donne une description très pauvre de contenu des manuscrits. Pour enrichir le catalogue nous proposons une solution qui consiste à annoter les images numérique des manuscrits pour associer à certaines de leurs parties des mots clés, des notes ou plus généralement, une représentation textuelle de leur contenu. L'indexation peut alors s'effectuer facilement sur ces représentations textuelles.

2. Méthodologie d'accès aux manuscrits Arabes numérisés

2.1 Le catalogue traditionnel des manuscrits (manuelle)

Un catalogue se définit comme une liste d'éléments appelés « notice ».Chaque notice Donne une idée générale du manuscrit, elle informe de ses richesses, de ses particularités.

L'absence de normalisation des notices est le principal obstacle de leur mise en œuvre. En effet plusieurs tentatives de catalogage ont été réalisées sans pour autant produire un modèle unique de catalogage, sur lequel pourrait s'appuyer l'ensemble des catalogueurs de manuscrits arabes [Soualah 2008]. Chaque catalogue est fait en fonction de l'intérêt et de la spécialité des catalogueurs [Kaileh 2004].

Une notice peut contenir que les informations bibliographiques minimales nécessaires, à savoir: l'auteur, le titre, la date de la copie… . Comme elle peut donner une description plus exhaustives et contient des informations plus détaillées sur l'apparence extérieure des manuscrits et sur la manière suivant laquelle ils ont été faits à savoir la décoration, l'encre, la reliure….

Le catalogue traditionnel des manuscrits a été sous forme d'un livre imprimé ayant ses propres caractéristiques et sa propre présentation [Kaileh 2004]. Il est propre à une bibliothèque précise et demande le déplacement des utilisateurs pour pouvoir y accéder.

2.2 Le catalogue informatisé (électronique)

L'informatisation des catalogues est une technologie qui a amélioré l'utilisation du contenu des anciens catalogues traditionnels [Kaileh 2004]. En utilisons la technique de numérisation, un manuscrit est considéré comme un ensemble d'images numériques,

Ces images sont accessibles grâce au catalogue électronique qui est considéré comme un outil d'indexation pour le manuscrit.

Le catalogue électronique est la première initiative menée par des institutions qui conservent les documents manuscrits, leur but est de faciliter l'accès distant à leurs fonds. Avec l'arrivé du l'Internet, le catalogue électronique peut être interrogé par des internautes du monde entier. Ce qui permet un accès universel au document manuscrit.

3. Mode d'accès aux manuscrits en utilisant le catalogue électronique

Rappelons que l'accès aux manuscrits numérisés se fait par l'intermédiaire du catalogue. L'accès au catalogue électronique des manuscrits arabes numérisés se fait selon deux modes :

3.1 Accès par entrée standard (vedette titre, vedette auteur,…..etc) : ce mode permet à l'utilisateur d'accéder au catalogue par un ou plusieurs critères bien définis tels que, l'auteur, le copiste, le titre du manuscrit, …etc [Soualah 2010].

3.2 Accès libre : Dans ce mode d'accès, l'utilisateur ne connait pas à priori le contenu du catalogue de ce fait, il est autorisé à émettre sa requête sans aucune restriction

[Soualah 2010].

4. Rôle du catalogage pour l'accès aux manuscrits numérisés

Un document non catalogué est un document inaccessible, il peut être considéré comme un document mort [Soualah 2008]. Le catalogue est destiné à faciliter la recherche, en effet il permet de repérer la disponibilité d'un document dans un premier temps, puis l'identification Et la localisation de manuscrit.

5. Métadonnées

Une métadonnée est une donnée à propos d'une donnée, traitant ainsi des informations sur le contenu d'un document pour faciliter sa recherche, sa localisation et son identification.

Les métadonnées d'un manuscrit est un ensemble structuré d'informations le concernant. Ces informations peuvent être l'auteur, le titre, le copiste … [El bannay 2009]. D'une manière générale les métadonnées d'un manuscrit recouvre tous les éléments traditionnels d'un catalogue.

6. Encodage du catalogue des manuscrits

Les formats informatiques actuellement utilisés pour la description des manuscrits médiévaux reposent sur le langage informatique XML (eXtensible Markup Language) (voir annexe), qui permet de définir des formats de documents, et peut s'appliquer à toutes les règles de catalogage.

XML présente de nombreux intérêts pour le codage des catalogues des manuscrits, que nous décrivons dans ce qui suit :

- XML peut être utilisé pour la description de tous types de document.
- XML rend possible des échanges de données entre des systèmes d'informations hétérogènes.
- XML n'est lié ni à un système d'exploitation ni à une famille de logiciel. Par défaut, les documents XML sont des documents texte dont le jeu de caractère est l'Unicode, c'est-à-dire un jeu de caractères très complet prenant en charge de nombreux systèmes d'écriture dont l'arabe.
- Il permet la consultation des données sous forme statique, après transformation (exemple : HTML).

Beaucoup de format informatique sont utilisés pour décrire des manuscrits, et il se base tous sur XML, nous citons entre autre : EAD (Encoded Archival Description), TEI (Text Encoding Initiative), MASTER (Manuscript Access through Standards for Electronic Records)…. etc

7. Choix du format d'encodage

Pour le codage des manuscrits arabes anciens, nous avons choisis d'étudier le format XML/TEI qui est un modèle de document spécifiquement utilisé dans le domaine des sciences humaines. Elle se présente sous une forme souple et adaptable. En effet, la structure modulaire de la TEI permet à l'utilisateur de choisir les outils qui lui conviennent.

La TEI fournit une bonne base à la description des manuscrits et s'adapte donc bien à nos besoins .La section qui suit permet décrire le formalisme XML/TEI

8. Le format d'encodage XML/TEI

TEI (Text Encoding Initiative) est un projet universitaire pluridisciplinaire visant à uniformiser autant que possible le codage de documents manuscrit ou imprimé en vue de leur échange et de leur publication en ligne ou hors ligne.

8.1 Définition

TEI est une norme de codage de textes qui repose sur XML. Il s'agit d'un format de codage dit «structuré», TEI permet une une représentation informatique d'un document (manuscrits ou imprimés) L'objectif du codage d'un document au moyen de balises est de rendre ses caractéristiques exploitable et traitable par des programmes informatiques. De plus il offre la possibilité d'usages multilingues [Pierrat 1999].

9. Structure globale d'un fichier XML/TEI

Comme nous l'avons définit auparavant, le catalogue est constitué d'un ensemble de notices, lesquelles sont formées à leur tour, d'éléments descriptifs d'un manuscrit. Avec la TEI, une notice à une structure de balisage décrivant le manuscrit.

L'encodage global d'un document manuscrit se présente ainsi :

<TEI>

<teiHeader>

<!-- ... métadonnées décrivant le manuscrit -->

</teiHeader>

<facsimile>

<!-- ... métadonnées décrivant les images numériques -->

</facsimile>

<text>

<!-- (optionnel) représente une éventuelle transcription -->

</text>

</TEI>

[Burnard 2008]

Voici la description des ces différents éléments :

9.1 <teiHeader>

L'élément principal de <teiHeader> est <msDescription>.

<msDescription> décrit un seul manuscrit à la fois, il est formé de six éléments dont chaque élément est formé d'un ensemble de sous éléments, qui à leur tour peuvent être constitués d'autres éléments. L'ensemble, donnant à la TEI une structure hiérarchique en forme arborescente.

- <msIdentifier> : permet l'identification d'un manuscrit particulier. Il contient un ensemble de sous éléments tels que :<idno> (identifiant), <country> (pays), <region> (région)…
- <msContents> : Décrit le contenu intellectuel du manuscrit. En trouve <author> (l'auteur), <title>(titre), <incipit>(début du manuscrit), <explicit>(fin du manuscrit), <textLang>(langue)...
- <physDesc> : Il permet la description physique du manuscrit, En trouve <support>, <handNote>, <decoDesc>….

- <history> : Il regroupe les éléments décrivant l'historique du manuscrit, Il contient les éléments suivants : <origin>, <provenance>, <acquisition>.
- <additional> : Il regroupe les informations additionnelles telles que la bibliographie du manuscrit, les informations administratives,

<adminInfo> (info administrative), <listbibl> : (liste des citations bibliographiques)

- <msPart> : Dans le cas où un manuscrit serait formé par l'assemblage de plusieurs manuscrits, cette élément permettrait de décrire chaque manuscrit comme un objet à part et totalement indépendant des autres volumes du manuscrit.

9. 2 <facsimile>

L'élément <facsimile> est employé pour décrire les images numériques d'un manuscrit .Il contient, les deux éléments < surface > et <graphique> pour chacune des images.

L'élément <graphique> peut avoir un ou plusieurs éléments <zone > qui permet de spécifier les coordonnées d'une partie de l'image numérique.

Exemple :

```
<facsimile>
    <surface ulx="00" uly="00" lrx="400" lry="280">
        <graphic url="graphic.png "/>
            <zone ulx="20" uly="40" lrx="500" lry="321"> </zone>
            <zone ulx="10" uly="00" lrx="45" lry="51"> </zone>
    </surface>
</facsimile>
```

9.3 <text>

<text> est un élément optionnel, il peut contenir un ou plusieurs éléments <seg>. Ces derniers contiendront des éléments du texte de manuscrit.

10. Limite de catalogue pour l'accès au manuscrit numérisé

Le catalogage est un outil qui facilite l'accès au manuscrit numérique, mais il présente quelques limites:

- Il est très sommaire car il ne décrit pas le manuscrit d'une manière efficace.
- Connaissance au préalable du contenu de catalogue c'est-à-dire l'utilisateur doit lancer sa requête par des mots clés qui ont été définit dans la catalogue à savoir le nom de l'auteur, titre du manuscrit, …etc. ce que qui n'est pas évident car le catalogage demeure la fonction du concepteur du système, pas celle d'utilisateur.

- le catalogue est statique, une fois construit pour le système ou qu'il est destiné il ne changera pas, on ne peut pas ajouter d'autre informations.

La solution d'accès aux manuscrits arabes numérisés en utilisant le catalogue qui décrit d'une manière sommaire le manuscrit a montré ses limites. En effet, le système reste insuffisant pour l'utilisateur, qui a besoin de Plus de liberté pour formulé sa requête.

Dans ce mémoire, notre but est d'enrichir le catalogue, et pour cela nous introduirons la notion d'annotions d'image numérique du manuscrit. En effet nous donnons à l'utilisateur, expert dans le domaine des manuscrits arabes (celui qui consulte l'image) la possibilité d'annoter l'image avec des remarques, description du contenu ou transcription du contenu, ce qui permet d'enrichir le catalogue et le système de recherche.

11. Les annotations

11.1 Définition

L'annotation est un objet attaché à un document, cette objet contient un bref commentaire sur le contenu de document et permet la mise en valeur de ce dernier et la sauvegarde des traces de l'activité de la recherche qui peut être exploité par des utilisateurs ultérieurs

[Sidhom 2011].

Une annotation regroupe essentiellement trois éléments principaux, à savoir :

- L'annotateur : la personne qui réalise l'annotation.
- Le document source concerné par l'annotation.
- L'objet annotation introduits sur le document.

L'objet annotation est caractérisé par son ancre et sa forme graphique affichée sur le document annoté, nous détaillons ci-dessous ces deux éléments :

1. L'ancre : est définie comme étant l'emplacement sur le document où l'annotation est placée ou la partie annotée du document, ces deux endroits sont souvent différents, pour les différencier, nous appelons une

- **Ancre physique** le lieu où est placée l'annotation sur le document.
- **l'ancre sémantique** le contenu annoté, ce point va être détaillé par la suite.

2. La forme graphique : L'annotation peut prendre plusieurs formes, les formes les plus utilisées sont les marques graphiques. Parmi ces formes nous pouvons citer le rectangle, la note marginale, la flèche, et l'astérisque.

11.2 L'activité de l'annotation

L'activité d'annotation représente le processus qui permet de poser l'objet d'annotation sur le document.

L'activité de l'annotation peut être décomposée en trois sous processus :

1. Choisir l'ancre et la forme de l'annotation.
2. Spécifier le document à annoter.
3. Choisir la cible de l'annotation (l'emplacement du document à annoter) [Ouadah 2008].

11.3 Intérêt de l'annotation

- introduire des éléments d'évaluation sur le document.
- permettre une prise de vue indépendante de celle de l'auteur.
- fournir une traçabilité d'exploitation du document.
- accumuler des commentaires explicites sur le contenu.
- favoriser le raisonnement critique.
- partager l'information.
- faciliter la compréhension et la relecture d'un document.

Les objectifs d'annotation ne sont pas toujours liés aux questions de collaboration et d'enrichissement, ils peuvent inclure d'autres visées pour la recherche d'information (indexation) [Sidhom 2011]. De ce fait L'annotation est considérée comme une nouvelle approche dans la conception d'un système de recherche d'information, son objectif est de faciliter l'appariement entre requêtes d'interrogation et sources documentaires dans un processus de recherche d'informations qui est l'objectif principale dans ce mémoire.

11.4 Annotation d'image

Annotation d'image est une association d'objet textuel à une image [Doumat 2008].

Exemple : association du contenu textuel secondaire à une portion de l'image d'un manuscrit.

11.5 Synthèse sur la sémantique de l'annotation

Dans cette partie nous mettons en évidence l'importance de la sémantique de l'annotation. En effet, l'étude de la sémantique d'une image ne peut pas être séparée de l'étude des besoins de l'utilisateur.

Plusieurs études ont montré l'importance de l'annotation pour ajouter des faits. Par conséquent, l'annotation d'une image constitue l'outil principal pour associer de la sémantique à une image. L'ajout de métadonnées sur une image permet d'enrichir sa description et permet la construction d'outils de consultation et de recherche plus performants.

Ainsi, une annotation peut être retrouvée sous forme :

- **Transcription :** une réécriture du contenu de l'image manuscrite sur l'annotation.
- **Avis personnelle :** représente une manière de voir, une opinion de l'utilisateur sur le contenu de l'image de manuscrit.
- **Remarque :** dégagement d'une idée après une observation ou une examinassions avec soin du contenu de l'image**.**

Ces deux derniers points sont des annotations sur le contenu et demande aux utilisateurs de consacrer du temps pour l'annotation.

12. Indexation des données

Comme nous l'avons définit et détaillé dans le deuxième chapitre l'indexation d'un document consiste à extraire des éléments censés représenter au mieux son contenu. En ce qui concerne l'indexation des manuscrits arabes, nous venons de voir les annotations comme instrument de recherche principale pour accès aux images numérisé des manuscrits.

