الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الاعلام الالي
Département d'informatique

En vue d'obtenir le diplôme de master

Domaine : Mathématique et informatique

Filière : Informatique

Spécialité : Informatique

Option : Traitement automatique de langue

Thème

Système de création d'unités éditoriales à partir des documents plein texte vers des documents semi structurés en XML

Présenté par :

- ben salah Mohamed amine
- ould larbi amine

promoteur :

-Nasri Ahlam

Encadreur :

-BOUGUERRA ELOUANES

Soutenu le :

Devant le jury :

-M cherif zahar	Président
-M .oueld aissa	Examinateur
-Mme Nasri Ahlam	Promoteur
-M.BOUGUERRA ELOUANES	Encadreur

Année universitaire : 2018/2019

Remerciements

Tout d'abord, nous remercieront Allah de nous avoir aidé et donné la force et la volonté de réaliser ce travail

Ensuite, nous tenons à exprimer nos plus vifs remerciements et gratitude à notre Encadreur Mr BOUGUERRA ELOUANES pour son encadrement continu, pour les remarques constructives Qu'il nous a fournies ainsi que pour ses précieux conseils durant toute la période de notre travail. On le remercie également pour la confiance qu'il nous a accordée et pour la grande liberté d'idées et de travail qu'il nous a donnée. Nous n'oublierons pas aussi de les remercier pour ses qualités humaines, son hospitalité et son soutien qui ont permis de bien mener bien cet ouvrage.

Nous tenons à remercier les membres du jury d'avoir bien voulu participer à l'évaluation de ce travail.

Quelques personnes ont contribué à la réalisation de ce travail et méritent des remerciements. Nous somme également redevable notre promotrice Mme ahlam nasri qui croyait en notre capacité à réaliser quelque chose de grand. En nous donnant l'opportunité de poursuivre ce rêve et de le réaliser. Avec un cœur plein de gratitude.

Enfin, nous tenant à remercier nos familles pour leur encouragement, leur aide et leur grande patience avec nous.

Résumé

La plupart des documents du patrimoine arabe sont des fichiers de texte soient composés de textes non organisés Ou bien ils ont une structuration (physique et logique) qui ne permet pas l'identification et l'exploitation.
Notre travail consiste à proposer une approche pour réaliser des conversions Les document de patrimoine arabe vers des documents semi structuré en XML.
Ce projet est réalisée en deux étape la première étape consiste à l'extraction des métadonnées titre auteur maison édition … par trois étape premier étape les règles de font, la deuxième les règles linguistique, et le dernier les expressions régulière
et la seconde étape de notre projet c'est la détermination la structure des documents de patrimoine arabe en fonction de leur table des matières par étude de plusieurs livre
avec création un DTD XML pour examiner les livre après la conversion
a la fin nous avons fait des expérimentation sur plusieurs livres de patrimoine arabe et les résultats obtenus sont meilleurs.

Mots clé : document semi structuré, métadonnée, XML, corpus CNDA, traitement automatique de langue ,la langue arabe.

Abstract

Most of the Arab heritage documents are text files that are composed of unorganized texts Or they have a structuring (physical and logical) that does not allow identification and exploitation.

Our job is to propose an approach to achieve conversions Arab heritage document to semi structured documents in xml.

This project is carried out in two stages The first step is to extract the metadata title author house edition ... by three step first step the rules of font, the second the rules language, and the last the regular expressions

and the second step of our project is the determination of the structure of the Arab heritage documents according to their table of contents by study of several books

with creating a xml DTD to examine the books after the conversion

in the end we did experimentation on several books of Arab heritage and the results obtained are better.

Keywords: semi structured document, metadata, xml, corpus CNDA automatic language processing, Arabic language.

ملخص

معظم كتب التراث العربي عبارة عن ملفات نصية تتكون من نصوص غير منظمة أو لديها بنية (جسدية ومنطقية) لا تسمح بتحديد الهوية والاستغلال.

مهمتنا هي اقتراح نهج لتحقيق تحويلات كتب التراث العربي إلى كتب شبه منظمة في XML

يتم تنفيذ هذا المشروع على مرحلتين. الخطوة الأولى هي استخراج البيانات الوصفية العنوان المؤلف المحقق ... بواسطة ثلاث مراحل الأولى قواعد الخط ، والثانية القواعد اللغة ، والأخيرة التعبيرات العادية

والخطوة الثانية من مشروعنا هي تحديد هيكل كتب التراث العربي وفقا لجدول محتوياتها من خلال دراسة العديد من الكتب

مع إنشاء DTD xml لفحص الكتب بعد التحويل.

في النهاية أجرينا تجارب على العديد من كتب التراث العربي والنتائج التي تم الحصول عليها جيدة.

الكلمات المفتاحية: وثيقة شبه منظمة ، بيانات وصفية ، الذخيرة العربية، XML,معالجة تلقائية للغة ، اللغة العربية ،

Sommaire

Liste des figures

Liste des tableaux

Liste des abréviations

DTD : document type définition
Rdf : Resource Description Framework

Introduction général

Le contexte

Bien que la plupart des fichiers de texte soient composés de textes non organisés, les documents du patrimoine arabe partagent souvent une structure inhérente, bien que non documentée. Afin de faciliter une recherche archivistique efficace et coordonnée et de permettre l'intégration d'informations dans des collections de textes contenant des sources de données pertinentes, cette structure inhérente doit être interprétée aussi précisément que possible. L'inférence avec la DTD et la conversion ultérieure du texte correspondant dans des documents XML sont un moyen efficace d'atteindre cet objectif.

Il est intéressant de convertir ces documents en un document semi-structuré format. Les motivations pour la conversion de documents sont diverses, comprenant généralement l'intention de réutiliser ou de réutiliser des parties des documents.

Le formalisme XML est devenu le standard industriel pour l'échange de données entre services et entre entreprises. Utiliser XML comme format d'échange pour la capture et la réutilisation d'informations est devenu critique pour les entreprises. Ce formalisme offre de nouvelles possibilités dans le domaine de la gestion documentaire, de la publication ou du multimédia. Le langage XML est devenu le standard le plus utilisé pour la représentation de la structure logique des documents. Une partie de cette structure est appelée métadonnées (le titre, auteur,...).

Dans notre mémoire on s'intéresse par proposer un approche qui transformer les documents de patrimoine arabe vers des documents semi structure en xml et extraire les métadonnées.

.

Problématique

Le processus de conversion nécessite souvent la définition d'un modèle de document cible exprimé par une grammaire XML. Cette grammaire peut être représentée
par exemple sous la forme d'un XML Schéma, d'une DTD (doucement type définition)
Elle définit les éléments structurels et sémantiques des documents de la patrimoine arabe,
Le langage XML est devenu le standard le plus utilisé pour la représentation de la structure logique des documents. Une partie de cette structure est appelée métadonnées (comme le titre, la date, auteur,...).
La reconnaissance de la structure logique d'un document textuel (en format Word) est une tâche complexe. Dans notre mémoire, nous allons étudier cette problématique en essayant de répondre à des questions comme :
Comment extraire les métadonnées ?
Comment reconnaitre auteur, enquêteur, maison édition …. ?
Comment identifier la structure logique et physique d'un document du patrimoine arabe?
Comment faire pour transformer les documents du patrimoine arabe .docx vers des documents semi structuré en XML ?

Contributions

Pour répondre aux questions précédemment posées, notre mémoire apporte les propositions suivantes :

-Proposition une approche globale bien détaillée décrit toutes les étapes de système

-détecter la structure de document du patrimoine après étude de plusieurs livres et faire un DTD (Document type définition) XML

-proposition une approche pour extraction des métadonnées par trois étapes

Premier étape les règles de fonte

La deuxième les règle linguistiques exploité par la académie algérienne de la langue arabe et dernier étape est expressions régulières

-Proposition une approche Structuration les documents en fonction de leur table des matières pour convertir les documents de la patrimoine arabe vers des documents semi structuré en XML basé sur les travaux précédents comme

X. Lin, «Navigation automatique dans les documents pour le remaniement du contenu numérique», Rapport technique HP, 2003,X. Lin, "Scission de journaux basée sur l'exploration de texte", Actes de la septième Conférence internationale sur l'analyse et la reconnaissance des documents, ICDAR'03, 2003..

Nos contributions ont été évaluées en utilisant des documents du patrimoine arabe

En format Word extraire par l'académie algérienne de la langue arabe

Les objective de notre projet :

Ce projet fait partie du projet le corpus de CNDA (اللجنة الوطنية للذخيرة العربية) projet corpus CNDA est un projet arabe à superviser par l'organisation de la langue arabe pour l'éducation, la culture et la science. Son objectif est de créer une banque électronique pour la langue arabe déjà utilisée et de créer un dictionnaire électronique avec l'arabe dans les mots correspondants en anglais et en français. Le projet est dirigé par le Dr. Abdul Rahman Haj Saleh.

Organisation du mémoire

Le mémoire est organisé en deux parties. La première partie décrit et représente l'état de l'art sur les documents numériques, et leur structure et aussi le langage XML (chapitre I, chapitre II). La deuxième partie est destinée à la représentation et l'évaluation de nos contributions (chapitre III, et chapitre IV).

Le premier chapitre représente l'état de l'art des documents numériques, leur, les déférents types de structure de documents, et les déférentes représentations aussi les métadonnées

Le deuxième chapitre composé en deux parties

Premier partie est un aperçu sur la langage XML et les différents extensions DTD **(**Document Type Définition) xpath Xslt xml schéma aussi Framework rdf (Resource Description Framework) et le norme de livre ISBN .

On deuxième partie on représenté traitement automatique de langue et leur application

Le chapitre III nous allons expliquer l'approche que nous avons utilisée pour la convertion des documents de patrimoine arabe .docx (en format word)vers des documents semi structuré en XML et extraire les métadonnées. Nous allons détailler toutes les étapes que nous avons suivies.

Chapitre IV est consacré à l'exposition de nos expérimentations et résultats des évaluations appliquées sur notre système. Nous terminons la mémoire par une conclusion et des perspectives.

Partie I- Etat De L'art

Chapitre 1 : Documents Numériques et structure

1.1 Introduction

Le concept «document numérique» est un concept très récent par rapport au document papier, Il est apparu avec l'apparition des nouvelles technologies de l'informatique. Mais il reste toujours difficile à le définir et à le référer.

Dans ce chapitre nous serons parlons sur c'est quoi le un document et définitions et représentation du document numérique.

Aussi le notion de métadonnée et les normes dublin core et onix

1.2Notion de document

Il est difficile de donner une définition complète et précise de la notion du document. on peut appeler «document » le contenu de ce que l'on produit, distribue, utilise ou garde lors d'un processus de communication écrite ou électronique. Nous limiterons toutefois id ce terme aux objets ainsi manipules lorsque la partie textuelle est prépondérante. (1)

1.3 Document numérique

C'est un ensemble composé d'un contenu, d'une structure logique, d'attributs de présentation permettant sa représentation, exploitable par une machine afin de restituer une version intelligible pour l'homme.

par le dictionnaire Larousse : « tout renseignement écrit ou objet servant de preuve ou d'information ».. Une telle définition est fournie par l'ISO : « Un document est l'ensemble d'un support d'information et des données enregistrées sur celui-ci sous forme en général permanente et lisible par l'homme et la machine». (2)

1.3.1 Documents non structurés, semi-structurés et structurés

1.3.1.1 Document non structuré

Les documents non structurés, appelés encore documents « plats », sont des documents qui n'intègrent aucune marque explicite d'élément de structure. Ainsi, dans ces documents, on ne retrouve pas la disposition et l'emplacement des informations. Le document est présenté comme une suite de caractères (plein texte). Selon (Bringay et al. 2004), un document plat est un document pour lequel ni le lecteur ni le système n'est capable de décrire ou détecter une structuration de son contenu (3).

1.3.1.2 Document semi structuré

Un document semi-structuré est un pont entre données structurées et non structurées Données non structurées (Également appelées données à plat) sont des données dont nous ne connaissons ni le contexte, ni la manière dont les informations sont corrigées. Il comprend documents contenant principalement du texte en langage naturel, tels que des fichiers de traitement de texte, des courriers électroniques et des champs de texte de bases de données ou applications. Les données semi-structurées apparaissent lorsque la source n'impose pas de structure rigide (telle que le Web) et lorsque les données sont combinées à partir de plusieurs sources de données hétérogènes (4)

```
<article>
  <titre> Remplacement d'équipement en période de changement technologique
</ titre>
  <journal>
  <nom> Nav. Res. Logist. </ nom>
  <volume> 41 </ volume>
  <numéro> 1 </ numéro>
  </ journal>
  <abstract> Pour les problèmes d'économie de remplacement à horizon infini, il est
courant
pratique de tronquer le problème à un horizon fini. Nous développons des limites sur
l'erreur
en raison d'une telle troncature. Les limites sont illustrées par un exemple numérique
(voir page
148) à partir d'un cas réel de remplacement de véhicule
. </ abstract>
</ article>
```

Figure 1.1 exemple de document semi structuré

1.3.1.3 Document structuré

'il existe beaucoup de documents structurés. Ils sont beaucoup plus lisibles par l'humain que les documents non-structurés. Le type le plus commun est le document organisé de manière hiérarchique, que ce soit un texte structuré en chapitres, sections et paragraphes, que ce soit un code source avec les fonctions et les objets, voir même un fichier, . . . Même une carte peut se voir comme un dessin où des nœuds relient des arêtes. Nous donnons dans la suite des types de documents structurés extrêmement importants en pratique (4).

1.4 Représentation de document

Les documents doivent être convertis en une forme de représentation qui est lisible par la machine et adéquate pour les logiciels.
On classe généralement les modèles de documents en trois grandes catégories : les modèles qui se réfèrent à la structuré logique décrit l'organisation du contenu sous forme d'éléments logiques et ceux qui se réfèrent à la structure physique qui permet de regrouper les caractéristiques visuelles du contenu.et la structure sémantique qui permet d'expliciter le sens d'un contenu.

