REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

N° d'ordre : .....

UNIVERSITE DE M'SILA

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

Département d'Informatique

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique
Filière : Informatique
Spécialité : Systèmes d'Informations Avancés

Par: RABHI Warda

SUJET

SYSTEME AUTEUR D'AIDE A LA CREATION DES OBJETS PEDAGOGIQUES TEXTUELS

Soutenu publiquement le : 06 /2015 devant le jury composé de :

……	**Université de M'sila**	**Président**
Mr .BRIK Mourad	**Université de M'sila**	**Rapporteur**
……	**Université de M'sila**	**Examinateur**
……	**Université de M'sila**	**Examinateur**

Promotion : 2014 /2015

Remerciement

Nos remerciement à Allah le tout puissant de nous avoir donné le courage jusqu'à l'achèvement de ce mémoire.

Au terme de ce travail, nous adressons notre profonde gratitude à Monsieur BRIK Mourad.

Nous vous remercions pour la gentillesse et la spontanéité avec lesquelles vous avez bien voulu diriger ce travail. Nous avons eu le grand plaisir de travailler sous votre direction ;

Un grande merci non moins reconnaissant à tous nos enseignants pour toutes les connaissances qu'ils nous ont inculquées tout au long des cinq années.

Nous remercions

A toute personne ayant participé de prés ou de loin dans l'élaboration de ce travail.

TABLE DES MATIÉRES

CHAPITRE 4 RÉALISATION ET EXPÉRIMENTATION

LISTES DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

CIMI : Computer Interchange for Museum Information

DCMI: Dublin Core Metadata Initiative

DL : Longueur du Document

DTD : Document Type Definition

EAD : Encoded Archival Description

EXIF : EXchangeable Image File

HTML: Hyper Text Markup Language.

IDF : Inverse Document Frequency

IPTC : International Press and Telecommunications Council

ISBD : International Standard Bibliographic Description

JPEG : Join Photographic Experts Group

LDA : Latent Dirichlet Allocation

LOM : Learning Object Metadata

MARC: Machine Readable Cataloging

MAS : Métadonnées d'Administration Systèmes

ME : Métadonnées Externes

MIE : Métadonnées d'indexation Explicites

MM : Métadonnées orientées Métiers

MODS : Metadata Object Description Schema

NCSA : National Center for Supercomputing Applications

NISO : National Information Standards Organisation

OCLC : Online Computer Library Center

RDF : Ressource Description Framework.

RI : Recherche d'Information

TF : Term Frequency

URI : Uniform Resource Identifier

OP : Objet Pédagogique

INTRODUCTION GENERALE

Les pratiques dans le monde éducatif ont beaucoup changé ces dernières années, notamment par l'utilisation des TIC[1] qui permettent le développement de ressources électroniques et leur utilisation lors d'activités d'apprentissage en présence ou à distance, mais l'usage des ressources pédagogiques pose des problèmes de recherche et de réutilisation de ce dernier.

Parmi les nombreux documents disponibles sur le web, nous nous intéressons au sous-ensemble de ceux que nous qualifions de pédagogiques parce qu'ils ont été créés ou qu'ils peuvent être utilisés pour l'enseignement (la formation, l'apprentissage.. .). La production de ces documents pédagogiques est longue et coûteuse et ne peut être rentabilisée qu'en les réutilisant le plus largement possible. Pour cela, il est nécessaire de les spécifier, de les structurer et de les l'indexation des objets pédagogiques pour une nécessité absolue si l'on veut les retrouver. Pour cela, il faut leur ajouter des informations de nature sémantique. Ces informations sont des métadonnées : données décrivant des données.

L'indexation des ressources via des métadonnées est une solution efficace des problèmes de réutilisation, visant à faciliter le partage et la réutilisation de celles-ci.

Problématique et objectif de notre travail

Compte tenu de la croissance exponentielle des ressources, documents et services disponibles sur le Web et de leur hétérogénéité, il est difficile de trouver les ressources pertinentes qui répondent à une requête posée par un utilisateur en vue de l'utiliser ou réutiliser.

Après l'étude faite sur l'état de l'art du sujet en question, nous avons fait ressortir les problématiques suivantes :

- Plusieurs travaux sont concentrés sur la composition des ressources pédagogiques, ces travaux ne prennent pas en compte l'aspect ressources pédagogiques, ils utilisent ces ressources comme des unités prêtes à être agrégées avec d'autres ressources, sans souci de leur préparation en vue de les retrouver et réutiliser.
- Peu de outils permettant la création des ressources pédagogiques plus au moins granulaires à partir des ressources web existant déjà (documents HTML). Nous

[1] TIC : Technologies de l'Information et de la Communication.

avons trouvé quelques outils de décomposition mais ils sont relatifs à des contextes particuliers.

Notre travail décrites les ressources pédagogiques par des métadonnées comme le standard de l'IEEE, le Learning Object Metadata (LOM). Ainsi, la description pourrait se faire avec une implémentation technique en fichiers XML, cela n'est possible qu'après avoir passé par certains étapes préliminaires telles que : les opérations d'analyse et de prétraitement sur les données textuelles visant à préparer les textes aux algorithmes d'apprentissage, alors notre travail regroupe toutes ces étapes y compris la classification proprement dite.

Organisation du mémoire :

Ce mémoire est structuré comme suit :

- **Introduction générale** : dans cette section nous présentons le contexte de nos études et la problématique traitée dans notre sujet de recherche, en l'occurrence le besoin d'un outil permettant l'indexation semi automatique des ressources pédagogiques.
- **Chapitre 1 Les métadonnées et le e-Learning :** Dans ce chapitre, nous avons évoqué d'une manière détaillée les notions générales liées aux métadonnées et le e-Learning.
- **Chapitre 2 Extraction automatique de termes-clés :** Ce chapitre présente les principales méthodes d'extraction automatique de termes-clés.
- **Chapitre 3 Processus d'indexation :** Dans ce chapitre, nous nous intéressons aux différentes notions liées à l'indexation
- **Chapitre 4 Réalisation et expérimentation :** La quatrième partie de ce mémoire porte sur l'étude conceptuelle.

1.1. Introduction

Le processus d'indexer des ressources pédagogiques via les métadonnées est une tache importante visant a facilité la recherche et la réutilisation de ces ressources. Les métadonnées doivent suivre cependant des normes et des standards bien précis pour permettre de retrouver facilement ces ressources pédagogiques. Il s'agit aussi de décrire les principales métadonnées (Dublin Core, Dublin Core Education, Learning Object Meta Data…) et d'évoquer rapidement les problèmes d'implémentation.

1.2. Généralités

1.2.1. Nécessité d'une indexation des ressources pédagogiques :

Selon la définition, très vaste, du LOM (Learning Object Metadata) d'IEEE2, une ressource pédagogique correspond à toute entité (numérique ou non) utilisée dans un processus d'enseignement, de formation ou d'apprentissage et qui est :

- disponible librement (web) ou vendue (consortium, campus virtuel…) ;
- réutilisable ;
- abordable, adaptable, composable, découvrable, durable, fiable, gérable ;
- interchangeable, évaluable, livrable, réutilisable ;
- décrite par des métadonnées.

1.2.2 Pourquoi indexer les ressources pédagogiques ?

Bien évidemment pour les retrouver , mais l'enjeu est plus important. En effet, le temps est révolu où un enseignant pouvait se glorifier de mettre son cours en ligne sur un site pour ses propres étudiants. Une ressource pédagogique est longue et coûteuse à produire. C'est pourquoi elle doit être développée en respectant des normes et des standards. La seule façon de la rentabiliser se trouve dans la vente et/ou la mutualisation. Cela nécessite de réfléchir à la fois au fond et à la forme, ce qui demande du temps. De telles ressources pédagogiques peuvent être disponibles librement sur le web, via des consortiums, des campus virtuels, des universités numériques thématiques. Leur réutilisation est importante, ce qui soulève un certain nombre de questions. Comment retrouver ces ressources ? Est-ce bien la dernière version de la ressource ? Est-ce que ce cours a déjà été produit ? Comment assembler plusieurs ressources automatiquement ? La présentation peut-elle être différente selon les utilisateurs ?

D’où l’utilité de la description de la ressource pédagogique, à travers l’indexation, qui se situe à plusieurs niveaux. [7]

1.3 .Les métadonnée :

Les métadonnées, ensembles de données structurées sur les données, représentent la "colonne vertébrale" d’un paysage numérique de qualité [9]. Elles donnent un sens structurel et cognitif à l’information, participent à l’interopérabilité des applications et à la pérennité des ressources, et sont un élément incontournable du Web de demain. Les acteurs du Web et de la gestion de l’information dans tous les domaines s’y impliquent ; des standards d’implémentation s’imposent, les réalisations, locales, nationales et internationales se multiplient, notamment dans le cadre des systèmes d’information scientifiques et techniques. De multiples questions stratégiques et techniques se posent en parallèle.

1.3.1. Qu'est-ce qu'une métadonnée ?

Le terme "meta" vient du grec et dénote quelque chose de nature plus élevée ou plus fondamentale. Les métadonnées sont littéralement "des données relatives à d’autres données" (data about data : données sur des données). Toutefois, l'importance des métadonnées aujourd'hui mérite quelques précisions dans leur définition. Nous citons ici la définition donnée par le National Information Standards Organisation (NISO), dans un article paru en 2004, intitulé "Understanding Metadata" : « Une métadonnée (du Grec, “méta”, ce qui dépasse, englobe) est une donnée à propos d’une autre donnée. En sciences de l’information, les métadonnées sont des ensembles de données structurées décrivant des ressources physiques ou numériques, ou, sur un plan plus fonctionnel, “de l’information structurée qui décrit, explique, localise la ressource et en facilite la recherche, l’usage et la gestion” ». [8]

1.3.2. Intérêts des métadonnées

Le terme « métadonnée » est utilisé depuis longtemps dans certains domaines d’activité, comme la description des documents géographiques, la gestion des ressources images et multimédias ou les bases de données. Ce concept est aussi au coeur de certains métiers comme ceux des bibliothèques et de l’archivistique. Il concerne aujourd’hui tous les acteurs de l’environnement numérique. Les fonctions des métadonnées peuvent en effet être déclinées en six groupes [9] :

1. Améliorer la recherche d'information et la « découverte » des ressources ; avec des métadonnées descriptives du contenu : titre, résumé, mots-clés, classement ... Cet objectif, premier dans l'histoire des métadonnées, peut inclure d'autres éléments comme la qualité du document.

2. Gérer les ressources, grâce à deux grands sous-ensembles de métadonnées :

- D'une part, des métadonnées administratives, portant sur la propriété intellectuelle, la responsabilité, les droits d'usage et les sources utilisées.
- D'autre part, des métadonnées instancielles, ou techniques et de structure, regroupant les caractéristiques physiques ou informatiques, telle que le format, les techniques détaillées de documents particuliers comme les images, dates significatives dans le cycle de vie, structure ou place dans une hiérarchie, logiciels de consultation...

3. **Gérer les « archives »**, au sens du Record management ; les archives sont ici l'ensemble des documents utiles à court et moyen terme. Dans ce processus, il s'agit « d'identifier, authentifier, localiser, et contextualiser les données, ainsi que les personnes, les processus et les systèmes qui les créent, les gèrent ou les utilisent et les politiques qui les régissent » pour garantir la qualité, la fiabilité, l'accessibilité et la pérennité des ressources [14].

4. **Faciliter le partage de données et leur réutilisation** : cette fonction, importante pour limiter les coûts, passe par l'amélioration de l'interopérabilité au travers de standards, ainsi que par la présence d'informations contextuelles pour guider l'interprétation des données. Dans le cadre du Web sémantique, l'interopérabilité entre ensembles de données concerne d'abord les machines.

5. **Participer à la pérennité des ressources numériques**, qui garantit que « l'essence du contenu est accessible pour toujours ». On utilise pour cela des métadonnées de préservation qui décrivent « entre autres les actions réalisées en vue d'assurer la pérennité et l'accès pérenne, telles les migrations ou contrôles d'intégrité des fichiers » [14].

6. **Décrire les utilisateurs pour gérer les accès**, leur permettre des personnalisations de consultation, analyser les comportements d'usage.

1.4. Typologie des métadonnées

Le terme « métadonnée » regroupe plusieurs typologies. Cette diversité est relative à plusieurs critères. Parmi ces derniers, "la source" constitue le critère le plus utilisé pour classifier les métadonnées. [18] distingue des métadonnées internes aux ressources qu'elles décrivent et d'autres externes.

- **Métadonnées internes** : Elles sont intégrées dans la ressource elle-même, de façon implicite ou explicite.
  - **Implicite** : Le logiciel génère automatiquement des informations sur le document.
  - **Explicite** : Sous forme de balisage de données : on inclut un ou plusieurs jeux de métadonnées dans la ressource.

Le document, s'il est déplacé, transporte automatiquement ses métadonnées avec lui. Cependant, il peut y avoir une différence de poids entre un fichier incluant ses métadonnées et un fichier leur faisant référence de manière externe [23].

Dans cette première catégorie, nous distinguons entre autre:

- **Les métadonnées d'administration systèmes (MAS),** destinées aux robots Pour la gestion des serveurs web et des sites web. Parmi les balises représentatives, nous citons <http-equiv>, <refresh>.
- **Les métadonnées d'indexation textuelles (MIT)**, insérées par le concepteur du site web ou par des indexeurs comprenant essentiellement les balises META et respectant les 15 champs bibliographiques du Dublin Core (DC).
- **Les métadonnées d'indexation iconographiques (MII**) portées par la balise <IMG> et ses attributs **"Src"**, **"Alt"**, **"Title"** et **"LongDesc"** pour la description des images.
- **Les métadonnées d'indexation explicites (MIE)** qui sont des commentaires (non affichés par les navigateurs) insérés dans le code source pour expliciter la démarche et qui servent aussi à l'indexation des pages web.

- **Les métadonnées externes (ME) :** Elles sont contenues soit dans une notice séparée du document, comme c'est le cas d'une notice d'un catalogue de bibliothèque ; soit dans un thesaurus ou une base de données externe via des outils pour constituer un réservoir de métadonnées pour le web sémantique. Le but est d'assister l'usager dans sa recherche d'information et sa navigation [18].

Si on bouge la ressource ou on l'utilise en dehors de son cadre de référence, on Perd les métadonnées. Dans le cas de conservation à long terme, on doit s'assurer de conserver les deux parties [18].

Une autre catégorie de métadonnées utilisées pour décrire un type particulier de ressources (audiovisuelles par exemple), ou bien liées à un domaine particulier tel que l'éducation, ce sont les " **Métadonnées dites « métiers » (MM)",** qui sont des références bibliographiques liées à :

- Un domaine scientifique, comme le LOM pour l'éducation.
- Une profession comme le MARC ou l'ISBD pour les bibliothécaires.
- Un secteur institutionnel comme l'EAD, pour les archives, Nous détaillerons dans ce qui suit ces différents types :

1.4.1. Métadonnées d'administration systèmes

Elles sont insérées dans l'entête d'une page HTML, pour déterminer l'activité des robots des moteurs de recherche sur le site visité, en lui ordonnant ou lui interdisant d'indexer des documents et de suivre les pages liées ou encore en lui spécifiant la fréquence de ses visites. Elles suivent l'une des syntaxes suivantes :

<META NAME ="Nom du tag" CONTENT="Attribut">…… (1)

<META HTTP-EQUIV="Nom du tag" CONTENT="Attribut"> …… (2)

Les métadonnées de type ***<META NAME>*** les plus utilisées sont les suivantes :

<META NAME ="Robot" CONTENT="Attribut">

L'élément ***CONTENT="Attribut"*** de cette balise peut prendre plusieurs valeurs :

<META NAME ="Robot" ***CONTENT="All">*** : Spécifie que le robot peut indexer toute la page.

<META NAME ="Robot" ***CONTENT=" index / noindex">*** : Elle spécifie que le robot peut ou ne doit pas indexer la page.

<META NAME ="Robot" ***CONTENT=" follow / nofollow">*** : Elle autorise ou pas au robot de suivre les liens de la page.

<*META NAME ="Robot"* ***CONTENT=" none"***> : Elle empêche le robot d'aller plus loin. Les balises **<*META HTTP EQUIV*>** sont moins nombreuses et n'apportent pas le même type d'information. Leur rôle est :

• D'indiquer les normes et les caractères utilisés dans la page afin d'éviter les problèmes de caractères accentués avec d'autres codes sur divers systèmes (Mac/Windows). La syntaxe utilisée est la suivante :

<META HTTP EQUIV = " Content-type" CONTENT = " Type de langage; Charte">

Exemple :

<META HTTP EQUIV = " Content-type" CONTENT = "Text/HTML; charset = iso-8859-1">

• Indiquer au logiciel de navigation de changer de page après une certaine durée d'affichage.

<META HTTP EQUIV = " Refresh " CONTENT=" X; URL = adresse ">

Ainsi, la page est rafraîchie après "x" secondes.

Si l'URL est absent, c'est cette même page qui est rafraîchie.

Si l'URL existe, alors la page indiquée dans *URL="adresse"* sera chargée.

Il existe d'autres balises **<*META HTTP EQUIV*>** destinées au navigateur et/ou robots d'indexation, parmi lesquelles nous citons [18] :

<META http-equiv = " Expires " CONTENT = "date"> :
Elle spécifie la "date à laquelle la page expire", après laquelle le navigateur recharge automatiquement le document depuis le serveur, s'il provient d'un serveur Proxy.
<META http-equiv = " Rev" CONTENT = "administrateur@email"> :
Indique l'e-mail du Webmaster du site ou l'e-mail de d'Administrateur du serveur.
<META http-equiv = "Revisit-after" CONTENT = " x days "> :
le nombre de jours.

1.4.2. Métadonnées d'indexation textuelles

Elles permettent de décrire n'importe quel type de document textuel. Elles sont insérées par le concepteur d'un site Web par exemple, ou par des indexeurs comme dans

le cas des documents papiers, respectant essentiellement l'ensemble des balises META définies par un standard de description de ressources. La description doit tenir compte de trois aspects :

- Le contenu du document (titre, sujet, description, …).
- La matérialisation du document (type, format, date de création et de mise à jours, …).
- La propriété intellectuelle du document (auteur, contributeur, …).

L'un des standards les plus importants couvrant ces trois niveaux de description est le "Dublin Core", qui sera décrit en détail dans la partie suivante (les standards existants).

1.4.3. Métadonnées iconographiques

Les métadonnées sont particulièrement importantes pour les ressources visuelles qui, sans elles, peuvent demeurer pratiquement inexploitables et impossibles à retrouver. Les utilisateurs dépendent en effet des informations ajoutées aux images ou vidéos pour effectuer des recherches pertinentes et précises. Les métadonnées aident alors les utilisateurs à découvrir l'existence de ressources et la nature de ce qu'ils recherchent. Les informations ajoutées à une ressource servent aussi à évaluer la ressource, à porter un jugement sur celle-ci, et à la comparer à d'autres ressources [23].

Dans le cas d'une page web, un certain nombre d'éléments peuvent ne pas être accessible au navigateur ou à l'agent utilisateur. Les images, par exemple, ne sont bien évidement pas "lues" par un navigateur vocal ou par un utilisateur en bas débit qui décide alors de désactiver les images. Cela peut-être le cas des applets, des objets, des images ou de tous les éléments graphiques. Pour que les utilisateurs ne perdent pas pour autant l'information, on fournit à l'agent utilisateur une alternative textuelle, sous forme d'une métadonnée, qui a pour but d'indiquer à l'utilisateur le contenu de l'applet, de l'objet ou de l'image. La syntaxe utilisée est la suivante :

```
<IMG SRC = "Source de l'image" ALT = "Texte alternatif" TITLE = "Titre de l'image">
```

L'attribut SRC : indique l'endroit où l'image est stockée.

L'attribut TITLE : est généralement affiché dans les navigateurs graphiques sous forme d' infobulle (tooltip).

L'attribut ALT : sert à fournir une description alternative destinée aux moteurs de recherche, La spécification 4.0 de HTML préconise d'afficher cette description

uniquement lorsque l'image ne peut pas être affichée (par choix de l'utilisateur, limitation du logiciel ou problème de chargement).

Exemple :

<**IMG SRC** ="images/logo.gif" **ALT** ="Logo de l'INI" **TITLE** = "Institut National de formation en Informatique"/>

Le contenu du ALT ne doit pas dépasser un certain nombre de caractères (60 caractères recommandé par la norme HTML). Si l'objet, par contre, fournit de nombreuses informations, qu'il est impossible de le résumer dans un ALT, on utilise alors en complément l'attribut LONGDESC qui fait appel à un fichier de description plus complet.

Exemple :

<**IMG SRC** ="images/logo.gif" **ALT** ="Logo de l'INI" **TITLE** = "Institut National de formation en Informatique" **LONGDESC** = "Description.txt"/>

Le balisage des images présente toutefois deux limitations majeures [23] :

- Il ne peut se substituer à la description à l'aide de métadonnées stockées dans une base de données. Seules les bases de données permettent de gérer d'importantes quantités d'images et de rechercher efficacement dans un vaste ensemble.
- Tous les programmes de manipulations d'images ne sont pas capables de lire ou même de préserver les métadonnées incluses. On peut considérer que le fait de ne pas être capable d'afficher les métadonnées incluses dans une image est supportable.