L'instrument de recherche constitué en format XML rend l'indexation implicite. En effet, les informations à indexer sont contenues dans l'ensemble des fichiers XML/TEI représentant les images. De ce fait, l'index renfermera un ensemble de mots clés que l'utilisateur saisira manuellement lors du processus d'annotation.

13. Conclusion

La numérisation des manuscrits arabes n'aura aucun sens sans la mise en place d'un moyen efficace pour accéder à ces ressources.

Le catalogue se voit comme un moyen qui est représenté dans une structure spécifique pour rendre les images numérique des manuscrits arabes plus accessibles et plus manipulables par les utilisateurs mais il reste très sommaire.

Dans ce chapitre nous avons introduit les annotations comme outil d'accès aux images numériques des manuscrits arabes anciens, et cela pour palier au manque rencontré en utilisant le catalogue.

Les annotations permettent de stocker un ensemble varié d'information sur une image et jouent le rôle d'index pour permettre un accès facile et rapide à une image numérique d'un manuscrit.

1. Introduction

Avant toute réalisation d'une application informatique, il convient de suivre une démarche méthodologique et rigoureuse pour planifier et concevoir l'application, en mettant en évidence tous les objectifs tracés pour la bonne élaboration du projet souhaité.

Ce présent chapitre comporte une description de la solution d'annotation pour l'accès aux images des manuscrits arabes anciens et un modèle entité association qui représente notre base de données, de fait que l'indexation des documents XML/TEI représentants le catalogue des manuscrits arabes anciens se fera grâce à une base de données.

2. Description de la solution pour la recherche d'image des manuscrits par annotation

Dans ce qui suit nous présenterons la solution pour l'accès aux images numériques des manuscrits. Cette solution est basée sur les annotations des utilisateurs affin de contribuer à un raffinage de la recherche d'information. En fait, Nous nous intéressons aux annotations dans un processus de recherche d'image de manuscrits arabes numérisés. Principalement, notre objectif s'oriente vers l'exploitation des annotations pour déterminer des sources informationnelles pertinentes (images bien spécifiques) .L'annotation dans ce cas sert d'index aux images des manuscrit arabe anciens.

Le système que nous réaliserons à la possibilité d'annoter une image ou de la chercher en utilisant soit la langue française soit la langue arabe.

Tous d'abord nous associerons à chacune des images deux documents XML/TEI, soient « doc1 » et « doc2 ».

- Doc1 permet le couplage entre l'image et ces annotations textuelles écrites en français.
- Doc2 permet le couplage entre l'image et ces annotations textuelles écrites en arabe.

La procédure d'annotation commence par la sélection d'une zone d'image d'un manuscrit à l'aide de la souris qui permet de déposer une trace graphique autour d'une portion d'image. Ensuite, le système génère une interface d'annotation qui apparait sur l'image du manuscrit visité. L'utilisateur choisi la langue puis saisi l'annotation, qui peut être une bref description (remarque, avis personnel, ou une transcription).

Après validation de l'annotation des éléments XML (balises) sont automatiquement créés et ajoutées à la représentation interne du document XML spécifique de l'image courante et de la langue correspondante, c'est-à-dire si l'utilisateur annote en français par exemple les nouvelles balises sont insérées dans doc1.

L'utilisateur à la possibilité d'annoter une image plusieurs fois, ainsi l'image est interactivement découpée en plusieurs parties. Chacune d'elle est associée à des balises XML/TEI. En effet, les balises XML ajoutées à chaque annotation d'une partie de l'image permettent de stocker un ensemble varié d'informations sur lesquelles il est possible d'effectuer divers recherches.

3. Quel élément XML/TEI à utiliser ?

Les annotations d'une image du manuscrit sont décrites au moyen des éléments XML/TEI suivant :

D'une part la balise <zone> est utilisée pour représenter la partie géométrique de l'annotation. A chaque élément <zone> est associé un attribut id qui sert à différencier les parties sélectionnées de l'image, et quatre variables x, y, dx, dy qui permettent de stocker les coordonnées du rectangle de sélection.

Exemple : <zone id="1" x="407" y="25" dx="166" dy="76" />

D'autre part l'annotation textuelle est représenter au moyen de la balise <seg> qui a comme attribut facs qui sert d'identifiant et correspond à l'identifiant id de la zone sélectionnée. En effet les annotations textuelles sont liées à leurs parties géométriques.

Exemple : <seg facs="1">note</seg>

Rappelons ici que :

- L'élément <zone> est un sous élément de <facimile> utilisé pour décrire les images numérique des manuscrits.
- L'élément <seg> est un sous élément de <text> utilisé pour décrire le texte manuscrit.

Ces deux éléments cités sont générés automatiquement à chaque nouvelle annotation et exactement au moment de sa validation.

4. La représentation physique des documents XML

La représentation logique des documents se fera suivant une approche basée sur les bases de données relationnelle. Commençant par les règles de gestion engendrant le modèle entité association.

4.1 Le modèle entité association

Partant de système étudié nous pouvons établir les règles suivantes :

-La base documentaire est un corpus parallèle. Chaque document (fichier XML/TEI) existe en deux exemplaires (arabe, français).

-Un document est formé de nœuds (balise, feuilles).

-Une balise peut contenir des attributs.

-Un attribut est caractérisé par sa valeur, son nom et la balise à laquelle elle appartient.

-les nœuds feuilles sont composés de mots (termes).

-Un terme (après lemmatisation) est présent une seule fois dans le vecteur des termes.

Les règles de gestions ci-dessus permettent d'établir le modèle entité association suivant :

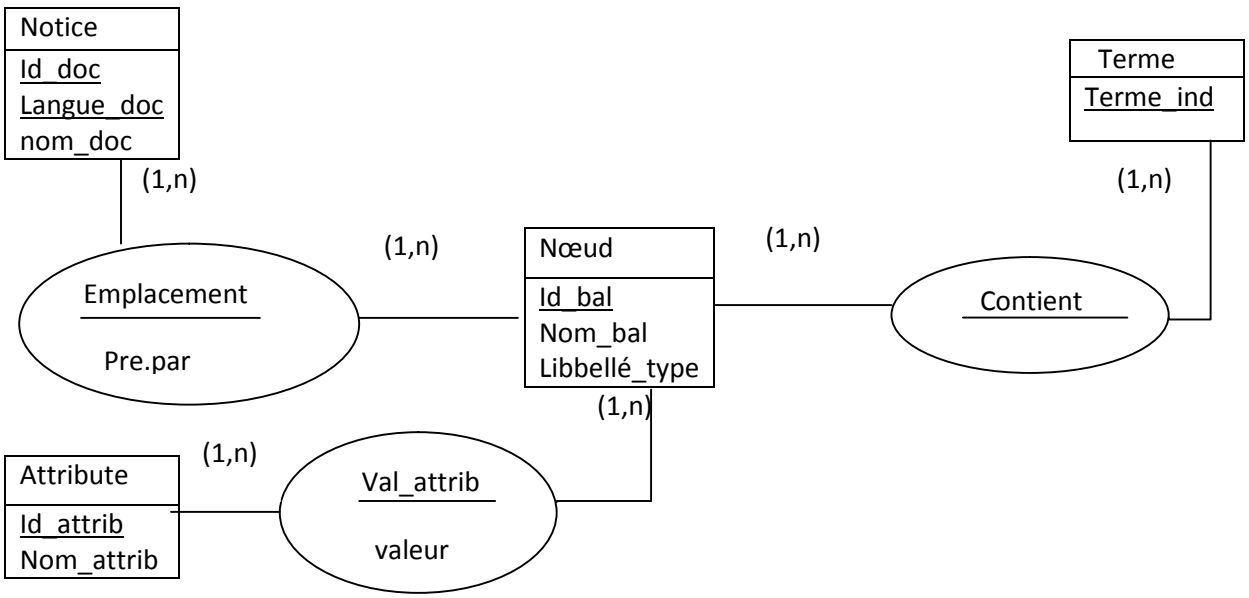


Fig.14. Modèle entité association

Pour l'implémenter nous passons du modèle entité association en modèle relationnelle en utilisant les règles qui permet ce passage ; on obtient ce tableau :

Table	**Role**	**Description**
Terme (terme_ind)	Contient l'ensemble des termes significatifs de la base documentaire	-terme_ind : représente un terme issus de l'indexation
Document (id_doc, nom_doc, langue_doc)	la table contenant les notices	-id_doc : identifiant de la notice -nom_doc : nom physique du document(cheminde stockage) -langue_doc : langue de la notice
Feuille (id_doc, langue_doc, id_noeud, val_pre, par)	La table contenant les nœuds feuilles de tous les documents de la base documentaire	-id_doc : identifiant du document contenant le nœud -id_noeud : identifiant de nœud -val_pre : valeur pré-ordre -val_post : valeur post-ordre Par : val_pre du nœud parent
Emplacement_terme(id_noeud, terme_ind)	Table de jointure entre les termes est leur emplacement	-id_noeud : identifiant de feuille contenant le terme -terme_ind : le terme indexé
Valeur_attribut(id_attr, id_noeud, valeur)	Table de jointure entre les balises et les valeurs d'attributs	-id_attr: identifiant de l'attribut le contenant -id_noeud : identifiant de nœud correspondant -valeur : valeur de l'attribut dans le nœud
Attribut (id_attr, nom_attr)	Table contenant les attributs des balises	-id_attr : identifiant de l'attribut -nom_attr : nom de l'attribut
Nœud (id_noeud, nom_noeud,	Table contenant les	-id_noeud : identifiant de nœud

libellé_typ)	nœuds des documents	correspondant -nom_neoud :nom de la balise -libellé_typ :type de neoud(balise ou feuille)

5. Conclusion

Le modèle conceptuel des données que nous avons conçu a permet une représentation schématique de l'ensemble des données de notre domaine d'étude ainsi que les relations qui existent entre ces données.

1. Introduction

Dans ce chapitre nous décrivons l'environnement et les outils qui ont servi au développement et à la réalisation de notre application, et nous terminons par la présentation de ses fonctionnalités à travers ses différentes interfaces.

2. Représentation du catalogue

Le système que nous réalisons manipule deux catalogues (arabes et français), chaque un est représenté dans un répertoire et contient l'ensemble des notices décrivant les images avec la langue correspondante, chaque notice est représenter dans un fichier XML/TEI et identifier par un numéro (<idno>).La figure suivante présente l'architecture le système de catalogage.

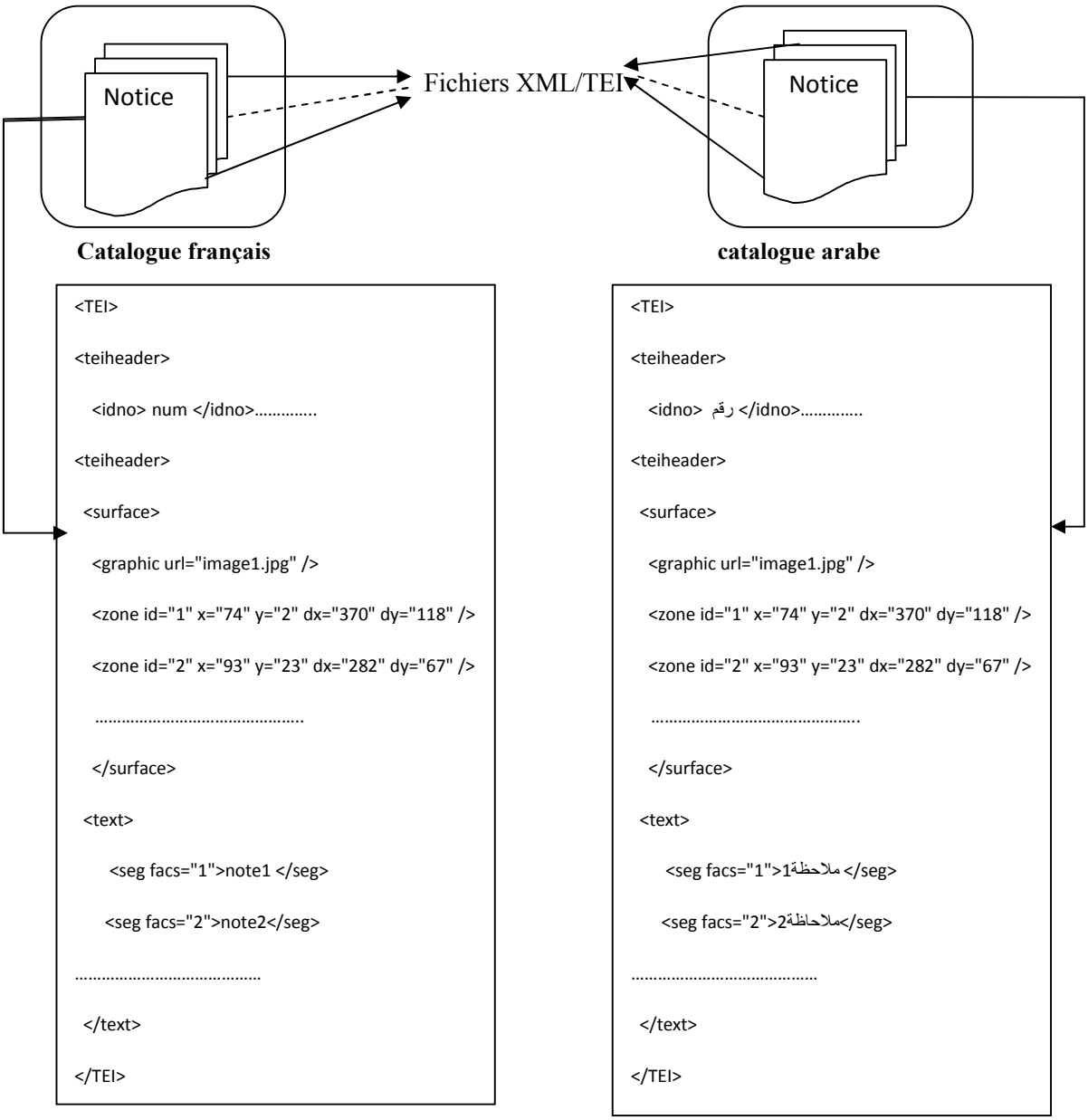


Figure 15 : Architecture générale du système de catalogage

3. Représentation des documents XML

L'indexation basée sur les arbres, est une approche d'indexation qui utilise la structure d'un graphe orienté (arbre) d'un document XML aussi bien pour l'identification des éléments constituent le document (attribut, balise et texte), que pour l'accès aux éléments.