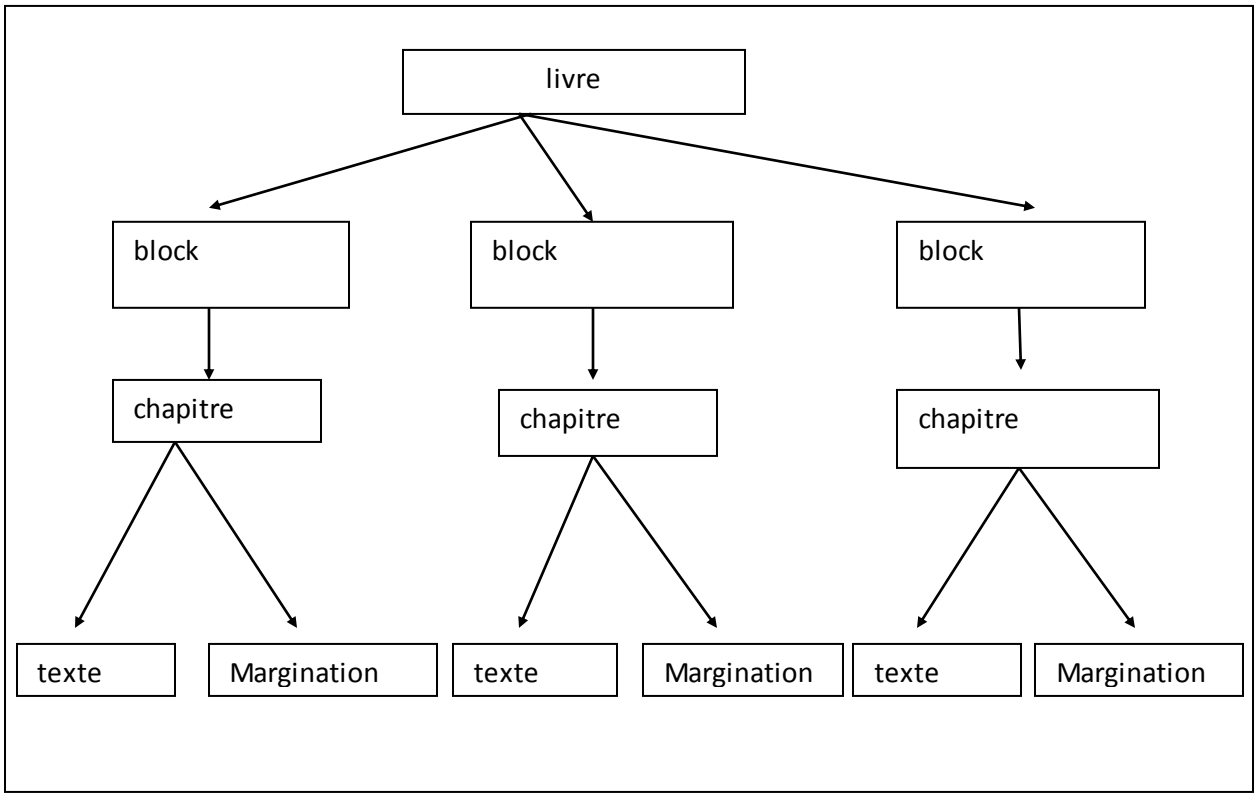


Figure 1.2 Représente d'un arbre de livre arabe

1.4.1 la structure physique d'un document.

La structure physique d'un document correspond à l'organisation du contenu par découpe le document en pages, qui déterminé les espaces et les zones de texte. Test la structure physique qui joue sur les polices, les corps et les attributs typographiques du texte. Même si, dans certains cas, comma dans Ia publicité par exemple, la structure physique a pour principal rôle d'attirer l'ceil et de faire passer une information ou un sentiment a travers le seul aspect graphique, le plus souvent elle traduit seulement Ia structure logique. La structure physique permet de décrire les relations Existantes entre les divers objets physiques (5).

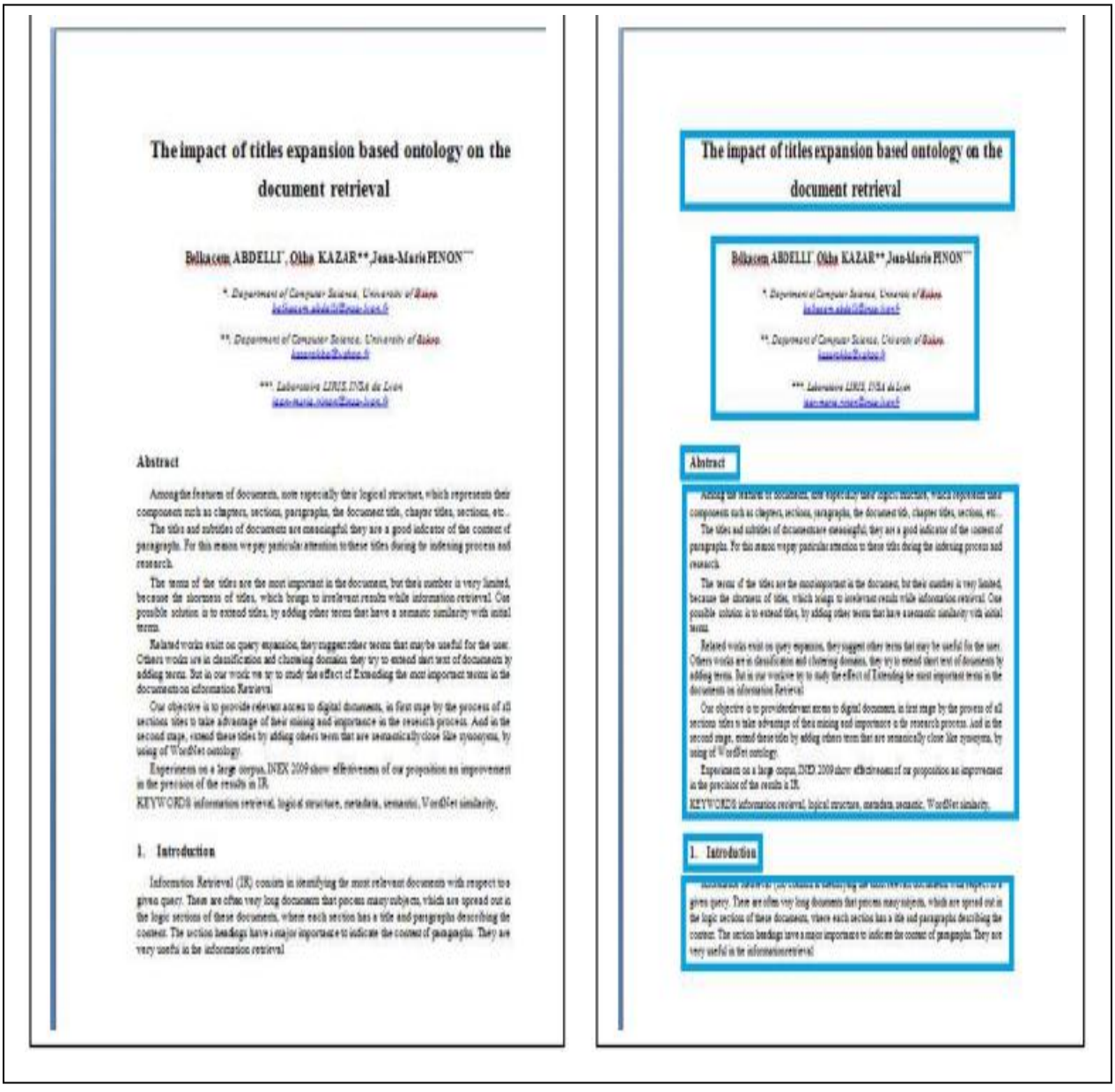


Figure 1.3 Exemple représente un bloc de text

1.4.2 La structure logique d'un document

La structure logique d'un document explicite la signification de chaque composant de la structure physique. Elle reflète la façon dont l'information est organisée en termes d'objets logiques, à savoir, chapitres, sections, sous sections, titres, paragraphes, figures, en-têtes, etc. La structure logique spécifie la fonction et la signification des objets physiques formant le document et les relations entre eux.

Les relations entre les objets logiques dans la structure logique sont généralement sous une forme hiérarchique.par exemple un article comprend En-tête, Résumé, Corps, Référence

Chaque corps contient plusieurs sections et aussi chaque En-tête contient un titre auteur affiliation (6).

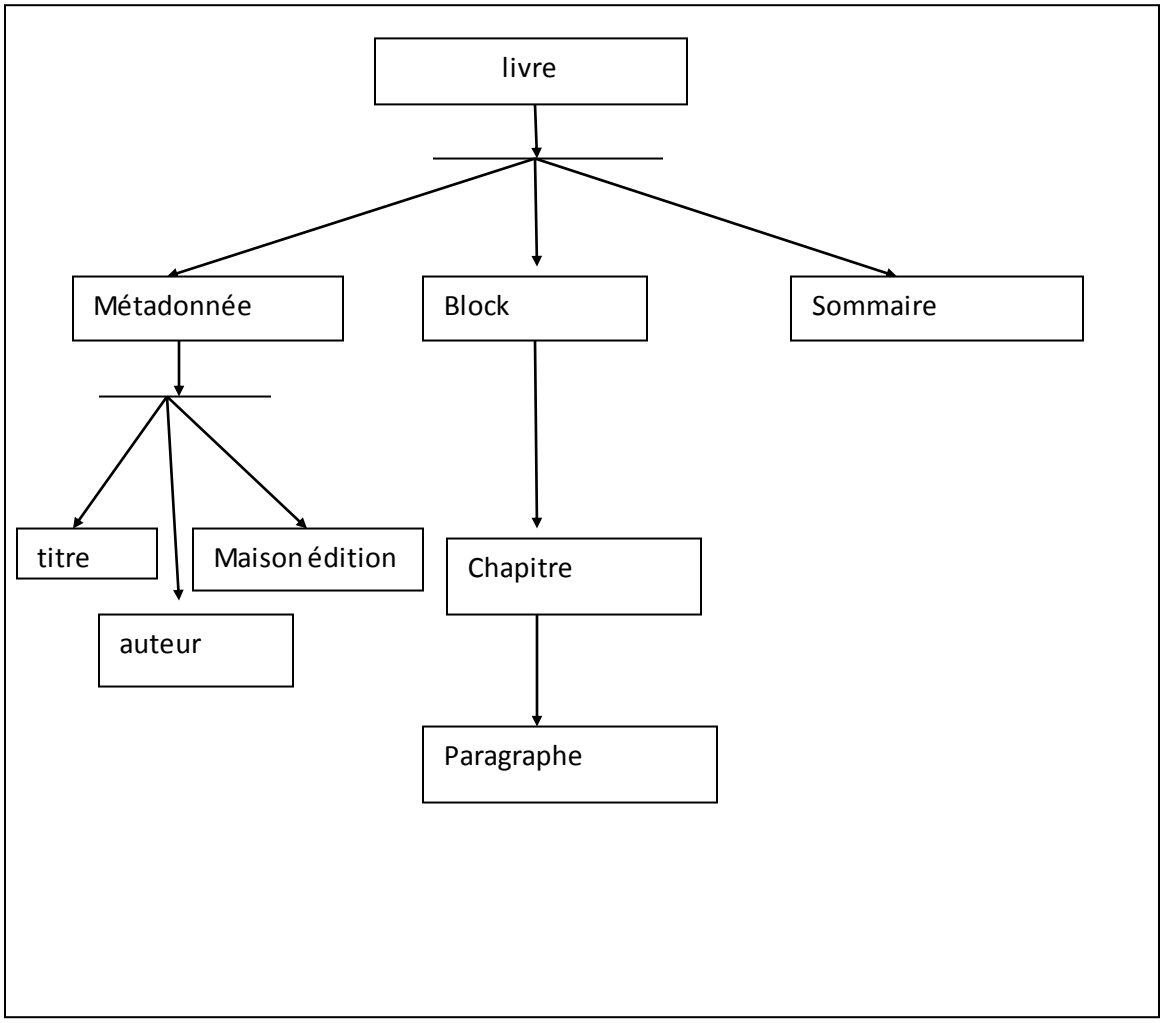


Figure 1.4 exemple représente Structure Logique d'un livre

Pour modéliser la structure logique du document, plusieurs outils de structuration documentaires sont apparus, parmi ceux-ci il faut citer le langage XML (Extensible Markup Language défini par W3C3), qui est bien standardisé et bien formalisé. XML est un langage défini pour faciliter la manipulation et l'échange de documents, grâce à ce langage de nouvelles tendances sont développés tels que ; la recherche d'information par le contenu, le web sémantique, le web service... etc. XML est connu par l'utilisation des marques spéciales appelée balises qui rendent la structure du contenu bien explicite, et bien délimité. Ce langage qui sera détaillé dans le deuxième chapitre)

```
<livre>

<métadonnée>

<titre>.....</titre>

<auteur>.....</auteur>

<maison_édition>

…

</maison_édition>

</métadonnée>

<block>

<chapitre>

</chapitre>

<block>

<sommaire>…

.</sommaire>

</livre>
```

Figure 1.5 La structure logique d'un livre (de la figure 1.4) représenté par XML

1.4.3 la structure sémantique d'un document

La structure sémantique est l'organisation des entités d'informations qui représentent des idées ou des connaissances décrites dans le document.
La structure sémantique permet de décrire le sens du contenu du document, et définir la relation sémantique entre les termes du contenu, la structure sémantique et décrite par un ensemble de concepts au lieu de mots. Pour extraire les concepts on doit utiliser les outils de traitement automatique de la langue (TAL).
pour comprendre bien la structure sémantique de on a fait un structure de la thèse nous nous somme les appuyés sur l'utilisation de métadonnées capables de catégorise de la thèse
a partir de ces métadonnées, la thèse est donc découpée en segments qui correspondent à l'expression d'un ou plusieurs concepts. Ces segments ne sont pas indépendants les uns des autres. Il existe des relations entre les différents segments (7).

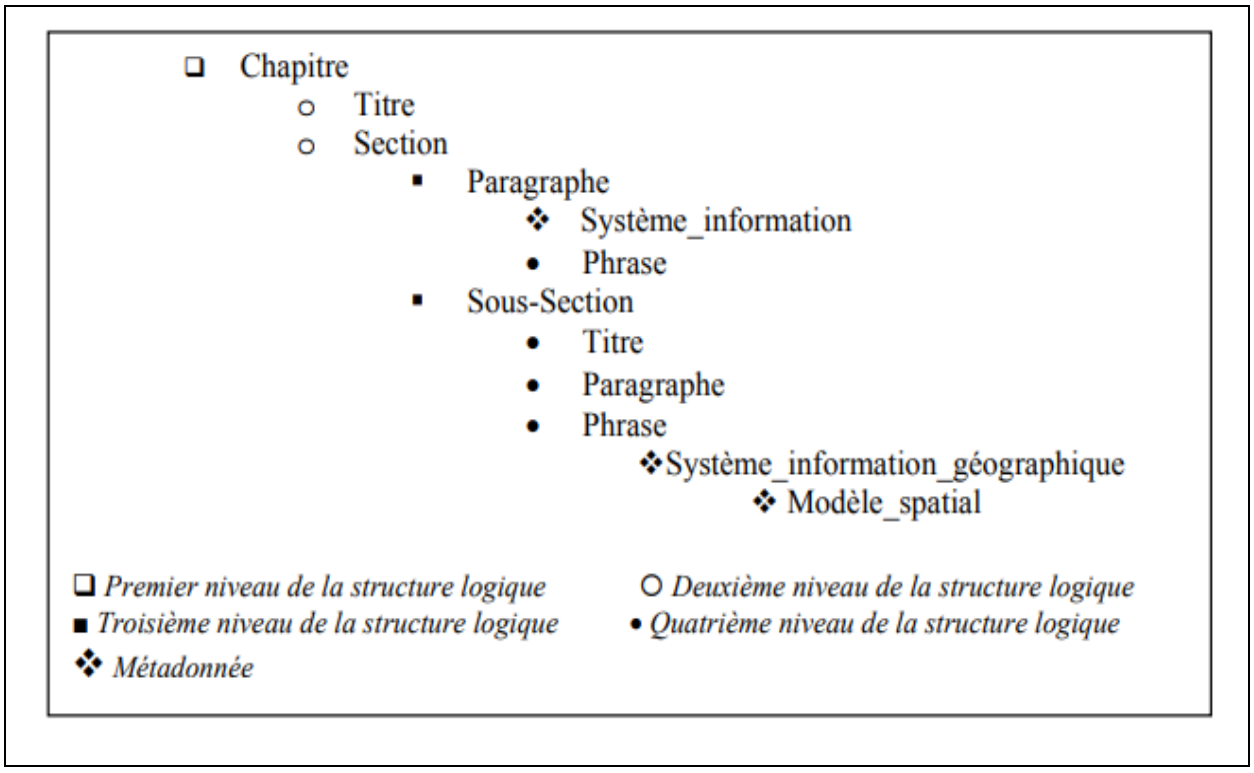


Figure 1.6 exemple de structure logique et sémantique d'une thèse

1.5 Les formats de fichiers

Un format de documents électroniques est un ensemble de règles ou conventions qui régissent l'interprétation de documents électroniques. On distingue la syntaxe et la sémantique d'un format. La syntaxe est un ensemble de règles auxquelles une séquence de caractères doit se conformer pour être reconnue comme un document valide; la sémantique est un ensemble de règles permettant de transformer un document électronique valide en un document "réel" (8).