1.4.4. Métadonnées d'annotations ou de commentaires

Une annotation est une information graphique ou textuelle attachée à un document [15]. C'est une note critique ou explicative accompagnant un document.

Desmontils & Jacquin présentent les différentes dimensions d'une annotation suivant son utilisation, le formalisme dans lequel elle est décrite et son rôle. Dans le cadre du Web sémantique, on parle d'annotations sémantiques (dans le sens formel). Dans ce cadre, on essaie de représenter le contenu du document par une description formelle. Annoter des documents, c'est les décrire par des méta-données. [27]. Dans le cas d'une page Web, en plus des éléments descriptifs (catalogage), en code HTML, tout les

caractères qui se trouvent derrière un point d'exclamation "**!**" sont invisibles sur un navigateur .. (donc non-visible pour l'internaute). Cela s'appelle une ligne de commentaire, elle est utilisée pour donner des indications dans le code d'une page afin d'en faciliter la mise à jour et expliquer le déroulement d'un programme informatique. C'est un endroit « idéal » pour insérer des mots clés supplémentaires, afin d'améliorer la densité, et le référencement. [17] Parmi ces métadonnées d'annotation, nous pouvons citer :

a. Titre de l'annotation
b. Nom de l'auteur
c. Son émail
d. Date de création
e. Catégorie de l'annotation (commentaire, correction, …etc.)
f. Typologie de l'annotation (privé ou public)
g. Mots clés associé

1.4.5. Métadonnées orientées "métiers"

Nombreuses sont les communautés qui s'intéressent aux métadonnées, à l'image des bibliothécaires, documentalistes, archivistes, conservateurs de musées, etc. Par ailleurs, les ressources décrites sont très variées : monographies, publications en série, articles, archives, pièces de musée, images, séquences audio ou vidéo, etc. On ne décrit pas toutes ces ressources de la même façon. Les standards concernant les métadonnées sont donc très nombreux et orientés "métiers". [18]

A titre d'exemple, on peut citer :

- .Le format MARC (Machine Readable Cataloging) : Pour la description des ouvrages.
- ISBD (International Standard Bibliographic Description), pour la description des publications en série
- DEWEY Decimal Classification System : Pour la classification décimale des ouvrages.
- EAD (Encoded Archival Description) : Pour la description des archives.
- CIMI consortium (Computer Interchange for Museum Information) : Pour la description des ressources museographiques et/ou iconographiques.

- RKMS (RecordKeeping Metadata Schema) : Pour la description des ressources audio.
- MPEG-7 (Multimedia Content Description Interface) : pour la description des objets multimedia.
- LOM (IEE – Learning Object Metadata) : Pour la description des ressources liées à l'éducation.

1.4.6. Métadonnées informatiques

Ce sont des métadonnées contenues de façon implicite dans les objets informatiques. Patrick Peccatte [23] donne plusieurs exemples de ce type de métadonnées, qu'il répartit comme suit :

- Métadonnées contenues dans l'URL d'une ressource : Comme exemple, considérons la ressource suivante : http://www.ini.dz/formation/iside.htm
  Cette ressource contient plusieurs métadonnées :
  - Protocole http.
  - Top level domain : com
  - Type de la ressource : page Web.
  - Sujet : à priori, elle traite de la formation ISIDE.

- Métadonnées contenues dans les noms informatiques des fichiers, et plus généralement, toutes les informations fournies par les systèmes d'exploitation, comme le chemin d'accès, nom, extension, taille, attributs, date de création, date de modification, propriétaire, droits d'accès, etc.
- Les champs <title>et <meta>des fichiers HTML ;
- Propriétés des documents MS Office (Word, Excel, etc.) : 25 éléments dont : Titre, Auteur, Sujet, Mots-clés, Commentaires, Responsable, Société, Catégorie, etc.
- Propriétés des documents OpenOffice.org : 25 éléments dont Titre, Description, Sujet, Mots-clés, Créateur initial, etc.
- Informations sur les documents PDF : 9 éléments, dont : Titre, Auteur, Sujet, Mots-clés, Créateur, Producteur, etc.
- Champs IPTCdes images JPEG/TIFF : 33 éléments, dont : Titre, Source, Crédit, Copyright, Statut éditorial, Priorité, Catégorie, Mots-clés, etc.

- Champs EXIF des images JPEG : 30 éléments dont : Fabricant de la caméra, Modèle, Orientation, Temps d'exposition, Résolution en largeur, Résolution en hauteur, etc.
- Champs ID3 des fichiers MP3 : 74 éléments organisés en frames, parmi lesquels nous citons : Titre, Compositeur, Auteur du texte, Durée, Copyright, etc.
- Métadonnées spécifiques à chaque plate-forme :
  - Macintosh : Famille (Essentiel, Important, En cours, Personnel, etc.) et Commentaires
  - Windows 2000 : Propriétés associées à un fichier quelconque (Titre, Sujet, Catégorie, Mots-clés, etc.)

- Estampillage électronique : Dans le but d'authentifier un document (garantie de non falsification) et prouver l'appartenance d'une oeuvre à son propriétaire, en utilisant du Filigrane, tatouage, estampillage, etc. Dans le cas d'un document électronique, il s'agit d'une insertion d'informations numériques dans les fichiers binaires que sont les images, sons, vidéo sous forme de métadonnées.
- Stéganographie : C'est la science qui consiste à cacher de l'information dans un quelconque medium de façon à ce que seul un utilisateur muni du secret adéquat puisse retrouver cette information.

1.4.7. Analyse et autres typologies

Selon la manière avec laquelle elles sont associées au document numérique, les métadonnées peuvent être classées en deux grandes catégories, comme décrit précédemment : les métadonnées internes et les métadonnées externes. D'autres typologies existent et prennent en compte d'autres critères pour classifier les métadonnées. Anne J. Gilliland-Swetland donne un résumé intéressant sur ces différentes typologies, dans son article : *Introduction to Metadata* [16], où elle répartie les métadonnées en s'appuyant sur sept critères : la source, le mode de création, la nature, le statut, la structure, la sémantique et le niveau. Emmanuël Colinet et Inge Alberts [8] ont repris ces différents attributs auxquels ils ont rajouté un huitième attribut : La granularité.

Les différentes typologies :

La source : Cet attribut peut prendre deux valeurs, comme mentionné au début de cette section (Typologie des métadonnées) :

- Métadonnée interne, générée au moment de la création du document ou sa numérisation par l'auteur ou le créateur de la ressource en général.
- Métadonnée externe, créée plus tard, par une personne autre que l'auteur du document [15].

Le mode de création : Nous distinguons principalement deux modes de création :

- Automatique : Les métadonnées sont générées automatiquement par le système informatique : indexation et condensation automatiques. Elles sont regroupées par exemple dans des fichiers logs.
- Manuel : Les métadonnées sont générées à partir de schémas de métadonnées maison ou standardisés, tel que le Dublin Core ou le LOM.

La nature :

- Non spécialisée : Les métadonnées sont créées par des non spécialistes du domaine.
- Spécialisée : Les métadonnées sont générées par des bibliothécaires, archivistes ou professionnels de l'information.

1.5. Schémas des métadonnées

1.5.1. Dublin core (DC)

1.5.1. 1.Origine du Dublin Core

En Mars 1995 s'est tenu un workshop sur les métadonnées, parrainé par Online Computer Library Center (OCLC) et le National Center for Supercomputing Applications (NCSA*)*, rassemblant 52 chercheurs et professionnels des bibliothèques, de l'informatique, et des spécialités connexes, pour faire avancer l'état de l'art dans le développement des descriptions de ressources électroniques. Les buts de ce workshop incluaient une compréhension commune des besoins, des points forts, des défauts, et des solutions à proposer; et l'atteinte d'un consensus sur un ensemble d'éléments de métadonnées pour décrire des ressources d'informations. Depuis qu'Internet contient plus d'information complète que de simples résumés professionnels, les indexeurs et catalogueurs tentent de gérer cela en utilisant les méthodes et systèmes existants: Il était évident qu'une alternative pour obtenir des métadonnées utilisables pour des ressources électroniques doit donner aux auteurs et aux fournisseurs d'information un moyen permettant de décrire les ressources eux-mêmes, sans formation intensive et spécifique

préalable. C'est pour atteindre ce but, que la tâche majeure du workshop sur les métadonnées, était d'identifier et de définir un ensemble simple d'éléments pour décrire des ressources électroniques diffusées (en réseau). La première version de la synthèse des travaux du workshop établissait un ensemble minimal de treize éléments de métadonnées qui fut nommé : ensemble d'éléments de core de métadonnées de Dublin (ou plus simplement Dublin Core). La liste actuelle a été établie en décembre 1996, et comporte 15 éléments. Aujourd'hui, le Dublin Core est maintenu par le **Dublin Core Metadata Initiative** (**DCMI)** et donne lieu à une conférence annuelle au contenu riche et varié. Le DCMI comporte de nombreux groupes de travail, comme le groupe « éducation » qui tient une conférence commune avec le groupe LOM. [5]

1.5.1.2. Intérêts et limites

Dublin Core fait l'objet d'un large consensus et d'une large utilisation aujourd'hui grâce aux atouts suivants :

- Sa création dans un contexte international et multidisciplinaire ;
- Sa sémantique simple et "commune", facilement compréhensible, particulièrement pour les éléments de base ;
- Son extensibilité (compatible avec d'autres jeux d'éléments, évolutivité) et sa flexibilité (grande souplesse d'implémentation) ;
- Son adoption dans différents domaines, métiers et pays, et dans des applications non prévues initialement ou des domaines industriels connexes ;
- Son évolutivité au travers de groupes de travail ouverts ;
- La volonté du DCMI, de diffuser et faire adopter ce modèle ; le site officiel, www.dublincore.org, est très riche en tutoriels et recommandations, exemples, modèles et outils, et reflète bien l'activité de ce groupe d'acteurs ;
- La normalisation des 15 éléments de base à partir de 2003 par l'ISO (norme 15836-2003). Cependant, on peut résumer au passif du Dublin Core :
- Son côté généraliste et incomplet, nécessitant souvent des extensions,
- Sa relative jeunesse et le fait qu'il évolue encore (bien que les éléments de base semblent être « gravés dans le marbre »).

1.5.1.3. Quelques principes du Dublin Cor

Les auteurs du Dublin Core ont établi une liste de principes qui devaient guider davantage le développement de l'ensemble des éléments de métadonnées. Ces principes

sont : propriété intrinsèque, extensibilité, indépendance de syntaxe, optionalité, répétabilité et modifiabilité. [5]

1.5.1.4. Liste des éléments du Dublin Core

Les éléments du DC peuvent être classés en trois groupes qui indiquent la classe ou le type d'information correspondante :

- Des éléments qui concernent principalement le contenu de la ressource : Titre, sujet, description, source, langue, relation et couverture.
- Des éléments liés à la propriété intellectuelle de la ressource : Créateur, éditeur, contributeur et droits.
- Des éléments correspondants surtout à la matérialisation de la ressource : Date, type, format et identifiant.

Élément	Élément (anglais)	Commentaire
1.(métadonnée)	Title	Titre principal du document
2. Créateur (métadonnée)	Creator	Nom de la personne, de l'organisation ou du service à l'origine de la rédaction du document
3. Sujet (métadonnée) ou mots clés	Subject	Mots-clefs, phrases de résumé, ou codes de classement
4. Description (métadonnée)	Description	Résumé, table des matières, ou texte libre. Raffinements : table des matières, résumé
5. Éditeur	Publisher	Nom de la personne, de l'organisation ou du service à l'origine de la publication du document
6.Contributeur	Contributor	Nom d'une personne, d'une organisation ou d'un service qui contribue ou a contribué à l'élaboration du document. Chaque contributeur fait l'objet d'un élément Contributor séparé
7. Date (métadonnée)	Date	Date d'un évènement dans le cycle de vie du document
8. Type de ressource	Type	Genre du contenu
9. Format	Format	Type MIME, ou format physique du document

10. Identifiant de la ressource	Identifier	Identificateur non ambigu : il est recommandé d'utiliser un système de référencement précis, afin que l'identifiant soit unique au sein du site, par exemple les URI ou les numéros ISBN. Raffinement : Is Available At
11. Source	Source	Ressource dont dérive le document : le document peut découler en totalité ou en partie de la ressource en question. Il est recommandé d'utiliser une dénomination formelle des ressources, par exemple leur URI
12. Langue (métadonnée)	Language	Langage(s) du contenu intellectuel de la ressource. Si approprié, le contenu de ce champ devrait correspondre à la norme RFC 1766.
13. Relation (métadonnée)	Relation	Lien avec d'autres ressources. De nombreux raffinements permettent d'établir des liens précis, par exemple de version, de chapitres, de standard, etc.
14. Couverture (métadonnée)	Coverage	Couverture spatiale (point géographique, pays, régions, noms de lieux) ou temporelle
15.Droits (métadonnée)	Rights	Droits de propriété intellectuelle, Copyright, droits de propriété divers

Table 1.1 Liste des éléments et raffinements

1.5.1.5. Dublin Core qualifié

Les quinze éléments de base sont considérés comme un dénominateur commun et constitue ce que l'on appelle « Dublin Core simple ». Dans la plupart des cas, ces éléments sont insuffisants pour décrire une ressource spécifique à un domaine particulier avec précision. C'est pour cela que dès le départ, les éléments de base ont été étendus (ou précisés) par un ensemble d'autres termes, parfois appelés « qualifiers ». Deux classes de « qualifiers » sont reconnues : [5]

- Les « raffinements d'éléments » qui rendent plus spécifique le sens d'un élément.
- Les « schémas d'encodage » ou vocabulaires contrôlés.

Les éléments supplémentaires et « raffinements d'éléments »

Le DCMI a défini trois éléments supplémentaires et une trentaine de raffinements d'éléments.

- Les éléments supplémentaires : **Audience, Provenance et RightsHolder** étendent l'ensemble des quinze éléments de base.
- Les raffinements d'éléments précisent le sens d'un élément existant :

Exemple :

L'élément **« is version Of »** qualifie l'élément de base **« Relation ».** Il spécifie que la ressource décrite est une version, édition, adaptation de la ressource référencée ici (modification du contenu intellectuel).

Les schémas d'encodage (vocabulaires contrôlés)

Chaque élément, ou raffinement d'élément, peut disposer d'un ou plusieurs schémas d'encodage. Le DCMI, soucieux de ne pas réinventer la roue, fait référence à des schémas existants et en a inventé un, DCMIType, pour décrire les types (logiques et non physiques) d'objets, d'autres, DCMIBox et DCMIPoint, pour délimiter géographiquement un objet ou
lieu et DCMIPeriod pour situer un événement dans le temps.

Le vocabulaire DCMI Type

Le vocabulaire DCMI Type contient une liste de différents types de ressources :

Type d'objet	Description
Collection	Groupe de documents (ou ressources)
Dataset	Base de données, liste, tableau
Event	Evénement ponctuel (conférence…)
Image	Image, photo, (tous formats, générique)
Interactive Resource	Objet qui demande l'interaction de l'utilisateur
Moving Image	Image animée, vidéo…
Physical Object	Un objet lui-même (et non son image)
Software	Logiciel disponible pour installation sur une machine

Table 1.2 Le vocabulaire DCMI Type. [5]

1.5.2. Learning Object Meta data (LOM)

IEEE a développé en 2002 un standard de description des ressources pédagogiques, qu'il a appelé LOM (Learning Object Metadata). Celui-ci permet de décrire les outils

éducatifs, notamment logiciels de e-Learning. Certains éléments de métadonnées sont générés au moment de la création de la ressource. Par ailleurs, l'utilisation de la ressource génère de nouvelles métadonnées sur le processus d'apprentissage (comportement de l'utilisateur, enregistrement des résultats de tests...) [16].

Le LOM propose 45 éléments descriptifs de premier niveau (au total, il compte 79 éléments), regroupés dans 9 catégories. Le tableau suivant décrit ces catégories :

Catégorie	Description
General	Regroupe l'information générale qui décrit la ressource dans son ensemble.
Life cycle	Décrit l'état passé et actuel de la ressource, et ceux qui ont modifié cette ressource durant son cycle de vie.
Metametadata	Décrit l'information spécifique aux métadonnées elles-mêmes. Cette catégorie décrit l'auteur, par exemple, des métadonnées.
Technical	Décrit les conditions techniques requises et les caractéristiques de la ressource tel que le format, la taille, etc.
Educational	Décrit les conditions techniques requises et les caractéristiques de la ressource tel que le format, la taille, etc.
Rights	Décrit les droits de propriété intellectuelle et les conditions pour utiliser la ressource.
Relation	Définit les relations entre la ressource et d'autres ressources précises. Pour définir plusieurs relations, chaque cible est décrite par une nouvelle instance de relation.
Annotation	Fournit des commentaires sur l'usage pédagogique de cette ressource, l'auteur et la date de création de l'annotation. Si plusieurs annotations sont associées à ressource, de multiples instances de cette catégorie peuvent être utilisées.
Classification	Décrit une ressource par rapport à un système de classification particulier. Pour utiliser de multiples classifications, il peut y avoir de multiples instances de cette catégorie.

Table 1.3 Les 9 catégories du LOM. [5]

L'analyse des éléments constitutifs du LOM permet de déduire que l'effort principal de ses concepteurs a porté essentiellement sur la détermination des métadonnées permettant une description efficace des objets, leur partage et leur réutilisation. Nous citons aussi, IMS-Metadata qui est un autre jeu de description des ressources éducatives. SCORM, Sharable Content Object Reference Model, une spécification permettant de créer des objets pédagogiques structurés, interopérables, durables et réutilisables, intègre ces deux jeux.

1.5.2.1 Exemple d'utilisation du schéma LOM

<**lom**>
<**general**>
<title>
<langstring xml:lang="fr"> Initiation à Microsoft Word </langstring>
</title>
<language> fr </language>
<description>
<langstring xml:lang="fr">Un support de cours présentant les fonctionnalités avancées de Microsoft Word, c'est une ressource pédagogique pour l'enseignement de l'informatique de base en ligne
</langstring>
</description>
</**general**>
</**lom**>

1.6. L'E-Learning

On parle de l'environnement de travail collaboratif pour le monde professionnel et le e-Learning pour le monde de la formation. Le e-Learning prend en considération un processus d'apprentissage à distance s'appuyant sur des ressources multimédias, qui permet à une ou plusieurs personnes de se former à partir de leur ordinateur. Notons aussi que le domaine du travail collaboratif compte le e-Learning comme l'un de ses moyens ou outils de collaboration.

Le but principal du e-Learning est bien d'améliorer la qualité de l'apprentissage et non de se substituer aux modes d'apprentissages traditionnels. La figure qui suit illustre les principaux composants d'un environnement d'apprentissage à distance utilisant les (TIC):

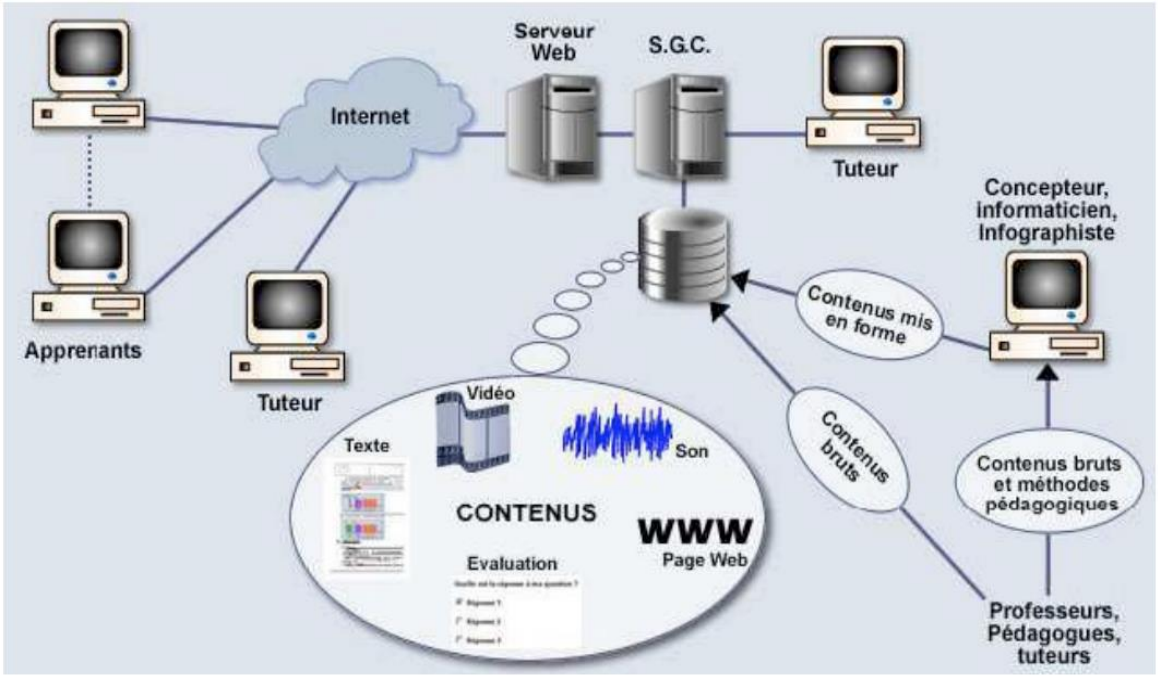

Figure 1.1 Le système d'apprentissage et ses composantes [2]

Le premier cours à distance fut le « cours par correspondance ». Ce fut l'avènement de l'enseignement à distance. Depuis, l'enseignement a connu un essor très important. Aujourd'hui, nous ne parlons plus que de « réseau, internet, web » et de « campus virtuel ». Le livre et la poste sont dorénavant remplacés par l'ordinateur multimédia qui permet la numérisation, le stockage et la restitution de l'information quelle que soit sa nature (texte,dessin, photo, vidéo et son).