3.1 La numérotation de Dietz

Un document XML est un arbre composé de nœuds. Afin de pouvoir naviguer et de déterminer les relations ancêtre-descendant dans l'arbre ainsi que de permettre l'accès rapide à un nœud nous avons utilisé la numérotation de Dietz.

Cette méthode est caractérisée par les deux valeurs : pré-ordre et post-ordre.

Les valeurs de pré-ordre et post-ordre sont assignées durant le parcours de l'arbre du fichier XML ce qui permet une gestion à la fois de la structure (la disposition des balises) et du contenu (informations contenues dans les nœuds feuilles).

La numérotation de Dietz procède à une identification des nœuds par un parcours pré-ordre et en post-ordre du fichier XML. Par ailleurs, le parcours post-ordre dans notre solution s'avère inutile du moment qu'un simple parcours pré-ordre suffit pour identifier un noeud, d'une manière univoque, l'ensemble des chemins du document XML et le parent de tout les nœuds.

Ainsi, nous n'avons retenu dans la procédure de numérotation que le parcours en pré-ordre. Ce dernier permet d'identifier à la fois, l'ensemble des nœuds et les parents associés. Ce qui optimise davantage notre solution.

La détermination du parent associé à chaque nœud, permet un accès rapide à la notice.

3.2 La représentation logique des documents XML

La méthode de numérotation utilisée permet donc de retrouver non seulement les nœuds des feuilles contenant les termes, mais aussi de déterminer rapidement les relations

ancêtre-descendant, ce qui permet un accès rapide aux nœuds.

Le modèle de représentation des documents permet la navigation dans la structure en arbre des documents XML, et la représentation du contenu et de la structure, afin de pouvoir les interroger et de récupérer les images qui répondent à la requête de l'utilisateur.

Soit le document XML/TEI suivante :

```
<TEI>
  <teiheader>
     <idno> num </idno>
  <teiheader>
  <surface>
     <zone id="1" x="74" y="2" dx="370" dy="118" />
   </surface>
   <text>
     <seg facs="1">note1 </seg>
   </text>
</TEI>
```

Voici l'arbre représentant le document XML/TEI avec les numérotations de Pré-ordre et de poste-ordre:

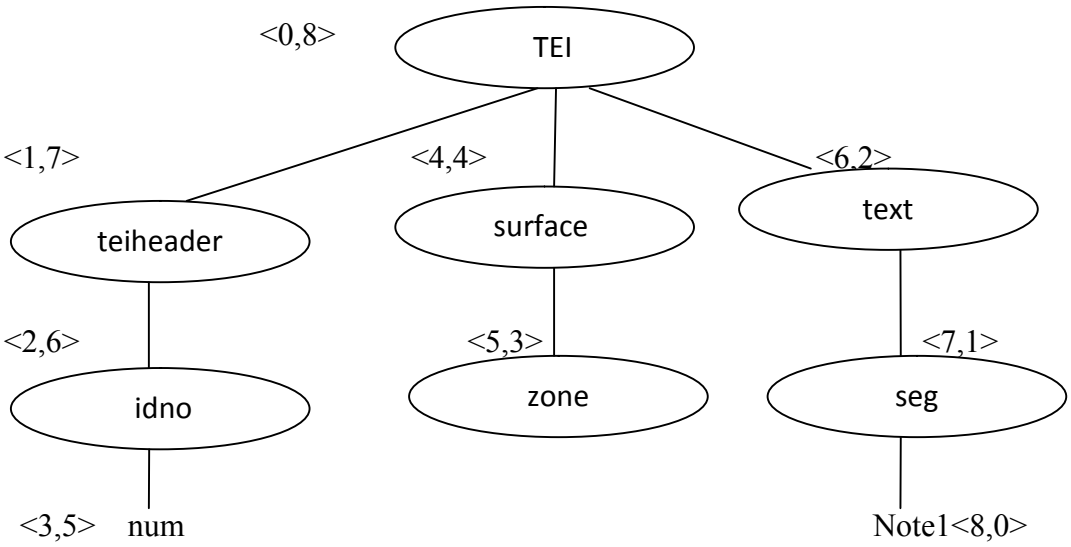


Figure 16 : Arbre représentant le document XML/TEI

3.3 La représentation physique des documents XML

L'indexation des images se fera grâce à une base de données. Leur représentation physique se fera suivant une approche basée sur les bases de données relationnelle.

4. Architecture générale de notre système

L'architecture générale de notre système comporte deux modules : module annotation et recherche d'images des manuscrits arabes anciens.

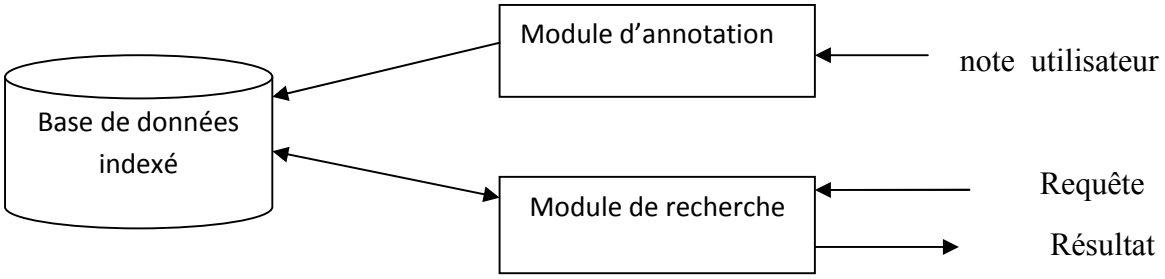


Figure 17 : Architecture générale de notre système

4.1. Module d'annotation

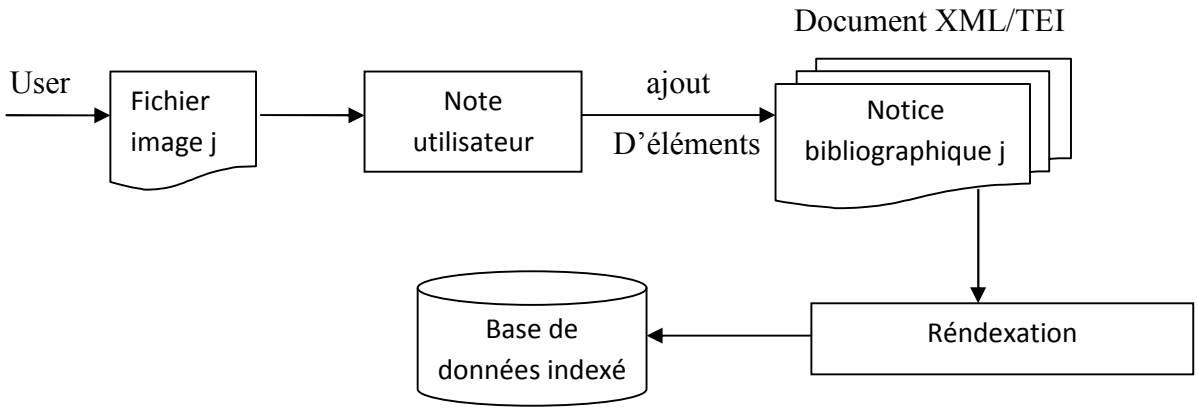


Figure 18 : Module Annotation

4.2 : Module de recherche

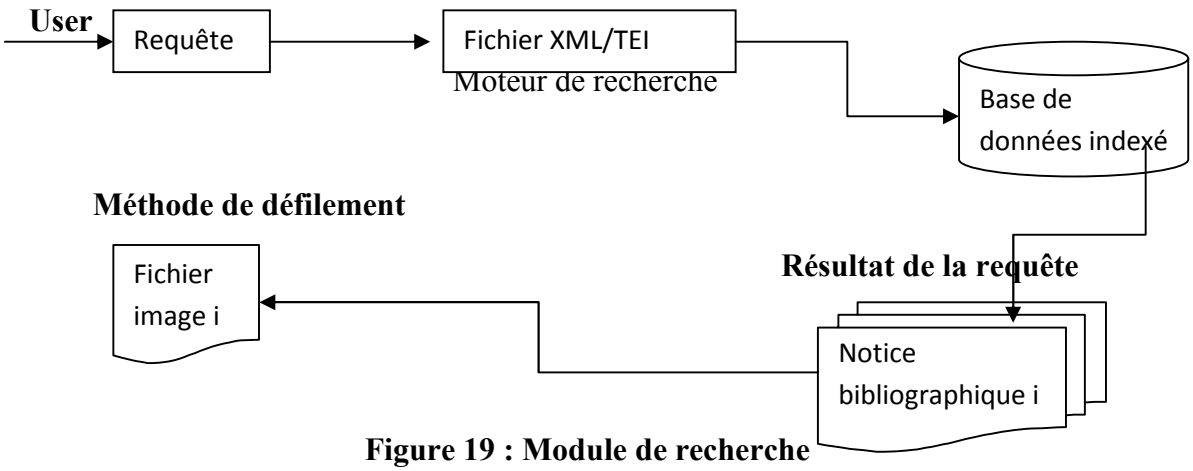


Figure 19 : Module de recherche

5. Système d'indexation

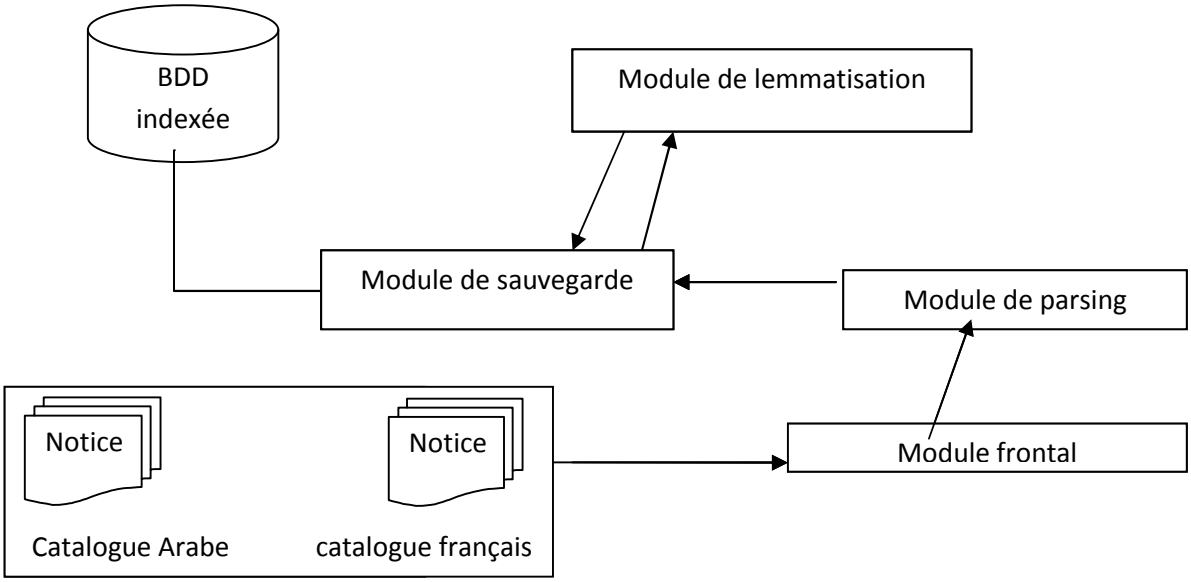


Figure 20 : système d'indexation

- **Module frontale**

Ce module permet de lister tout les chemins d'un répertoire du catalogue, Ces derniers seront passés pour indexation pour la prise en compte des nouvelles données.

- **Module de parsing :**

Le module de parsing permet de séparer le contenu de la structure des documents XML. Ce module s'appui sur l'API SAX (Simple API for XML), notre but ici est d'identifier les différents éléments du fichier XML afin de permettre une meilleur manipulation des différentes informations contenues dans ce derniers.

- **Module de lemmatisation :**

Ce module reçoit en entrée du texte (contenu dans une chaine de caractère). Ce texte est issu des nœuds des feuilles d'un fichier, mais aussi il reçoit la langue du fichier, ainsi il élimine les mots vides qui correspondante a la langue du document (français, arabe), est retourne un vecteur contenant la liste des mots clés.

- **Module de sauvegarde**

Entrée :

- Liste des chemins des notices
- Résultat de lemmatisation
- Résultat du parsing

Sortie :

- Actionner la sauvegarde de la BDD

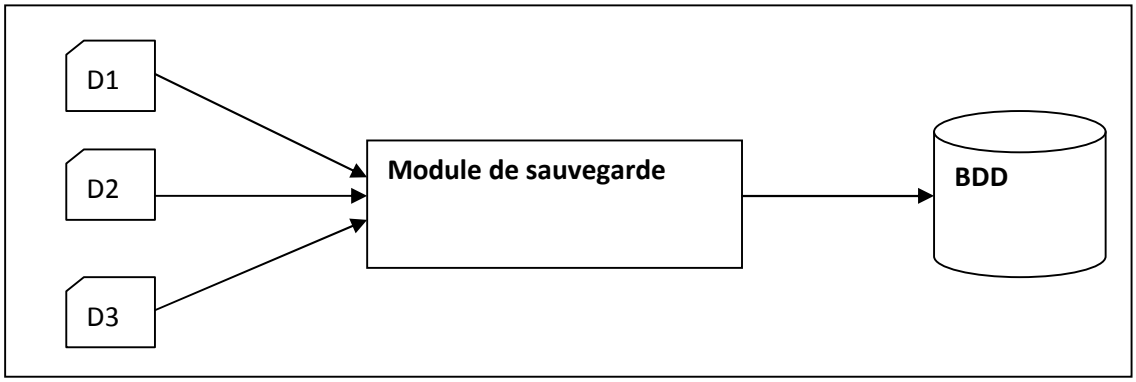


Figure 21: représentation du module de sauvegarde

6. Outils de développement

6.1 Langage java

Afin de réaliser notre application qui permet aux utilisateurs de rechercher des images à partir des annotations nous avons choisi java pour les raisons suivantes :

- Java est un langage orienté objet ;
- Une de ses plus grandes forces est son excellente portabilité : le programme créé, il fonctionnera sous Windows, Mac, Linux, etc.
- Permettant de créer de nouvelles classes à partir des classes existantes (héritage).
- Java permet de développer des applications d'interface graphique (fenêtres, menus graphisme, boite de dialogue, …) ,
- La JDK (Java Développent Kit) regroupe l'ensemble d'éléments permettant le développement, la mise au point et l'exécution des programme ;
- Java doté d'une riche bibliothèque de classe (le regroupement des traitements concernant une même donnée, au sein d'une même entité logicielle) ce qui permet la réutilisabilité de ces classes dans des contextes applicatifs différents.
- On peut faire de nombreuses sortes de programmes avec Java :
    - des applications, sous forme de fenêtre ou de console ;
    - des applets, qui sont des programmes Java incorporés à des pages web ;
    - des applications pour appareils mobiles, avec J2ME …

6.2 Eclipse

Eclipse est un environnement de développement intégré(IDE) dont le but est de fournir une plateforme de développement modulaire pour permettre de réaliser des développements informatiques. Il a était développer par I.B.M.