1.5.1 Les formats de fichiers pour documents textuels

1-Doc :

Cette extension de fichier est utilisée pour représenter beaucoup de formats de textes. Le logiciel Microsoft Word, en version Windows, utilise l'extension .doc pour ses fichiers. Ces derniers sont assez répandus dans Internet.

. Utiliser : Windows , Logiciels de traitement de texte.

2-PDF (Portable Document Format)

Le format PDF est à présent un format standard d'échange des documents via Internet, grâce à sa compacité et à sa visualisation robuste fonctionnalités. Malgré ces capacités importantes, les documents PDF sont difficiles à indexer d'extraction d'informations car leur contenu est le plus souvent désorganisé du fait de la d'optimisation et ne respectent donc pas l'ordre de lecture. Ainsi, existant Les systèmes d'indexation doivent toujours prétraiter les documents PDF afin d'extraire et structurer le contenu (9).

3--RTF (Rich Text Format) :

Est reconnu par la majorité des logiciels de traitement de **t**exte et permet de conserver les accents, les styles et la mise en page du document,
Il est utilisé par défaut dans l'éditeur texte dit de Mac os, dans Word pad de Windows, et dans le traitement de texte *Ted*, courant sous les systèmes de type Unix (10).

4-ODF (Open Document Format):

Le format ODF (Open Document Format) est un fichier basé sur XML langage de balisage. ODF est un système ouvert, non exclusif format englobant les documents texte, les présentations et les feuilles de calcul.
Un an plus tard, le format a été approuvé pour publication en tant que
Norme internationale ISO et CEI.
ODF a l'avantage d'être le format par défaut pour OpenOffice (version 2.0 et supérieure).
OpenOffice est une suite bureautique open source gratuite et peut exporter des fichiers vers de nombreux formats différents, y compris MS Word, format RTF, HTML, PDF et autres. D'autres applications, telles que Koffice, utilisent Open Format du document. (11).

1.6 Les métadonnées

Les métadonnées sont des informations structurées qui décrivent, expliquent, localisent une ressource d'information, ces ressources deviennent plus facile à utiliser, sont un maillon essentiel pour l'interopérabilité de l'information et sa gestion. Des applications sont développées par de nombreux acteurs de différents domaines pour tous les types de ressources numériques; à côté du jeu d'éléments Dublin Core, central pour l'interopérabilité, il existe des ensembles complémentaires tout aussi importants, qui répondent chacun à des besoins particuliers. Leur implémentation se fait aujourd'hui principalement en XML, et l'adoption du format RDF permettra d'atteindre une réelle interopérabilité entre ces ressources. Ce document est un essai de synthèse sur le contexte d'apparition des métadonnées, les principaux jeux d'éléments et leurs formats d'implémentation, avec des exemples d'applications. (12)

1-Rattachement des métadonnées au document :

Comme les métadonnées sont une information à propos des données qui composent le document, il importe de pouvoir les rattacher au document ou de créer un lien entre elles. Il existe cinq façons d'effectuer un rattachement entre une « ressource » (le document proprement dit) et les métadonnées qui y sont associées:

· par insertion : l'énoncé est contenu dans la ressource ;

· par accompagnement : l'énoncé est externe, mais véhiculé avec la ressource

· par attache lâche à un service (service bureau) : l'énoncé est externe, obtenu séparément, indiqué simplement avec un URL (Uniform Resource Locator) ;

· par attache insécable à un service (service bureau) : l'énoncé est externe, obtenu séparément, indiqué par un lien insécable grâce aux techniques cryptographiques

par enveloppement : l'ensemble des énoncés contient la ressource. (13)**.**

1-Métadonnées externes aux documents

Dans la plupart des systèmes informatisés, les métadonnées sont stockées dans une base de données spécifique et reliées à la ressource avec un lien hypertexte. C'est la technologie utilisée habituellement dans les systèmes documentaires pour retrouver les documents recherchés au sein d'un ensemble de ressources et avec la souplesse nécessaire (recherches sur plusieurs critères, recherche sur plusieurs structures, etc.) (14) .

2-Métadonnées internes aux documents

Pour inclure un ou plusieurs jeux de métadonnées dans une ressource, il faut ajouter un balisage à cette dernière. Les métadonnées sont alors insérées dans la ressource sous forme d'informations textuelles renseignées directement à partir de la ressource. (14).

2-Importance de métadonnée

L'objectif principal des métadonnées est de faciliter l'exploration des informations pertinente, en plus de la découverte de nombreuses autres tâches qu'on décrit dans les points suivants (6)
- Découverte des ressources Les métadonnées permettent la découverte de ressources électroniques par le biais de :
  - Diagnostiquer et identifier les ressources
  - Combiner ensemble les ressources similaires
  - Distinguer les ressources qui ne ressemblent pas.
  - Donner des informations de localisation

-Identification numérique C'est un numéro standard pour identifier de manière unique
La ressource ou l'objet pour que les métadonnées se réfèrent.

3-Métadonnées pour les documents numériques

outils de recherche d'informations opérationnels actuellement, attribuent généralement à chaque document une simple liste de mots clés pour permettre leur recherche. Les travaux étudiant la création, la structuration et la modélisation des descripteurs soutiennent par contre que c'est en organisant les descripteurs dans une structure que des requêtes complexes pourront être posées. Ainsi la recherche d'informations portant sur ces documents numériques devient plus riche. Les documents numériques contiennent de nombreuses métadonnées implicites et/ou explicites, classées selon trois principaux types (14) :

- ✓ **Les métadonnées descriptives ou sur le contenu** : elles donnent de l'information sur le contenu d'un document (nom de l'objet, titre, matériaux, dates, description physique, etc.).
- ✓ **Les métadonnées techniques** : il s'agit de métadonnées sur les média utilisés eux mêmes, et non sur les contenus (auditifs ou visuels) véhiculés. Ces métadonnées donnent de l'information sur les processus techniques à employer pour la saisie, la manipulation ou la restitution des média : images, audio, couleurs, bande passante, formats de fichier, etc. Une partie des données techniques enregistrées sur une image (exemple : le type de fichier d'image) doit être sous forme numérique (et respecter des formats techniques précis tel que le format jpeg) afin qu'un système informatique puisse afficher correctement l'image.
- ✓ **Les métadonnées administratives** : elles contiennent l'information liée à la gestion des documents (exemple : la gestion des droits).

4-Normes et standards de métadonnées

1-La norme onix

ONIX est une norme internationale permettant de représenter et de communiquer des informations sur les produits de l'industrie du livre sous forme électronique Il fournit un format de message XML pour l'échange d'informations entre systèmes, qui peuvent utiliser en interne différents systèmes de métadonnées. Les éléments de données ONIX ont été définis pour les informations sur le produit - en-têtes de message, numéros de référence et de produit, auteur, sujet, éditeur, etc. - pour les livres, les vidéos et les produits multimédias associés. ONIX est principalement gérée par éditeur une organisation composée de membres qui se concentre sur les normes de commerce électronique dans les industries du livre et des publications en série

ONIX est une norme complexe qui se présente sous forme d'un fichier XML lourd avec une syntaxe et des règles d'applications rigoureuses et difficiles à maitriser: (15).

```
<product>
<a001>9438000062</a001>
<a002>03</a002>
<productidentifier>
<b221>02</b221>
<b244>2765406553</b244>
</productidentifier>
<b012>BC</b012>
<title>
<b202>01</b202>
<b203 textcase = "02">Traité pratique d'édition</b203>
</title>
<contributor><b035>A01</b035>
<b037>Schuwer, Philippe</b037>
<b044>Philippe Schuwer a été secrétaire de rédaction dans la presse, sousdirecteur de fabrication aux PUF, directeur aux éditions Tchou, directeur de département chezHachette, Nathan et Larousse. Diplômé du British Institute et de l'Ecole des hautes études ensciences sociales, il a créé les premiers cours d'édition à l'Université Paris VIII.</b044></contributor>
```

Figure 1.7 exemple de norme onix

2-Dublin core

Beaucoup de formats de métadonnées sont élaborés dans une variété d'environnements et de disciplines utilisateur. Le plus connu est Dublin Core .
L'ensemble des éléments de métadonnées du noyau de Dublin (DCMES) est né d'une réunion de 1995 à Dublin, dans l'Ohio.
Qui était axé sur les métadonnées pour les informations électroniques en réseau. Les participants ont été chargés de l'identification d'un ensemble de fonctionnalités de base communes à la plupart des types d'informations numériques. Dans cette première
Lors de la réunion, 13 éléments de base ont été définis, qui sont rapidement passés aux 15 éléments connus sous le nom de DCMES.
Aujourd'hui. Ce sont: contributeur, couverture, créateur, date, description format, identifiant, langue, éditeur, relation, droits, source, sujet, titre et type. Cet ensemble, également appelé (simple Dublin) (16).

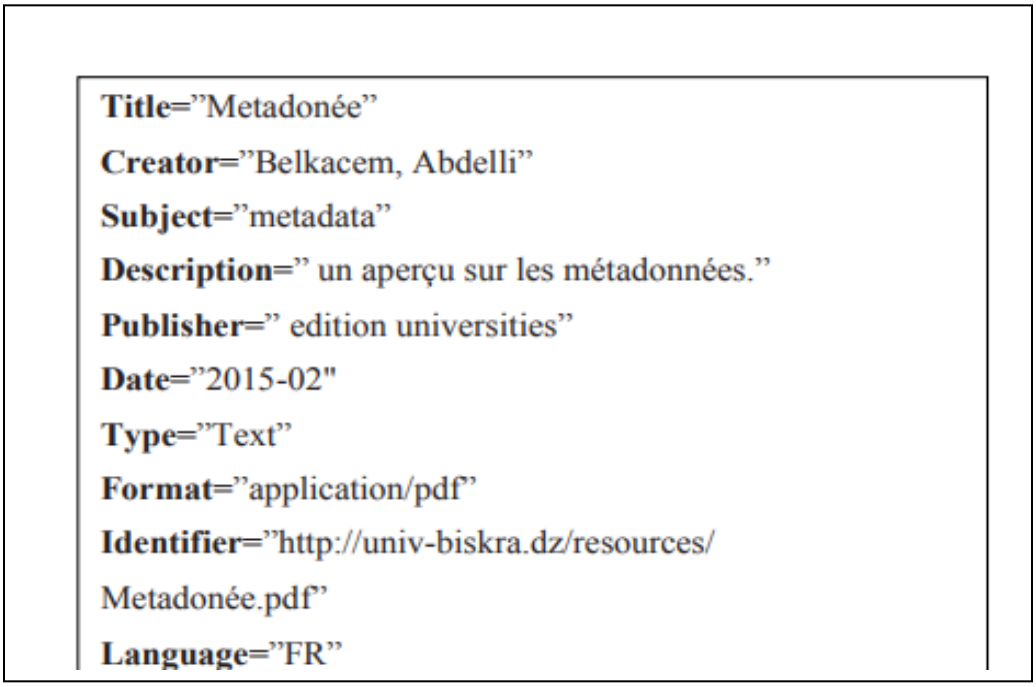


Figure 1.8 exemple de Dublin core

1.7 Conclusion

L'objectif de ce chapitre est de présenter un état de l'art sur la notion de document numérique Nous avons présenté les notions et les concepts préliminaires liés au document numérique. Aussi les trois modèles de représentation, ainsi que la structure qui le forment.
et sur celle de métadonnées associées à ce document
, Dans le chapitre suivant nous allons décrire le langage XML Pour modélisé les documents Semi structure.

Chapitre 2 : langage XML et traitement automatique de langue

2.1.1 Introduction

Le traitement des documents numériques en particulier les documents semi structurés peut être associés au langage XML (extensible Mark up Language).

XML est un langage hiérarchique simple, il est constitué d'une structure type et d'un ensemble de règles qui permet de définir une structure associée à un document, il permet également la description et l'échange des documents semi-structurés sur le Web qui se présente sous forme de balises de début et de fin encadrant un fragment du document.

Le traitement automatique des langues (TAL) s'intéresse aux traitements informatisés mettant en jeu du matériau linguistique : analyse de textes, génération de textes, les relations syntaxiques entre les mots d'une phrase, aussi pour l'accès au contenu des documents.

2.1.2 langage XML

Extensible Markup Language [XML] est un format de fichier destiné à représenter sous forme textuelle des données arborescentes. On peut considérer XML comme une syntaxe concrète pour un modèle abstrait d'arbre. Il n'y a pas de consensus sur la nature exacte de ce modèle de document, mais dans une vision simplifiée, les arbres XML sont des arbres d'arité variable, dans lequel les nœuds (appelés éléments) sont étiquetés et possèdent chacun un ensemble fini d'attributs textuels, et dont les feuilles sont des caractères. (18)

La Figure 1 affiche un document XML contenant des informations sur un livre. Dans cet exemple, il existe un élément book qui a deux sous-éléments, booktitle et author. L'auteur L'élément a un attribut id avec la valeur "dawkins" et est en outre imbriqué pour fournir le nom et l'adresse.

```
<livre>

<titre_livre>  Le gène égoïste  </ titre_livre>

<auteur  id = "dawkins"  –">

<prénom>  Richard  </ prénom>

<nom>  Dawkins  </ nom>



<adresse>

<cité>  Tombouctou  </ cité>

<zip>  99999 </ zip>

</ adresse>

</ auteur>

</livre>
```

Figure 2.1 Exemple d'un document XML

2.1.3 Modèles de règles syntaxiques

Dans la gestion documentaire, les Document Type Définition (DTD) ont une grande importance.
Ce sont des modèles qui définissent des règles syntaxiques supplémentaires. Ce procédé permet de définir une structure de document XML, qui pourra être réutilisée. Ainsi tous les documents XML qui définissent dans leurs prologues la référence à une DTD, devront tous suivre l'ensemble des règles syntaxiques définies dans cette DTD. Cela permet de définir des règles communes à un ensemble de documents. Une DTD en XML permet de définir d'une part les balises qui délimitent les éléments d'un document dans leur ordre logique de succession et d'autre part les attributs de ces éléments. Une DTD contient donc la liste des éléments et des attributs qu'un document XML peut contenir. De plus, une DTD définit le type de contenu d'un élément, elle indique également si un document doit contenir des renvois, ou bien si l'ensemble des éléments doit être numéroté de façon absolue. Les notions de métadonnée et d'élément sont les mêmes. L'élément donc est la représentation « physique » (dans un document XML) d'une métadonnée. Nous pouvons de plus ajouter une nuance, qui donne à un élément son titre de métadonnée quand celui-ci donne réellement une indication sur les informations qu'il renferme. (18).

```
<!DOCTYPE Article [

<!ELEMENT article (titre, auteur, année, résumé, section )>

<!ELEMENT titre (#PCDATA)>

<!ELEMENT auteur (#PCDATA)>

<!ELEMENT année (#PCDATA)>

<!ELEMENT résumé (#PCDATA)>

<!ELEMENT section (titre, paragraphe)>

<!ELEMENT titre (#PCDATA)>

<!ELEMENT paragraphe (#PCDATA)>])
```