Le développement des technologies de l'information et de la communication (TIC) offre une dimension importante de diffusion des savoirs et des connaissances distribués par le réseau d'Internet.

L'apprenant est aussi doté d'un espace d'apprentissage individualisé prenant en considération ses exigences, ses contraintes et ses spécificités personnelles. Nous ne saurons parler d'e-Learning sans que mention ne soit faite de ses différents constituants. Aussi nous vous entretiendrons, après les différentes définitions de ce concept, de ses caractéristiques, et de ses composantes. Ces dernières sont éloquentes du fait qu'elles embrassent tous les éléments du « système » e-Learning, en l'occurrence « les acteurs » et les nouveaux métiers auxquels il donne naissance ; « les contenus » et leur scénarisation, « les normes et standards » auxquels ce concept fait appel ; ainsi que « les plateformes » de formation à distance qu'il utilise pour la distribution des cours en ligne [20].

1.6.1. Positionnement du e-Learning dans une organisation [19]

D'un point de vue système d'information, une institution d'enseignement est une organisation pourvue d'une mission d'éducation ou formation, d'un ensemble d'acteurs ayant des responsabilités, représentant ses « compétences » et « son savoir-faire » donc son « métier », ainsi que d'un ensemble de moyens lui permettant d'accomplir ses tâches. Par analogie, un dispositif de formation à distance utilisant les TIC(s) peut être vu comme une institution virtuelle de formation faisant intervenir un ensemble d'acteurs coopérants entre eux dans le but d'assurer des enseignements.

1.6.2. Définitions du e-Learning

« L'e-Learning est l'utilisation des nouvelles technologies multimédias de l'Internet pour améliorer la qualité de l'apprentissage en facilitant d'une part l'accès à des ressources et à des services, d'autre part les échanges et la collaboration à distance ». (Union Européenne).

- **Définition selon [22]:**

« Le e-Learning est un mode d'apprentissage basé sur l'utilisation des nouvelles technologies, permettant l'accès à des formations en ligne, interactive et parfois personnalisées, diffusés par l'intermédiaire d'Internet, d'un Intranet ou toute autre média électronique, afin de développer les compétences, tout en rendant le processus d'apprentissage indépendant de l'heure et de l'endroit ».

Le e-Learning est supporté par des plates- formes logicielles qui lui sont dédiées communément appelées « Plate-forme de e-Learning ». Un environnement de e-Learning est alors formé par trois communautés d'acteurs interagissant entre eux via la

plate- forme, formant ainsi un SIC. Ces communautés sont les enseignants dont la mission est d'enseigner, les étudiants dont la mission est d'étudier et les managers (ou administrateurs) dont la mission est d'assurer toutes les fonctions de gestion liées aux communautés adhérentes, aux contenus et ressources et au système technologique qui supporte le e-Learning. Les liens de coopération peuvent être retrouvés à différents niveaux et pour divers objectifs.

1.6.3. Processus de e-Learning

Deux principales catégories sont liées à un environnement d'enseignement : Les processus liés à l'*activité d'enseignement* en général et ceux liés à l'*activité de* management . Comme exemple de la première catégorie, nous pouvons citer le déroulement d'une session de formation dans une matière ou domaine donné engageant des étudiants et un staff pédagogique ou encore la préparation collective d'un cours par une équipe d'experts. Comme exemple de la deuxième catégorie, les processus de gestion des adhérents à une plate- forme ou encore celui de la gestion de contenus.

1.6.4. Les caractéristiques du e-Learning

a. **L'accessibilité :** Le problème d'accès difficiles aux publics est résolu par la proposition de situations d'apprentissage-enseignement qui tiennent compte des contraintes individuelles des apprenants: telles que les contraintes spatiales, temporelles, technologiques, psychosociales, et socioéconomiques qui bloquent l'accès au savoir [14].

b. **La contextualisation :** L'e-Learning permet à l'apprenant d'apprendre dans son contexte immédiat. Le contact direct, permanent avec les différentes composantes de l'environnement est ainsi maintenu. Ceci facilite l'intégration des savoirs scientifiques aux savoirs pratiques [25].

c. **La flexibilité :** L'e-Learning offre la possibilité d'assouplir les organisations de formation et d'enseignement en utilisant des approches qui permettent à l'apprenant de planifier dans le temps et dans l'espace ses activités d'étude et son rythme d'apprentissage. Il permet de concevoir des activités offrant à l'apprenant des choix dans les contenus, les méthodes et les interactions et ainsi prendre en considération les caractéristiques individuelles de chaque apprenant telles que : L'individualisation ou l'adaptation des offres de formation à des besoins individuels.

d. La diversification des interactions : En rapprochant le savoir des apprenants, le e-Learning reconnaît que l'apprentissage ne résulte pas essentiellement de l'interaction entre le professeur et l'élève ou entre ce dernier et d'autres élèves mais aussi entre l'apprenant et l'ensemble des individus qui l'entourent.

e. La diversification du savoir : Dans tout processus d'enseignement, les contenus sont formalisés de sorte à transmettre des connaissances, des cognitions et une connaissance affective qui semblent s'imbriquer dans la situation elle-même. Mais la distance ne permet pas à l'enseignant de s'adapter aux représentations, à la pensée et aux démarches de l'apprenant, on parle ainsi de désaffectation du savoir.

1.6.5. Les standards d'e-Learning

Il s'agit de rendre accessibles des cours à partir d'environnements technologiques différents (des plateformes d'apprentissage à distance, par exemple) de manière à faciliter la mutualisation des ressources pédagogiques [20]. Tout l'intérêt du e-Learning repose sur le fait de permettre, aux différents acteurs du monde de l'éducation et de la formation :

- d'accéder, d'utiliser et de manipuler des objets pédagogiques.
- de partager et d'échanger ces objets entre différents environnements pédagogiques.
- de manipuler le contenu éducatif et les résultats des apprenants de façon standardisée et indépendante du contenu lui-même.
- Permettre à différents objets pédagogiques de travailler ensemble dans un environnement pédagogique ouvert et distribué.
- Intégrer les notions de sécurité et d'authentification nécessaires à la distribution et à l'utilisation des objets pédagogiques.

1.6.6. Les standards dédiés au domaine éducatif

Les normes existant dans le domaine pédagogique sont classé selon leur :
Caractérisation : L'approche documentaliste centrée sur l'objet pédagogique (cours, leçon..) qui s'intéresse à sa production et à son indexation et l'approche centrée « activité », initiée par Rob Kopper qui considère que l'activité doit être au centre du processus d'apprentissage et non pas la ressource. Nous avons choisi de travailler avec

la norme SCORM dans ce travail, et les autres normes seront données en annexe du mémoire.

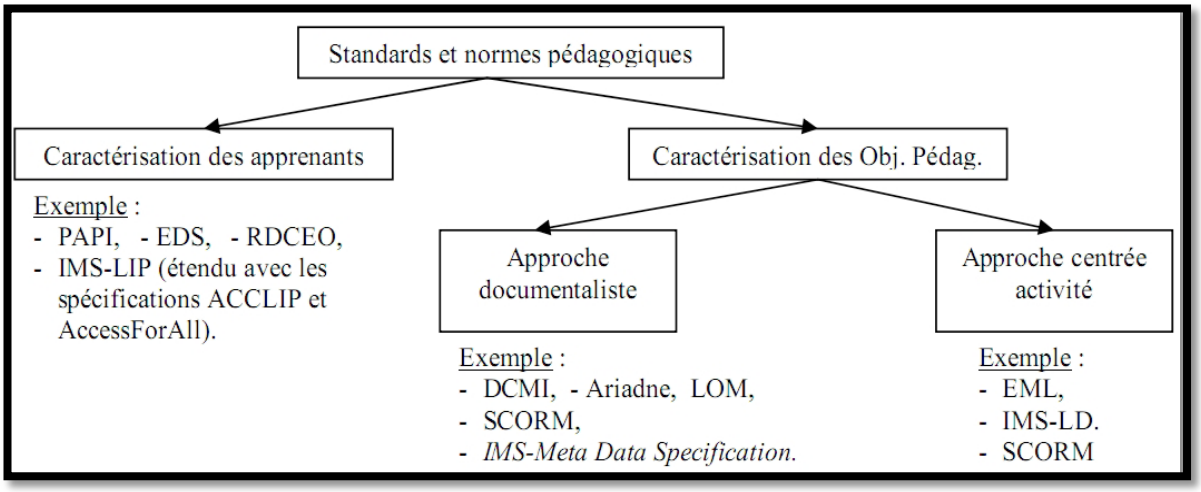

Figure 1.2 Schéma présentant les différentes normes et standards du e-Learning

1.7.Conclusion

Dans ce premier chapitre, nous avons évoqué d'une manière détaillée les notions générales liées aux métadonnées. Nous avons traité les différents types et standards de métadonnées existants et les outils permettant leur création, leur extraction et leur transfert.

Le e-Learning a tendance à spécifier les rôles et à bien les définir car il n'y a pas cette interaction directe qui n'est d'ailleurs remplaçable par aucun média. De plus, chaque rôle est susceptible d'être joué par une personne différente d'où la difficulté d'assurer un enseignement cohérent. On s'intéresse alors à la recherche d'information dans le domaine e-Learning en utilisant les métadonnées externes, car toute métadonnée informatique liée à une ressource représente de fait un index pour cette ressource.

2.1. Introduction

Les termes-clés sont les mots ou les expressions poly lexicales qui représentent le contenu principal d'un document. Ils sont utiles pour diverses applications, telles que l'indexation automatique ou le résumé automatique, mais ne sont pas toujours disponibles. De ce fait, nous nous intéressons à l'extraction automatique de termes-clés et, plus particulièrement, à la difficulté de cette tâche lors du traitement de documents appartenant à certaines disciplines scientifiques. L'extraction automatique de termes-clés consiste à extraire du contenu d'un document les unités textuelles les plus importantes, celles qui permettent de le résumer. Parmi les méthodes d'extraction automatique de termes-clés existantes, nous distinguons deux catégories : les méthodes supervisées et les méthodes non-supervisées.

2.2. Les méthodes d'extraction automatique de termes-clés

L'extraction de termes-clés est une tâche qui consiste à analyser un document et à en extraire les aspects importants. Alors que les méthodes de résumé automatique utilisent des phrases pour construire une vision synthétique du document, l'extraction de termes-clés se focalise sur les unités textuelles qui composent ces phrases. Un ensemble de termes-clés peut donc être perçu comme un résumé dont les points clés sont exprimés sans liaisons entre eux. Les unités textuelles sur lesquelles travaillent les systèmes d'extraction automatique de termes-clés sont appelées termes candidats. Ces derniers sont des mots ou des multi-mots (phrasèmes) pouvant être promus au statut de terme-clé. L'extraction de termes candidats est une étape préliminaire de l'extraction de termes-clés, que ce soit pour les méthodes non-supervisées ou supervisées. Cette étape est importante, car si certains termes-clés du document analysé ne sont pas présents dans l'ensemble des termes candidats, alors ceux-ci ne pourront pas être extrait. Hulth (2003) étudie trois méthodes d'extraction des termes candidats. L'une consiste à extraire les chunks nominaux 1, tandis que les deux autres extraient tous les n-grammes et les filtrent, soit pour retirer les termes contenant des mots outils dans le premier cas, soit pour ne retenir que les termes respectant certains patrons syntaxiques dans le second cas (usage des parties du discours). Dans ses expériences Hulth (2003) montre que l'extraction de termes-clés à partir de n-grammes filtrés avec les mots outils donne les meilleurs résultats parmi les trois méthodes qu'elle propose. Les travaux de Hulth

(2003) sont évalués avec un corpus dont les documents sont des résumés d'articles scientifiques. Cependant, dans d'autres domaines tels que la bio-médecine, la nature des termes à extraire n'est pas la même. En effet, ce sont les acronymes et les entités nommées (noms de protéines par exemple) qu'il est nécessaire d'extraire en tant que termes-clés (Nobata *et al.* 2008). Pour cela, l'extraction de termes candidats est spécifique au domaine d'application. Les méthodes d'extraction de termes-clés présentées dans cet article traitent des documents supposés sans spécificités particulières, les méthodes d'extraction de termes candidats sont donc les mêmes que celles expérimentées par Hulth (2003), mais il est envisageable de les adapter à des domaines présentant des spécificités particulières. Utilisés avec les méthodes non-supervisées, les termes candidats sont ordonnés selon un score d'importance obtenu soit à partir d'eux-mêmes, soit à partir de l'importance des mots qui les composent. Si une méthode s'appuie uniquement sur les mots, alors le score d'un terme candidat est généralement calculé en faisant la somme des mots qui le composent. Cependant, ceci n'est pas toujours juste, c'est donc un inconvénient important des méthodes travaillant sur les mots pour extraire les termes-clés. En effet, la sommation peut privilégier des termes qui contiennent beaucoup de mots non-importants vis-à-vis de termes contenant très peu de mots, mais importants. Utilisés dans les méthodes supervisées, les termes candidats sont classés en tant que termes-clés ou non termes-clés grâce à des méthodes de classification.

2.2.1. Méthodes non-supervisées

Les méthodes non-supervisées d'extraction de termes-clés ont la particularité de s'abstraire du domaine et de la langue des documents à analyser. Cette abstraction est due au fait que les termes candidats sont analysés avec des règles simples déduites à partir de traits statistiques issus seulement du texte analysé, ou bien d'un corpus de référence non annoté. De nombreuses approches sont proposées. Certaines se fondent uniquement sur des statistiques alors que d'autres les combinent avec des représentations plus complexes des documents. Ces représentations peuvent aller de groupes de mots sémantiquement similaires à des graphes dont les noeuds sont des unités textuelles (mots, expressions, phrases, etc.) liées par des relations de recommandation.

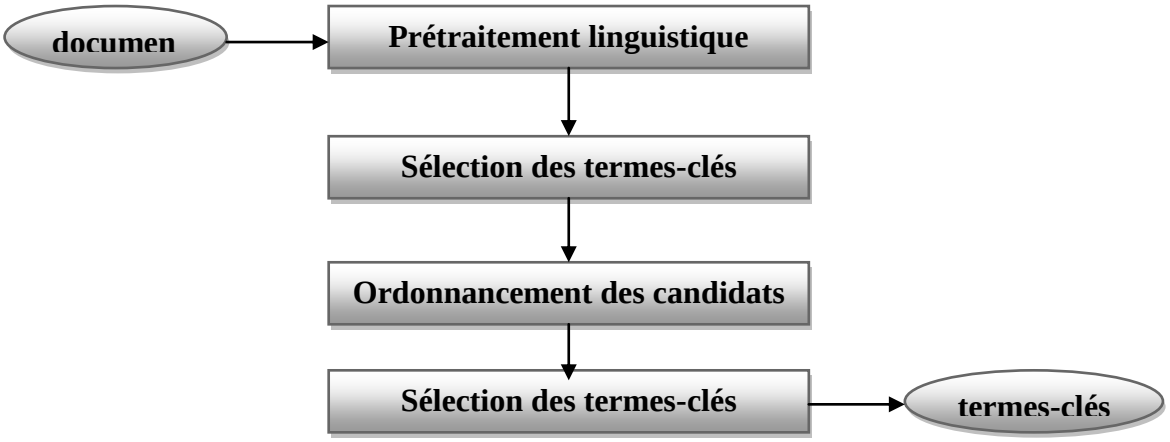

Figure 2.1 Les quatre principales étapes de l'extraction automatique de termes-clés

2.2.1.1. Approches statistiques

Plusieurs approches cherchent à définir ce qu'est un terme-clé en s'appuyant sur certains traits statistiques et en étudiant leur rapport avec la notion d'importance d'un terme candidat. Plus un terme candidat est jugé important vis-à-vis du document analysé, plus celui-ci est pertinent en tant que terme-clé.

TF-IDF (cf. équation 1) de Jones (1972) et Likey (cf. équation 2) de Paukkeri et Honkela (2010) sont deux méthodes qui comparent le comportement d'un terme candidat dans le document analysé avec son comportement dans une collection de documents (corpus de référence). L'objectif est de trouver les termes candidats dont le comportement dans le document varie positivement comparé à leur comportement global dans la collection. Dans les deux méthodes ceci s'exprime par le fait qu'un terme a une forte importance vis-à-vis du document analysé s'il y est très présent, alors qu'il ne l'est pas dans le reste de la collection.

$$\boldsymbol{TF - IDF(terme) = TF(terme) \log\left(\frac{N}{DF(\text{terme})}\right)} \qquad \textbf{(1)}$$

$$\boldsymbol{Likey = \frac{\text{rang}_{\text{document}}(\text{terme})}{\text{rang}_{\text{corpus}}(\text{terme})}} \qquad \textbf{(2)}$$

Dans TF-IDF, TF représente le nombre d'occurrences d'un terme dans le document analysé et DF représente le nombre de documents dans lequel il est présent, *N* étant le nombre total de documents. Plus le score TF-IDF d'un terme candidat est élevé, plus celui-ci est important dans le document analysé. Dans Likey, le rang d'un terme candidat dans le document et dans le corpus est obtenu à partir de son nombre d'occurrences, respectivement dans le document et dans le corpus de référence. Plus le rapport entre ces deux rangs est faible, plus le terme candidat évalué est important dans le document analysé.

Okapi (ou BM25) (Robertson *et al.* 1999) est une mesure alternative à TF-IDF. En Recherche d'Information (RI), celle-ci est plus utilisée que le TF-IDF. Bien que l'extraction automatique de termes-clés soit une discipline à la frontière entre le TAL et la RI, la méthode de pondération. Okapi n'a, à notre connaissance, pas été appliquée pour l'extraction de termes-clés. Dans l'article de Claveau (2012), Okapi est décrit comme un TF-IDF prenant mieux en compte la longueur des documents. Cette dernière est utilisée pour normaliser le *TF* (qui devient TF_{BM25}) :

$$\mathbf{Okapi(terme)} = \boldsymbol{TF_{BM25}}\mathbf{(terme)} * \mathbf{log}\left(\frac{\boldsymbol{N} - \boldsymbol{DF}\mathbf{(terme)} + \mathbf{0,5}}{\boldsymbol{DF}\mathbf{(terme)} + \mathbf{0,5}}\right) \qquad (\mathbf{3})$$

$$\boldsymbol{TF_{BM25}} = \frac{\boldsymbol{TF}\mathbf{(terme)} \times (\boldsymbol{k1} + \mathbf{1})}{\boldsymbol{TF}\mathbf{(terme)} + \boldsymbol{k1} \times \left(\mathbf{1} - \boldsymbol{b} + \boldsymbol{b} \times \frac{\boldsymbol{DL}}{\boldsymbol{DL}\mathbf{moyenn}}\right)} \qquad (\mathbf{4})$$

Dans la formule (4), $\boldsymbol{k1}$ et $\boldsymbol{b}$ sont des constantes fixées à 2 et 0, 75 respectivement. $\boldsymbol{DL}$ représente la longueur du document analysé et *DLmoyenne* la longueur moyenne des documents de la collection utilisée. Barker et Cornacchia (2000) estiment que les grands phrasèmes sont plus informatifs et qu'ils doivent être privilégiés. Pour cela, leur approche est très simple : plus un groupe nominal est long et fréquent dans le document analysé, plus il est jugé pertinent en tant que terme-clé de ce document. Cependant, pour éviter la répétition dans le texte, les auteurs des documents utilisent les même expression sous des formes alternatives (plus courtes, par exemple). La fréquence d'une expression ne reflète donc pas forcément sa fréquence réelle d'utilisation, car celle-ci est répartie dans les différentes alternatives. De ce fait, Barker et Cornacchia (2000) repèrent dans les groupes nominaux la tête nominale et utilisent en plus la fréquence de celle-ci. Tomokiyo et Hurst (2003) tentent de vérifier deux propriétés, en utilisant des

modèles de langue uni-grammes et n-grammes et en calculant leur divergence (Kullback-Leibler). Les deux propriétés qu'ils tentent de vérifier sont les suivantes:

- La grammaticalité : un terme-clé doit être bien formé syntaxiquement.
- L'informativité : un terme-clé doit capturer au moins une des idées essentielles exprimées dans le document analysé.

Pour un terme candidat donné, plus sa probabilité en passant du modèle uni-gramme généré à partir du document vers le modèle n-gramme généré à partir du même document augmente, plus il respecte la propriété de grammaticalité. De même, plus sa probabilité en passant du modèle n-gramme généré à partir d'un corpus de référence vers le modèle n-gramme généré à partir du document analysé augmente, plus le terme candidat est informatif.