Voici l'interface de travail sous eclipse.

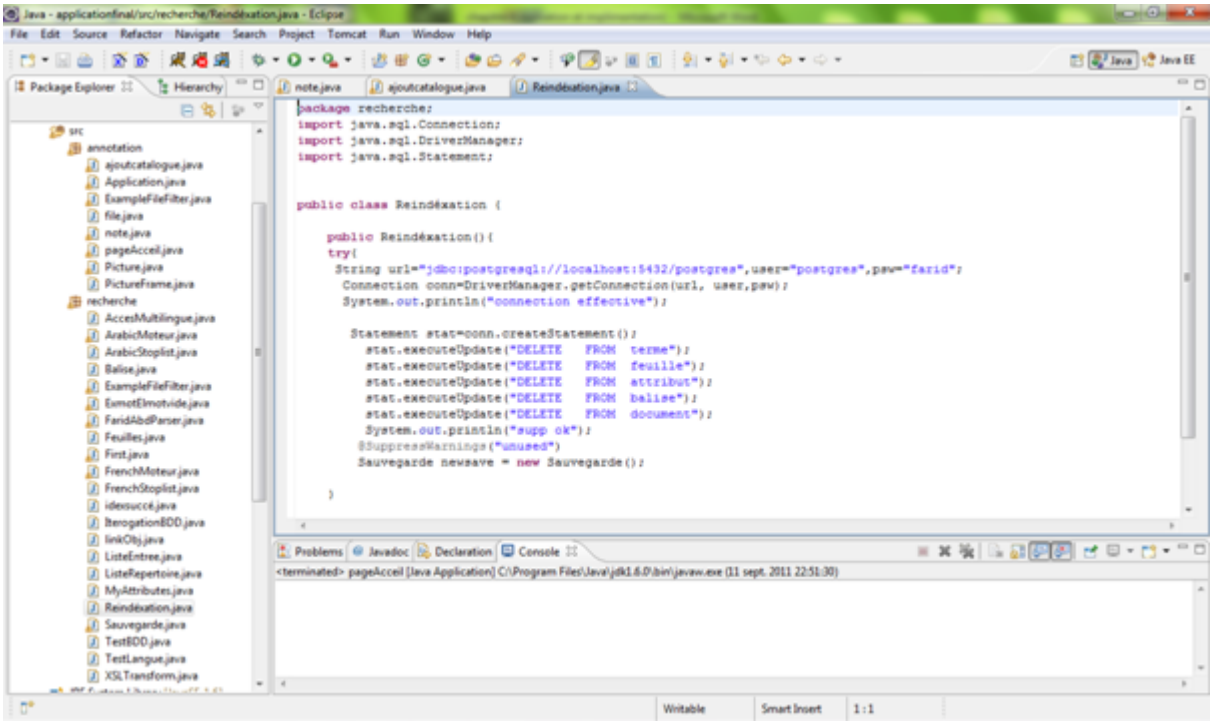


Figure 22 : interface de travail d'éclipse.

6.3 Postgresql

Postgresql est un système de gestion de bases de données relationnelles objet fondé sur Postgres. Ce dernier a été développé à l'université de Californie au département des sciences informatiques de Berkeley.

Postgres est à l'origine de nombreux concepts qui ne seront rendus disponible au sein de systèmes de gestion de bases de données commerciales que bien plus tard.

PostgreSql est un descendant OpenSource du code original de Berkeley. Il supporte une grande partie de standard SQL tout en offrant de nombreuses fonctionnalités modernes :

- Requetes complexes .
- Clés etrangéres .
- Triggers.
- Intégrité des transactions.

De plus PostgreSql est extensible par l'utilisateur de plusieurs façons. En ajoutant par exemple :

-de nouveaux types de données.

-de nouvelles fonctions.

-de nouveaux operateurs.

Voici l'interface de travail sous PostgreSql.

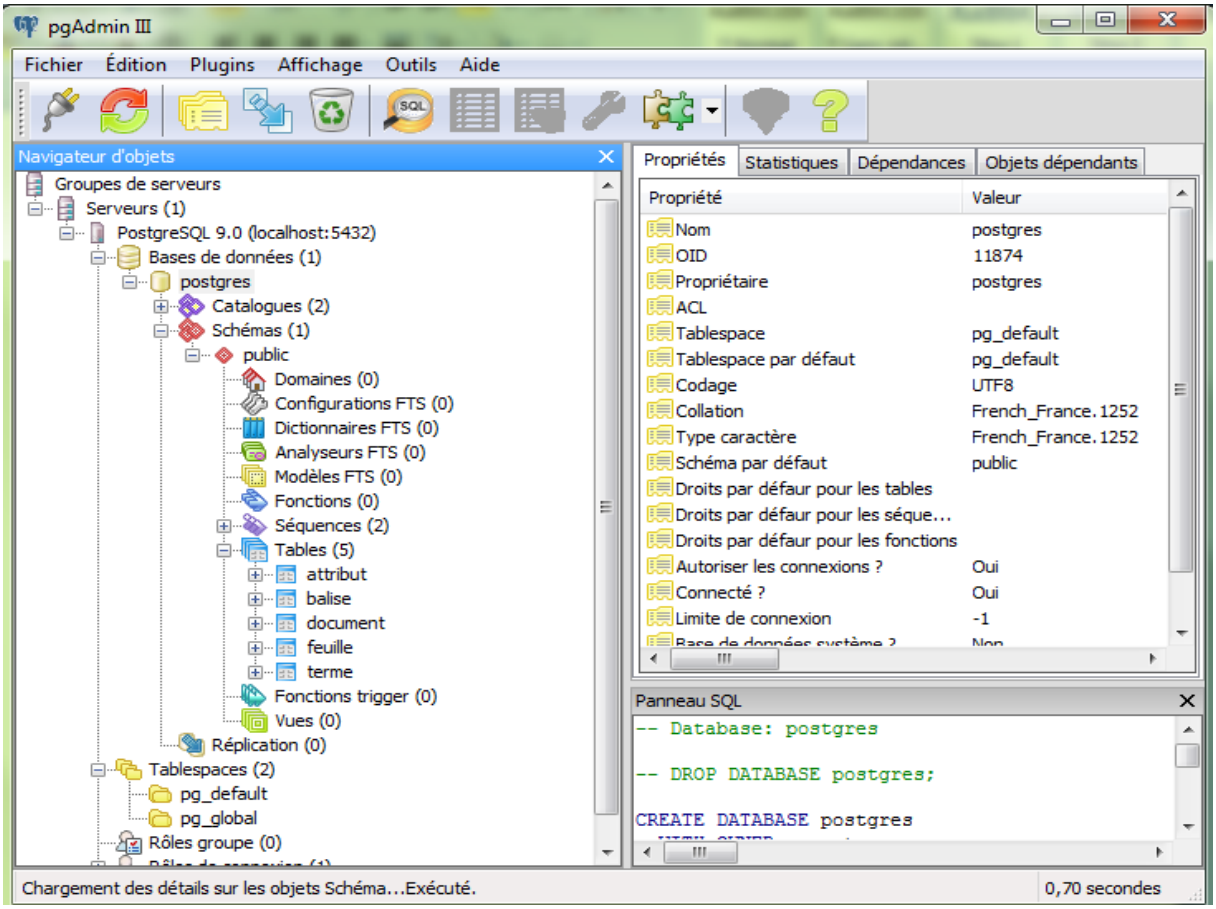


Figure 23 : interface de travail de PostgreSql

6 .3.1 Accès aux bases de données

Grace à sa licence libérale, PostgreSql, peut être utilisé, modifier est distribué librement, quel que soit le but visé, qu'il soit privé, commercial ou académique.

JDBC (Java Data Base Connectivity) désigne l'API défini par SUN Microsystems pour permettre un accès aux bases de données avec Java. Pour pouvoir utiliser JDBC, il faut un pilote qui est spécifique au SGBD qui contient la base avec laquelle on veut se connecter.

Pour ce qui PostgreSql on utilise « $JDBC_3$ PostgreSql pilote » disponible sur le site de Sun Microsystems. Les classes de JDBC sont regroupées dans le package java.sql et sont inclus dans la JDK depuis la version1.1 de cette API.

Ce package contient les quatre classes suivante, chacune correspondant à une étape d'accès à la base.

- **Driver Manager :** permet de changer et de configurer le driver de la base de données
- **Connection :** établie la connexion et l'authentification de la base de données
- **Statement : (**PreparedStatement**)** permet la transmission de la requête à la base de données
- **ResultSet :** les objets de cette classe sont destinés à contenir les informations retournées par la base dans le cas d'un Select.

7. Présentation des interfaces graphiques

- **Page d'accueil de l'application :**

C'est la première page qui apparaît en cliquant sur le lien de l'application responsable, elle conduit l'utilisateur à choisir entre deux taches: annotation d'image ou recherche d'image **(Figure 24).**

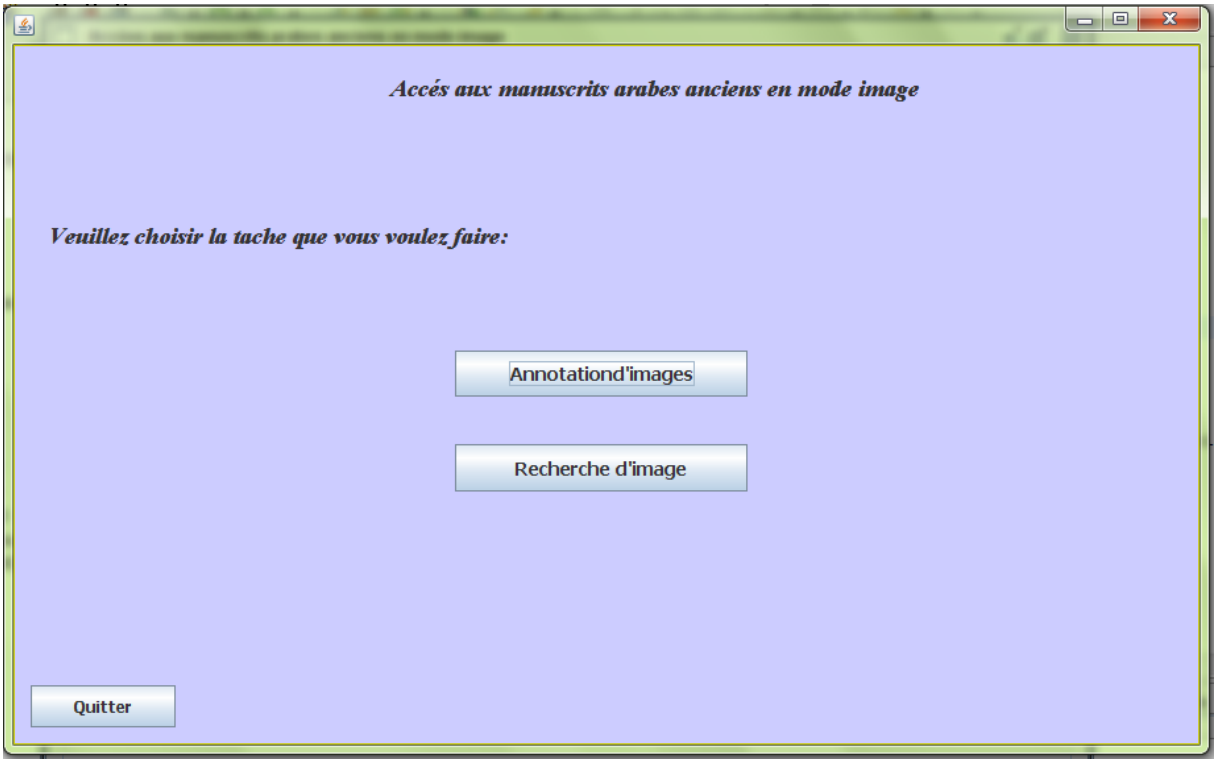


Figure 24: page d'accueil de l'application

- **Page d'accueil annotation :**

C'est la première page visualisée pour l'annotation, elle représente la porte principale de cette tâche. Son objectif est d'entrer de nouvelles métadonnées au système (figure 25).