Figure 2.2 Exemple d'un DTD

2.1.4 XML Schéma

Le Schéma XML est une norme W3C pour spécifier le contenu d'un document XML. La syntaxe est moins lisible que celle des DTD, car ils sont écrits en XML, au contraire des DTD. Voici un exemple de schéma xml : (19)

```
<?xml version="1.0"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">
<xsd:element name="reference" type="ElemReference" />
<xsd:complexType name="ElemReference">
<xsd:sequence>
<xsd:element name="titre" type="xsd:string" />
<xsd:element name="auteur" type="xsd:string" />
<xsd:element name="ISBN" type="xsd:string" />
</xsd:sequence>
</xsd:complexType>
 </xsd:schema>
```

1-Association entre un document et un schéma local

Pour attribuer un schéma de validation local à un document XML, on peut ajouter un attribut situé dans un « name space » spécifique :

```
<?xml version="1.0"?>
<reference
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="reference.xsd">
<titre>Comprendre XSLT</titre>
<auteur>Bernd Amann et Philippe Rigaux</auteur>
<ISBN>2-84177-148-2</ISBN>
</reference>
```

2-Association entre un document et un schéma public

Lorsque le schéma est public, mis sur un serveur, c'est un peu différent car il faut définir un 'namespace' et un l'URL d'accès :

```
<? Xml version="1.0"?>
<reference
xmlns="http://www.iut-lannion.fr"
xmlns: xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi: schema Location="http://www.iut-lannion.fr reference.xsd">
<titre>Comprendre XSLT</titre>
<auteur>Bernd Aman et Philippe Rivaux</auteur>
<ISBN>2-84177-148-2</ISBN></référence>
```

2.1.5 XSLT

Le langage XSL (XML Style sheet Language) a été conçu pour transformer des documents XML en d'autres formats comme PDF ou des pages HTML. Au cours de son développement, il a été scindé en deux unités distinctes XSLT et XSL-FO. Le langage XSLT (XML Style sheet Langage Transformation) est un langage de transformation de documents XML. Le langage XSL-FO (XML Style sheet Language - Formatting Objets) est un langage de mise en page de document. Le processus de transformation d'un document XML en un document imprimable, au format PDF par exemple, est donc découpé en deux phases. Dans la première phase, le document XML est transformé en un document XSL-FO à l'aide de feuilles de style XSLT. Dans la seconde phase, le document FO obtenu à la première phase est converti par un processeur FO en un document imprimable. (20)

2.1.6 XPath

XPath est le résultat d'un effort visant à fournir une syntaxe et une sémantique communes pour les fonctionnalités partagées entre XSL Transformations et XPointer. L'objectif principal de XPath est d'adresser des parties d'un document XML . À l'appui de cet objectif principal, il fournit également des installations de base pour la manipulation de chaînes, de nombres et de booléens. XPath utilise une syntaxe compacte non XML pour faciliter l'utilisation de XPath dans les valeurs d'attributs URI et XML. XPath fonctionne sur la structure abstraite et logique d'un document XML plutôt que sur sa syntaxe de surface. XPath tire son nom de l'utilisation d'une notation de chemin d'accès, comme dans les URL, pour naviguer dans la structure hiérarchique d'un document XML.

XPath modélise un document XML sous la forme d'une arborescence de nœuds. Il existe différents types de nœuds, y compris les nœuds d'élément, les nœuds d'attribut et les nœuds de texte. XPath définit un moyen de calculer une valeur de chaîne pour chaque type de nœud. Certains types de nœuds ont également des noms. XPath prend entièrement en charge les espaces de noms XML [noms XML]. Ainsi, le nom d'un nœud est modélisé comme une paire composée d'une partie locale et d'un URI d'espace de nom éventuellement nul; c'est ce qu'on appelle un nom développé. (21).

2.1.7 The Ressource Description Framework (RDF)

Le cadre de description des ressources est un mécanisme pour dire quelque chose sur les données. Comme son nom l'indique, ce n'est pas une langue mais un modèle de représentation des données sur «les choses sur le Web. "Ce type de données sur les données est appelé métadonnées. Les «choses» sont des ressources dans le vocabulaire RDF. RDF Schéma est un type simple système pour RDF. Il fournit un mécanisme, il définit des propriétés spécifiques à un domaine et classes de ressources auxquelles on peut appliquer ces propriétés. Les primitives de modélisation de base en RDF ; Les schémas sont des définitions de classe et des instructions de sous-classe Avec ces primitives, on peut construire un schéma pour un domaine spécifique. (22)

2.1.8 La norme 2108 ISBN

Le numéro international normalisé du livre (ISBN) est un numéro d'identification unique lisible par la machine, défini dans la norme ISO 2108. À la suite de l'édition électronique, le secteur de l'édition a connu des changements et des modifications qui ont rendu la capacité de numérotation du système ISBN est consommée plus rapidement que prévu initialement lorsque la norme avait été conçue à la fin des années 1960. Des plans sont déjà en cours pour fournir une solution avant que le point de crise ne soit atteint. Cependant, puisque la solution consiste à restructurer l'ISBN, cela aura un impact à des degrés divers sur tous les utilisateurs de l'ISBN: éditeurs, distributeurs, libraires, bibliothèques et fournisseurs de systèmes et de logiciels pour la communauté de l'information et le livre. (23).

2.2 Traitement automatique de langue

Traitement automatique de langue en tant que discipline se développe depuis de nombreuses années. Il a été créé en 1960 en tant que sous-domaine de l'intelligence artificielle et de la linguistique, dans le but d'étudier les problèmes de la génération automatique et de la compréhension du langage naturel.

Au cours des dernières années, les contributions à ce domaine se sont considérablement améliorées, permettant ainsi de traiter d'énormes quantités d'informations textuelles avec un niveau d'efficacité acceptable. Un exemple en est l'application de ces techniques en tant que composant essentiel des moteurs de recherche Web, des outils de traduction automatique ou des générateurs de résumé. (25)

2.2.1 Les applications de TAL

1-La tokenisation

La tokenisation, c'est l'identification de chaque "atomique" unité, représente la toute première opération à effectuer en traitement des documents: néanmoins, il est souvent négligé en raison de sa nature supposée fondamentale

la tokénisation peut être définie comme la tâche de fractionnement un flux de caractères en mots. Cependant, très souvent, il est associé à des processus de niveau inférieur ou supérieur. Même s'il existe une tendance à regrouper les deux tâches sous l'étiquette vague: «prétraitement», la tokénisation diffère néanmoins des «procédures de nettoyage» préliminaires telles que

- supprimer les balises inutiles de la définition de type informations dans les archives de journaux;

- supprimer les éléments «non textuels» tels que les objets horizontaux balises de saut de ligne et de page dans les documents HTML ou simleys de courrier électronique - :-) ou :-(- et représentation de la citation (comme au début de la ligne)

-éliminer des parties qui n'appartiennent pas à des langues naturelles: formules mathématiques ou chimiques, programmes (24).

2-Post Tagging

POST est le processus par lequel une balise spécifique est assignée à chaque mot d'une phrase pour indiquer la fonction de ce mot dans le contexte spécifique. La poste arabe (APOST) n'est pas une tâche facile en raison de la grande ambiguïté qui résulte de l'absence de signes diacritiques et de la complexité de la morphologie arabe.
Nous avons étudié les principaux aspects de la morphologie et de la grammaire arabes. Ce qui suit est un bref aperçu de ces aspects. Les structures verbales arabes sont composées de trois classes: nom, verbe et que nous appellerons particule (25).

a-Nom

C'est soit un nom, soit un mot qui décrit une personne, une chose ou une idée. Il peut être défini ou indéterminé et peut être subdivisé en plusieurs catégories: personne (narrateur, interlocuteur et absent), nombre (singulier, duel, pluriel), sexe (masculin, féminin) et grammatical

b-Verbe

C'est un mot qui dénote une action et qui peut être combiné avec des particules. En terme de temps,

le verbe peut être passé (impératif), présent (imparfait) ou impératif. Un futur verbe existe, mais c'est un dérivé du présent que vous réalisez en attachant un préfixe au présent du verbe. Des particules peuvent être ajoutées en tant que préfixes et / ou suffixes indiquant le nombre, le sexe et la personne du sujet, comme par exemple يقولان ,يقولون

c- Particule

Cette classe comprend tout ce qui n'est ni un verbe ni un nom. Il contient les prépositions "jarr", les prépositions pour la coordination et les mots fonctionnels tels que "inna wa akhawatuha "…..

3-Racinisation

La racinisation est le processus d'extraction des racines des mots. Une définition courte, certes, mais qui dissimule beaucoup de complexité lorsqu'il s'agit de la langue arabe.

Par ses propriétés morphologiques et syntaxiques, la langue arabe est considérée comme une langue difficile à maîtriser dans le domaine du traitement automatique des langues . A la différence des autres langues comme, le français ou l'anglais, dont les étiquettes grammaticales proviennent d'une approche distributionnelle caractérisée par une volonté "d'écarter toute considération relative au sens", les étiquettes de l'arabe viennent d'une approche où la sémantique côtoie le formel lié à la morphologie du mot, sans référence à la position de ce dernier dans la phrase. Ce phénomène est matérialisé par la notion de schèmes et de fonctions qui occupent une place importante dans la grammaire de l'arabe (26).

2.3 Conclusion

Nous avons présenté dans ce chapitre en premier partie tout sur le langage XML et les différents.

Extensions Dtd xml schéma, xpath , xslt aussi nous avons mentionné le langage rdf ainsi le norme iso qui utilise dans les livres

Pour la deuxième partie on a définir les différentes applications de traitement automatique de langue.

Dans le chapitre suivant nous allons détailler notre approche pour migration des documents du patrimoine arabe vers document semi structuré en XML.

Partie II. Contribution

Chapitre 3 : vers un système de migration des documents textuels non structuré en format semi structuré en XML

3.1 Introduction

Avec l'avènement des technologies XML .Cela est devenu particulièrement important dans la disposition et la structure des documents, en particulier des documents anciens tels que les livres du patrimoine arabe.

Dans ce chapitre, nous allons présenter nos contributions et l'architecture générale de notre approche qui consiste extraction des métadonnées, et les différentes étapes qui nous avons fait Ensuite analyse et structure des documents du la patrimoine arabe, et extraire structure a partir table de matière pour la migration des documents vers des documents semi structure en XML.

3.2 Architecture globale du système

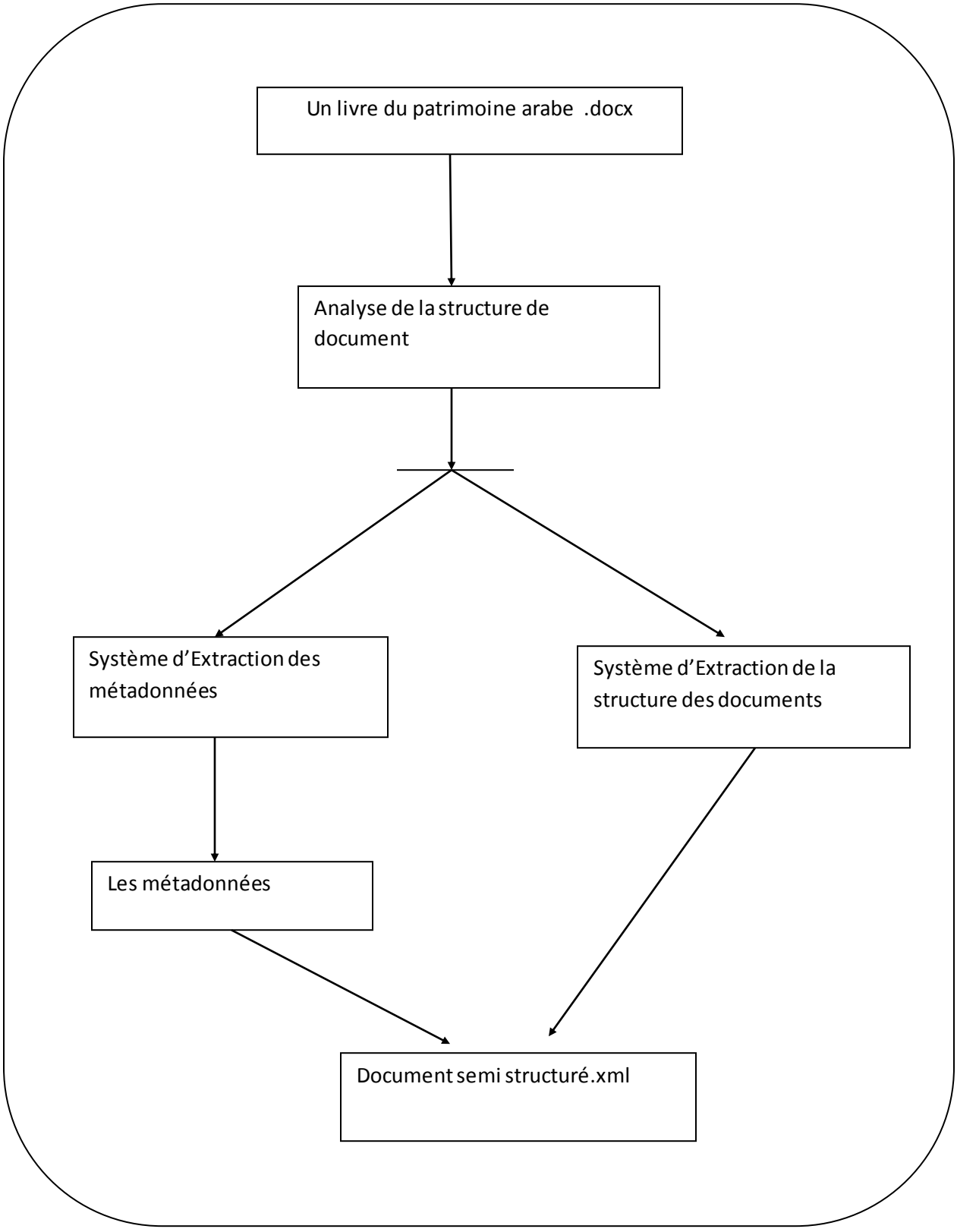


Figure 3.1 Approche générale du système

3.3 Prétraitement des documents

Nos données d'entrée correspondent à un document segmenté en pages.
Chaque page est composée d'une séquence ordonnée de blocs de texte, ou des chapitres, Section...
L'ordre des blocs doit correspondre à l'ordre de lecture humain du Document. Dans de nombreux cas, l'ordre de lecture approprié doit être Déterminé, cela la segmentation fournit des éléments correspondant approximativement à la Notion de paragraphes.
Les en-têtes et les pieds de page méritent également d'être Identifiés, en particulier pour les documents où les en-têtes ou les pieds de page Correspondent aux en-têtes de section.