La méthode que propose Ding et al. (2011) utilise TF-IDF comme indicateur de l'importance d'un terme-clé. Dans un ensemble, cette importance doit être maximisée pour chaque terme-clé, mais les auteurs estiment que ceci n'est pas suffisant. Comme Tomokiyo et Hurst (2003), ils définissent deux propriétés qui doivent être respectées :

- La couverture : un ensemble de termes-clés doit couvrir l'intégralité des sujets abordés dans le document représenté.
- La cohérence : les termes-clés doivent être cohérents entre eux.

La propriété de couverture est évaluée avec le modèle *Latent Dirichlet Allocation* (LDA) qui donne la probabilité d'un terme candidat sachant un sujet. La cohérence est évaluée pour chaque paire de termes-clés de l'ensemble avec la mesure d'information mutuelle. Ces deux propriétés sont définies comme contraintes que les auteurs utilisent avec une méthode de programmation par les entiers (technique d'optimisation), la maximisation de la pertinence de chaque terme-clé étant l'objectif à atteindre. Les traits statistiques utilisés dans les méthodes précédentes sont uniquement utilisés pour déterminer un score de pertinence des termes candidats en tant que termes-clés. Une donnée statistique non citée précédemment, mais pourtant récurrente dans les méthodes d'extraction de termes-clés, est la fréquence de co-occurrences entre deux phrasèmes (termes). Deux phrasèmes co-occurrent s'ils apparaissent ensemble dans le même contexte. La co-occurrence peut être calculée de manière stricte (les phrasèmes doivent être côte-à-côte) ou bien dans une fenêtre de mots. Compter le nombre de co-occurrences entre deux termes permet d'estimer s'ils sont sémantiquement liés ou non. Ce lien sémantique à lui seul ne peut pas servir à extraire des termes-clés, mais il permet

de mieux organiser les termes d'un document pour affiner l'extraction (Matsuo et Ishizuka, 2004; Liu *et al.* 2009; Mihalcea et Tarau, 2004).

2. 2.1.2. Approches par regroupement

L'objectif des approches par regroupement est de définir des groupes dont les unités textuelles partagent une ou plusieurs caractéristiques communes. Ainsi, lorsque des termes-clés sont extraits à partir de chaque groupe, cela permet de mieux couvrir le document analysé selon les Caractéristiques utilisées. Dans la méthode de Matsuo et Ishizuka (2004), ce sont les termes (phrasèmes) qui sont regroupés. Parmi ceux-ci, seuls les plus fréquents sont concernés par le regroupement. Celui-ci s'effectue en fonction du lien sémantique 4 entre les termes. Après le regroupement, la méthode consiste à comparer les termes candidats du document analysé avec les groupes de termes fréquents, en faisant l'hypothèse qu'un terme candidat qui co-occurre plus que selon toute probabilité avec les termes fréquents d'un ou plusieurs groupes est plus vraisemblablement un terme-clé. Dans l'algorithme Key Cluster, Liu *et al.* (2009) utilisent aussi un regroupement sémantique, mais dans leur cas ils considèrent les mots du document analysé et ils excluent les mots outils. Dans chaque groupe sémantique, le mot qui est le plus proche du centroïde est sélectionné comme mot de référence. L'ensemble des mots de référence est ensuite utilisé pour filtrer les termes candidats en ne considérant comme termes-clés que ceux qui contiennent au moins un mot de référence (tous les mots de référence devant être utilisés dans au moins un terme-clé).

2.2.1.3. Approches à base de graphe

Les approches à base de graphe consistent à représenter le contenu d'un document sous la forme d'un graphe. La méthodologie appliquée est issue de Page Rank (Brin et Page, 1998), un algorithme d'ordonnancement de pages Web (nœuds du graphe) grâce aux liens de recommandation qui existent entre elles (arcs du graphe). Text Rank (Mihalcea et Tarau, 2004) et Single Rank (Wan et Xiao, 2008b) sont les deux adaptations de base de Page Rank pour l'extraction automatique de termes-clés . Dans celles-ci, les pages Web sont remplacées par des unités textuelles dont la granularité est le mot et un arc est créé entre deux nœuds si les mots qu'ils représentent co-occurrent dans une fenêtre de mots donnée.

Le graphe est noté $\boldsymbol{G = (N, A)}$, où $\boldsymbol{N}$ est l'ensemble des nœuds du graphe et où A est l'ensemble de ses arcs entrants et sortant : $\boldsymbol{A_{entrant} \cup A_{sortant}}$. Pour chaque nœud du graphe, un score est calculé par un processus itératif destiné à simuler la notion de recommandation d'une unité textuelle par d'autres 7 (cf. équation 5). Ce score à chaque

nœud n_i permet d'ordonner les mots par degré d'importance dans le document analysé. La liste ordonnée des mots peut ensuite être utilisée pour extraire les termes-clés.

$$S(ni) = (1-\lambda) + \lambda \times \sum_{n_j \in A\text{entrant}(n_i)} \frac{pj,i \times S(n_j)}{\sum_{n_k \in A\text{sortant}(n_j)} pj,k} \quad (5)$$

λ est un facteur d'atténuation qui peut être considéré ici comme la probabilité pour que le nœud *ni* soit atteint par recommandation. *pj,i* représente le poids de l'arc allant du noeud n_j vers le noeud n_i, soit le nombre de co-occurrences entre les deux mots *i* et *j* . Dans leurs travaux, Wan et Xiao (2008 b) s'intéressent à l'ajout d'informations dans le graphe grâce à des documents similaires (voisins) et aux relations de co-occurrences qu'ils possèdent (Expand Rank). L'objectif est de faire mieux ressortir les mots importants du graphe en ajoutant de nouveaux liens de recommandation ou bien en renforçant ceux qui existent déjà. L'usage de documents similaires peut cependant ajouter ou renforcer des liens qui ne devraient pas l'être. Pour éviter cela, les auteurs réduisent l'impact des documents voisins en utilisant leur degré de similarité avec le document analysé. Une alternative à Expand Rank, Collab Rank, également proposée par Wan et Xiao (2008a), fonctionne de la même manière, mais certains choix des auteurs rendent impossible l'usage du degré de similarité pour réduire l'impact des documents voisins. Les résultats moins concluants de Collab Rank tendent à confirmer l'importance de l'usage du degré de similarité. Dans l'optique d'améliorer encore TextRank/SingleRank, Liu *et al.* (2010) proposent une méthode qui cherche cette fois-ci à augmenter la couverture de l'ensemble des termes-clés extraits dans le document analysé (Topical Page Rank). Pour ce faire, ils tentent d'affiner le rang d'importance des mots dans le document en tenant compte de leur rang dans chaque sujet abordé. Le rang d'un mot pour un sujet est obtenu en intégrant à son score Page Rank la probabilité qu'il appartienne au sujet (cf. équation 6). Le rang global d'un terme candidat est ensuite obtenu en fusionnant ses rangs pour chaque sujet.

$$S_{\text{sujet}}(Ni) = (1-\lambda) \times p(\text{sujet}|i) + \lambda \times \sum_{N_j \in A\text{entrant}(\text{N}_\text{j})} \frac{pj,i \times S(N_J)}{\sum_{N_k \in \text{A}_\text{sortant}(\text{N}_\text{j})} pj,k} \quad (6)$$

Les approches à bases de graphe présentées ci-dessus effectuent toutes un ordonnancement des mots du document analysé selon leur importance dans celui-ci. Pour extraire les termes-clés il est donc nécessaire d'effectuer du travail supplémentaire à partir de la liste ordonnée de mots. Dans la méthode TextRank, les *k* mots les plus importants sont sélectionnés et retournés (après que ceux apparaissant en collocation dans le document aient été concaténés). La technique utilisée dans les autres méthodes consiste à ordonner les termes candidats en fonction de la somme du score des mots qui les composent. Cependant, puisque l'un des avantages du graphe est que les noeuds peuvent avoir une granularité contrôlée, Liang *et al.* (2009) décident d'utiliser des mots et des multi-mots au lieu de simples mots et de tirer profit de traits supplémentaires, la taille du terme ou encore sa première position dans le document analysé.

2.2.2. Méthodes supervisées

Les méthodes supervisées sont des méthodes capables d'apprendre à réaliser une tâche particulière, soit ici l'extraction de termes-clés. L'apprentissage se fait grâce à un corpus dont les documents sont annotés en termes-clés. L'annotation permet d'extraire les exemples et les contres exemples dont les traits statistiques et/ou linguistiques servent à apprendre une classification binaire. La classification binaire consiste à indiquer si un terme candidat est un terme-clé ou non. De nombreux algorithmes d'apprentissage sont utilisés dans divers domaines. Ils peuvent potentiellement s'adapter à n'importe quelle tâche, dont celle de l'extraction automatique de termes-clés. Les algorithmes utilisés pour celle-ci construisent des modèles probabilistes, des arbres de décision, des Séparateurs à Large Marge (SVM) ou encore des réseaux de neurones. KEA (Witten *et al.* 1999) est une méthode qui utilise une classification naïve bayé sienne pour attribuer un score de vraisemblance à chaque terme candidat, le but étant d'indiquer s'ils sont des termes-clés ou non 10. Witten *et al.* (1999) utilisent trois distributions conditionnelles apprises à partir du corpus d'apprentissage. La première correspond à la probabilité pour que chaque terme candidat soit étiqueté *oui* (terme-clé) ou *non* (non terme-clé). Les deux autres correspondent à deux différents traits qui sont le poids TF-IDF du terme candidat et sa première position dans le
document :

$$\boldsymbol{P}(\mathbf{terme}) = \frac{\boldsymbol{P}\mathbf{oui}(\mathbf{terme})}{\boldsymbol{P}\mathbf{oui}(\mathbf{terme}) + \boldsymbol{P}\mathbf{non}(\mathbf{terme})} \qquad (\mathbf{7})$$

$$\boldsymbol{P}\mathbf{oui}(\mathbf{terme}) = \boldsymbol{P}(\mathbf{terme}|\mathbf{oui}) \times \prod_{\mathbf{trait}\in\{\mathbf{TF-IDF},\mathbf{position}\}} \boldsymbol{P}\mathbf{trait}\,(\mathbf{trait}(\mathbf{terme})|\mathbf{oui})$$

$$\boldsymbol{P}\mathbf{non}(\mathbf{terme}) = \boldsymbol{P}(\mathbf{terme}|\mathbf{non}) \times \prod_{\mathbf{trait}\in\{\mathbf{TF-IDF},\mathbf{position}\}} \boldsymbol{P}\mathbf{trait}\,(\mathbf{trait}(\mathbf{terme})|\mathbf{non})$$

L'un des avantages de la classification naïve bayé sienne est que chaque distribution est supposée indépendante. L'ajout de nouveaux traits dans la méthode KEA est donc très aisé.

Parmi les variantes de KEA proposées, Frank *et al.* (1999) ajoutent un troisième trait : le nombre de fois que le terme candidat est un terme-clé dans le corpus d'apprentissage. L'ajout de ce trait permet d'améliorer les performances de la version originale de KEA, mais uniquement lorsque la quantité de données d'apprentissage est très importante. Une autre amélioration de KEA, proposée par Turney (2003), tente d'augmenter la cohérence entre les termes candidats les mieux classés. Pour ce faire, une première étape de classification est effectuée avec la méthode originale. Cette première étape permet d'obtenir un premier classement des termes candidats selon leur score de vraisemblance. Ensuite, de nouveaux traits sont ajoutés et une nouvelle étape de classification est lancée. Les nouveaux traits ont pour but d'augmenter le score de vraisemblance des termes candidats ayant un fort lien sémantique avec certains des termes les mieux classés après la première étape. Enfin, Nguyen et Kan (2007) proposent l'ajout des informations concernant la structure des documents. En effet, certaines sections telles que l'introduction et la conclusion dans les articles scientifiques sont plus susceptibles de contenir des termes-clés qu'une section présentant des résultats expérimentaux, par exemple. Dans leur version modifiée de KEA, ils proposent aussi l'usage de traits linguistiques tels que les parties du discours qui ont prouvées jouer un rôle non-négligeable pour l'extraction de termes-clés (Hulth, 2003). En même temps que KEA (Witten *et al.*, 1999), Turney (1999) met au point l'algorithme génétique GenEx. GenEx est constitué de deux composants. Le premier composant, le géniteur, sert à apprendre des paramètres lors de la phase d'apprentissage. Ces paramètres sont utilisés par le second composant, l'extracteur, pour donner un score d'importance à chaque terme candidat.

Plus les paramètres sont optimaux, meilleure est la classification des termes. Pour ce faire, les paramètres sont représentés sous la forme de bits qui constituent une population d'individus que

le géniteur fait évoluer jusqu'à obtenir un état stable correspondant aux paramètres optimaux. Dans son article présentant GenEx, Turney (1999) discute une autre méthode pour l'extraction de termes-clés. Cette méthode utilise de nombreux traits qui servent à entraîner 50 arbres de décision C4.5 (technique de *Random Forest*). Dans un arbre de décision, chaque branche représente un test sur l'un des traits d'un terme candidat. Les tests permettent un routage du terme candidat vers la feuille de l'arbre qui détermine sa classe. Grâce à la technique de *Random Forest*, soit l'usage de plusieurs arbres entraînés sur un échantillon différent du corpus d'apprentissage, l'extraction automatique de termes-clés est réduite à un vote de chaque arbre pour chaque terme candidat.

Cela permet un classement des termes candidats en fonction de leur nombre de votes positifs. Les termes-clés extraits correspondent aux termes candidats les mieux classés. La même année que les travaux de Hulth (2003) sur le bien fondé d'utiliser des traits linguistiques pour l'extraction automatique de termes-clés, Sujian *et al.* (2003) proposent une méthode utilisant un modèle d'entropie maximale (cf. équation 8) dont l'un des traits repose sur les parties du discours des mots qui composent les termes candidats. Un modèle de maximum d'entropie consiste à trouver parmi plusieurs distributions, une pour chaque trait, laquelle a la plus forte entropie. La distribution ayant la plus forte entropie est par définition celle qui contient le moins d'informations, ce qui la rend de ce fait moins arbitraire pour l'extraction des termes-clés.

$$\mathbf{Score(terme)} = \frac{\boldsymbol{P}(\mathbf{oui}|\mathbf{terme})}{\boldsymbol{P}(\mathbf{non}|\mathbf{terme})} \qquad (\mathbf{8})$$

$$\boldsymbol{P}(\mathbf{classe}|\mathbf{terme}) = \frac{\exp(\sum_{\mathbf{trait}} \boldsymbol{\alpha}_{\mathbf{trait}} \times \mathbf{trait}(\mathbf{terme}, \mathbf{classe}))}{\sum_{\boldsymbol{c}\in\{\mathbf{oui},\mathbf{non}\}} \exp(\sum_{\mathbf{trait}} \boldsymbol{\alpha}_{\mathbf{trait}} \times \mathbf{trait}(\mathbf{terme}, \boldsymbol{c}))}$$

Le paramètre $\boldsymbol{\alpha}_{\mathbf{trait}}$ définit l'importance du trait auquel il est associé. Les Séparateurs à Large Marge sont aussi des classifieurs utilisés par les méthodes d'extraction automatique de termes-clés. Ils exploitent divers traits afin de projeter des exemples et des contres exemples sur un plan, puis ils cherchent l'hyperplan qui les sépare. Cet hyperplan sert ensuite dans l'analyse de nouvelles données. Dans le contexte de l'extraction de termes-clés, les exemples sont les termes-clés et les contres exemples sont les termes candidats qui ne sont pas des termes-clés. Ce mode de fonctionnement des SVM est utilisé par Zhang *et al.* (2006), mais un autre type de SVM est plus largement utilisé dans les méthodes supervisées d'extraction de termes-clés. Il s'agit de SVM qui utilisent de multiples marges représentant des rangs. Ces classifieurs permettent donc d'ordonnancer les termes-clés lors de leur extraction (Herbrich *et al.*1999; Joachims, 2006; Jiang *et al.*, 2009). La méthode KeyWE de Eichler et

Neumann (2010) utilise ce type de SVM avec le trait TF-IDF ainsi qu'un trait booléen ayant la valeur vraie si le terme candidat apparaît dans un titre d'un article Wikipedia (un terme candidat apparaissant dans le titre d'un article de Wikipedia a une plus forte probabilité d'être un terme-clé). L'ordonnancement des termes candidats par le SVM permet ensuite de contrôler le nombre de termes-clés à extraire (choix des *k* termes candidats les mieux classés). Tout comme Turney (1999), Ercan et Cicekli (2007) utilisent eux aussi une forêt d'arbres C4.5 dans leur méthode d'extraction de termes-clés. Ils utilisent des traits classiques et leur contribution se situe au niveau de l'utilisation d'un trait calculé à partir de chaînes lexicales. Une chaîne lexicale lie les mots d'un document selon certaines relations telles que la synonymie, l'hyponymie ou la méronymie. Ces relations permettent de calculer un score qui sert de trait. Cette approche est intéressante, mais du fait de limitations des chaînes lexicales actuellement disponibles elle présente l'inconvénient de ne retourner que des mots (aucun multi-mot). Cependant, l'usage d'une forêt d'arbre C4.5 permet un classement des mots à partir de leur nombre de votes positifs. Il est donc envisageable de déduire les termes-clés à partir de la liste ordonnée et pondérée des mots clés (voir les méthodes non-supervisées à bases de graphe – section 2.1). Une autre méthode pour l'extraction automatique de termes-clés consiste à utiliser un perceptron multicouches (Sarkar *et al.* 2010). Un perceptron multicouche est un réseau de neurones constitué d'au moins trois couches, chaque couche étant composée de neurones. Dans les deux couches extrêmes les neurones représentent respectivement les entrées et les sorties. Les couches centrales sont des couches cachées qui permettent d'acheminer les valeurs des entrées vers les sorties, où de nouvelles valeurs sont obtenues grâce à la pondération des transitions d'un neurone d'une couche vers un neurone de la couche suivante. Les entrées correspondent aux traits d'un terme candidat (ici TF-IDF, la position, la taille, etc.) et les sorties représentent les classes qu'il peut prendre (terme-clé ou non terme-clé). La valeur obtenue pour chaque sortie (classe) permet d'obtenir une probabilité pour que le terme candidat analysé soit un terme-clé ou non. Dans leur méthode, Sarkar *et al.* (2010) utilisent cette probabilité pour ordonner les termes candidats afin de mieux contrôler le nombre de termes-clés à extraire. Dans leurs travaux, Liu *et al.* (2011) proposent une méthode d'extraction de termes-clés basée sur un modèle génératif. Leur méthode est très différente de celle de Witten *et al.* (1999) puisqu'ils décident d'utiliser une approche de traduction automatique. L'usage original de cette approche est justifié par le fait qu'un ensemble de termes-clés doit décrire de manière synthétique le

document. Leur hypothèse est donc qu'un ensemble de termes-clés est une traduction d'un document dans un autre langage. Le modèle est appris à partir de paires de traductions dont l'un des termes est issu des titres ou des résumés des documents du corpus d'apprentissage et dont l'autre terme est issu des corps de ces mêmes documents. Les titres et les résumés sont utilisés comme langage synthétique et les corps des documents comme le langage naturel de ceux-ci.

2.3.Conclusion

Ce chapitre présente les principales méthodes d'extraction automatique de termes-clés. La tâche d'extraction automatique de termes-clés consiste à analyser un document pour en extraire les expressions (phrasèmes) les plus représentatives de celui-ci. Les méthodes d'extraction automatique de termes-clés sont réparties en deux catégories : les méthodes supervisées et les méthodes non supervisées. Les méthodes supervisées réduisent la tâche d'extraction de termes-clés à une tâche de classification binaire (tous les phrasèmes sont classés parmi les termes clés ou les non termes-clés). Cette classification est possible grâce à une phase préliminaire d'apprentissage, phase qui n'est pas requise par les méthodes non-supervisées. Ces dernières utilisent des caractéristiques (traits) extraites du document analysé (et parfois d'une collection de documents de références) pour vérifier des propriétés permettant d'identifier ses termes-clés.

3.1. Introduction

Pour que le coût d'une recherche d'information dans une base documentaire soit acceptable, il convient d'effectuer une étape primordiale sur la base. Cette étape consiste à analyser chaque document de la collection afin de créer un ensemble de mots-clés : on parle de l'étape d'indexation. Ces mots-clés seront plus facilement exploitables par le système lors du processus ultérieur de recherche.

Dans ce chapitre, nous nous intéressons aux différentes notions liées à l'indexation. L'intérêt que nous portons pour cette étape du processus de filtrage d'information se justifie par le fait que c'est dans cette étape où une structuration de documents basée sur les métadonnées intervient.

3.2. Processus d'indexation

L'indexation est l'opération qui vise à construire une structure d'indexe qui permet de retrouver très rapidement les documents incluant des mots demandés.

Cette étape consiste à analyser chaque document de la collection afin de créer un ensemble de mots-clés. Son objectif est de trouver les concepts les plus importants du document (ou de la requête), qui formeront le descripteur du document.