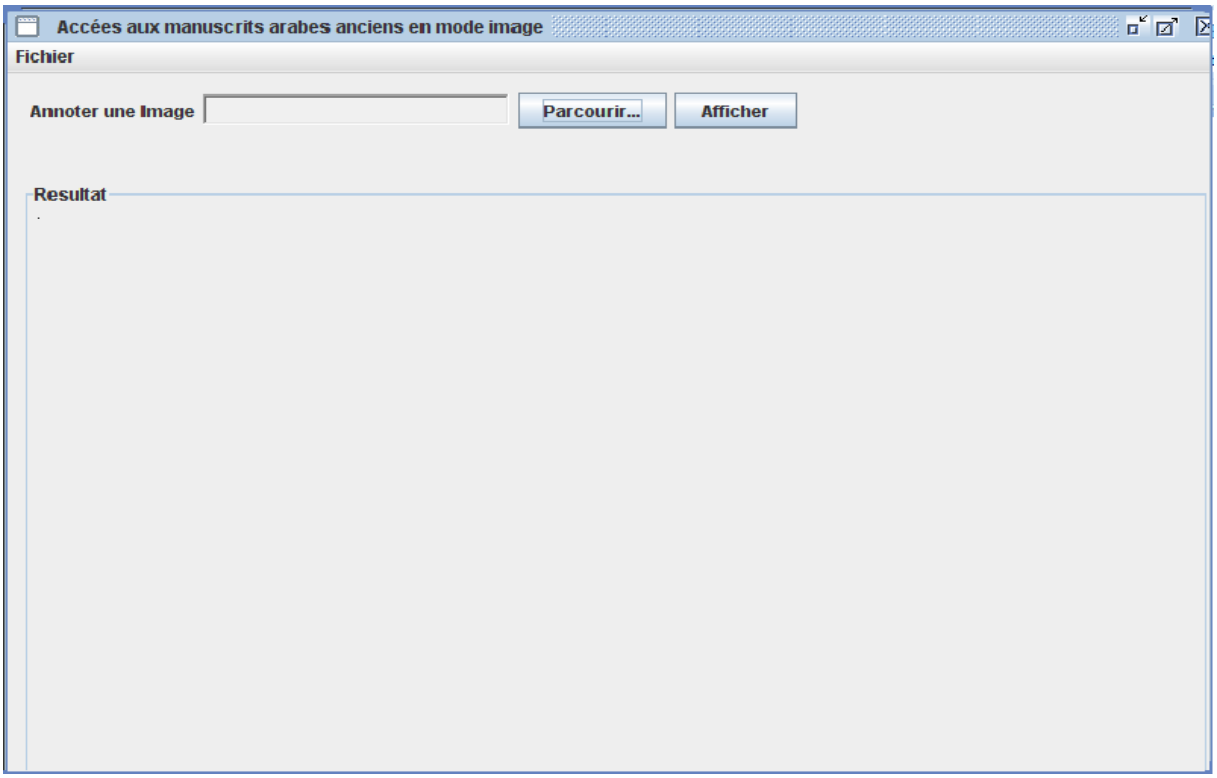


Figure 25 : page d'accueil annotation

- **Formulaire de sélection des images :**

En cliquant sur le Botton « parcourir » (dans la page d'accueil annotation), un formulaire de sélection des images s'affiche.

En cliquant sur le Botton « selectionner » la fenêtre disparaît, et le lien de l'image sélectionner s'affiche dans la page d'accueil annotation (**Figure 26**).

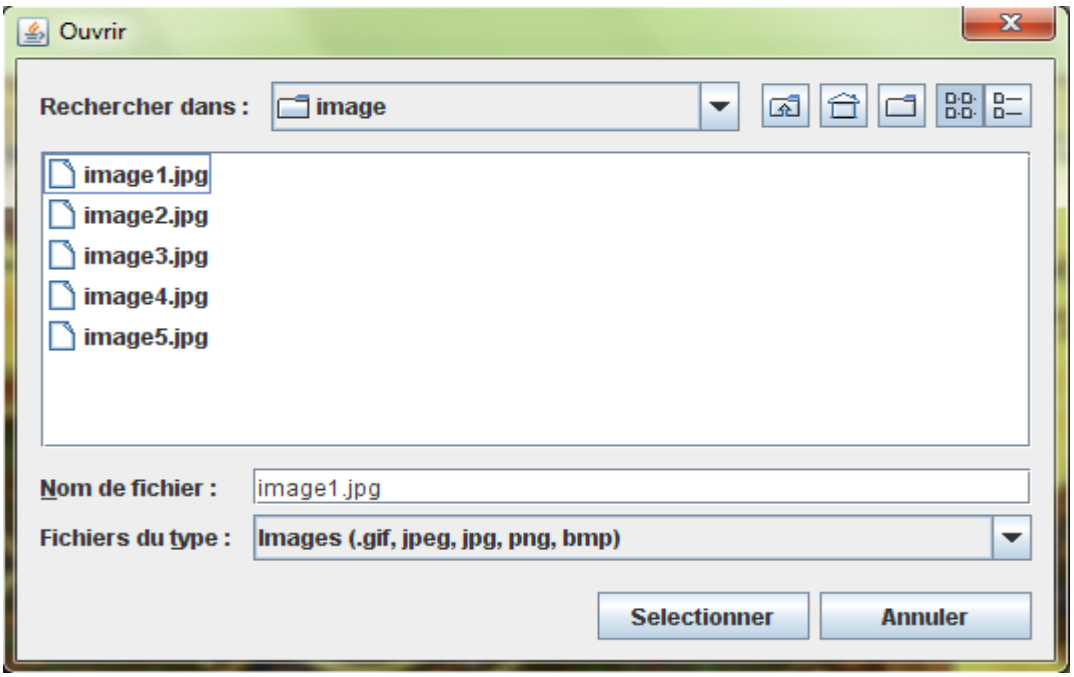


Figure 26 : formulaire de sélection

- **Affichage de résultat de sélection:**

En cliquant sur le bouton « Afficher», l'image s'affiche dans l'interface, dotée des barres de défilement afin de parcourir toute l'image. L'utilisateur peut alors lire ou analyser l'image pour produire des annotations (figure 27).

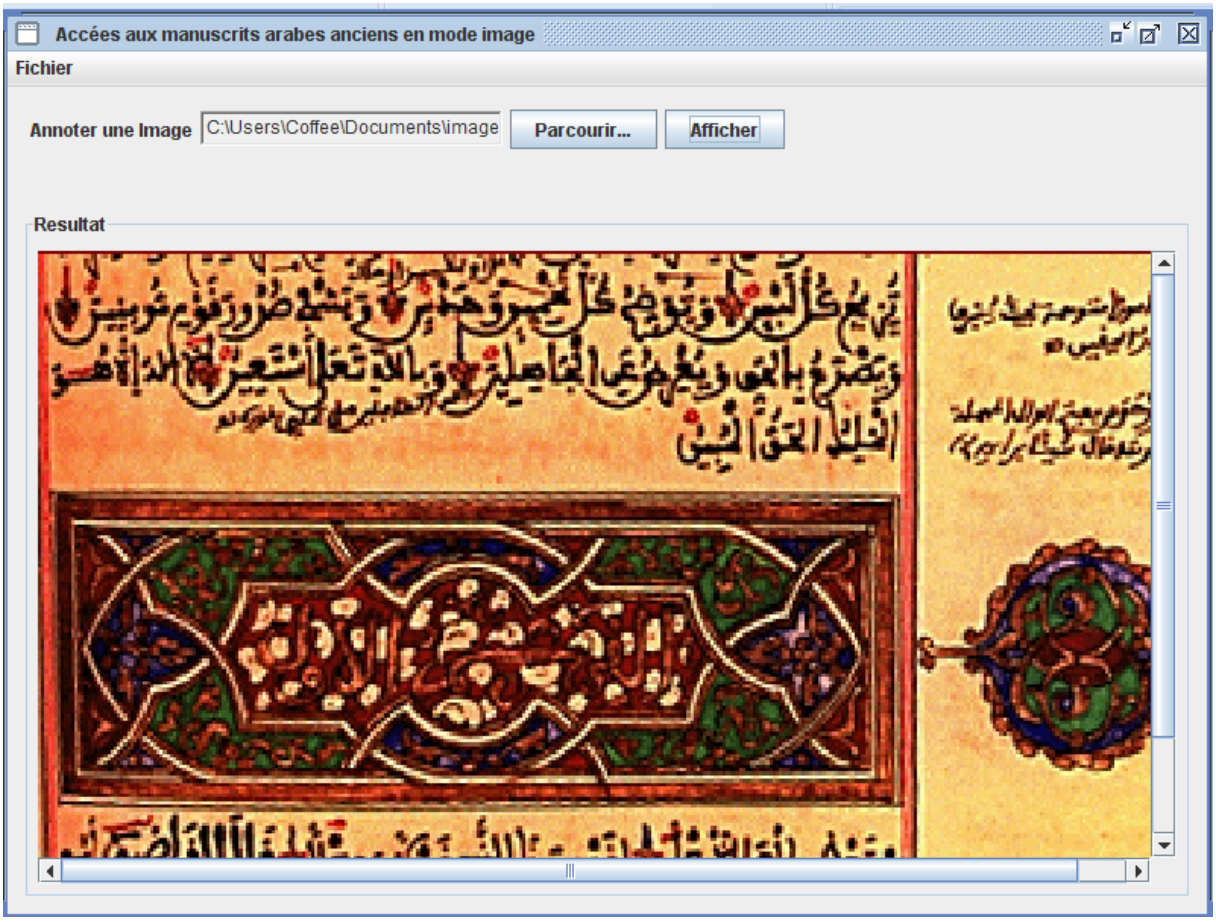


Figure 27 : Affichage de résultat de sélection

- **Production des annotations :**

En cliquant et maintenir le Botton de la souris enfoncé puis tirer, un rectangle transparent s'affiche sur la partie désiré. En lâchant la souris une fenêtre d'annotation s'affiche doté d'une zone de texte sur la quelle les notes seront rédigées (figure 28).

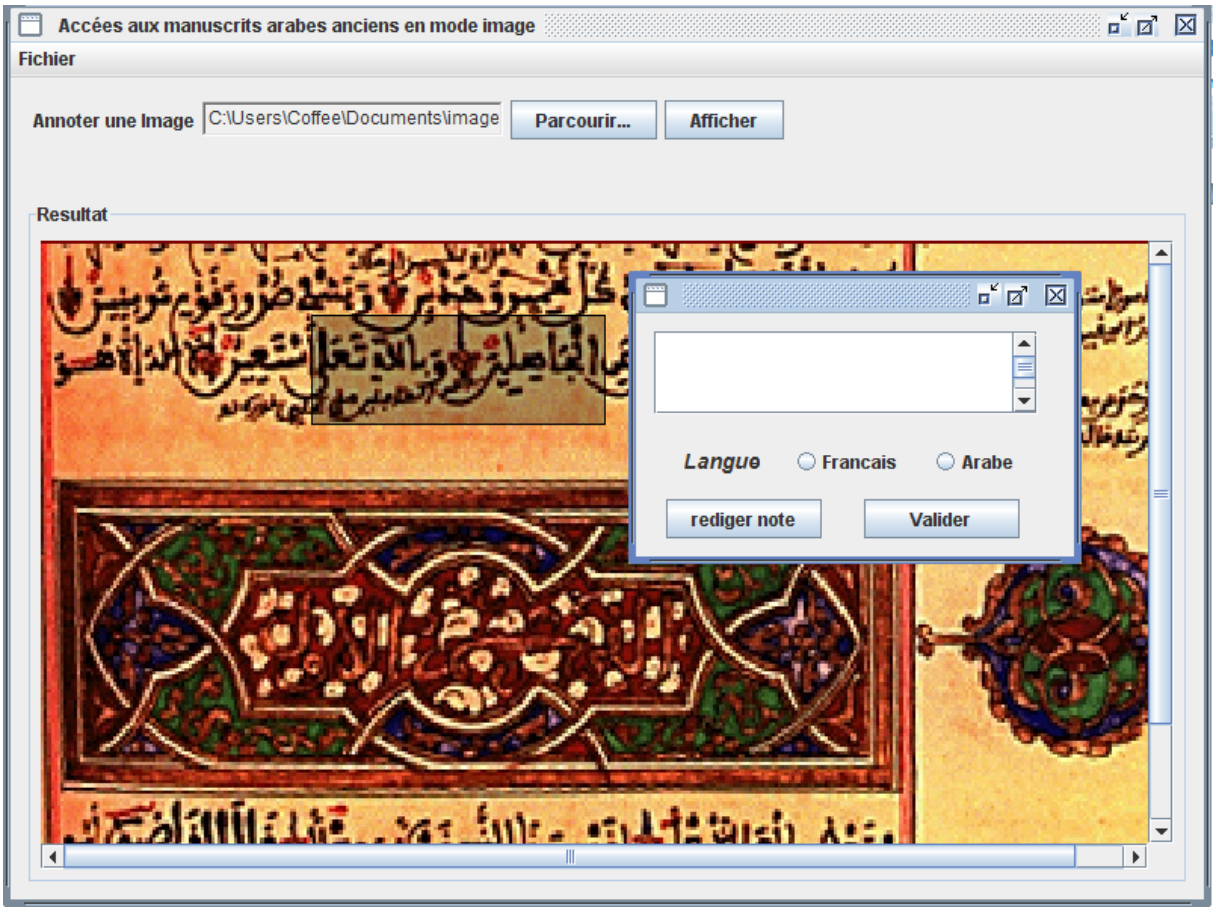


Figure 28 : Production des annotations

- **Fenêtre d'annotation en langue Française :**

Pour que l'utilisateur puisse rédiger une note en langue française ,il doit coché le Botton « français » puis cliqué sur le Botton « rédiger note » ce qui permet d'activé la zone de texte avec des entrées latins (a,b,c…..,) .Après avoir saisit la note ,sa validation (mise à jour du catalogue français) se fait cliquant sur le Botton « valider » (figure 29).