3.4 Analyse de la structure des documents du patrimoine arabe

Pour faire la structure du document du patrimoine arabe :
Premièrement nous avons voir et étudier plusieurs livres arabes pour obtenir une structure générale comprend Presque tout le livre arabe, dans cette étape on a étudié environ 20 livre :

- شرح التسهيل المسمى تمهيد القواعد بشرط تسهيل الفوائد
- إرشاد السالك إلى حل ألفية ابن مالك
- المقدمة الجز ولية في النحو-
- النحو والصرف عند ابن عمار المهدي من خلال كتابه التحصيل لفوائد التقصي
-حاشية شرح القطر في علم النحو
-شرح كتاب سيبويه من باب الندبة إلى نهاية باب الأفعال في القسم
- شرح التصريف
- شرح ابن طولون على ألفية ابن مالك
- النحو وكتب التفسير
- أوزان الشعر
- الاقتراح في أصول النحو
- البسيط في شرح جمل الزجاجي
- شرح كتاب سيبويه من باب الندبة إلى نهاية باب الأفعال في القسم
- حاشية شرح القطر في علم النحو-
- الصرف العربي أحكام ومعان
- الانتصار لسيبويه على المبرد
- دروس في المذاهب النحوية
- في أصول النحو
- الإصباح في شرح الاقتراح

Figure 3.2 représente une partie de document du patrimoine arabe

Après nous avons conclu que la plupart du livre contient une structure comme suite

-Des métadonnée : titre de livre, les auteurs, année édition, numéro d'édition, numéro de partie,Maison édition

-Le contenu : introduction, introduction enquêteur, introduction superviseur, chapitres, blocks,, section, ressources, annexe, sommaire

Il ya des livres a l'intérieurs des blocks existe des chapitre, section

Aussi certain document à l'intérieur des chapitres contient des sections

Ci –dessus on fait Dtd (document type définition) xml, il bien explique la structure des documents du patrimoine arabe.

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT livre (metadonnée,
introduction,introduction_L'enquêteur?,introduction_superviseur?,(chapitre*
| block*  |section* ) ressource?, Annexe?, sommaire)>

<!ELEMENT métadonnées (titre, auteurs, année?, partie?, édition ?, maison Edition)
(#PCDATA)>

<!ELEMENT  auteurs (auteur, L'enquêteur, superviseur) (#PCDATA) >


<!ELEMENT chapitre (titre, introduction?,(section*),conclusion?)>
<!ELEMENT section (paragraphe*, marginalisation*)>
<!ELEMENT chapitre (titre, introduction?,(section*),conclusion?)>
<!ELEMENT block (titire ,introduction?,(chapitre*|section*),conclusion?)>
<!ELEMENT sommaire (type sommaire*)>



<!ELEMENT titre (#PCDATA) >
<!ATTLIST introduction_ L'enquêteur page ID #REQUIRED  >

<!ATTLIST introduction_superviseur page ID #REQUIRED  >
<!ATTLIST section page ID #REQUIRED  >
<!ATTLIST annexe page ID #REQUIRED  >
<!ATTLIST chapitre page ID #REQUIRED  >
<!ATTLIST block page ID #REQUIRED  >
<!ATTLIST Introduction page ID #REQUIRED >
<!ATTLIST sommaire page ID #REQUIRED>
```

Figure 3.3 représente DTD XML des documents du patrimoine arabe

Dictionnaire pour signifier les mots dans DTD XML en livre de patrimoine arabe

Titre	عنوان
Année	سنة الإصدار
Partie	الجزء
maison édition	دار النشر
Édition	طبعة
Auteur	مؤلف
L'enquêteur	محقق
Superviseur	ناشر
Introduction	مقدمة
Introduction l'enquêteur	مقدمة المحقق
Introduction superviseur	مقدمة الناشر
Paragraphe	فقرة
Section	فروع أو مسائل أو ,باب
Chapitre	فصل
Block	قسم
Conclusion	خاتمة
Ressource	مراجع
Annexe	ملحق
Marginalisation	التهميش
Sommaire (type sommaire)	فهرس (أنواع الفهارس)
*	[0.....n]
+	[1.....n]
?	0 OU 1

Tableau 3.1 représente détermination les mots dans DTD (document type définition) précédent

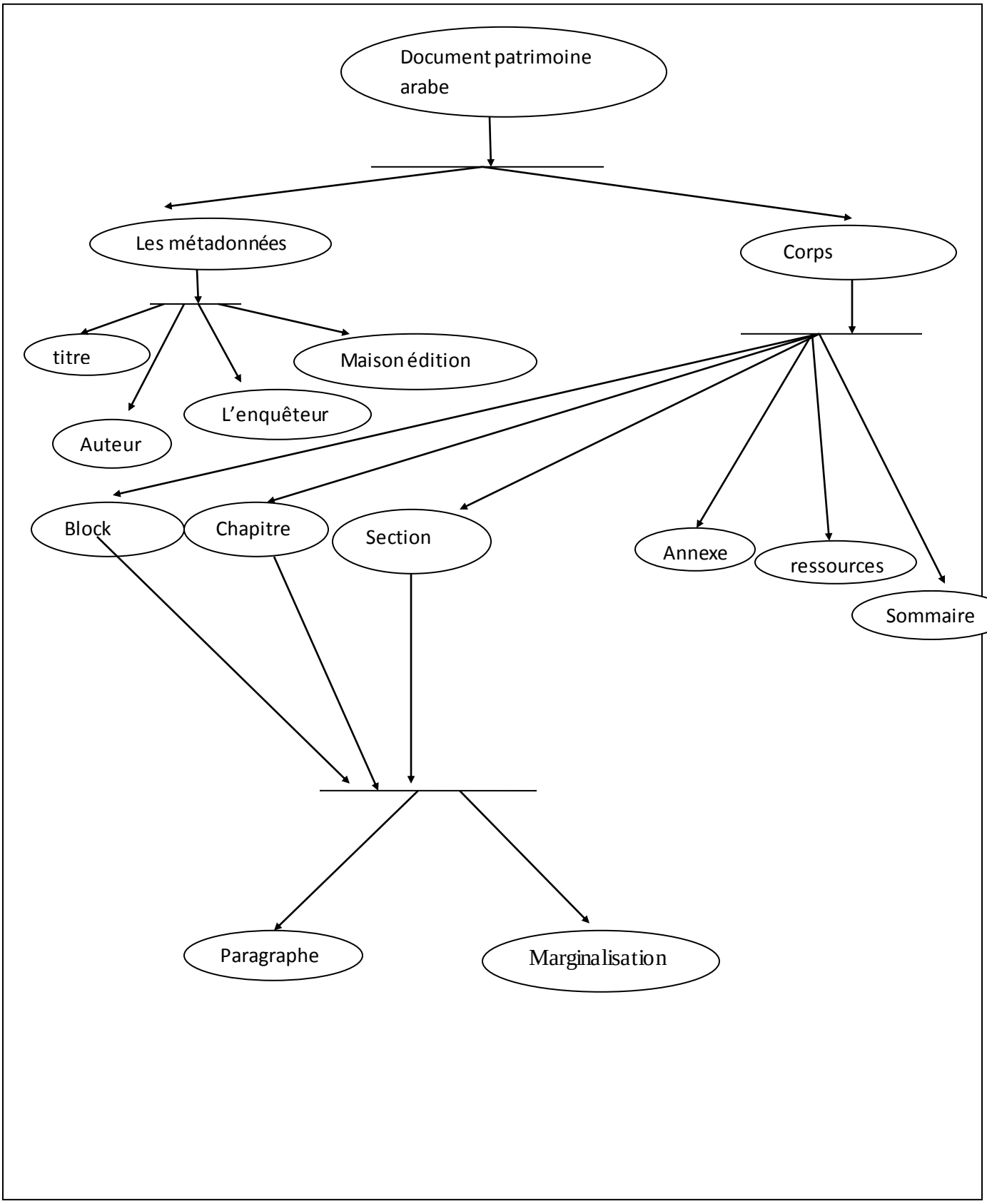


Figure 3.3 arbre des composants les documents patrimoine arabe

En raison de la difficulté de déterminer la structure exacte de livre patrimoine arabe, et pour simplifier les choses on a résumé un DTD XML pour examiner notre projet.