Ainsi en pratique on cherche plutôt des représentants des concepts. Ces représentants peuvent être de formes différentes: des mots simples ou de groupes de mots (mots composés). Les descripteurs des documents (mots, groupe de mots) sont rangés dans une structure appelée dictionnaire constituant le langage d'indexation.

Ce langage peut être de deux types :

- ***Langage libre :*** est construit à partir des termes extraits du document analysé.
- ***Langage contrôlé :*** est construit à partir d'un ensemble des termes préalablement définis et organisés généralement dans un thésaurus.

Lorsqu' un document est analysé on ne grade que les mots clés qui appartiennent à ce thésaurus.

L'indexation peut être :

- ✓ ***Manuelle***

Chaque document est analysé par un spécialiste du domaine ou par un documentaliste, mais elle nécessite assez du temps pour sa réalisation, en plus, des termes différents peuvent être présentés par deux documentalistes différents pour

représenter un même document, et un indexeur, à deux moments différents peut présenter deux termes distincts pour représenter le même concept.

- ✓ ***Automatique***

L'indexation automatique consiste à simuler par une machine cette opération d'indexation, que ce soit dans sa méthode ou sur tout dans ses resultants. En effete, si un ordinateur ne peut trouver un terme réellement descripteur d'un concept d'un document, il pourra caractériser celui-ci de façon à le retrouver. Le processus d'indexation est dans ce cas entièrement informatisé.

L'indexation automatique repose sur un ensemble des méthodes automatisées sur un document comme l'extraction automatique des mots des documents, l'élimination des mots vides, la lemmatisation (radicalisation), la pondération des termes et la création de l'index.

- ✓ ***Semi-automatique***

Le choix final revient au spécialiste ou au documentaliste, qui intervient souvent pour choisir d'autres termes significatifs. Cette méthode est une combinaison des deux méthodes précédentes, elle appelé aussi indexation supervisée.

Qu'elle soit manuelle, semi-automatique ou automatique, l'indexation répond aux deux problèmes suivant : le choix des mots qui représente chaque document et l'évaluation de leur pouvoir de représentation.

Enfin l'indexation peut être caractérisée par sa fonction de pondération.

Voice la suite des opérations traditionnellement effectuées sur les documents textuelles lors de l'indexation, avec l'illustration des étapes d'indexation dans la figure 1.2 :

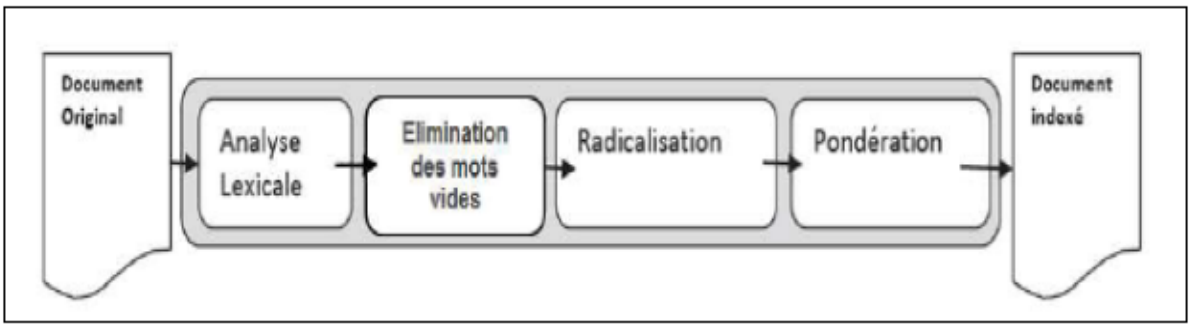

Figure 3.1 Suite des traitements lors de l'indexation

3.3. Traitements linguistiques

On a choisi d'utiliser deux techniques pour la représentation de textes à savoir: la représentation par mot et la représentation par lemme.

3.3.1. Représentation par mot :

Cette technique consiste à représenter chaque document sous forme d'un vecteur de mot (unité lexicale).

Cette représentation nécessite les prétraitements suivants :

- Tokénisation des documents.
- Elimination des majuscules.
- Elimination des mots vides.

a- <u>Tokénisation des documents</u>

Cette étape consiste a transformé un texte en un ensemble de termes, on a choisi de considérer un mot comme une suite de caractères situés entre deux séparateurs.

L'algorithme suivant illustre cette étape.

'\n' , ' '
',' , ' ;' , ' :' , '.' , ' ?' , ' ! ' : La ponctuation
'=' , '+' , '-' ,'*' , '<' , '>' : les opérateurs arithmétiques et de comparaison
'(' , ')' , '[' , ']' , '{' , '}' : Les parenthèses
' « ' , '\' , '/' , ' , '%'
'&' , '~' , '#' , '|' , '_' : les opérateurs logiques
'@' , '§' , '$' , '£'
'0' , '1' , '2' , '3' , '4' , '5' , '6' , '7' , '8' , '9' : les chiffres.

Figure 3.2 La liste des séparateurs

A la fin de cette opération chaque document sera représenté sans séparateurs (espaces de séparation des chiffres, des mots, des ponctuations, etc.) la figure 3.3 présente un exemple de cette tokénisation .

La pondération; est la Phase la Plus Importante dans la (SRI), le {SRI} 20 répond à un@ Besoin

La	pondération	est	la	Phase	la	Plus	Importante	dans
la	SRI	le	SRI	répond	à	un	Besoin	

Figure 3.3 Tokénisation de document

b- <u>Elimination des majuscules :</u>

Afin de pouvoir réduire la taille du tableau il est nécessaire de formater les mots avec majuscules et les mots minuscules comme étant un seul mot. Par exemple les mots « phase, Phase, pHAse, PHASE » considèrent comme un seul mot. Voilà la figure 3.4 illustre l'exemple précédent en éliminant les majuscules.

La	pondération	est	la	Phase	la	Plus	Importante	dans
la	SRI	le	SRI	répond	à	un	Besoin	

la	pondération	est	la	phase	la	Plus	Importante	dans
la	sri	le	sri	répond	à	un	besoin	

Figure 3.4 Elimination des majuscules

c- Elimination des mots vides :

Les mots vides ou mots outils sont les mots non significatifs trouvés dans les documents. En effet, ces mots ne traitent pas le sujet du document mais ils permettent de lier entre les mots d'une phrase pour la structurer comme les articles, les conjonctions de coordination, les verbes auxiliaires, etc.

Chaque langue a sa propre liste des mots vides. Dans notre application nous avons utilisé un fichier qui comporte 124 mots vides de la langue française et 582 mots de la langue anglaise. Ces mots ne portent pas de sens**.**

Voila un extrait des mots vides de la langue française:

Alors, au, aucuns, aussi, autre, avant, avec, avoir, bon, car, ce, cela, ces, ceux, chaque, ci, comme, comment, dans, des, du, dehors, depuis, deux, devrait, doit, donc, dos, droite, elle, elles, en, Il, ils, la, le …….

```
Entrée : tableau de mots, la liste des mots vides.
Sortie : tableau de mot sans mots vides.
Début
 Pour chaque mot de tableau faire
   Si le mot figure dans la liste des mots vides
alors
          Supprimer le mot du tableau
   Finsi
 Finpour
Fin
```

Cette étape permet d'analyser et de réduire la taille de l'index.la figure 3.5 montre l'exemple précédent après élimination les mots vides.

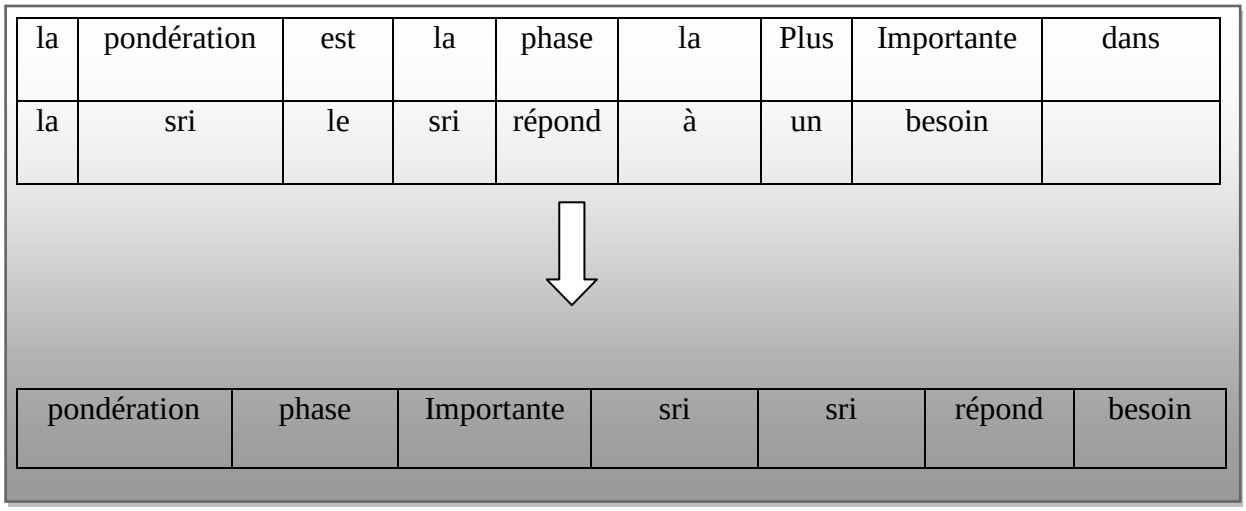

Figure 3.5 Elimination des mots vides.

3.3.2. Représentation par lemme :

La technique de lemmatisation consiste à remplacer chaque mot par sa forme canonique (lemme), en effet elle consiste à remplacer les verbes par leur forme infinitive et les noms par leur forme au masculin singulier. Voici quelques exemples de lemmatisation :

- écologie, écologiste, écologique ------» écolog.
- Informatique ------» informat.
- Petits, petite, petites ------» petit.
- Joue, jouer ------» jou.
- malade, malades, maladie, maladies, maladive ------» malad.

Cette phase consiste à indexer un ensemble de mots par un seul mot qui représente le même concept.

Il y a plusieurs algorithmes de lemmatisation telle que « l'algorithme de carry », « algorithme de paice/husk ».Dans notre système on a choisi d'utiliser « l'algorithme de porter », ce dernier est un algorithme de normalisation des mots. Il permet de supprimer les affixes des mots pour obtenir une forme canonique du mot. Cet algorithme a été proposé par Martin Porter en 1980, il est utilisé pour la langue anglaise, mais son efficacité est limitée pour la langue française où les flexions sont plus importants et plus

diverses. Il se présente comme un ensemble de règles dont l'application successive à un mot de l'anglais produit la racine de ce mot. Il reste toutefois un algorithme fondamental couramment enseigné en TALN (compréhension du texte).

```
Entrée : tableau de mot.
Sortie : tableau des lemmes
Début
  Pour chaque mot du tableau faire
  Lem ← porter (mot)
   Insère lem dans le tableau des lemmes
  Finpour
Fin
```

Cette phase est utilisable car elle est consacré à réduire le nombre de terme dans le tableau et permet de représenter par un même descripteur des mots qui ont le même sens. Enfin le cas de notre exemple comme indiqué dans la figure 3.6, on remarque que la taille du tableau a été réduite.

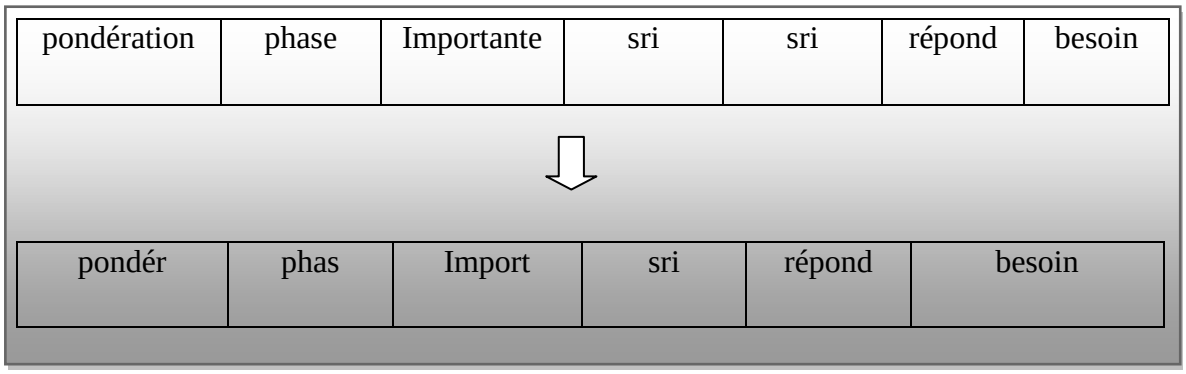

Figure 3.6 Lemmatisation des mots

3.4. La pondération :

La pondération consiste à donner aux termes de l'index un poids mesurant leur importance dans les documents qui les contiennent. En effet, la pondération permet d'affecter à chaque terme d'indexation une valeur qui mesure son importance dans le

document ou il apparait c'est-à-dire le mot est pondéré en fonction de sa rareté sur la toile, plus un mot est rare plus l'importance qui lui est accordé sur le site analysé sera grande et inversement. Ainsi, l'objectif est de trouver les termes qui représentent le mieux le contenu d'un document.

Plusieurs méthodes de pondération ont été proposées pour calculer les poids des termes de façon automatique, Les méthodes les plus utilisées dans ces domaines sont:

- ✓ **le facteur Tf (Term Frequency)** est basé sur la fréquence d'un terme dans le document, Plus un terme est fréquent dans un document plus il est important dans ce dernier.
- ✓ **Le facteur IDF (Inverse Document Frequency)** se base sur le nombre de documents contenant un terme donné. En effet, un terme apparaissant dans tous les documents n'est pas important. Sa formule est la suivante :

Idf =Log (N/df(i)), ou :

- **N** est la taille de la collection
- **df(i)** le nombre de documents contenant le terme ti

- ✓ **TF*IDF** (terme frequency * inverse of document frequency):

On combinant les deux techniques précédentes, elle donne une bonne approximation de l'importance du terme dans le document relativement à une collection selon cette pondération, pour qu'on mot soit important dans un document il ne suffit pas qu'il soit fréquent dans le document mais aussi absent dans les autres documents.

La formule de **TF*IDF** est la suivante :

- **TFIDF (T, D) :** $Tf(t, d) * log\frac{N}{df(N)}$

Ainsi un terme qui a une valeur de **TF*IDF** élevé doit être à la fois important dans le document et aussi il doit apparaitre peu dans les autres documents.

Voici notre corpus :
D1 :" Une organisation retenue pour nos travaux dans une organisation ".
D2 :" Un document renvoie à un ensemble formé par un support et une information ".
D3: " Une approche subjective élabore une distinction entre « document par attribution» et « document par intention», document ".
Calculons la pertinence du terme t1=" document " pour les 3 documents en utilisant Tf * Idf **:**
Tf ("document", D1) = 0 ; Tf ("document", D2) = 1 ; Tf ("document", D3) = 3
Idf ("document") = log10 (n/df ("document")) = log10 (3/2) = 0.1761
Ce qui fait:
Tf Idf ("document", D1) = 0
Tf Idf ("document", D2) = 0.1761
Tf Idf ("document", D3) = 0.5283

Figure 3.7 La pondération par la mesure « TFIDF »

3.5. Objets de l'indexation

Le but restant d'apparaître parmi la première vingtaine des réponses qui sont affichées, il est nécessaire d'effectuer une indexation performante et pertinente. Pour réaliser cela les administrateurs ont mis en place un système qui permet de donner un poids à chacun des critères, il revient à chaque moteur de choisir la configuration qu'il trouvera optimale, afin d'offrir aux utilisateurs quelque chose de performant et qui fournit des réponses en rapport direct avec leurs requêtes.

3.5. 1.Titre du document

Le titre d'un document, donné par son auteur (ou concepteur), apparaît comme étant le critère le plus important pour quasiment tous les moteurs de recherche. C'est la première chose qui va intéresser l'indexeur. Selon les moteurs de recherche, il va y avoir une limitation ou non de la taille. Par exemple, pour Excite, cette taille est limitée à 50 caractères alors que pour Altavista, elle est limitée à 80.

3.5. 2.Métadonnées de contenu

Elles ne sont pas utilisées par tous les moteurs, mais peuvent avoir une importance équivalente, voir supérieur par rapport au corps du texte. Certains moteurs conservent une confiance par rapport aux informations que peuvent donner les concepteurs de sites.

- **La balise <META Description>**

Cette balise, est prise en compte par la majorité des moteurs de recherche. Elle permet de décrire le contenu de la page sous forme d'un résumé. Elle est affichée par les moteurs dans la page de résultats pour donner un aperçu du contenu du document retourné. Lorsqu'une page web ne contient pas de balise <*META Description*> certains moteurs affichent les premiers mots visibles sur la page. Comme pour le titre, une limitation de taille est appliquée par certains moteurs de recherche tels que Altavista (1024 caractères) ou Voilà (400 caractères).

- **La balise <META Keywords>**

Cette balise contient un ensemble de mots clés, séparés par une virgule, en rapport avec le contenu d'une page Web. Les mots clés doivent être choisis judicieusement. Comme pour les autres métadonnées, celle-ci est prise en compte par la majorité des moteurs et subit une limitation en taille : 100 mots-clés, ou 1000 caractères. Au-delà, la balise est considérée comme du *spamming* et éventuellement pénalisée.

3.5. 3. Corps du texte

Le corps du teste est devenu maintenant, et pour une raison évidente de qualité d'indexation, un champ incontournable. Là encore, les techniques varient d'un moteur à un autre. Il est quand même possible de faire ressortir des points communs. Ils vont tous commencer par analyser le début du texte. Ils vont considérer que leur analyse est assez pertinente lorsqu'ils auront atteint une certaine taille au niveau du fichier d'indexation. D'une manière générale, ils privilégient les pages de petites tailles.

Par exemple, si on prend le cas d'*Altavista*, celui-ci va indexer tout le texte de la page jusqu'à atteindre 100 Ko, au-delà, seuls les liens seront indexés jusqu'à atteindre 4Mo, après plus rien ne sera indexé. Parallèlement à ça, *Infoseek* va indexer tout le texte mais en considérant que les mots clés importants sont dans la première moitié du document. *Google* indexe de manière égale le texte en entier.

3.5. 4. Les frames

Ce sont des pages web divisées en cadres, constituées :

- D'un fichier "mère", appelé aussi "fichier principal" qui sert uniquement à la description des zones, souvent appelé cadre.htm, ou frame.htm. C'est un fichier vide de données.

- De fichiers "filles" ou *"cadres"* formant les différents cadres de la page : cadre du haut, de gauche et central, etc.

Tous les moteurs traitent le fichier principal, certains vont plus loin et traitent le contenu de chacun des cadres qui composent la page.

Les moteurs de recherche traitent différemment ce type de pages. Les solutions qu'ils adoptent sont les suivantes :

- Page web avec frames ignorée : aucune indexation ; situation la plus répandue.
- Indexation seulement du fichier "mère" et ignorance des fichiers "filles" ; situation assez courante ; Résultat : le cadre vide est indexé seul et non les données contenues. -
- Indexation des fichiers "mère" et "filles" comme des fichiers distincts, sans indexation des liens entre eux. Résultat : perte du contexte des frames et affichage des fichiers isolément.
- Indexation des fichiers "mère" et "fille" avec leurs liens : solution idéale, respectant l'organisation des frames. MAIS pratiquement aucun moteur ne peut faire cette indexation.

Enfin, il apparaît évident qu'un site utilisant avec abondance les cadres risque d'être mal référencé. Pour éviter cela, il lui sera nécessaire de bien remplir sa page principale.

3.5. 5. Autres objets

• Les commentaires

Les commentaires ne sont jamais pris en compte par les moteurs de recherche.

• L'attribut « ALT »

L'attribut « ALT » de la balise IMG permet d'associer un texte à l'image. Lors de l'indexation des images, ce texte est pris en compte dans l'index pour ce document.

• L'URL

Presque tous les moteurs aujourd'hui indexent l'URL des pages web, qu'ils considèrent comme un champ de recherche interrogeable.

3.6. Module d'indexation

Comme cité précédemment, l'indexation se fait par l'utilisation d'un profil d'application LOM adapté aux spécifications de nos objets .En effet, vu le flou des définitions des indicateurs LOM et l'insuffisance de leur caractère pédagogique, le passage à ce format à partir des balises prédéfinies dans la DTD se fera de tels sorte que

la définition des éléments des différents catégories, en particulier la catégorie éducationnel puissent exprimer les informations du parcours. La figure suivante décrit le processus de génération des fichiers XML d'indexation.