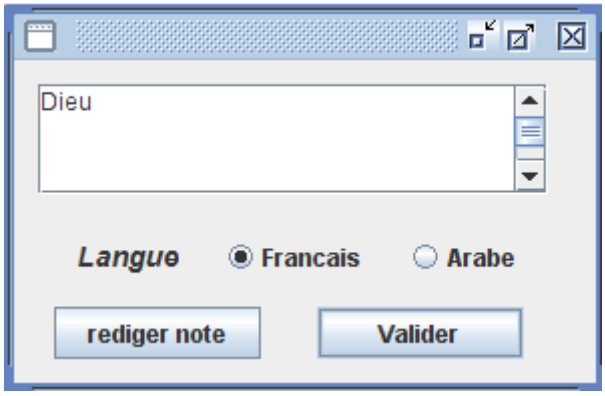


Figure 29 : Fenêtre d'annotation en langue Française :

- **L'interface d'annotaion en langue Arabe :**

Pour que l'utilisateur puisse rédiger sa note en langue Arabe ,il doit coché le Botton « arabe » puis cliqué sur le Botton « rédiger note » pour que la zone de texte soit activé avec des entrées arabe (ا, ب, ت...) .Après avoir saisit la note ,sa validation (mise à jour du catalogue arabe) se fait cliquant sur le Botton « valider ». (Figure 30)

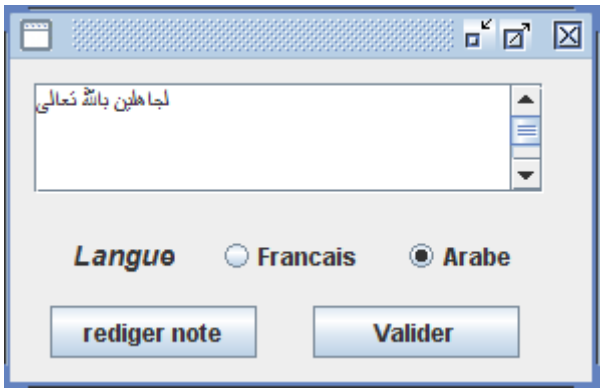


Figure 30: L'interface d'annotaion en langue Arabe :

- **Interface de recherche :**

Cette fenêtre s'affiche en cliquant sur le lien « recherche d'image » (dans la page d'accueil de l'application), elle permet la recherche en langue arabe ou la langue française (Figure 30)

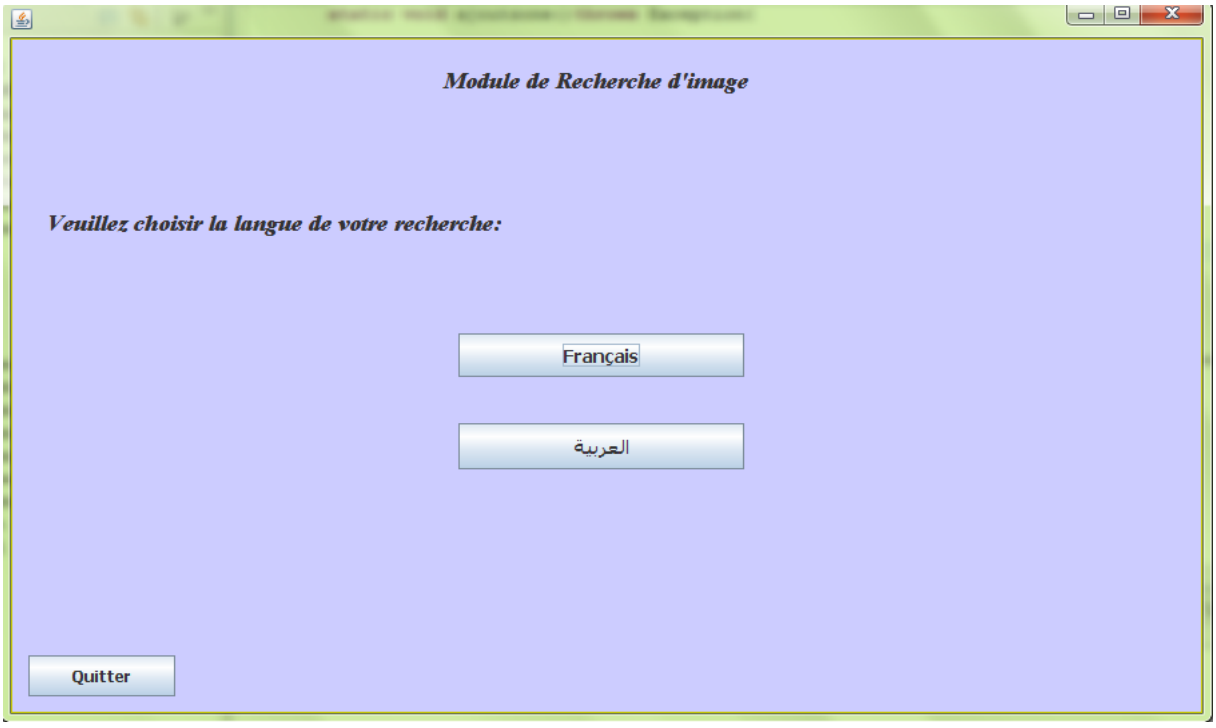


Figure 30 :Interface de recherche

- **Fenêtre de recherche en langue arabe:**

Cette fenêtre est composée d'une petite zone de texte activé en entrées arabe pour saisir les mots clés de la requête de l'utilisateur, Un botton «ابحث » pour lancer la recherche, et deux anglets, l'un est utilisé pour contenir les liens satisfaisant la requête, l'autre pour l'affichage des images correspondantes aux liens (Figure 31).

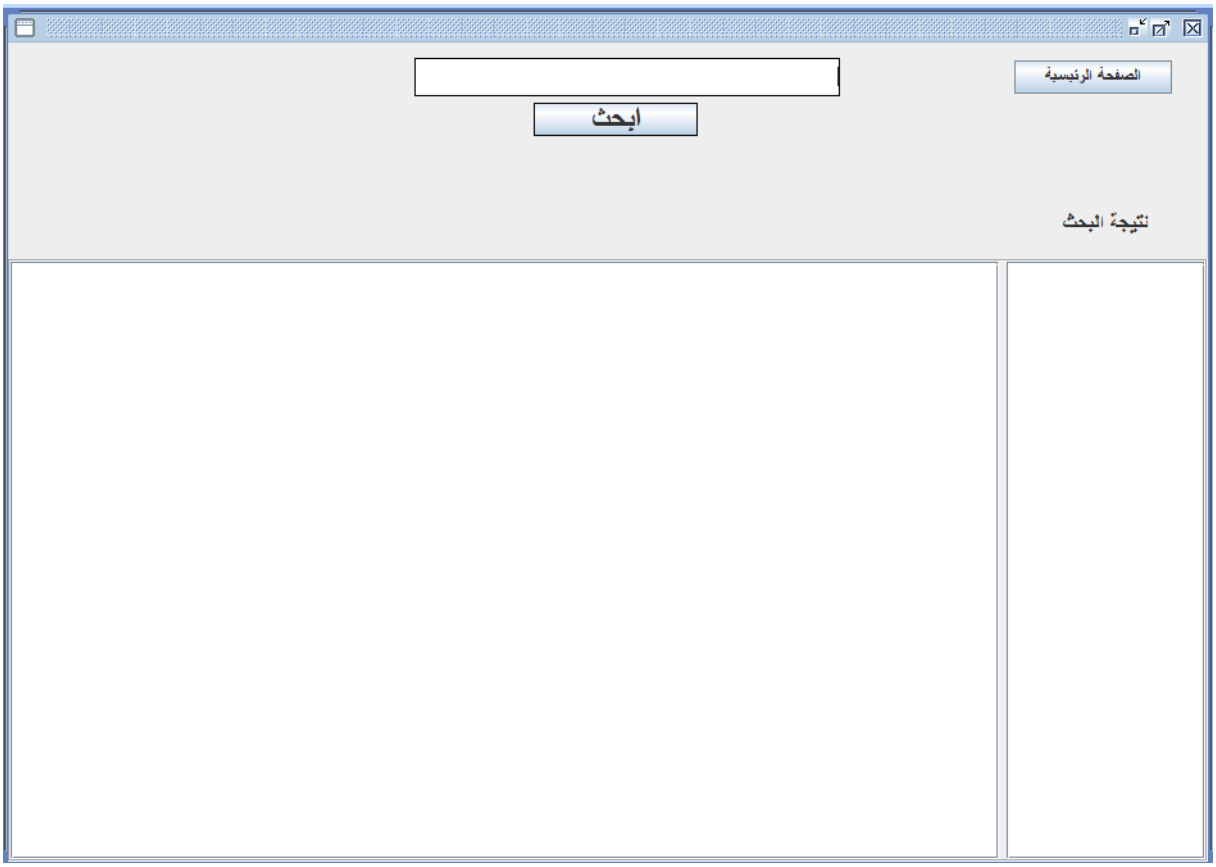


Figure 31 : Fenêtre de recherche en langue arabe

En saisissant des mots déjà introduit dans les notes des utilisateurs,les liens des images annoter aves ces mots seront affichées dans l'anglet gauche,en cliquant sur un lien, l'image correspondante sera affiché dans l'anglet droit (figure 32)

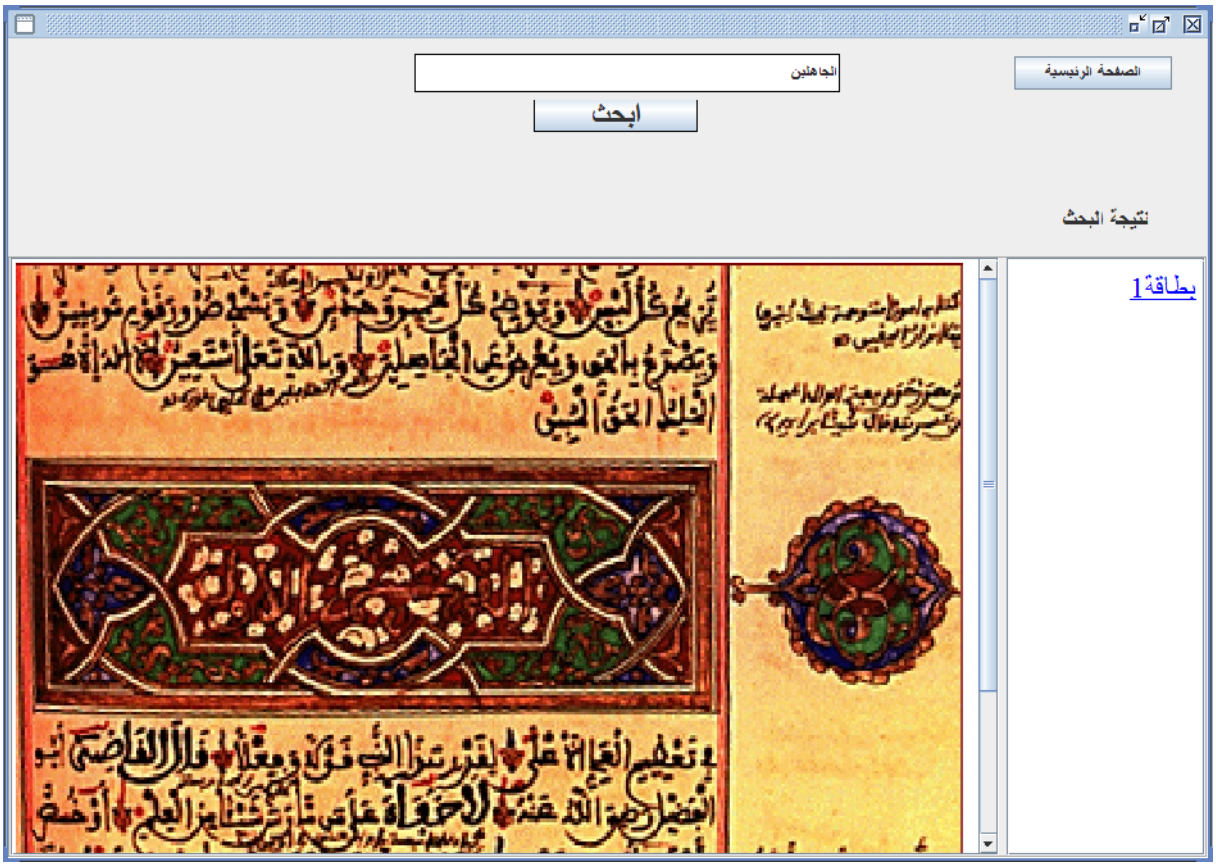


Figure 32 : Résultat de la recherche en langue arabe

- **Fenêtre de recherche en langue française:**

Dans cette fenêtre, La zone de texte destiné pour contenir la requête utilisateur est activée en entrées latins et le Botton «rechercher » permet de lancer la recherche, Les résultat sont affiché dans les anglets en dessous (figure 33).

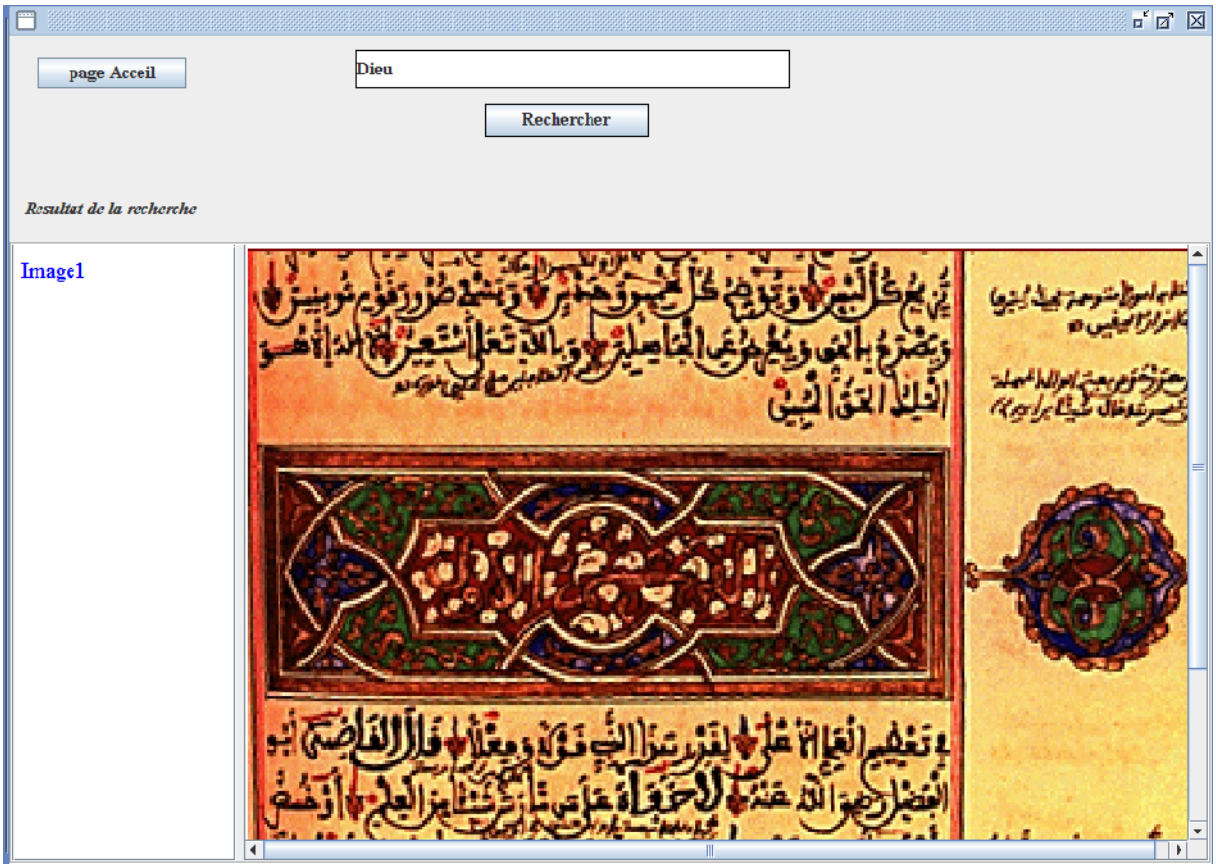


Figure 33 :fenetre de recherche en langue français

8. Conclusion

Dans ce chapitre nous avons présenté le résultat de notre implémentation ainsi que l'environnement que nous avons utilisé pour cette réalisation.