```
<?XML version="1.0" encoding="UTF-8"?>

<!ELEMENT livre (métadonnées, introduction?, section+, sommaires)>

<!ELEMENT métadonnées (titre, auteur, enquêteur ?, maison_édition)>

<!ELEMENT introduction (#PCDATA)>

<!ELEMENT  section (titre, chapitre+)>

<!ELEMENT sommaires (type, sommaire+)>

<!ELEMENT chapitre (titre, paragraphe*, marginalisation*)>

<!ELEMENT titre (#PCDATA)>

<!ELEMENT auteur (#PCDATA)>

<!ELEMENT enquêteur (#PCDATA)>

<!ELEMENT maison_édition (#PCDATA)>

<!ELEMENT paragraphe (#PCDATA)>

<!ELEMENT marginalisation (#PCDATA)>

<!ELEMENT sommaire (#PCDATA)>

<!ATTLIST section page ID #REQUIRED>

<!ATTLIST chapitre page ID #REQUIRED>

<!ATTLIST sommaire type CDATA #REQUIRED>
```

Figure 3.4 représente DTD pour examiner notre projet

3.5 Système d'Extraction des métadonnées

3.5.1 Architecture détaillée de système

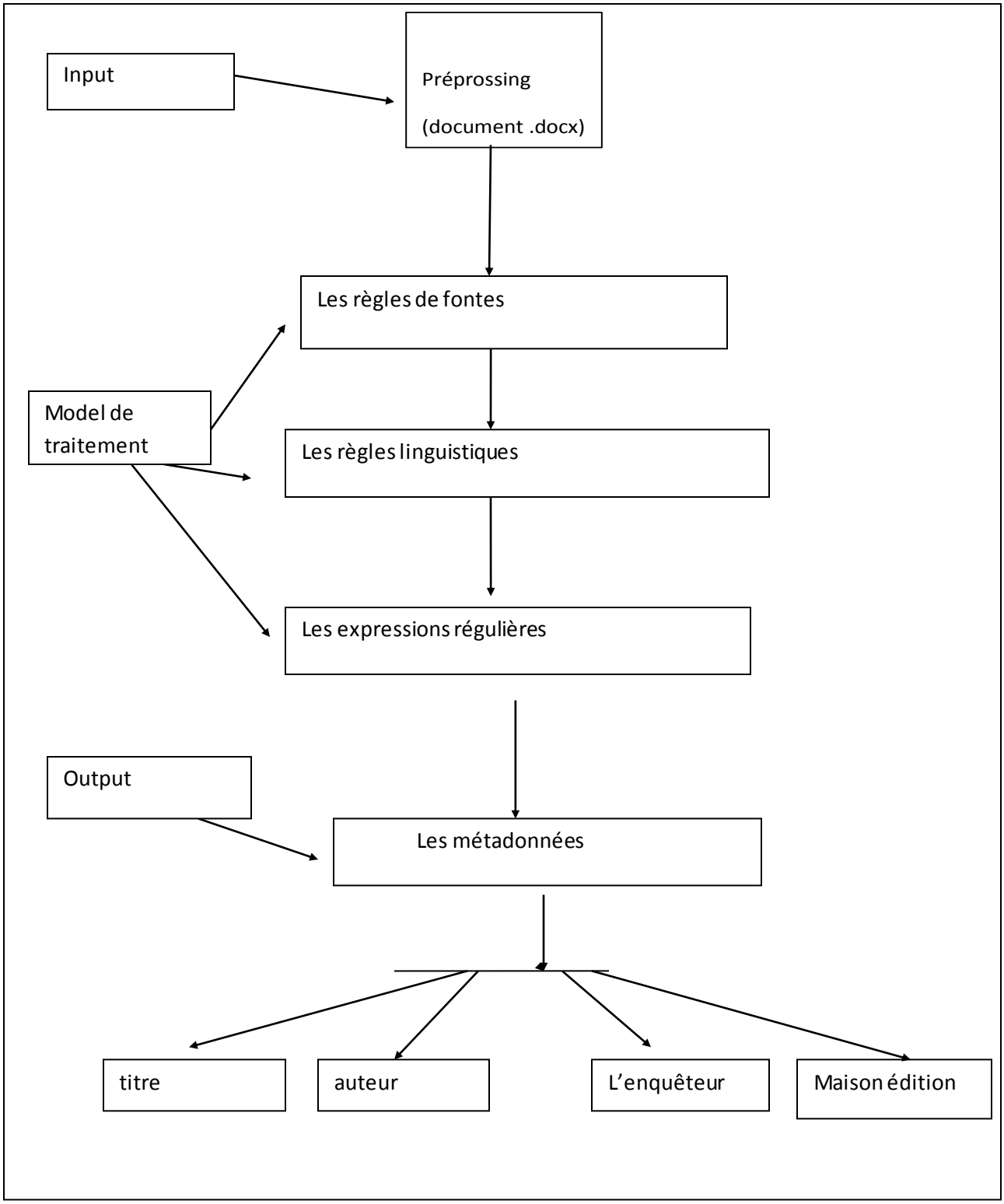


Figure 3.5 représente toutes les étapes d'extraction les métadonnées

3.5.2 Notre Approche

Les métadonnées sont des informations structurées qui décrivent, expliquent, localisent une ressource d'information, ces ressources deviennent plus facile à utiliser, La métadonnée dans les documents du patrimoine arabe sont tous existe dans le premier page selon structure suivant :

Titre de livre

L'auteur

L'enquêteur

Maison édition

Notre approche pour extraction lés métadonnée basé sur 3 étapes

1 -Les règles de font

2 - Les règles linguistiques

3 -Les expressions régulières

Première étape les règle de fonte la fonte c'est la taille utilisé pour les textes

Les règles de fonte pour extraire le titre

Après on deuxième étape vérifier le titre qui nous avons détecté en étape précédent par lés règles linguistiques

Troisième étape Les expressions régulières pour détecter et extraire auteur l'enquêteur maison édition.

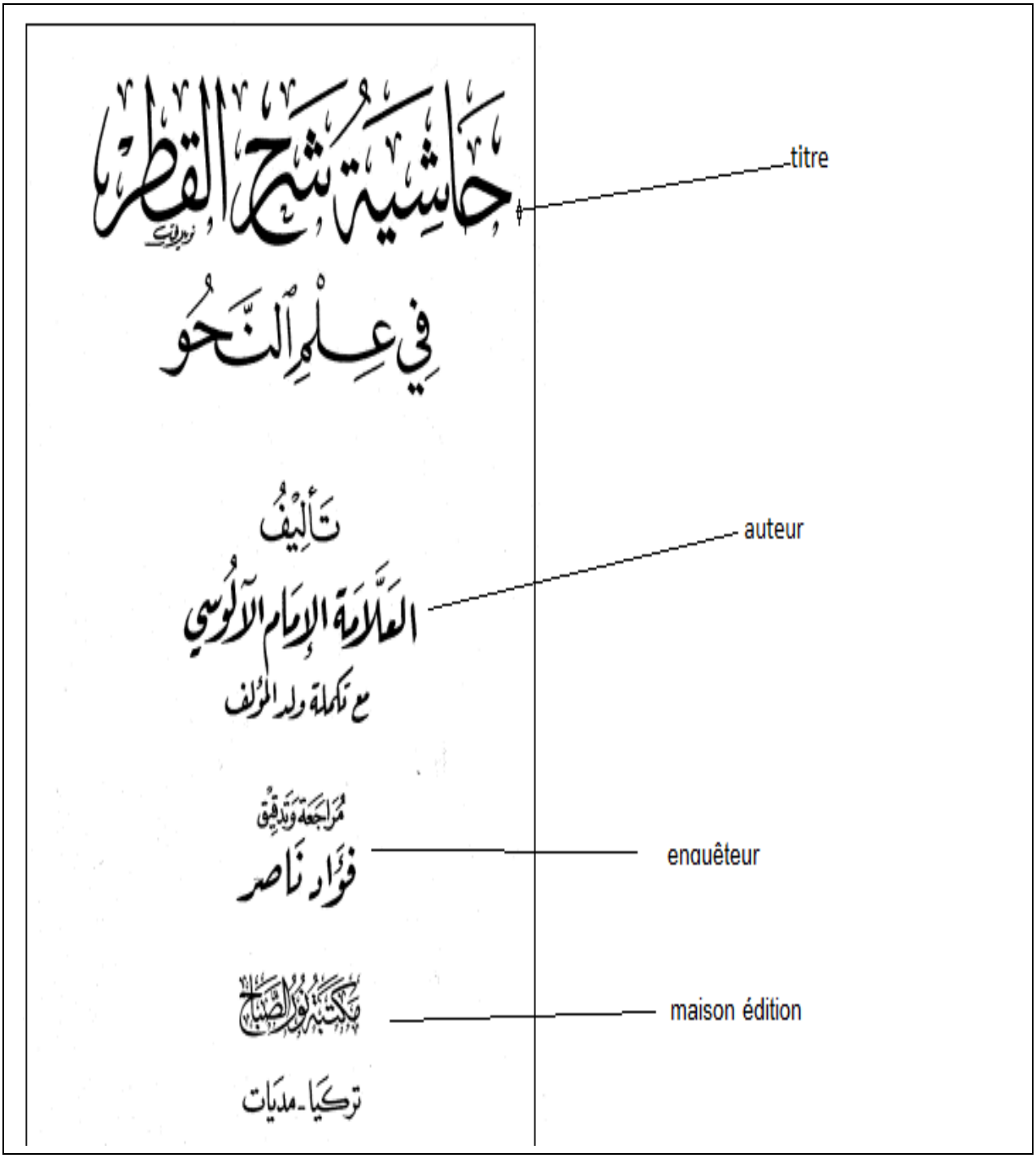


Figure 3.6 Exemple des métadonnées d'un document du patrimoine arabe

1 - les règle extraire font

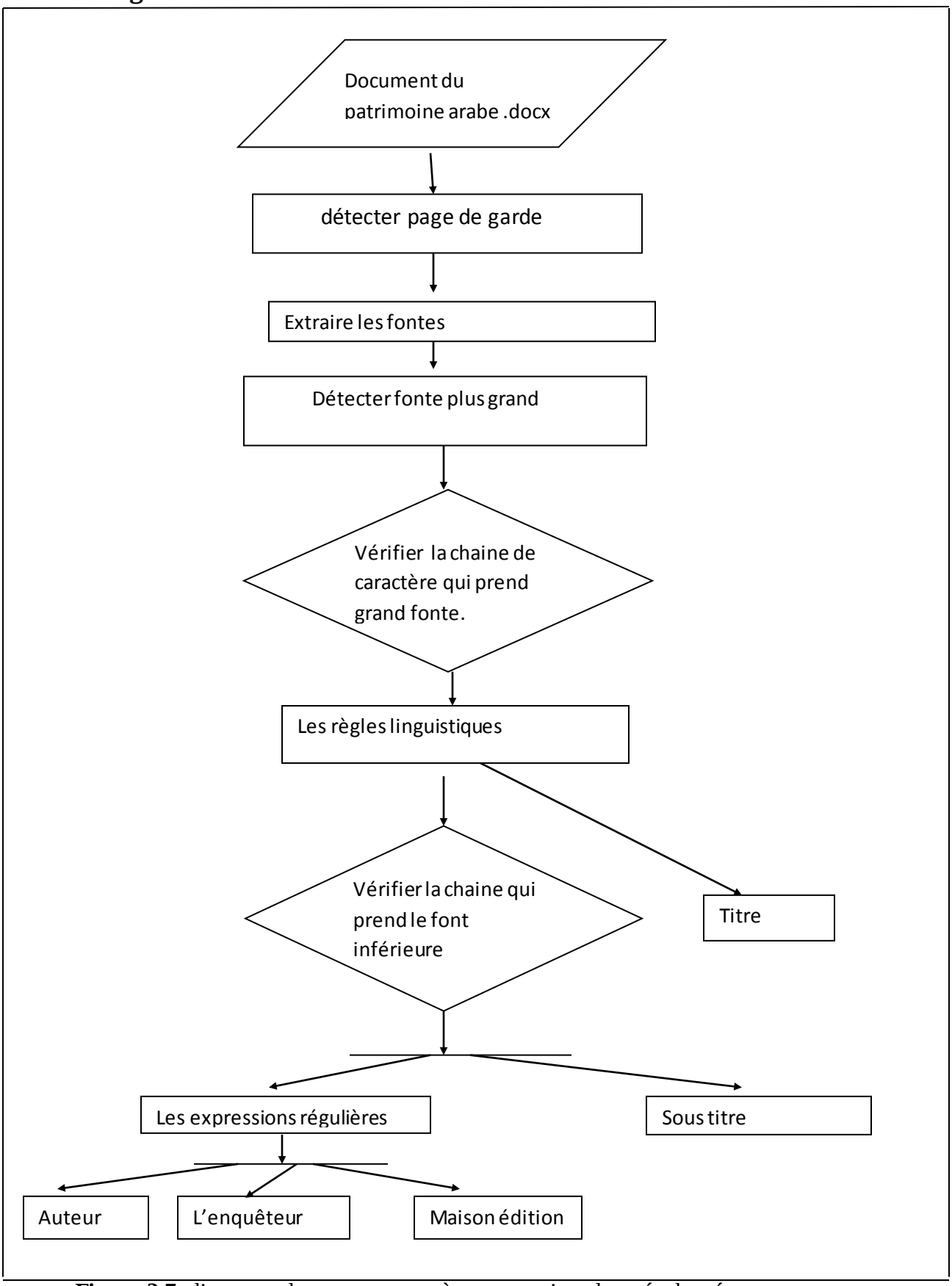


Figure 3.7 diagrame de processus système extraction les métadonnées

2- Les règles linguistiques

on Général Il n'y a pas de règles spécifiques pour extraire un titre de livres arabes donc a la aide un chercheur en académie algérienne de la langue arabe madame arab ouiza a pu identifier un ensemble de règles nous permettant de vérifier le titre dans les livres

Voila ci dessus les règles linguistiques de titre dans les livres du patrimoine arabe

اسم + فعل = الولد يتحدث:

فعل + اسم = قراءة الجدول:

اسم + اسم =الكلام الجامع:

اسم + اسم + اسم = السياسة اللّغوية الجزائرية

فعل + اسم + اسم = رمز الزمرة الدّمويّة

فعل + اسم + اسم + اسم = تركيب المثال النحوي المصنوع

فعل + اسم + اسم + اسم + اسم =استعملت الدراسات السابقة الموجودة.

اسم + جملة = علاقة المدرسة البصرية بالمدرسة الكوفية

حرف + حرف + اسم فعل = من على الشرفة رأيته

Les règles linguistiques arabes	détermination
اسم + فعل	{ (<NN>|<DTNN>) (<VBP>|<VBD>) }
فعل + اسم	{ (<VBP>|<VBD>) (<NN>|<DTNN>) }
اسم + اسم	{ (<NN>|<DTNN>) (<NN>|<DTNN>) }
اسم + اسم + اسم	{ (<NN>|<DTNN>) (<NN>|<DTNN>) (<NN>|<DTNN>) }
فعل + اسم + اسم	{ (<VBP>|<VBD>) (<NN>|<DTNN>) (<DTNN>|<NN>) }
فعل + اسم + اسم + اسم	{ (<VBP>|<VBD>) (<NN>|<DTNN>) (<DTNN>|<NN>) (<DTNN>|<NN>) }
فعل + اسم + اسم + اسم + اسم	{ (<VBP>|<VBD>) (<NN>|<DTNN>) (<DTNN>|<NN> (<DTNN>|<NN>) (<DTNN>|<NN>) }
اسم + جملة	{ (<NN>|<DTNN>) * (<VBD>|<VBP>) * (<IN>) * (<NN>|<DTNN>) *}
حرف + حرف + اسم فعل	{<IN><IN> (<DTNN>|<NN>) (<VBD>|<VBP>) }

Tableau 3.2 Déterminations les règles linguistiques arabes

3- Les expressions régulières

Les expressions régulières est une chaine de caractère, qui décrit, selon une syntaxe précise, un ensemble de chaînes de caractères possibles.
Pour connaitre l'auteur, l'enquêteur, maison édition, extraire les expressions régulières par la recherche la racine d'un mot voir le tableau ci-dessus un tableau :

La métadonnée	l'expression régulière qui rencontre La métadonnée
Auteur (مؤلف)	[.....ألف... | ل. ...]
L'enquêteur (محقق)	[.... أعد... |....حقق .....]
Maison édition (دار النشر | مكتبة)	[.... كتب.... | .....دار النشر .....]

Tableau 3.3 représente les expressions régulières pour extraire les métadonnées

3.6 Système d'Extraction de la structure des documents

3.6.1 Architecture détaillée du système

La conception de cette méthode a été guidée par l'intérêt porté à développer une méthode générique qui utilise très générale et très intrinsèque pour transformer le document vers document semi structuré en XML.

Notre approche basée sur les travaux précédents, dépend en général sur la structuration des documents du la patrimoine arabe selon table de matière.

(29) (30)[(31) Chacun a proposé une méthode de détecter les pages de la table des matières dans un document. (29) (30) La méthode de Lin s'applique à revues et tire parti d'une combinaison de correspondance de texte, mise en page et numéros de page pour déterminer les pages contenant la table des matières ainsi que la page de départ de chaque document référencé.

(31) La méthode de Mandale repose sur une heuristique basée sur le numéro de page et fonctionne sur la image de la page, avant de segmenter le contenu de la page.

3.6.2 Notre Méthode

Pour structurer un document en fonction de sa table des matières,
Les étapes suivantes sont effectuées:
1- Identifier la table des matières.
2 -Détermination de la structure de document par la table de matière.
3- Transformation vers document semi structuré en XML

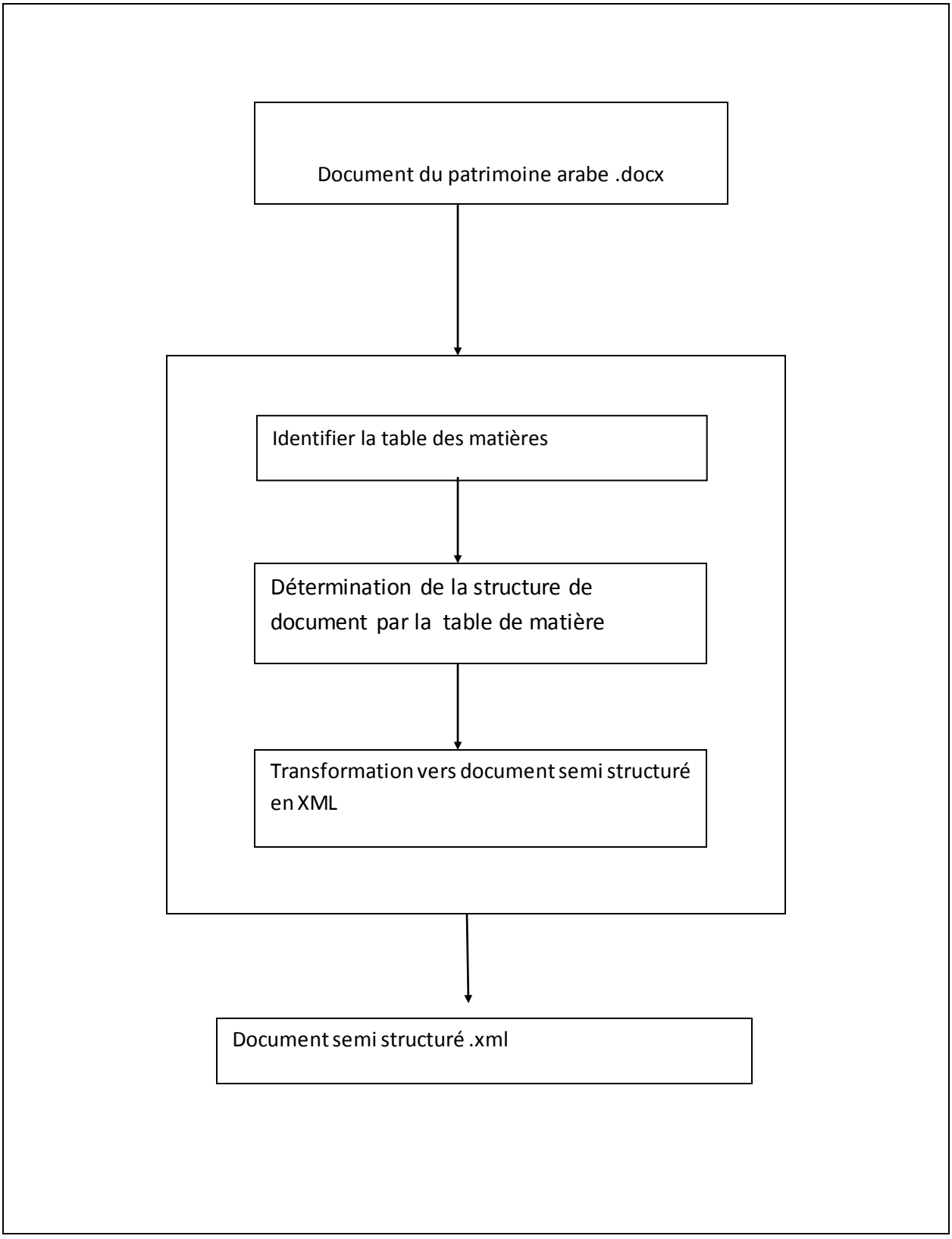


Figure 3.8 schéma représente système extraction de document et transformer vers Document semi structuré en XML

1- Identifier la table des matières

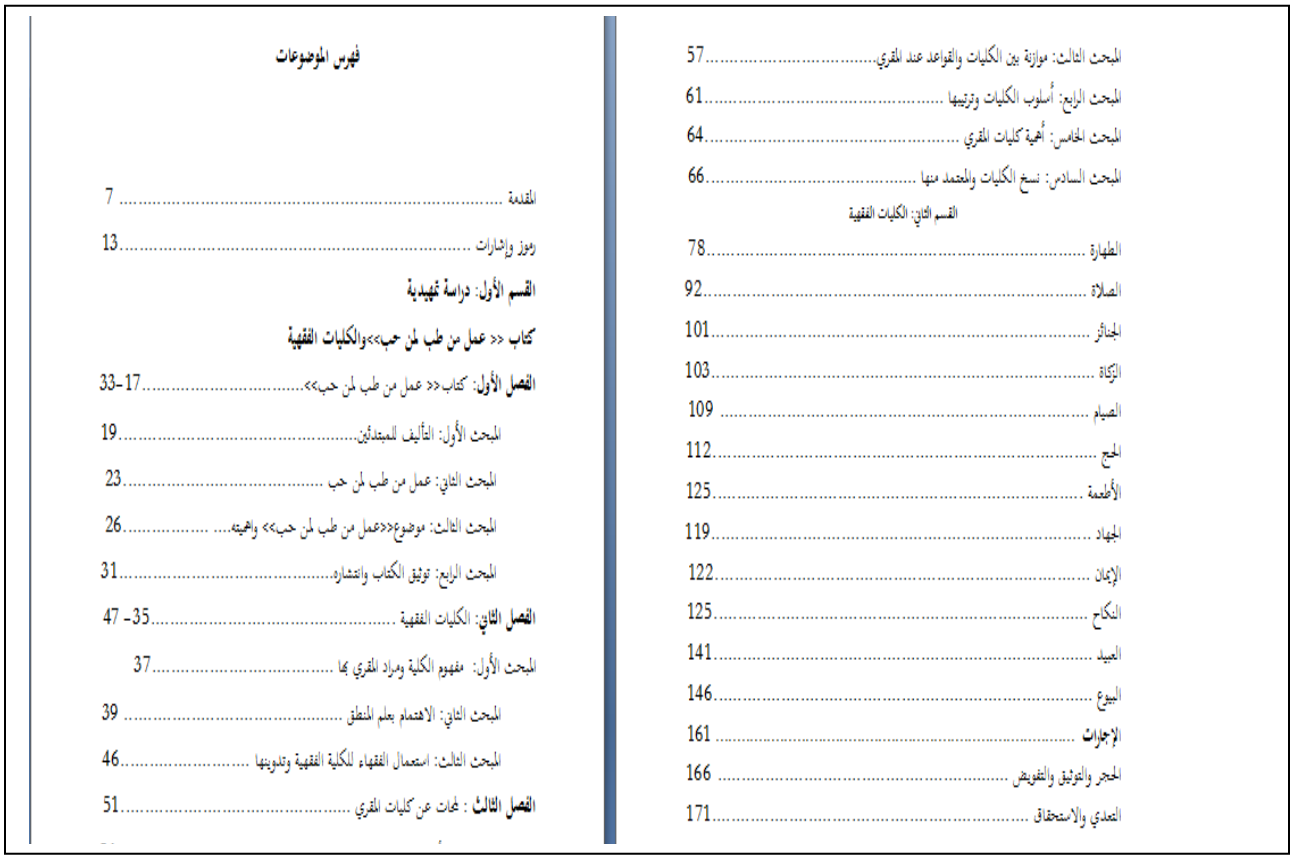
Les tables des matières sont des documents courants et simples à utiliser Construit, mais, en raison de la grande variabilité de leur disposition et Contenu à travers les classes de document et les langues, leur La détermination est difficile.

On Tout les documents du patrimoine arabe la table de matière Est situé à la fin

Aussi avec plusieurs types

Par exemple :

فهرس الأشعار,فهرس الرجاز ,فهرس المحتويات



فهرس الموضوعات

Figure 3.9 représente une table de matière de document du patrimoine arabe

2- Détermination de la structure de document par la table de matière

Cette étape est importante consiste à trouver l'organisation hiérarchique de document par la Table de matière, chaque document de la patrimoine arabe peut avoir des blocks, des chapitres section, branches….etc

Elle est caractérisée par un début (un nom tel que « table des matières »), et une ligne qui détermine la fin. Entre ce début et cette fin, il y a une succession de ligne ou chaque ligne représente le titre

Et chaque titre peut être contient des sous titre,

Chaque ligne de titre est composée de trois champs titre, ensemble des point et numéro de page

Titre : représente le titre l'objet logique (chapitre, block.. section…….)

Nous extrayons les lignes, jusqu'a la fin (ne trouve aucun titre avec page)

Ou à la ligne qui contient le titre qui Représente la « Conclusion Générale », avec toutes ses variations (conclusion Générale, conclusion et perspectives, …etc.)

Parfois, la conclusion générale a comme titre «conclusion », ce qui confond avec les « conclusions » des chapitres, alors si C'est le cas on s'arrête à la ligne références bibliographiques.

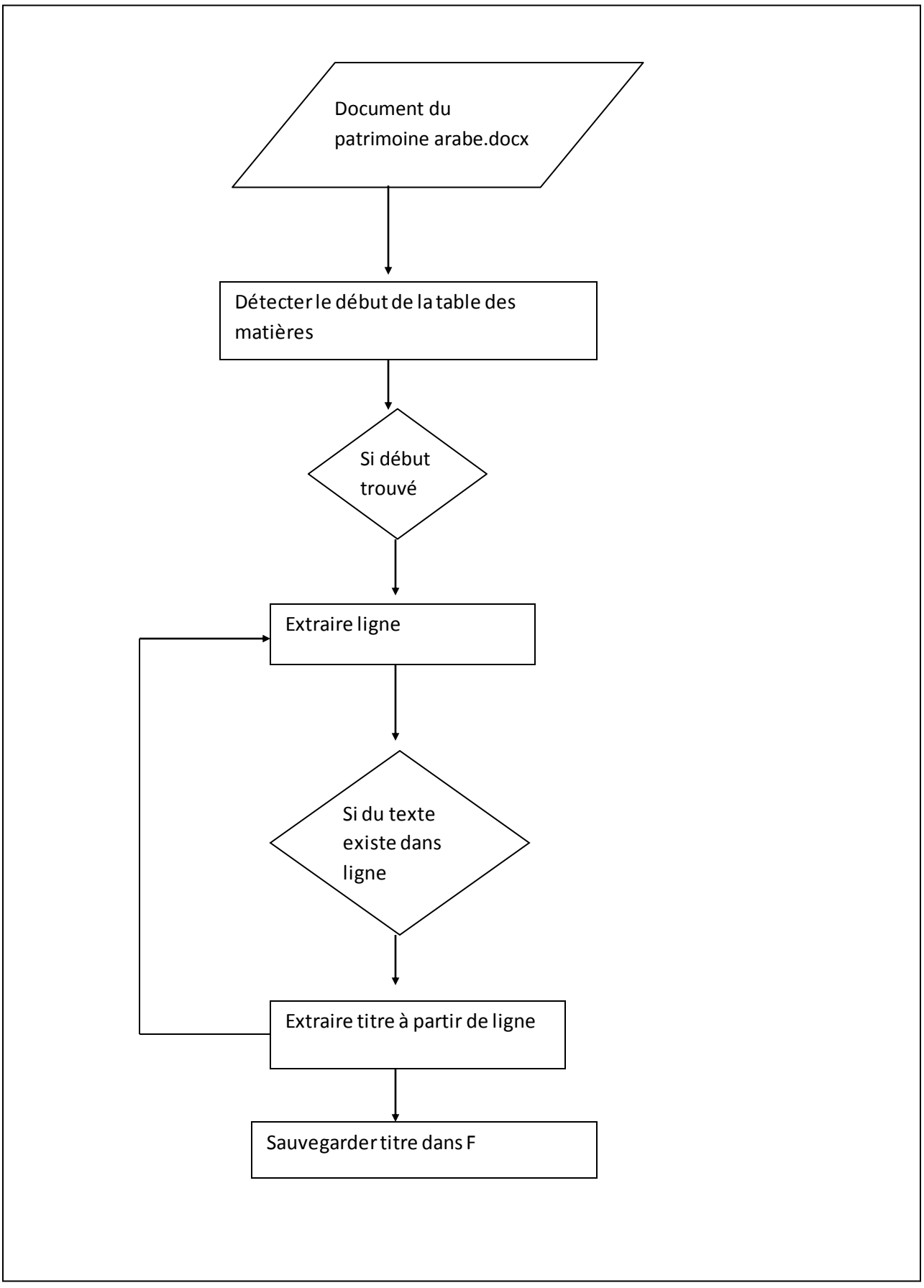


Figure 3. 10 diagrame de processus système extraction de la structure des documents

3- Transformation vers document semi structuré en XML

La dernière étape consiste simplement à appliquer une transformation par détecter le contenu de chaque titre par numéro de page Si titre en Gras , ou commencer par mot باب ajouter balise section et les titres qui son l'intérieur avec balise chapitre de chaque titre

et extraire le contenu jusqu'un titre suivant fermer la balise chapitre ainsi de suite jusqu'un trouvé un nouveau section (titre en gras ou commencer par mot باب) fermer balise section et ainsi de suite.

3.7 Conclusion

Dans ce chapitre, nous avons proposé contribution principale sur notre approche qui nous avons fait sur ce projet, cette approche compose deux étape principal
Première étape consiste à extraire les métadonnées en les premier pages ainsi on a décrit Tout les méthodes qui utilisé.
La deuxième étape structuration des documents
Cette étape base sur extraire les titres par la table de matière ensuite pour obtenir la structure Des documents et transformer vers document semi structure en XML.
Dans le chapitre suivant nous allons décrire sur les résultats des test après des Expérimentation de plusieurs livres du patrimoine Et les différents outils qui nous avons utilisé.

Chapitre 4 Expérimentation et évaluation

4.1 Introduction

Dans ce chapitre, nous présentons une mise en œuvre de nos contributions citées précédemment, afin de juger nos propositions et d'évaluer l'efficacité de notre système de
. Nous étudions les deux propositions principales de notre travail à savoir :
- Extraction les métadonnées
- L'extraction des titres et les sous-titres titres qui se trouvent dans les tables de matières des documents du patrimoine arabe non structurés, puis la transformation de documents en format semi structuré en xml.