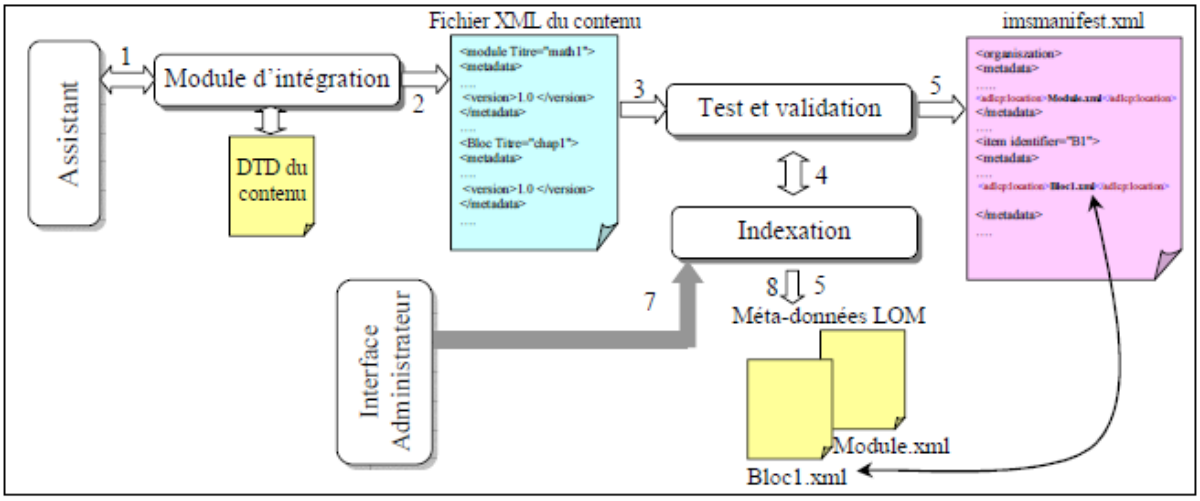

Figure 3.8 Processus d'indexation

3.6.1. Le standard LOM

Le LOM ou Métadonnées pour les objets pédagogiques, est un standard élaboré en 2002 par le consortium IEEE qui définit la structure d'une instance de métadonnée pour un objet pédagogique. Il est constitué d'un ensemble de 80 éléments répartis dans neuf catégories (général, life cycle, méta-metadata, éducationnel, technical, right, relation, annotation, classification) accomplissant chacune une fonction différente.

Les descripteurs du LOM peuvent être utilisés dans la conception des systèmes d'elearning pour l'indexation d'objets pédagogiques. Pour cela ils doivent être implémentés dans un langage structuré.

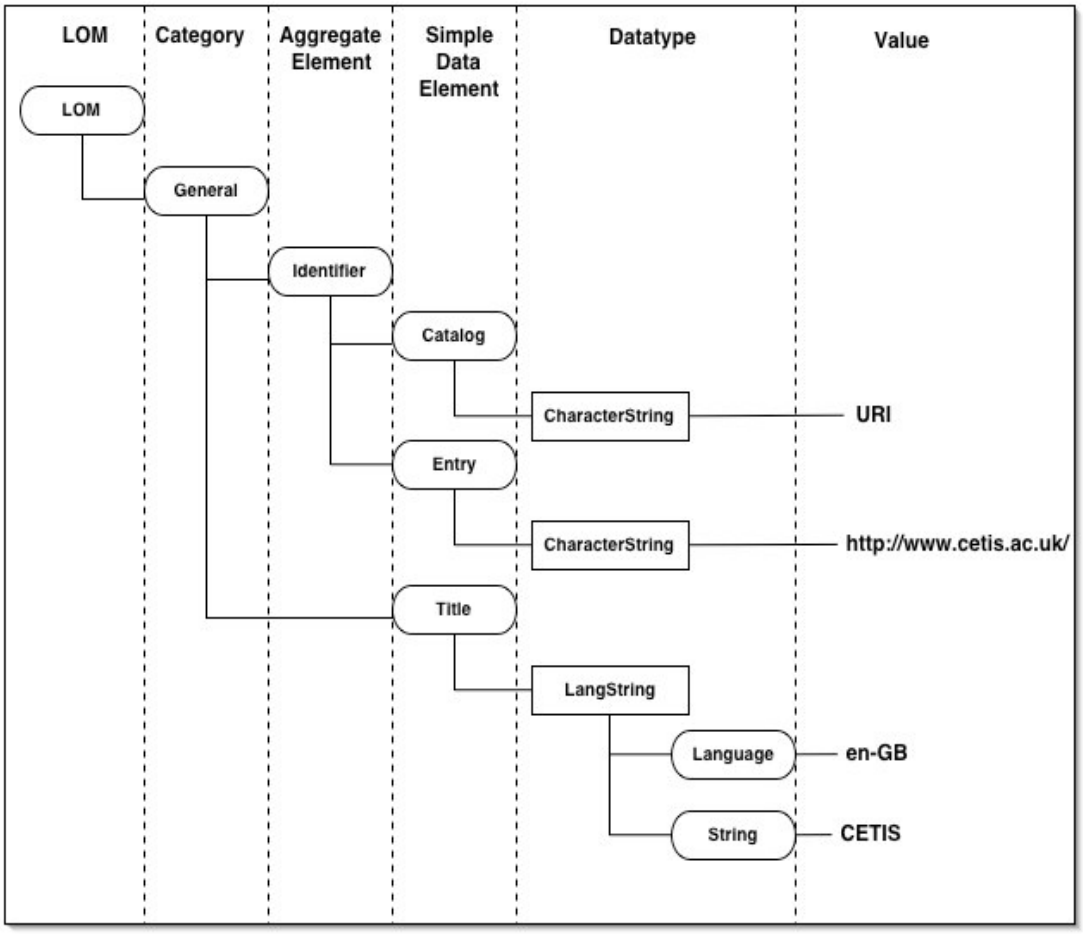

Figure 3.9 LOM Conceptuel Data Schéma Structure

3.6.2. Implémentation du LOM

La représentation du modèle abstrait dans un format spécifique est appelé "binding ". Pour les métadonnées du LOM, il en existe : le XML binding.

L'utilisation d'un langage tel que XML permet de percevoir la structure du LOM et facilite l'échange des métadonnées entre différents systèmes. On peut par exemple à travers XML binding discerner le titre ou la description d'un objet pédagogique, déduire que ces éléments appartiennent à la catégorie "General" mais on ne peut forcément deviner la signification du niveau d'agrégation ou de la structure d'un OP qui sont deux éléments figurants dans la catégorie "General" du LOM.

Si le XML binding du LOM a le mérite d'être facile à implémenter, il reste insuffisant pour la représentation des éléments du LOM puisque il ne permet pas d'exprimer la sémantique de ces éléments.

```
<lom xmlns="http://ltsc.ieee.org/xsd/LOMv1p0">
  <general>
    <title>
      <string xml:lang="en">Ecologues</string>
      <string xml:lang="la">BUCOLICA</string>
    </title>
    <language>la</language>
  </general>
  <technical>
    <location type="URI">
   http://classics.mit.edu/Virgil/eclogue.html
    </location>
  </technical>
</lom>
```

3.7. Conclusion

L'indexation, une opération préalable et indispensable à tout processus de recherche/filtrage d'information. Il est donc difficile de parler de recherche d'information sans parler d'indexation, au sens procédural du terme.

De nombreux traitements (statistiques et/ou linguistiques) doivent être effectués pour une meilleure caractérisation du contenu des documents à indexer.

4.1. Introduction

Dans ce chapitre, nous essayons de donner un environnement approprié pour le traitement de grandes quantités d'informations représenté par des textes français en utilisant par plusieurs outils, tels que le choix d'un langage de programmation simple et répandu, , un matériel relativement sophistiqué . Nous présentons également les principales fonctions de notre indexeur, ainsi que sa structure générale**.**

4.2. Présentation des corpus utilisés

Les textes du corpus sont enregistrés sous l'encodage UTF-8 sous forme de fichiers textes (.txt) (voir Table 4.1**)**

Catégorie	**Sport**
Nombre de documents d'apprentissage	10
Nombre de documents de test	4

Table 4.1 Représentation des corpus utilisés

4.3. L'approche utilisée pour la représentation des textes

Pour la représentation des textes, nous avons utilisé l'approche « mots » qui est la plus utilisée dans le domaine et vue sa simplicité et son efficacité. Cette approche a été présentée avec plus de détail dans le chapitre 3.

4.4. Le langage et l'environnement de programmation

4.4.1. Le langage de programmation (JAVA)

Le langage Java est un langage de programmation informatique orienté objet créé par James Gosling et Patrick Naughton, employés de Sun Microsystems, avec le soutien de Bill Joy (cofondateur de Sun Microsystems en 1982), présenté officiellement le 23 mai 1995 au SunWorld.

4.4.2. L'environnement de programmation (NetBeans IDE)

C’est un environnement de développement intégré (IDE) pour Java, placé en open source par Sun en juin 2000 sous licence CDDL (Common Développement and Distribution License). En plus de Java, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, XML et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages web). [42]

4.4.3. Extensible Markup Language (XML)

L'Extensible Markup Language (XML, « langage à balise extensible » en français) est un langage informatique de balisage générique qui dérive du SGML. Cette syntaxe est dite « extensible » car elle permet de définir différents espaces de noms, c'est-à-dire des langages avec chacun leur vocabulaire et leur grammaire, comme HTML, XSLT, RSS, SVG… Elle est reconnaissable par son usage des chevrons (< >) encadrant les balises. L'objectif initial est de faciliter l'échange automatisé de contenus complexes (arbres, texte riche…) entre systèmes hétérogènes (interopérabilité). Avec ses outils et langages associés, une application XML respecte généralement certains principes :

- La structure d'un document XML est définie et validé par un schéma .
- Un document XML est entièrement transformable dans un autre document XML.

4.5. Schéma illustratif de l’architecture du prototype

Prototype est constitué des étapes décrites par le schéma suivant :

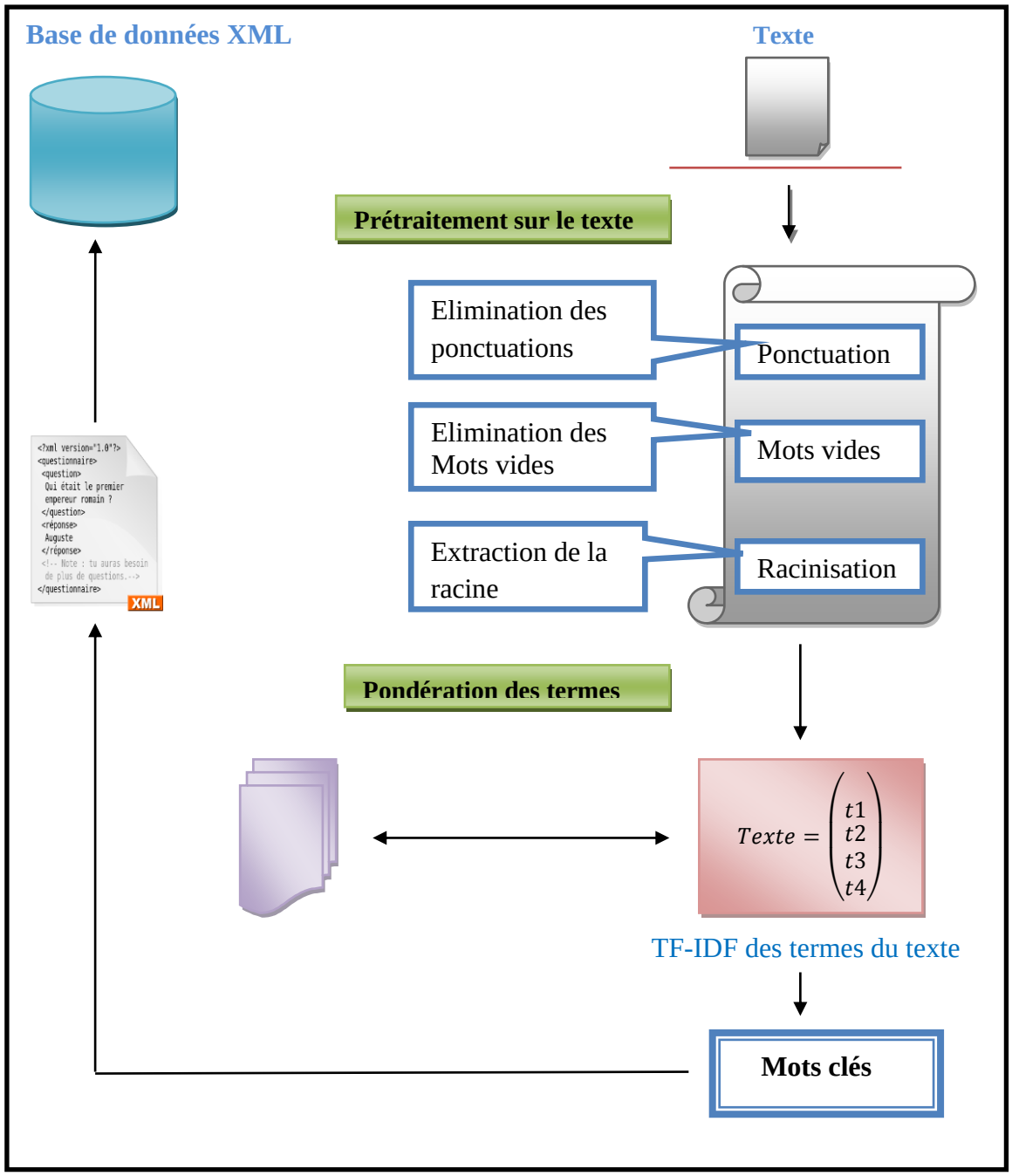

Figure 4.1 Schéma de prototype

4.5. Structure et fonctionnement de notre Outil

Notre indexeur est composé d'une interface principale peut être définie comme étant un interprète qui assure et facilite le dialogue entre l'utilisateur et l'application. Grace à cette interface principale que l'utilisateur peut effectuer les opérations ou des

traitements désirés en sélectionnant un élément du menu ou en cliquant sur un bouton.Cette interface est représentée dans la figure suivante :

4.5.1. Interface principale

L'utilisateur ouvre l'interface principale de l'application, cette interface est composée d'une barre d'outil et d'une zone de travail où l'utilisateur peut construire son procédé