Nous avons décrit quelques fenêtres de notre application ainsi que le fonctionnement générale.

Nous estimons avoir atteindre notre objectif en réalisant une application qui répond aux objectifs fixés.

Conclusion

La numérisation des manuscrits sous forme d'image numérique répond non seulement aux besoins du monde de la recherche, mais aussi aux préoccupations et aux principales missions des bibliothèques qui sont en occurrence, la conservation, la préservation et la communication des documents. C'est dans cette optique que se situent les projets de numérisation des manuscrits arabes anciens.

L'accès au contenu du manuscrit numérisé en mode image constitue un challenge de taille, auquel est confrontée la communauté de recherche d'information. En effet, il nécessite une étape qui se positionne comme un problème difficile, qui est reconnaissance des caractères arabes manuscrits.

Par conséquent, le catalogue se propose comme un moyen d'accès efficace aux images numériques des manuscrits numérisés.

Néanmoins, la recherche de nouveaux outils qui permettraient d'atténuer les faiblesses sémantiques et contextuelles du catalogue constitue une solution palliative. Ainsi, l'annotation des manuscrits numérisés fait partie de ces outils.

Notre travail a consisté à mettre en place un tel outil, dont le rôle principal est d'enrichir le contenu du catalogue, de ce fait son indexation.

Ainsi, nous avons montré tout d'abord que la technique interactive d'annotation permet d'associer à des parties d'une image du texte structuré, codé en XML/TEI. Des traitements automatiques étendus peuvent alors être effectués sur les balises et leur contenu. Il est ainsi possible, entre autre, d'accéder aux images par indexation.

Un document inaccessible est un document mort. La mise en œuvre de notre solution est une réponse à cette problématique, qui permettra d'une part une meilleure longévité des manuscrits Arabes anciens et d'autre part leur valorisation en les rendant accessibles par une plus large communauté.

[Abbaci2003] Abbaci.F.Méthodes de selection de coolectios dans un environnement de Recherche d'Information distribuées. Thèse pour l'obtention du titre de docteur de L'université Jean Monnet de St-Etienne et de l'Ecole nationale supérieure des Mines de St-Etienne et du titre de Docteur Science de l'université de Neuchatel,2003

[Burnard 2008] : Lou Burnard, James Cummings, Matthew Driscoll, Sebastian Rahtz, Documentation and Training Materials for use with TEI P5 Specification for ENRICH, Octobre 2008,

[Calabretto2003] Calabretto S, *Recherche d'Information*, LIRIS, INSA Lyon. 2003.

[Chevalier2011] Chevalier M. Usagers & Recherche d'Information. Thèse pour l'obtention de l'Habilitation à Diriger de Recherches, Institut de Recherche en Informatique de Toulouse (UMR 5505) ; l'université Paul Sabatier (Toulouse III) ,2011

[Dahak2006] Dahak F. Indexation des documents Semi-Structurés. Thèse de Magister de l' Institut National de formation en Informatique I.N.I. Alger .2006

[Doumat 2008] : Reim Doumat, Elöd Egyed-Zsigmond, Jean-Marie Pinon, Un modèle d'une bibliothèque numérique collaborative – ARMARIUS, LIRIS –INSA de Lyon, 2008.

[El bannay 2009] : Omar El bannay, Rachid Benslimane, Noureddine El makhfi, Badraddine Aghoutane, Noureddine Rais, Searching in Arab Manuscripts Using Metadata and Annotation, Mars 2009, accessible sur : http://www.eurojournals.com/ejsr.htm

[Fellag2006] Fellag .Indexation de documents Structurés

[Idrissi2008] Idrissi N. La navigation dans les bases d'images. Thèse de Doctorat de l'université de Nantes. 2008

[Jalam2002] Jalam R, Chauchat JH. *Pourquoi les n-grammes permettent de classer des textes? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques*. 6es Journées internationales d'Analyse statistique des Données Textuelles, Laboratoire ERIC – France, 2002.

[Jerome2005] Jérôme L. Analyse multirésolution pour la recherche et l'indexation par le

contenu dans des bases de données Image-Application à la base d'image paléontologique Trans'Tyfipal. Thèse de doctorat de l'université Bourgogne. 2005

[Kaileh2004] Kaileh H. L'accès à distance aux manuscrits arabes numérisés en mode image, Thèse de Doctorat de l'université Lumière Lyon II.2004

[Karinne2003] Karinne S.B. Numérisation de patrimoine écrit et graphique, Bouchard/ABCD. 2003

[Lancaster1998] Lancaster F, *Indexing and abstracting in theory and practice*. Library Association Publishing. London, 1998.

[Mehadi2010] Mehadi F. Moteur de recherche dans un catalogue multilingue sous format XML, Thèse de licence de l'université Mouloud Mammeri de Tizi Ouzou (UMMTO). 2010

[Mokdem2010] Mokdem H. Mémoire pédagogique à base d'annotations pour l'apprenant, Thèse de Magister d'Ecole nationale Supérieur d'Informatique (ESI).2010

[Monique2000] Monique E. Développement du fonds angevin numérisé de la bibliothèque, projet professionnel d'Ecole nationale Supérieur des Sciences de l'Informatique et des Bibliothèque, université d'Angers. 2000

[Meylan2001] Meylan Ed, *Introduction théorique à la gestion de données textuelles*, Haute Ecole Spécialisée de Suisse Occidentale, 2001.

[Ouadah2008] Ouadah A. Mémoire d'annotation adaptative pour l'enseignant, Thèse de Magister d'Ecole nationale Supérieur d'Informatique (ESI).2008

[Paradis1996] Paradis F, *Un modèle d'indexation pour les documents textuels structurés*, Thèse de doctorat de l'Université Joseph Fourier - Grenoble 1, 1996.

[Petra2010] Petra B. Contributions à l'indexation et à la reconnaissance des manuscrits Syriaques, Thèse de doctorat de l'institut National des Sciences Appliquées de Lyon. 2010

[Pierrat Marie1999] Pierrat Marrie J. Approche critique de la pratique de la Text Encoding Initiative, TEI, pour la constitution d'une bibliothèque virtuelle en Sociologie. Cyberdocumentaliste. Service de formation et de

documentation Internet. Institut de Recherche sur les Sociétés Contemporaines IRESCO-CNRS 59-61, rue Pouchet 75849 Paris Cedex 17. Octobre 1999.

[Piwowrski2003] Piwowarski B. *Techniques d'apprentissage pour le traitement d'informations structurées : application à la recherche d'information*. Thèse de doctorat de l'université de PARIS 6, 2003.

[Rudolf2005] Rudolf P. Indexation d'images par une loi puissante, Thèse de doctorat de l'université de Paris 5René Descartes.2005

[Sahbi&Charles&Amos] Sahbi.S , Charles.R et Amos.D ,Analyse automatique de textes comme point de départ d'un processus d'annotation. Loria- université Nancy2.2005

[Soulah&Hassoun] Soualah M & Hassoun M. Accès multilingue en ligne aux manuscrits arabes numérisés,2010

[Soualah2008] Soualah M. Numérisation des manuscrits arabes :Catalogage et accès Multilingue, Thèse pour l'obtention de Magister l'Institut National de Formation I.N.I. Alger.2008

[Sidhom 2011] : Sahbi SIDHOM, Charles ROBERT, Amos DAVID, Analyse automatique de textes comme point de départ d'un processus d'annotation, janvier 2011, LORIA - Université Nancy2.

[Sauvagant2005] Sauvagnat K. *Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structurés*, Thèse en vue de l'obtention du Doctorat de l'Université Paul Sabatier, 2005.

[Weigel2004] Weigel F, Meuss H, Bry F and Schulz K.U, *Content and Structure in Indexing and Ranking XML*. Seventh International Workshop on the Web and Databases (WebDB) France, 2004.

[Zargayouna2005] Zargayouna H. *Contexte et sémantique pour une indexation de documents semi-structurés*. LIMSI/CNRS-Université Paris 11, 2004.

1 .Aperçus sur Java

Java est un langage de programmation à usage général, évolué et orienté objet dont la syntaxe est proche du C. Il existe 2 types de programmes en Java : les applets et les applications. Une application autonome (stand alone program) est une application qui s'exécute sous le contrôle direct du système d'exploitation. Une applet est une application qui est chargée par un navigateur et qui est exécutée sous le contrôle d'un Plug in de ce dernier.

Les programmes écrits en java sont portable sur plusieurs systèmes d'exploitation tels que Mac, UNIX, Microsoft Windows ou Linux.

Les programmes Java exécutés localement sont des applications, ceux qui tournent sur des pages Web sont des applets. Les principales différences entre une applet et une application sont : les applets n'ont pas de méthode main() : la méthode main() est appelée par la machine virtuelle pour exécuter une application. les applets ne peuvent pas être testées avec l'interpréteur mais doivent être intégrées à une page HTML, elle même visualisée avec un navigateur disposant d'un plug in sachant gérer les applets Java, ou testées avec l'applet viewer.

Java a donné naissance à un système d'exploitation, il dispose d'un environnement de développement (eclipse/JDK), d'une machine virtuelle(JRE)applicatives multi-plates-formes(JVM), une bibliothèque Java(J2ME) avec interface graphique(AWT/SWING) et des applications Java(servlet, applets, application). La machine virtuelle la traduction et l'exécution de bytecode qui généré lors de la compilation d'un programme écrit en java.

L'API Java standard pour la manipulation du format XML est **JAXP** (Java API for XML Processing). Cette API permet la lecture, la transformation et l'écriture de fichiers ou flux XML. C'est cette API que nous allons étudier dans la partie XML de cette FAQ.

JAXP n'est pas la seule API disponible pour travailler avec XML.

2. JAXP

JAXP (Java API for XML Processing) est composée de quatre packages. Cette API met à la disposition du développeur trois ensembles de fonctionnalités (la modélisation, le parsing et la transformation) regroupées en quatre packages distincts.

- **javax.xml.parsers** : Ce package contient un ensemble d'interfaces devant être implémentées par les différents parseurs (SAX ou DOM
- **org.w3c.dom** : Ce package contient l'ensemble des classes et interfaces nécessaires pour travailler avec **DOM** (*modélisation*).
- **org.xml.sax** : Ce package contient l'ensemble des classes et interfaces nécessaires pour travailler avec **SAX** (*parsing*).
- **javax.xml.transform** : Ce package contient l'ensemble des classes et interfaces nécessaires pour travailler avec **XSLT** (*transformation*).

3. Bref historique de Java

Les principaux événements de la vie de Java sont les suivants :

1995 mai : premier lancement commercial

1996 janvier : JDK 1.0

1996 septembre : lancement du JDC

1997 février : JDK 1.1

1998 décembre : lancement de J2SE et du JCP

1999 décembre : lancement J2EE

2000 mai : J2SE 1.3

2002 J2SE 1.4

2004 J2SE 5.0

4. Caractéristiques du java

Voici quelques caractéristiques :

- **Orienté objet :** La P.O.O. (programmation orientée objet) possède de nombreuses propriétés universellement reconnues désormais. Notamment, elle ne renie pas la programmation structurée (elle se fonde sur elle), elle contribue à la fiabilité des logiciels et elle facilite la réutilisation de code existant. Elle introduit de nouveaux concepts, en particulier ceux d'objets, d'encapsulation, de classe et d'héritage.
- **Héritage :** Il permet de définir une nouvelle classe à partir d'une classe existante (qu'on réutilise en bloc !), à laquelle on ajoute de nouvelles données et de nouvelles méthodes.
- **La réutilisabilité :** des classes peuvent être appelées par d'autres sont les crées à chaque fois.
- **Package :**c'est le regroupement d'un ensemble de classes qui ont les mêmes caractéristiques.
- **Multithreads :**pour permettre l'exécution de plusieurs taches à la fois(afin d'évite le blocage de certaines tache),il est nécessaire de mettre en œuvre des threads qui permet de gérer l'exécution de plusieurs taches au sien d'une même application.

5. La JDK

Ecrire un programme en java ne nécessite pas d'outil spécifique : un éditeur de texte tel Notepad suffit. Le programme java ; qui porte l'extension « .java » doit être compilé afin de générer un fichier « bytecode » qui l'extension « .classe »,qui sera interprété par la JVM.

Pour ce faire il faut télécharger la JDK(www.javasoft.com), il existe plusieurs plate-forme.

Ce kit de développement comprend plusieurs outils, ne disposant pas tous d'interface graphique, dont les principaux :

Javac :c'est le compilateur java ;il traduit un fichier d'extension .java en fichier d'extension .classe

Java :c'est un interpréteur java, c'est-à-dire implémentation de la JVM. Il traduit en langage machine les fichiers déjà compilés par javac après avoir effectué des contrôles.

Applet viewer :ce programme est un interpréteur java qui possède la spécificité d'exécuter que les applets. Disposant de sa propre interface graphique .il permet de tester ces derniers, sans recours à un navigateur web.

JRE : c' est un interpréteur allégé de java qui n'a pas besoin d'installer tout le JDK pour exécuter un programme compilé par javac. Il est destiné à des plates-formes à des clients qui ne développe pas d'application java mais qui les utilisent.

JDB :ce débogueur permet de détecter les erreurs de programmation

Javadoc : cet utilitaire permet de construire à partir des commentaires insérés dans des sources java et sources condition que ces derniers respectent une certaine syntaxe, des fichiers HTML documentant les classes ,les méthodes,… développées dans les sources

Javap :il permet de désassembler un fichier compilé

Jar : permet d'archiver plusieurs classes java.

6. la machine virtuelle Java (JVM)

Pour exécuter une application java, la présence sur la plate-forme d'exécution de l'implémentation d'une machine virtuelle java est obligatoire. Cette dernière aura la charge de traduire le bytecode en instructions exécutables par le processus de la machine hôte (on trouve de ce fait JVM, une pour Microsoft Windows, une pour Lunix,..). Mais la JVM détient aussi d'autres responsabilité que nous allons rencontrer en détaillant ses principaux composants :

- Le chargeur de classe dynamiques (Class Loader).
- Le vérificateur de bytecode (Bytecode verifier).
- Le gestionnaire de sécurité.
- Le moteur d'exécution.