Le critère principal pour la création d'un environnement expérimental était d'obtenir Une base de documents au format MS Word qui a est fournie par l'Académie algérienne de langue arabe.
En d'autres termes, nous utilisons un ensemble de test.

4.2 Environnement Technologique

1- al-Maktaba al-Shamela (المكتبة الشاملة)

À ce jour, le logiciel qui nous donne le plus de satisfaction, et qui semble la plus populaire dans le monde arabe, se nomme al-Maktaba al-Shamela (المكتبة الشاملة).

Bibliothéque elshamela nommée shamela parce que contient Presque tout les livre arabe dans tout les domaine ,on a utlisé pour Analyse de la structure des documents du patrimoine arabe.

2-Oxygen XML Editor

Oxygen XML Editor est le meilleur éditeur XML disponible et fournit une suite complète d'outils de création et de développement XML. Il est conçu pour accueillir un grand nombre d'utilisateurs, allant des débutants aux experts en XML. Il est disponible sur plusieurs plates-formes, sur tous les principaux systèmes d'exploitation, et en tant qu'application autonome ou plug-in Eclipse. Vous pouvez utiliser Oxygen XML Editor avec toutes les technologies basées sur XML. Il comprend une grande variété d'outils puissants pour la création, la modification et la publication de documents XML.

3-Langage python

Il existe un très grand nombre de langages de programmation, chacun avec ses avantages et ses inconvénients, dans notre application on a utilise le langage python , Python est un langage portable, dynamique, extensible, gratuit, qui permet (sans l'imposer) une approche modulaire et orientée objet de la programmation. Python est développé depuis 1989 par Guido van Rossum et de nombreux contributeurs bénévoles.

4-Bibliotheque nltk

NLTK est une plate-forme utilisée pour créer des programmes Python qui utilisent des données de langage humain pour une application dans le traitement statistique du langage naturel (PNL).

Il contient des bibliothèques de traitement de texte pour la création de jetons, l'analyse syntaxique, la classification, la création de liens, le balisage et le raisonnement sémantique. Il comprend également des démonstrations graphiques et des exemples de jeux de données, ainsi qu'un livre de recettes et un livre expliquant les principes sous-jacents des tâches de traitement du langage prises en charge par NLTK.

5-Bibliotheque Beautiful Soup

Beautiful Soup est une bibliothèque Python permettant d'extraire des données de fichiers HTML et XML. Il travaille avec votre analyseur préféré pour vous fournir des moyens idiomatiques de naviguer, de rechercher et de modifier l'arbre d'analyse. Cela permet généralement aux programmeurs d'économiser des heures ou des jours de travail.

6-Stanford corlenp server

On a utilise Stanford corlenp server , CoreNLP inclut un serveur API Web simple pour répondre à vos besoins en compréhension de langage humain (à partir de la version 3.6.0). Le serveur est démarré directement quand l'appelant avec des command java
java -Xmx600m -cp "*" edu.stanford.nlp.pipeline.StanfordCoreNLPServer -stanford-arabic-corenlp-2018-10-05-models.jar -port 9000 -timeout 15000
pour analyse chaine de caractère qui entré par utilisateur puis annoté chaque mot avec tags
Par exemple
la chaine entre est الولد يتحدث

résultat : [('VBP' ,'يتحدث') ,('DTNN' ,'الولد')]

Bibliothèque re

Les expressions régulières (re) sont essentiellement un petit langage de programmation hautement spécialisé embarqué dans Python et dont la manipulation est rendue possible par l'utilisation du module re. En utilisant ce petit langage, vous définissez des règles pour spécifier une correspondance avec un ensemble souhaité de chaînes de caractères ; ces chaînes peuvent être des phrases, des adresses de courriel.

Bibliothèque lxml

lxml est une bibliothèque Python qui permet de manipuler facilement les fichiers XML et HTML et peut également être utilisée pour le nettoyage Web. Il existe de nombreux analyseurs XML standard, mais pour obtenir de meilleurs résultats, les développeurs préfèrent parfois écrire leurs propres analyseurs XML et HTML. C'est à ce moment que la bibliothèque lxml entre en jeu.

4.3 Présentation de Data set

La collection textuelle est compose de 14 documents textuelle en format (docx) classés en une seule catégorie (Patrimoine algérien), et respecte la représentation du support physique à partir duquel elle a été rédigée.

4.4 Méthodologie de test

Afin d'évaluer notre système en terme d'efficacité, nous nous sommes basés sur des mesures de performances définies dans les systèmes de recherche d'information et l'adapte pour notre cas. Dans ces derniers, deux mesures, la précision et le rappel, permettent de déterminer l'efficacité du système pour retrouver les documents pertinents et ignorer les documents non pertinents.

Rappel : La capacité d'un système à sélectionner tous les documents pertinents de la collection.

Rappel = R/M

R : Nombre de documents pertinents retrouvés

M : Nombre de documents pertinents dans la collection

Précision : La capacité d'un système à sélectionner que des documents pertinents.

Précision = R/S

R : Nombre de documents pertinents retrouvés

S : Nombre de documents retrouvés

Pour évaluer nous système nous avons inspire par ces deux mesure, et tombe sur le choix de rappel, en modifiant sons formule pour adapte à notre cas d'étude.

4.5 L'évaluation de System d'extraction des métadonnées

En Calcule le rappel pour chaque une des composants des métadonnées (titre, auteur, enquêteur, Maison d'édition) puis en calcule le rappel total par la division de la somme de ses dénier par le nombre des composent des métadonnées, dans cette cas c'est 4.

Exemple : pour le titre

R : le nombre des documents dans lequel l'extraction de titre est correcte.

M : Nombre Total des documents dans la base dans lequel le titre est présent.

Rappel (titre) = R/M

Rappel(Total) = (Rappel (titre) +Rappel (auteur)+Rappel(enquêteur)+Rappel(Maison d'édition)) /4

Test et Résultats

	Titre	**auteur**	**enquêteur**	**Maison d'édition**	**Total**
Rappel	**14 /14**	**13/14**	**10/14**	**9/14**	**81%**

Tableau 4.1 résultat de test les métadonnées

Voila exemples de test sur un livre de patrimoine arabe :


ذِكْرُ الفَرْقِ
بَيْنَ الأَحْرف الخَمسَة
وَهي
الظَّاءَ وَالضَّادْ الذَّال وَالصَّاد وَالسِّيْن
تأليف
أبي محمد عَبْد ٱلله بن محمد
ابن السّيد البَطَلْيَوْسِي
المتوفى 521 هـ
تحقيق وتعليق
الدكتور حمزة عَبْد الله النشرتي
منشورات
محمد علي بيضون
ناشر كتب السنة والجماعة
دار الكتب العلمية
بيروت - لبنان

Figure 4.1 example de livre pour test de extraction des métadonnée

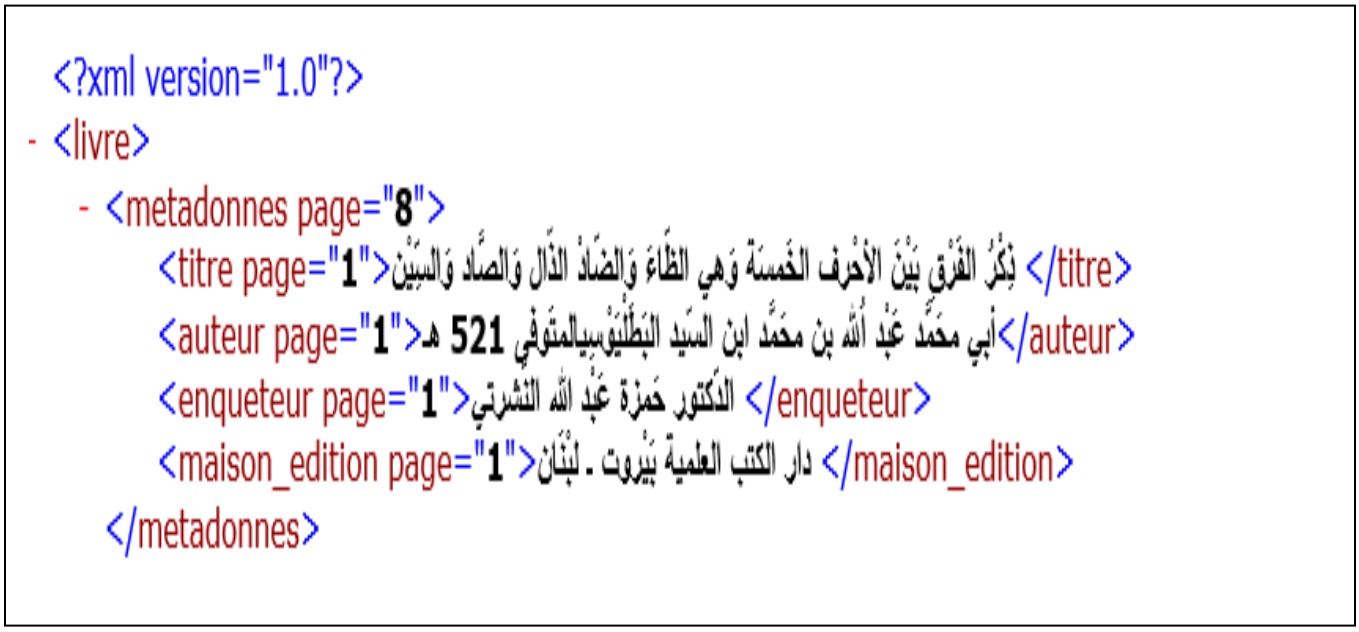


Figure 4.2 résultat d'extraction lés métadonnées

4.6 L'évaluation de System de détermination de la structure de document

Pour chaque Livre « i » dans notre collection **:**

R : le nombre des éléments de table de matière dans lequel l'extraction de titre est correcte

M : Nombre Total des éléments dans la table de matière

Rappel (Livre i) = R/M

Fin.

Le rappel total est calculé par la division des sommes des rappels des documents par le Nombre Total des documents dans la base.

Rappel(Total) = SUM (Rappel (livre i)) / Nombre Total des documents dans la base.

Test et Résultats

Document	Nome de livre	Rappel
1	رسالة غريب إلى الحبيب	31/31
2	المقدمة الوغليسية	30/34
3	الكليـــــــــــاتُ الفِقْهيـــــــــــهّ	35/40
4	زَهر الشماريخ	40/47
5	منشور الهدَايَــة	159/170
6	ذكر الفرق بين الأحرف الخمسة كامل	500/654
7	أسنى المتاجر	35/47
8	التحفة المرضية	80/94
9	الرحلة الورثيلانية	18/22
10	تفسِير ابن باديسَ	400/496
11	رَوْضَة الاسّ	50/59
12	شرف الطالب في أسنى المطالب	230/236
13	رسالة الشرك ومظاهره	40/40
14	منامات الوهراني	58/58
Total		87%

Tableau 4.2 résultat de test de structuration les documents

Exemple de test Text de livre de patrimoine arabe **كتاب ذكر الفرق بين الأحرف الخمسة كامل**

التَّعْظِيبُ والتَّعْضِيبُ والتَّعْذِيبُ

التعظيبُ بالظاء خشونة اليد من العمل. يقال: عَظْبَتْ يَدُهُ. أنْشَدَ أبو زَيْد:

لـو كنـتُ مـن زَوْفَـنَ أو بنيهـا قبيلـة قـد عَظْبَـتْ أيديهـا(3)

مُحَوِّديـن الحفْـرَ حفّـارِيهَـا لقـد حَفَـرَتْ نُبْتَـةُ تُرْوِيهـا

روى(4) أبو على البغدادى(5) عن ابن دُرَيد: زوفن بالزاى، ورواه غيره دَوْفَن بالدالِ غَيْرَ معجمة.

والتَّعْضِيبُ بالضاد كثرةُ القطع أو الكسر، **والتَّعْذِيبُ** بالذال كثرةُ العَذَاب، وقياس هذا الباب قياس الذى قبله.

---

(1) البيت من قصيدة مطلعها: خليلى عوجا بارك الله فيكما. المقروع: المختار، والعذف: الأكل. والعاذب: القائم الرافع رأسه لا يأكل وقوله: ندى بفتح النون مقصور: الصوت الضعيف يسمع بعيدًا وهو هناك شديد. والشاعر: غيلان بن عقبة يكنى أبا الحرث، وذو الرمة: لقبته به مية. انظر سمط اللآلىء (726/2)، والديوان (ص 86).

(2) سقط في الأصل. وقد جاء في اللسان: عذبة شراك النعل: المرسلة من الشراك.

(3) فى الأمالى: وأنشد أبو زيد برواية من زوفن: النبتة. الركية التى تخرج نبتتها. وفى الأصل «روفن» بالراء. وحفاريها بالجيم الأمالى (152/1).

(4) فى ب (وروى).

(5) هو أبو على القالى: إسماعيل بن القاسم: نسبة القالى إلى «قالى قلا». ونسبه البغدادى إلى بغداد وبهذا كان أهل المغرب يلقبونه توفى عام 356 هـ. انظر: نفح الطيب (85/2)، بغية الملتمس (216).

19

Figure 4.3 texte de livre de patrimoine arabe pour test de transformation

Voila un exemple de résultat