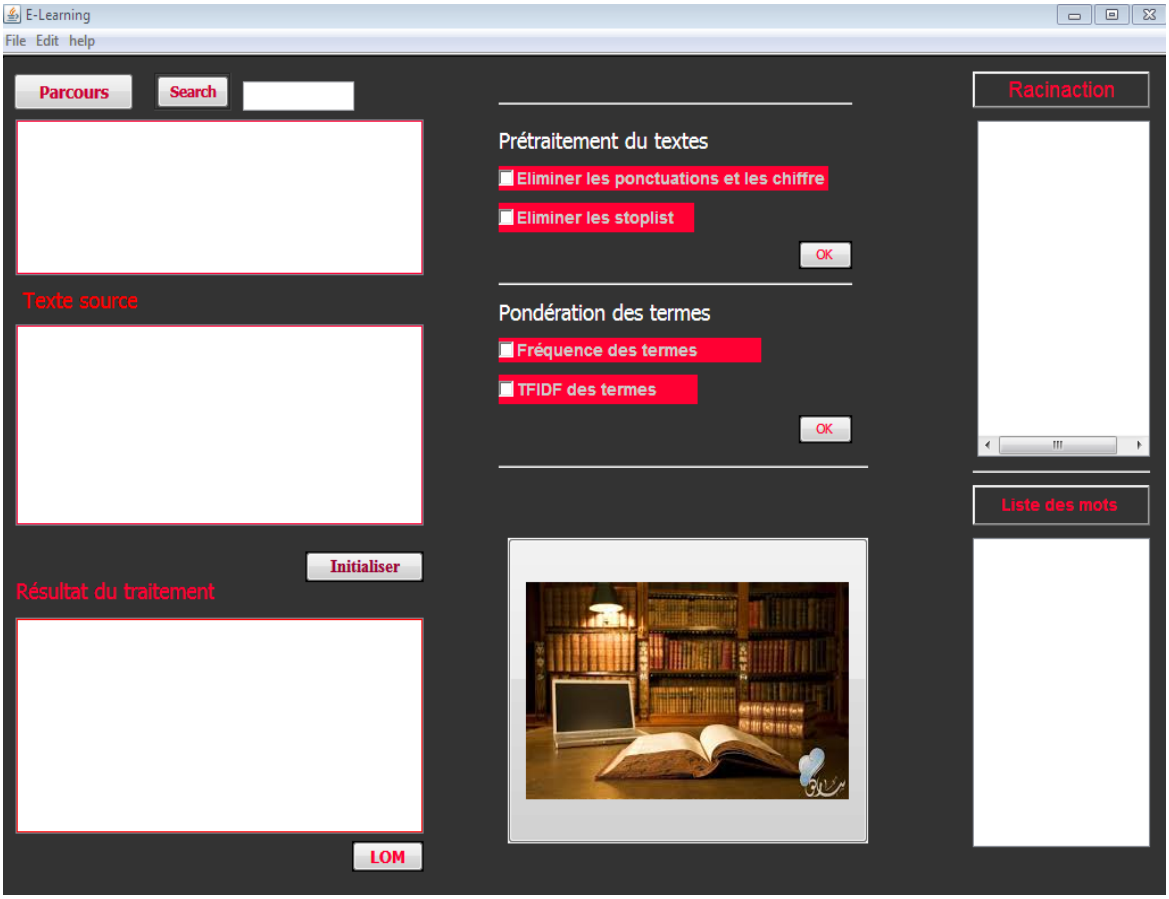

Figure 4.2 Interface principale

4.5.2. Prétraitements sur un fichier

- Faire un parcours pour obtenir les fichiers du corpus

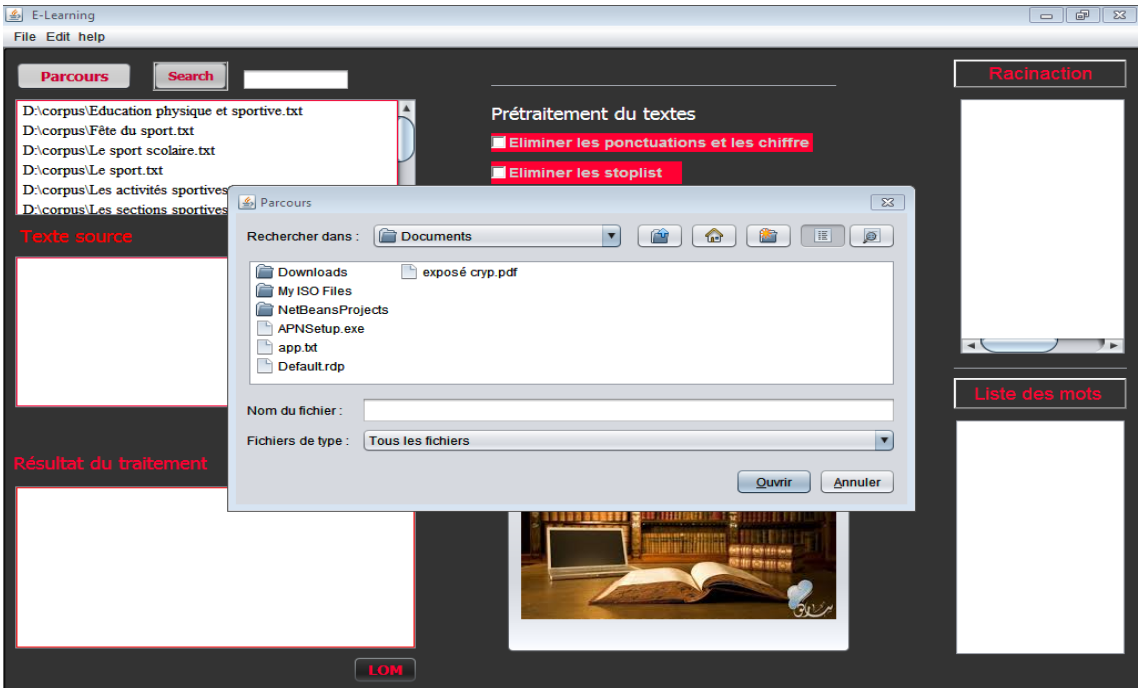

Figure 4.3: Parcourir des fichiers

- Sélectionner un fichier

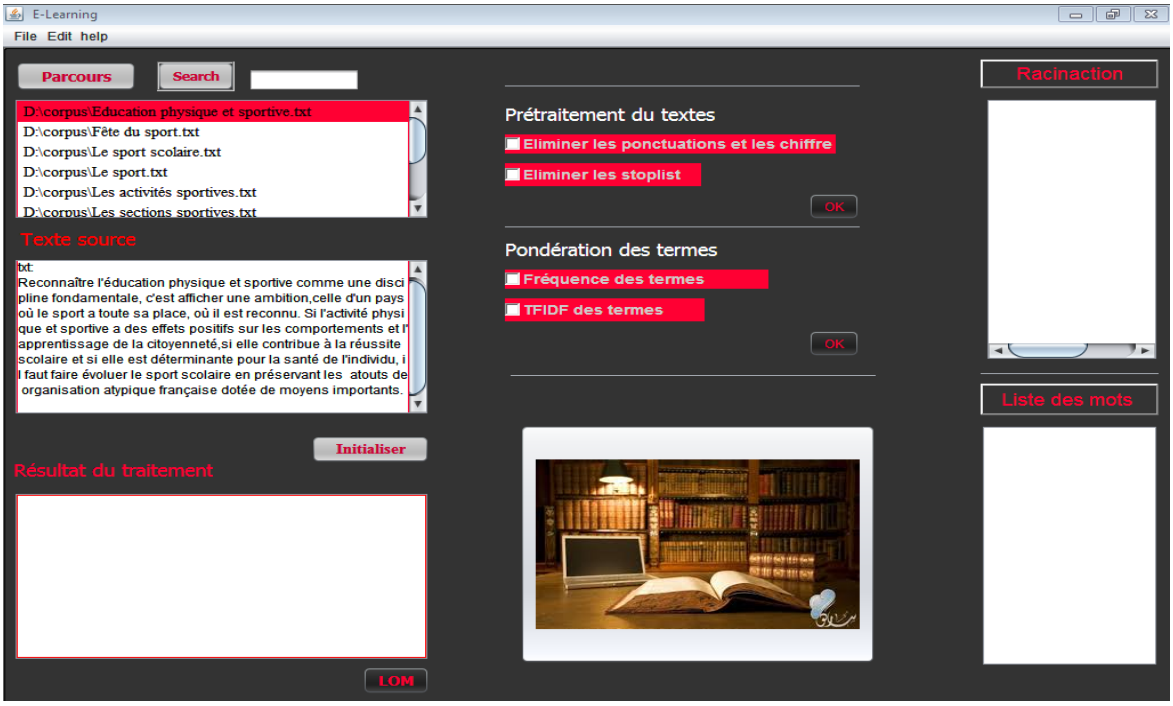

Figure 4.4 Sélection d'un fichier

❖ Elimination des ponctuations, chiffres et caractères spéciaux

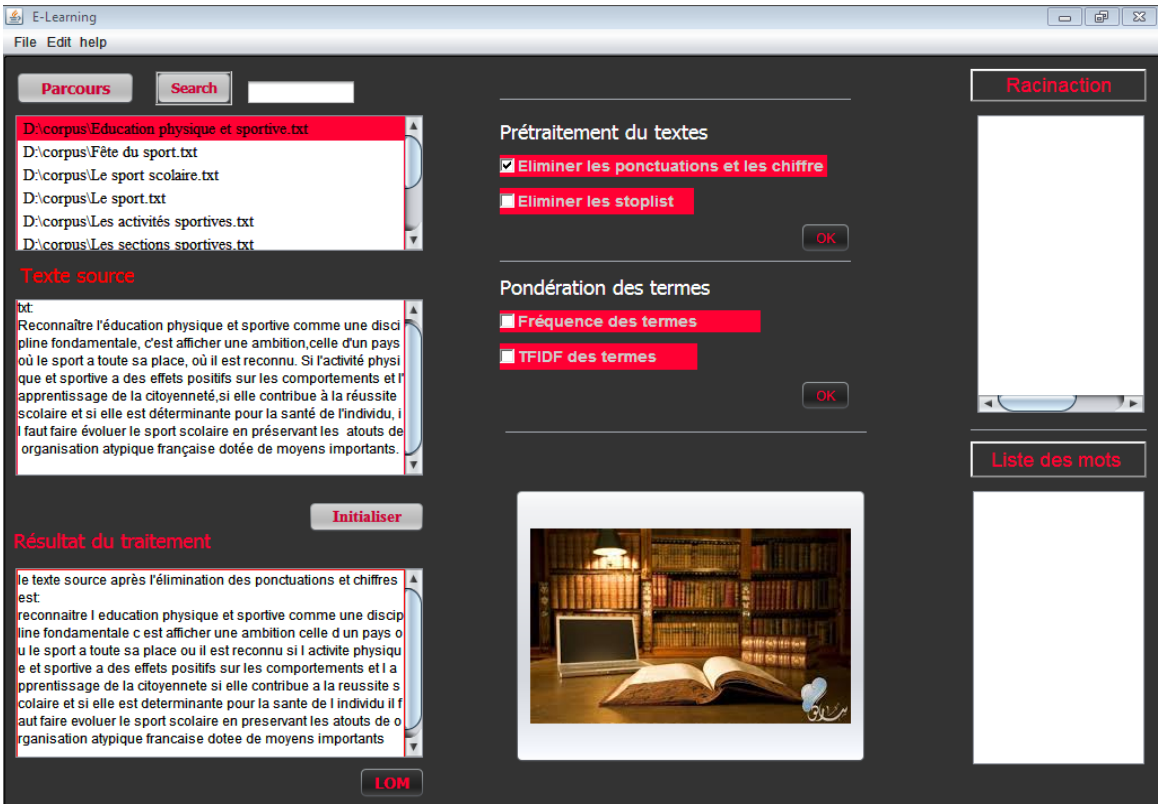

Figure 4.5 Elimination des ponctuations, chiffres et caractères spéciaux

❖ Elimination des Mots vides

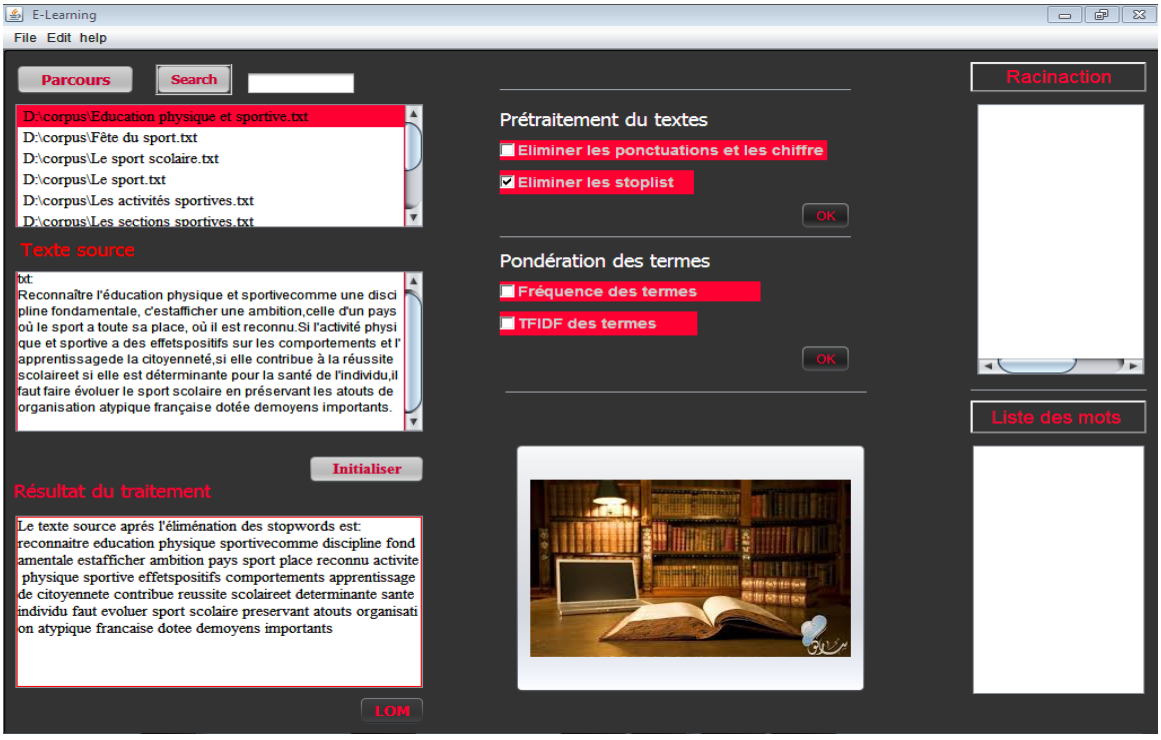

Figure 4.6 Elimination des Mots vides

❖ Extraction des racines

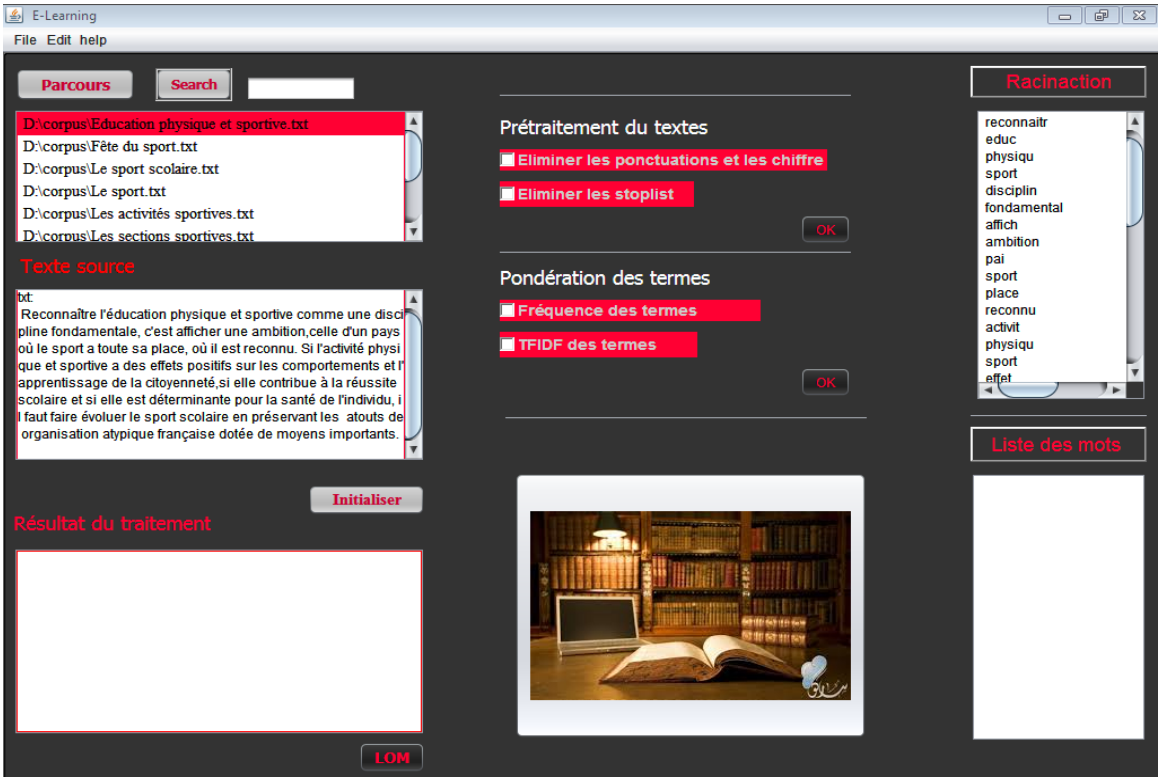

Figure 4.7 Extraction des racines

4.5.3. Pondération des termes du fichier

❖ Fréquences des termes du fichier

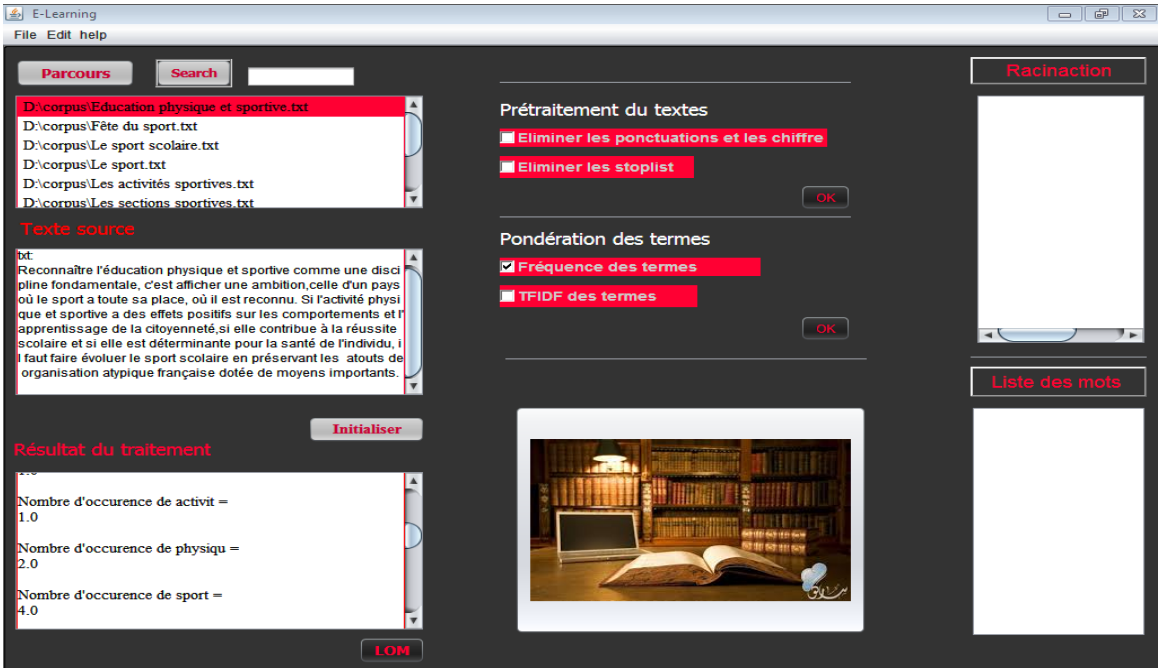

Figure 4.8 Fréquences des termes du fichier

❖ TF-IDF des termes du fichier

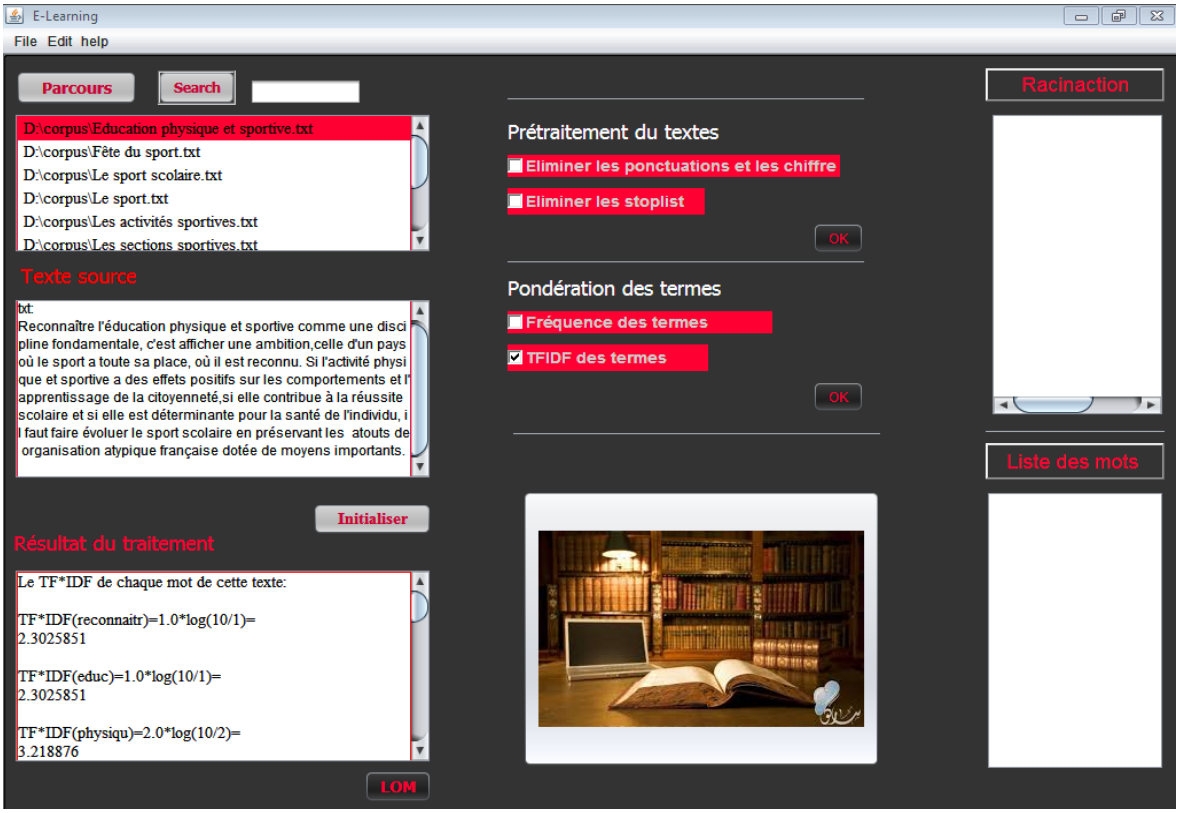

Figure 4.9 TF-IDF des termes du fichier

❖ Liste des mots clés

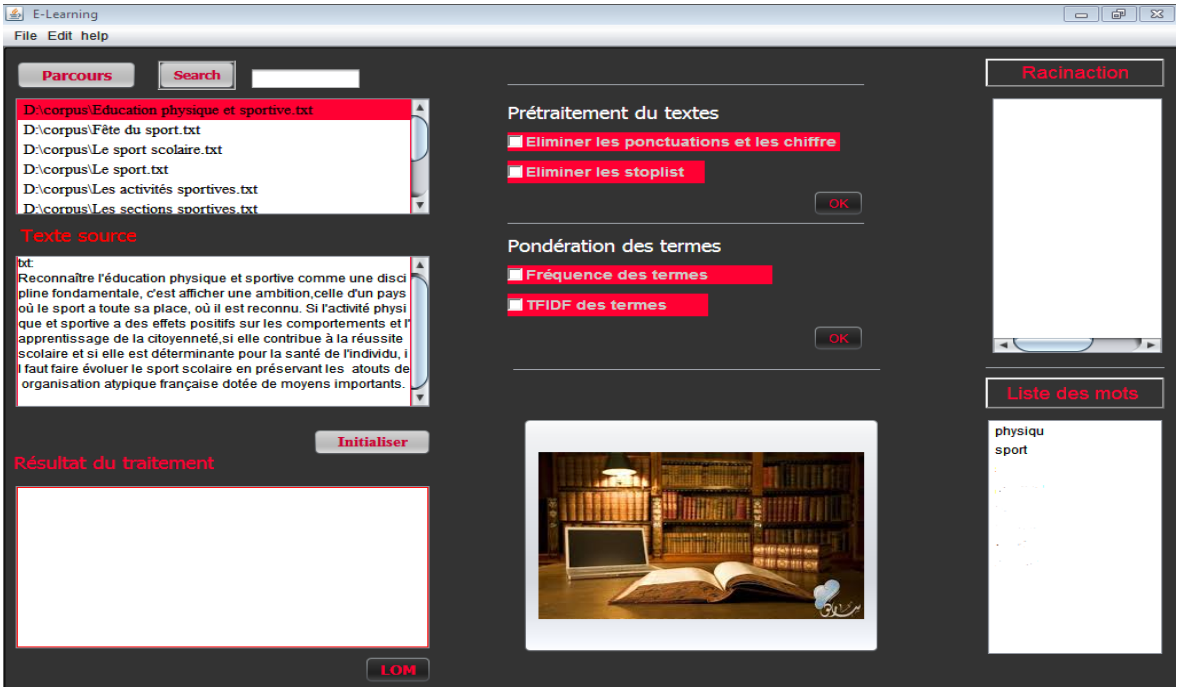

Figure 4.10 Liste des mots clés

4. 5.4. Profile de LOM

Le LOM définit un objet pédagogique comme « toute entité numérique ou non, utilisée dans un processus d'enseignement, de formation ou d'apprentissage ». Le LOM propose 45 éléments descriptifs de premier niveau regroupés dans 9 catégories, dans le cette travaille utilisé la catégorie général pour décrire des ressources numériques de texte.