1. Présentation du XSLT

XSL (Langage de Feuille de Style eXtensible) est le langage utilisé pour transformer et afficher des documents XML. Il n'est pas encore dans sa version définitive, donc méfiez-vous ! C'est un langage de formatage de document complexe qui est en lui-même un document XML. Il peut être subdivisé en deux parties : transformation (XSLT) et formatage d'objets (quelquefois comme FO, XSL:FO ou simplement XSL). Pour plus de simplicité je traiterai seulement de XSLT ici.

2. Transformations XSL (XSLT)

Le 16 novembre 1999, le World Wide Web Consortium a annoncé la publication de XSLT comme Recommandation W3C. Ce qui signifie essentiellement que XSLT est stable et ne changera pas dans l'avenir.

- **Le langage de transformation des données** (**XSLT**, *eXtensible Stylesheet Transformation*) permettant de transformer la structure des éléments XML.

Un document XML peut être représenté comme une structure arborescente. Ainsi XSLT permet de transformer les documents XML à l'aide de feuilles de style contenant des règles appelées **template rules** (ou *règles de gabarit* en français).

Le processeur XSLT (composant logiciel chargé de la transformation) crée une structure logique arborescente (on parle d'**arbre source**) à partir du document XML et lui fait subir des transformations selon les *template rules* contenues dans la feuille XSL pour produire un **arbre résultat** représentant, par exemple, la structure d'un document HTML. Les composants de l'arbre résultat sont appelés *objets de flux*.

Chaque *template rule* définit des traitements à effectuer sur un élément (noeud ou feuille) de l'arbre source. On appelle "**patterns**" (en français *motifs*, parfois "*éléments cibles*") les éléments de l'arbre source.

L'arbre source peut être entièrement remodelé et filtré ainsi qu'ajouter du contenu à l'arbre résultat, si bien que l'arbre résultat peut être radicalement différent de l'arbre source.

1. Présentation d'XML

XML (pour eXtensible Markup Language) est un langage de balisage (comme le HTML, par exemple), son développement a commencé en 1996 avec le XML Working Group du W3C[6] qui en développe la recommandation en Février 1998.

XML est un langage permettant de séparer le contenu des documents des instructions de présentations (décrire le contenu plutôt que la présentation ainsi sépare le contenu de son apparence(le contenant)) . il permet aussi de représenter et d'assurer l'échange de document semi-structurés. Les documents XML sont dits semi-structurés car, ils possèdent une structure qui n'est pas imposée par une norme (contrairement à HTML), mais une structure que le créateur peut déterminer lui-même au moment de la conception du document.[fellag2006][Mihadi2010]

XML est aussi un métalangage c'est-à-dire un langage pour écrire d'autre langage ; [Fellag2006]il permet l'élaboration de balisages spécialisés. C'est-à-dire qu'en fonction du contenu qu'on souhaite publier, on définit nous propre balises tout en respectant les critères XML.

2. Structure d'un document XML

Un document XML est structuré comme suit :

- **La première partie :** appelée Prologue, dont la présence est facultative mais conseillée. Il contiendra un certain nombre de déclaration.
- **La deuxième partie :** constitue de commentaires et d'instructions de traitement ;
- **La troisième partie :**constitue d'un fichier XML ,est l'arbre d'éléments qui forme le contenu de document.

2.1. Le prologue :

Il peut contenir une déclaration XML, des instructions de traitement et une déclaration de type de document.

- **Déclaration XML**
  Syntax <?xml version="1.0" encoding="ISO-1 "standalone="yes"?>
  Elle indique:
  1. **Version:** la version de XML utilisée dans ce document,1.0 en ce que nous concerne ;
  2. **Encoding :** le jeu de caractère utilisé par défaut, l'attribut encoding a la valeur **UTF-8 ;**
  3. **Standalone :** dépendance de document par rapport à une déclaration de type document(DTD). Si standalone a la valeur « yes », le processus de l'application n'attend aucune DTD extérieure au document ; Sinon, le processus attend une référence de déclaration de type de document. La valeur par défaut est « No ».

Cette déclaration est facultative, mais il est préférable de l'utiliser, les trois attributs s'ils apparaissent, doivent être dans cet ordre :version, encogin et standalone.

- **Instruction de traitement :**
  Est une instruction interprétée par l'application servant à traiter le document XML. Elle ne fait pas totalement partie du document. Les instructions de traitement qui servent le plus souvent sont la déclaration XML ainsi que la déclaration de feuille de style [Fellag2006].
  - **Définition de type de document (DTD) :**
    La DTD Document Type Définition de type de document permet de définir la structure de document. Elle peut être déclarer au sien de document ou sous forme de document externe.

Exemple de déclaration de type de document :

Syntaxe :< !DOCTYPE nom_type[déclaration]>
nom_type est un nom choisi arbitrairement et qui sert à indiquer la portée des déclaration définies entre crochet. Elles seront valides pour tout document dont l'élément racine sera du type.
La DTD peut aussi être définie dans un fichier externe.
Syntaxe :< !DOCTYPE nom_type SYSTEM"fichier.dtd"[déclarations]>
Il n'est pas obligatoire pour un document XML de se conforme à une DTD. Cependant, il doit être *bien formé*, c'est-à-dire qu'il doit respecter les régles générales de syntaxe d'XML. Lorsqu'une DTD est associe à un document XML, on dit qu'il est *valide.*

2.2. Les commentaires

Des commentaires peuvent être insères dans les documents.ils commencent par < !..et se termine par..>. ils peuvent etre placé à n'importe quel place tant qu'ils se trouvent à l'extérieur d'une autre balise.

Exemple :

< !..ceci est correct..>

<elt>< !..ceci est correct aussi..> Un peut de text</**elt>**

2.3.L'arbre d'élément :

La partie essentielle d'un document XML sera toujours formée d'une hiérarchie d'éléments qui dénote la sémantique de son contenu :c'est l'arbre d'éléments. Cette arborescence comporte une racine(unique), des branches et des feuilles.

Exemple d'un document XML :

```
< ?xml version= ''1.0''encoding=''ISO.8859-1 '' ?>
<Mémoire>
<Titre>Mémoire d'annotation adaptative pour l'enseignant</Titre>
<Auteur>
<nom>Abdelaziz</nom>
<prenom>Ouadah</prenom>
</Auteur>
<Encadreur>
<nom>Azouaou</nom>
<prenom>Faiçal</prenom>
</Encadreur>
</Mémoire>
```

- Elément racine : L'élément racine c'est la base de document XML.il est unique et englobe tous les autres éléments, il s'ouvre juste après le prologue, et se ferme à la fin du document. Dans l'exemple :**la racine est Mémoire.**
- **Les éléments :** les éléments forment la structure de document : ce sont les branches et les feuilles de l'arbre. Ils peuvent contenir du texte ou bien d'autres éléments, qui sont alors appelés « élément enfant », l'élément contenant étant quant à lui appelé « élément parent ».
  L'élément se compose d'une balise d'ouverture, d'un contenu d'élément et d'une balise de fermeture.

Exemple : d'élément contenant du texte : <Titre>Mémoire d'annotation adaptative pour l'enseignant</Titre>

Exemple :d'élément contenant d'autres éléments :

```
 <Encadreur>
<nom>Azouaou</nom>
<prenom>Faiçal</prenom>
</Encadreur>
```

Il ne contient pas d'élément vide

Exemple d'élément vide :<nom/>

- Les attributs : la balise d'ouverture d'un élément peut inclure des attributs sous la forme de paires nom='valeur'. La valeur d'un attribut est une chaine de caractère encadrée par des apostrophes(' ') ou par des guillemets (« »).
  Exemple :<Encadreur paye=' alg'>Alger</Encadreur>.
- **Les section CDATA :** une section **CDATA** est une section pouvant contenir toutes sortes de chaine de caractère. Une section **CDATA** permet de définir un bloc de caractères ne devant pas etre analysés par le processeur XML. Cela permet entre autre de garder dans un bloc de texte un exemple de code à afficher tel quel
  Exemple d'utilisation de **CDATA :**
  < ![CDATA[Une balise commence par un <et ce termine par un>.]]>

3. Manipulation des documents XML

XML est uniquement un langage de structuration et de représentation des données. Il ne comporte pas d'instruction de contrôle et ne permet pas d'exploiter directement les données. Pour traiter ces données, il faut disposer d'un analyseur. Un analyseur (ou parser en anglais), permet de récupérer dans une structure XML, des balises, leur contenu, leurs attributs et de les rendre accessibles.XML dispose de deux types de parser :

- Les parseurs validant (validating) permettant de vérifier qu'un document XML est conforme à sa DTD.
- Les parseurs non validant (non-validating) se contentant de vérifier que les documents XML est bien formé(c'est-à-dire respectant le syntaxe XML de base).

Les analyseurs XML sont également divisés en deux selon l'approche qu'ils utilisent pour traiter le document.

-Parseurs de type arbre : les analyseurs utilisant cette technique construisant une structure hiérarchique contenant des objets représentant les éléments du document, la principale API utilisant cette approche est **DOM**(document object model) orienté hiérarchie.

-Les parseurs de type orienté événement : permettant de réagir à des événements(début d'un élément, fin d'un élément) et de renvoyer le résultat de l'application utilisant cette API. **SAX** (simple API of XML)est la principale interface utilisant l'aspect événement.

3.1.DOM(Document Object Model) :

DOM est l'acronyme de document object model. C'est une spécification du W3C pour proposer une API qui permet de modéliser, de parcourir et de manipuler un document XML.

Le principale rôle de DOM non seulement le parcoure de document XML mais en fait il permet de fournir une représentation en mémoire de ce document.
Cette représentation est un arbre, que nous pourrons manipuler (parcourir, recherche et mise à jour).

A partir de cette représentation (le modèle), DOM propose de parcourir le document et aussi de pouvoir modifier. Ce dernier aspect est un des aspects les plus intéressent de DOM.

DOM est défini pour être indépendant du langage dans lequel il sera implémenté par un éditeur tiers.

Il exits deux versions de DOM nommées « niveau »

- DOM Core level$_1$: cette spécification contient les bases pour manipuler un document XML(document, élément et nœud).
- DOM level$_2$:ajoutant de nouvelles fonctionnalités comme la prise en compte des feuilles de style CSS dans la hiérarchie d'objetz
  Exemple de document XML et l'arbre DOM associe

.

```
< ?xml version=''1.0''encoding=''ISO-8859-1'' ?>
<Cinema>
<nom>Epée De Bois </nom>
<Adresse>100.Rue Mouffetrad</Adresse>
<Metro>Censier-Daubenton<</Metro>
< /Cenima>
```

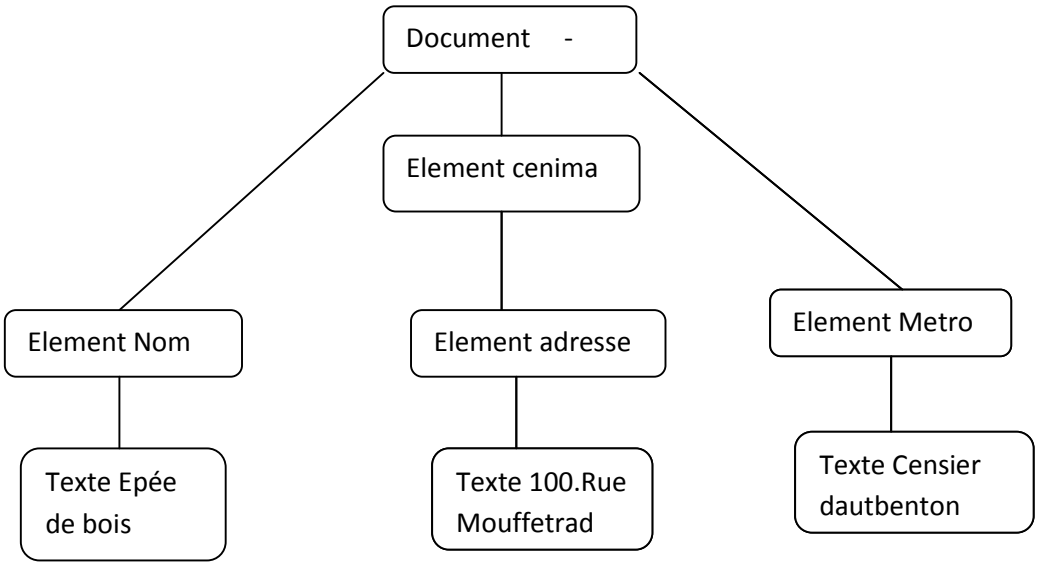


Figure : exemple d'arbre DOM

3.2. SAX (Simple API XML) :

SAX fournit une interface événementielle, cela signifie que SAX permet de déclencher des événements au cours de l'analyse de document XML. Une application utlisant SAX implémente généralement des gestionnaires d'événements, lui permettant d'effectuer des opérations selon le type d'élément rencontré.

Soit le document XML suivant :

```
< ?xml version=''1.0''encoding=''ISO-8859-1'' ?>
<Cinema>
<nom>Epée De Bois </nom>
<Adresse>100.Rue Mouffetrad</Adresse>
 < /Cenima>
```

Une interface événementielle telle que l'API SAX permet de créer des événements à partir de la lecture du document ci-dessus. Les événement générés seront :

Start document

Start element cineme

Start element nom

Characters Epée de bois

End element nom

Start element adresse

Characters 100.Rue Mouffetard

End element adresse

End document

Ainsi, une application basée sur SAX peut générer uniquement les éléments dont elle a besoin sans avoir à construire en mémoire une structure contenant l'intégralité du document.

L'API SAX définit les quatre interfaces suivantes :

- **DocumentHandler** possédant des méthodes renvoyant des événements relatifs au document :
    - startDocument() renvoyant un événement lié à l'ouverture du document

- startElement() renvoyant un événement lié à la rencontre d'un novel élément
- characters() renvoyant les caractères rencontrés
- Endelement()un événement lié à la fin d'un élément
- endDocument()renvoyant un événement lié à fermeture du document

- **ErrorHandler** possédant des méthodes renvoyant des événements relatifs aux erreurs ou aux avertissements.
- **DTDHandler** renvoie des événements relatifs à la lecture de la DTD du document XML
- **Entityresolver** permet de renvoyer une URL(Uniform Resource Locator) lorsqu'une URL est rencontrée.