```
- <section page="18">
    <titre>الظاء والضاد والذال باتفاق اللفظ واختلاف المعنى</titre>
  - <chapitre page="19">
      <titre>التعظيب والتعضيب والتعذيب</titre>
      <paragraphe>True*** ____ *</paragraphe>
      <paragraphe>التعظيبُ بالظاء خشونةُ اليد من العمل. يقال: عَظَبَتْ يَدُهُ. أنْشَدَ أبو زَيْد:</paragraphe>
      <paragraphe>لو كنتُ من زَوْفَنَ أو بنيها (3) قبيلة قد عَظَبَتْ أيديها</paragraphe>
      <paragraphe>مُعَوِّدِين الحفْرَ حفَارِيَهَا لقد حَفَرْتُ نُبْثَةً تُرْوِيهَا</paragraphe>
      <paragraphe>روى (4) عن ابن دُرَيد: زوفَنٍ بالزاى، ورواه غيره دَوْفَنٍ بالدالِ غَيْرَ معجمة (5) أبو على البغدادى.</paragraphe>
      <paragraphe>والتَّعْضِيبُ بالضاد كثرةُ القطع أو الكسر، والتَّعْيِيبُ بالذال كثرةُ العَذَاب، وقياس هذا الباب قياس الذى قبله.</paragraphe>
      <paragraphe>.....</paragraphe>
      <paragraphe>(1) البيت من قصيدة مطلعها: خليلى عوجا بارك الله فيكما. المقروع: المختار، والعذف: الأكل. والعاذب: القائم الرافع رأسه لا يأكل وقوله: ندى بفتح النون مقصور: الصوت الضعيف
        يسمع بعيداً وهو هناك شديد. والشاعر: غيلان بن عقبة يكنى أبا الحرث، وذو الرمة: لقبته به مية. انظر سمط اللآلىء (2/726)، والديوان (ص 86).</paragraphe>
      <paragraphe>(2) سقط فى الأصل. وقد جاء فى اللسان: عذبة شراك النعل: المرسلة من الشراك.</paragraphe>
      <paragraphe>(3) فى الأمالى: وأنشد أبو زيد برواية من زوفن: النبثة. الركية التى تخرج نبثتها. وفى الأصل «روفن» بالراء. وحفاريها بالجيم الأمالى (1/152).</paragraphe>
      <paragraphe>(4) فى ب (وروى).</paragraphe>
      <paragraphe>(5) هو أبو على القالى: إسماعيل بن القاسم: نسبة القالى إلى «قالى قلا». ونسبه البغدادى إلى بغداد وبهذا كان أهل المغرب يلقبونه توفى عام 356 هـ. انظر: نفح الطيب (2/85)،
        بغية الملتمس (216).</paragraphe>
      <paragraphe>20.....الظاء والضاد والذال</paragraphe>
      <paragraphe>العَظْم والعَضْم والعَذْم</paragraphe>
    </chapitre>
```

Figure 4.4 résultat de transformation document patrimoine vers document semi structuré en XML

4.7 Interprétation

A travers les tests que nous avons effectués, nous remarquons que

Dans système l'extraction les métadonnées le résultat de test est bonne Dans tout livre Il y a des livres qui ont extraits tout.

Et pour système de détermination la structure de document Quand les livres contiennent plusieurs > 100 titre dans tableau matière le résultat sont faible

Par apport les autre livre presque bien convertir.

4.8 Conclusion

Nous avons réalisé un système qui permet extraction les métadonnées et extraire les titres de table de matière pour les documents de patrimoine arabe, puis de l'utiliser dans les phases transformation.

Après expérimentations de extraction les métadonnées et détermination de la structure de document, on constate que les résultats obtenus sont meilleurs.

Conclusion général et perspective

La structuration logique des documents de patrimoine arabe est porteuse de sens, elle indique une représentation du contenu. Pour cette raison nous avons fait attention particulières à cette représentation au cours du processus de conversion.

Notre objectif est transformer les documents de patrimoine arabe vers les documents semi structuré en XML, et cela par l'extraction des parties les plus importantes de la structure des documents, particulièrement les titres et les sous-titres.

Des expérimentations sur 14 livre patrimoine arabe sur deux étape la première étape

Extraction les métadonnées et la deuxième étape c'est structuration les documents .

Dans ce projet on a travailler sur deux niveau section chapitre La première perspective à tirer de notre travail est améliorer la convertion enquouvant plus sur le contenu des paragraphe des chapitre par balisage les marginalisions, coran ,hadith,

La deuxième perspective c'est de proposer une méthode pour extraire les références de créer des liens et des relations avec les documents cités.

Référence

1. *Jacques André, Vincent Quint Structures et modèles de documents 1990.*
2. *Contribution à la Modélisation des Métadonnées Associées aux Documents.* 2013.
3. *Karim Djemal,DE LA MODELISATION A L'EXPLOITATION DES DOCUMENTS A STRUCTURES MULTIPLES.*
4. *Semi-structured documents mining: a review and comparison.* 2013.
5. *Stéphane Martin Édition collaborative des documents semi-structurés.* 2011.
6. *Jacques André, Vincent Quint Structures et modèles de documents 1990.*
7. *une approche semantique pour les document numerique .* 2016 .
8. *Nouveau modèle de documents pour une bibliothèque numérique de thèses accessibles par leur contenu sémantique .* 2005.
9. *Yves MARCOUX ,LES FORMATS NORMALISÉS DE DOCUMENTS ÉLECTRONIQUES.*
10. *Jean-Luc Bloechle, Maurizio Rigamonti, Karim Hadjar,XCDF: A Canon ical and Structured Document Format .*
11. *D Gilbert ,Guide de conception pédagogique et graphique de sites W3 éducatifs 1999.*
12. *Paul Murrell ,The Newsletter of the R Project 2006.*
13. *Catherine Morel-Pair ,Panorama : des métadonnées pour les ressources.*
14. *Richard Parent , Nicole Boulet les composants d'un document numerique 1999.*
15. *Anis JEDIDI ,MODÉLISATION GÉNÉRIQUE DE DOCUMENTS MULTIMÉDIA PAR DES MÉTADONNÉES.* 2005.
16. *Thomas Baker, Makx Dekkers Identification des éléments de métadonnées à l'aide d'URI 2003.*
17. *Comprendre les métadonnées. Organisation nationale de normalisation de l'information.* 2004.
18. *Denis Diderot,Théorie, conception et réalisation d'un langage de programmation adapté à XML2004.*
19. *María del Rocío ABASCAL MENA ,Nouveau modèle de documents pour une bibliothèque numérique de thèses accessibles par leur contenu sémantique.* s.l. : L'institut national des sciences appliquées de Lyon, 2005.
20. *Pierre Nerzic ,Documents et outils XML.* s.l. : université rennes 1.
21. *Olivier Carton ,L'essentiel de XML.* 2015.
22. W3C Recommendation . *W3C Recommendation.* [En ligne] 16 novembre 1999. https://www.w3.org/TR/1999/REC-xpath-19991116/.
23. *Michel Klein,XML, RDF, and Relatives.*
24. *Ann chapman ,New number on the block 2004.*
25. *Mari Vallez , Rafael Pedraza-Jimenez ,Natural Language Processing in Textual Information Retrieval and Related Topics 2007.*
26. *B. Habert, G. Adda, M. Adda-Decker, P. Boula de Mareuil, ¨ S. Ferrari, O. Ferret, G. Illouz, P. Paroubek ,Towards Tokenization Evaluation.*
27. *Y.O. Mohamed El Hadj, I.A. Al-Sughayeir, A.M. Al-Ansari ARABIC PART-OF-SPEECH TAGGING USING THE SENTENCE STRUCTURE.*
28. *El Younoussi Yacine ,Doukkali Sdigui Abdelaziz & Belahmer HabibLa racinisation de la langue arabe par les automates à états finis (AEF) .*
29. *X. Lin, "Automatic Document Navigation for Digital Content Re-mastering", HP technical report, 2003. .*

30. *[2] X. Lin, "Text-mining Based Journal Splitting", Proceedings of the Seventh International Conference on Document Analysis and recognition, ICDAR'03, 2003.*
31. *S. Mandal, S.P. Chow bury, A.K. Das, B. Chandra. Automated Detection and Segmentation of Table of Contents Pages from Document Images, ICDAR 2003.*
32. *Semi-structured documents mining: a review and comparison .* 2013.
33. *La Gestion Electronique des documents .* 2008.
34. *Mari Vallez , Rafael Pedraza-Jimenez , Natural Language Processing in Textual Information Retrieval and Related Topics.* 2007.