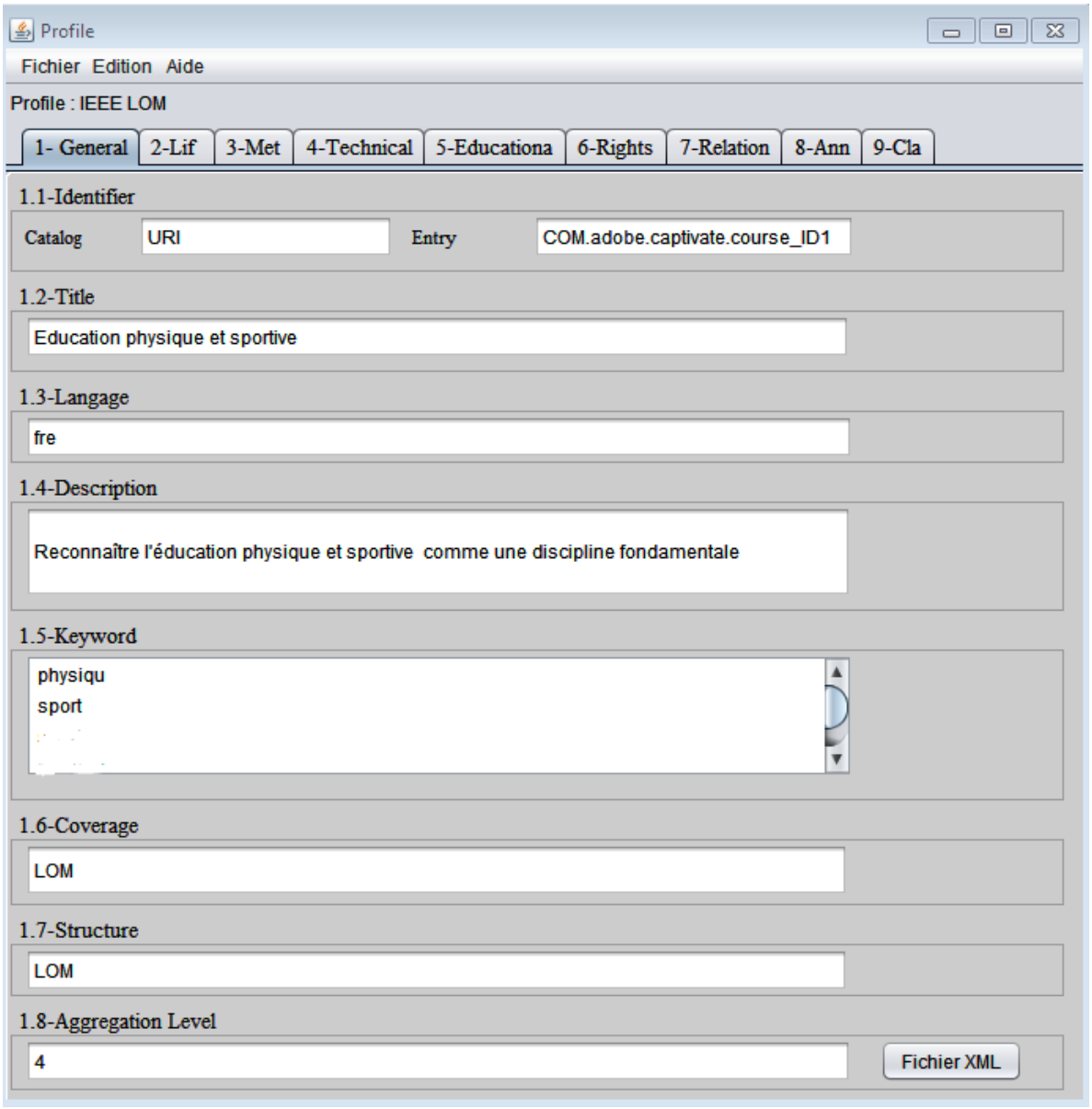

Figure 4.11 Profile d'IEEE LOM

4. 5.4.1. Le fichier XML de LOM

La représentation de fichier xml binding . Pour les métadonnées du LOM

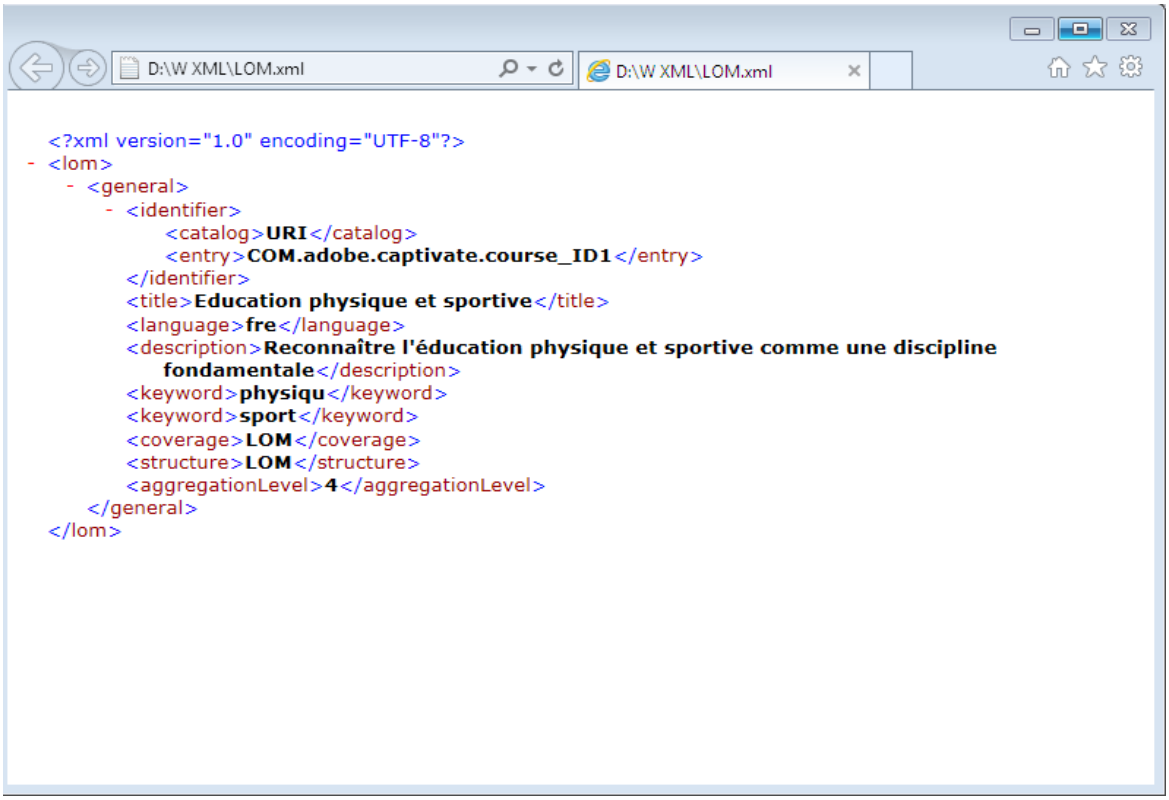

Figure 4.12 Fichier XML

4. 5.4.2. La base de données XML

Pour stocker les documents XML, BaseX utilise une représentation tabulaire de la structure arborescente XML. La base de données gère le stockage soit d'un seul document soit d'une collection de documents. On s'est inspiré du schéma d'encodage XPath Accelerator et de l'algorithme Staircase Join Operator pour accélérer les étapes de localisation XPath . De plus, BaseX fournit de nombreux types d'index pour améliorer la performance des opérations de requête sur l'arborescence, sur les attributs, de comparaisons de texte et de recherche plein-texte.

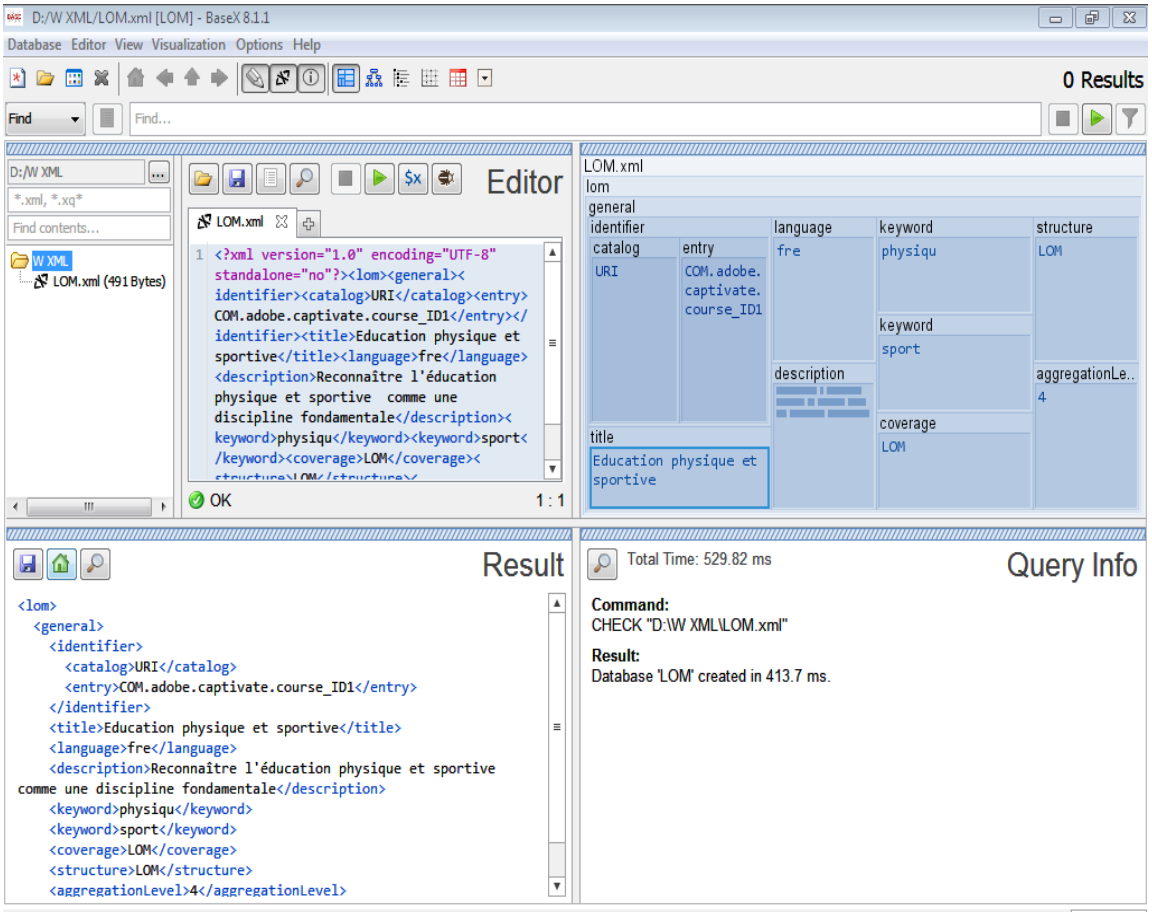

Figure 4.13 La base de données XML

4.7. Évaluation des résultats

✓ **Documents :**

Reconnaître l'éducation physique et sportive comme une discipline fondamentale, c'est afficher une ambition, celle d'un pays où le sport a toute sa place, où il est reconnu. Si l'activité physique et sportive a des effets positifs sur les comportements et l'apprentissage.

✓ **Prétraitement de documents :**

reconnaitr , educ, physiqu, sport, disciplin, fondamental, affich , ambition, sport reconnu, activit, comport, apprentissag

✓ **pondération de termes**

mots	TF	IDF	TFIDF
reconnaitr	1.0	0.33978952727983707	0.33978952727983707
educ	1.0	0.17884573603642702	0.17884573603642702
physiqu	2.0	0.35047480922384253	0.70094961844768506
sport	2.0	0.4734562936319301	0.9469125872638602
disciplin	1.0	0.33978952727983707	0.33978952727983707
fondamental	1.0	0.33978952727983707	0.33978952727983707
affich	1.0	0.33978952727983707	0.33978952727983707
ambition	1.0	0.33978952727983707	0.33978952727983707
reconnu	1.0	0.27047480922384253	0.27047480922384253
activit	1.0	0.27047480922384253	0.27047480922384253
comport	1.0	0.27047480922384253	0.27047480922384253
apprentissag	1.0	0.27047480922384253	0.27047480922384253

Table 4.2 Représentation des TFIDF

✓ **Le fichier XML**

```
<?xml version="1.0" encoding="UTF-8"?>
<lom>
  <general>
    <identifier>
      <catalog>URI</catalog>
      <entry>COM.adobe.captivate.course_ID1</entry>
    </identifier>
    <language>fre</language>
    <keyword> sport </keyword>
    <keyword> physiqu </keyword>
    <structure>LOM</structure>
  </general>
</lom>
```

4.8. Conclusion

Dans ce chapitre nous avons présenté certains outils utilisés dans la réalisation de notre projet , commençant par la présentation du corpus utilisé, et représentation d'une interface principale peut être définie comme étant un interprète qui assure et facilite le dialogue entre l'utilisateur et l'application.

CONCLUSION GENERALE

Nous avons présenté dans ce mémoire une généralisation sur les approches d'analyse intelligente de documents qui utilisent les techniques d'apprentissage automatique, et utilisateurs ont souvent besoin d'informations pédagogiques pour les intégrer dans leurs ressources pédagogiques, ou pour les utiliser dans un processus d'apprentissage. Une indexation de ces informations s'avère donc utile en vue d'une extraction des informations pédagogiques pertinentes en réponse à une requête utilisateur. La plupart des systèmes d'extraction d'informations pédagogiques existants proposent une indexation basée sur une annotation manuelle ou semi-automatique des termes pédagogiques, tâche qui n'est pas préférée par les utilisateurs.

Comme notre travail les ressources pédagogiques peuvent être décrites par des métadonnées comme le standard de l'IEEE, le Learning Object Meta data (LOM). Mais quel que soit le système retenu, leur implémentation n'est pas décrite par le standard. Ainsi, la description de plusieurs chansons pourrait se faire avec une implémentation technique en XML.

Malgré tout ça, nous avons entamé ce sujet en espérant qu'on a apporté une contribution significative, et qu'on a pu proposer des solutions plus au moins faisables aux problèmes précités et c'est exactement l'objectif d'une telle recherche. Par la même occasion le travail que nous avons réalisé nous a permis d'apprendre beaucoup des choses qui sont nouvelles pour nous telles que : les techniques et les méthodes de classification de documents, les algorithmes d'apprentissage, et les outils de programmation en langage Java NetBeans. Mais puisque rien n'est parfait, notre travail pourra subir des améliorations dans des projets futurs, notamment : l'utilisation de corpus français plus professionnels (de grandes tailles), l'application d'autres approches de représentation de textes, et pourquoi pas l'utilisation d'autres algorithmes d'apprentissage et d'établir une comparaison entre elles.

BIBLIOGRAPHIES

[1] Y. AMEROUALI, Métadonnées basées sur l'association d'éléments de description de ressources et d'éléments de profil d'utilisateur ,2005.

[2] A. BALLA, Un modèle générique d'environnement de développement des hypermédias adaptatifs et dynamiques générant des activités pédagogiques , Thèse de Doctorat d'état en Informatique, INI 2004.

[3] Mr. BENGHELIMA **,** Réalisation d'un système de recherche d'information , Thèse présentée devant université Abou Bakr Belkaid– Tlemcen, pp .30-36, 2013-2014

[4] A. BOUGOUIN, État de l'art des méthodes d'extraction automatique de termes-clés , LINA - UMR CNRS 6241, Université de Nantes, France, 17-21 Juin 2013

[5] Y. BOURDA ,Pourquoi indexer les ressources pédagogiques numériques ?, L'indexation des ressources pédagogiques, enssib, Villeurbanne,2004.

[6] A. BOUGOUIN, F. BOUDIN, TopicRank : ordonnancement de sujets pour l'extraction automatique de termes-clés, LINA - UMR CNRS 6241, Université de Nantes, UFR de Sciences et Techniques, 2 rue de la Houssinière, 44322 Nantes, France,2013

[7] A. BOUGOUIN, F. BOUDIN, B. DAILLE ,Influence des domaines de spécialité dans l'extraction de termes-clés, Traitement Automatique des Langues Naturelles, Marseille, 2014

[8] E. COLINET, I. ALBERTS, Les métadonnées nécessaires à la préservation de l'information numérique, École de bibliothéconomie et des sciences de l'information, Colinet, 2000

[9] E. CHERHAL, Présentation des standards : le Dublin Core ; Journée , Indexation des ressources pédagogiques , Cherhal, 2004

[10] D. DANIELEWSKI,Le LOM-FR, Educnet : nouvelles technologies de l'information et de la communication pour l'éducation – TICE , Danielewski 2005

[11] M. DAY, Installment on metadata , Digital Curation Manual, version 1.1, Day 2005

[12] E. DESMONTILS et C. JACQUIN,Annotation sur le web , Notes de lecture, Journées de l'AS-CNRS Web sémantique, 2002

[11] Mr. GERARD , Contribution théorique et méthodologique a l'élaboration d'un environnement de foad(formation ouverte et à distance), Thèse présentée devant

Institut National de formation en Informatique (INI) Oued Smar – Alger ,pp.92-93, 2004-2005

[12] O.GHEBGHOUB , LOMonto : Une ontologie pour l'indexation d'objets pédagogiques , Université de technologie de Compiégne, France,1998

[13] S. HIGGINS, What are metadata standards , Digital Curation Center (DCC), Higgins, 2007

[14] G. JACQUINOT, Apprivoiser la distance et supprimer l'absence? Ou les défis de la formation à distance ». Revue française de pédagogie, 102,pp . 55-67 , 1993

[15] A. J, Difining metadata dans Introduction to metadata : Pathways to Digital information , Edité par Murtha Baca ,Los Angeles ,Getty Information Institute , 2000.

[16] A. J, Introduction to metadata : Setting the stage ,Los Angeles ,Getty Information Institute,2000

[17] O. LAROUK, Cours sur l'Indexation Web et le catalogage des ressources numériques ,Master SIMIL, Unité d'Enseignement : SDN1, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB) Lyon, 2005.

[18] O. LAROUK, Typologie des métadonnées électroniques : Etat de l'art, définitions et Applications , Systèmes d'Information et Interfaces (SII), Ecole Nationale Supérieure
des Sciences de l'Information et des Bibliothèques (ENSSIB) Lyon, Larouk , 2007.

[19] L .MAHDAOUI, Proposition d'une Infrastructure Statique et dynamique pour le Support du Travail Collaboratif : Application au tutorat dans le e-Learning , Mahdaoui, 2008

[20] M. MIELNIKOFF , Qu'est-ce que l' e-Learning? , CRIITI-TTI,11.07.2005.

[21] C. MOREL, Métadonnées, pour quoi faire ? , Morel-Pair 2007

[22] D .PERAYA , N.DESCHRYYER , Staf17- Réalisation d'un dispositif de formation entièrement ou partiellement à distance, Notes pour le cours Staf17, TECFA ,2004

[23] P. PECCATTE, Les métadonnées : un élément clé de la gestion de contenu ; ATICA Deuxième journée de la réutilisation des données ; IPA Systems S.A. ; Soft Experience, Peccatte, 2002.

[24] P. PECCATTE, Métadonnées : une initiation - Dublin Core, IPTC, EXIF, RDF, XMP,etc , Soft Experience, Peccatte 2006

[25] Y.PEPIN, Savoirs pratiques et savoirs scolaires : une représentation constructiviste de l'éducation ,Revue des sciences de l'éducation , pp . 63-86 ,1994.

[26] S. YAHIAOUI , Structuration de métadonnées en vued'un filtrage d'information sur le Web , Thèse présentée devant Institut National de formation en Informatique (INI) Oued Smar – Alger, pp. 82 -84 ,2007-2008

[27] H. ZARGAYOUNA, Indexation sémantique de documents XML , Université Paris XI- UFR scientifique d'Orsay, Zargayouna 2005

Webographies

[28] http://www.enssib.fr/bibliotheque/documents/theses/amerouali/amerouali.pdf, Consultée le : 2/03/2015

[29] http://www.enssib.fr/pdf/Formist/journeeindexation/CHERHAL_presentation DCresume.pdf, Consultée le : 4/03/2015

[30] http://www.esi.umontreal.ca/~albertsi/INU1030/Cours/cours10.html, Consultée le : 05/03/2015

[31] http://www.dcc.ac.uk/resource/standards-watch/what-are-metadatastandards/, Consultée le : 6/03/2015

[32] http://www.esi.umontreal.ca/~albertsi/INU1030/Cours/cours10.html, Consultée le : 10/03/2015

[34] http://hal.ccsd.cnrs.fr/docs/00/04/04/73/PDF/Metas_panorama_CMO.pdf, Consultée le : 16/03/2015

[35] http://artist.inist.fr/article.php3?id_article=384 ; Consultée le : 17/03/2015.

[36] http://peccatte.karefil.com/software/Metadata.htm, Consultée le : 20/03/2015

[37] http://www.getty.edu/research/institute/standards/intrometadata/, Consultée le : 24/03/2015

[38] http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/metadata.pdf, Consultée le : 25/03/2015

[39] http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/metadata.pdf, Consultée le : 2/04/2015

[40] http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/metadata.pdf, Consultée le : 12/04/2015

[41] http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/metadata.pdf, Consultée le : 22/04/2015

[42] http://www.educnet.education.fr/articles/lom-fr.htm, Consultée le : 27/04/2015

ملخص:

البيانات الوصفية لوحدات التعلم (LOM) هو المعيار الأفضل لفهرسة المصادر التعليمية . وغالبا ما تنجز فهرسة هذه الموارد LOM يدويا من قبل أمناء المكتبات، في هذا العمل فإننا نقترح أداة للتعلم شبه لآلي لفهرسة وحدات التعلم تشير إلى توثيق مجموعة من الكلمات الرئيسية المرتبطة بهذه الموضوعات بناء على تحليل ومعالجة النصوص.

الكلمات المفتاحية : **الفهرسة ، استخراج كلمات ، Tf Idf ، خوارزمية التصنيف Porter.**

Abstract:

The Learning Object Metadata (LOM) is the undisputed standard for indexing éducational resources. Indexing of thèse resources LOM is often accomplished manually by librarians, In this work we propose to realize a tool for semi- automatic indexing of learning objects to suggest documentalists a set of keywords associated with thèse themes based on textual analysis of the resource.

Keywords: Indexing, Extraction keywords, Tf Idf, Algorithm porter.

Résumé:

Le Learning Object Metadata (LOM) est le standard incontestable pour l'indexation des ressources pédagogiques. L'indexation de ces ressources en LOM est souvent accomplie manuellement par des documentalistes, Dans ce travail nous proposons de réaliser un outil pour l'indexation semi-automatique des objets pédagogiques permettant de suggérer aux documentalistes un ensemble de mots-clefs associés à ces thématiques en se basant sur l'analyse textuelle de la ressource.

Mots clés : Indexation, Extraction des mots-clés, Tf Idf, Algorithme de porter.