

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE  
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ MOULOUD MAMMARI DE TIZI-OUZOU  
FACULTÉ DE GÉNIE ÉLECTRIQUE ET D'INFORMATIQUE  
DÉPARTEMENT D'INFORMATIQUE



# Mémoire de fin d'études

*En vue de l'obtention du diplôme de Master en Informatique  
Option : Conduite de Projet Informatique*

## Thème

Mise en place d'une plate-forme d'interopérabilité  
d'accès aux manuscrits arabes numérisés  
basée sur Dublin Core

**Soutenu devant le jury composé de:**

**Président :** M. SI-MOHAMMED Malik

**Examinatrice :** Mme. AIT ADDA Samia

**Examineur :** M. RAMDANE Mohammed

**Promoteur :** M. SOUALAH Mohammed Ourabah

**Réalisé par :**

M<sup>elle</sup> DJEFEL Ouzna

M<sup>elle</sup> OUAGUENOUNI Nacera

Promotion 2011/2012

## *Remerciements*

*Tout d'abord nous remercions le bon DIEU de nous avoir donnée les moyens et l'énergie de réaliser ce travail.*

*Nous tenons à exprimer notre profonde gratitude et nos sincères remerciements pour notre promoteur M.SOUALAH Mohammed Ourabah, qui nous a fait l'honneur de diriger ce travail du début jusqu'à la fin avec ses précieux conseils qui furent d'un apport considérable. Nous tenons aussi à lui reconnaître le temps précieux qu'il nous a accordé. De plus, nous nous laisserons jamais de le remercier pour nous avoir fait découvrir un domaine que le moins que l'on puisse dire est extraordinaire.*

*Nous remercions le président de jury Monsieur SIMOHAMMED Malik professeur à l'université Mouloud Mammeri de Tizi-Ouzou, d'avoir accepté d'honorer par son jugement notre travail. Nous sommes très honorés.*

*Que Madame AIT ADDA Samia trouvent ici nos remerciements les plus vifs pour avoir accepté d'évaluer ce travail en tant que membre de jury.*

*Nous tenons également de remercier Monsieur RAMDANE Mohammed pour qui a accepté de lire, d'évaluer notre travail et de participer au jury.*

*Nos sincères sentiments vont à tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce mémoire. En particulier nos chères familles et nos amis (es).*

## *∞ Dédicaces ∞*

*Je dédie ce modeste travail :*

*A ma très chère mère et mon très cher père, sans  
laquelle jen'aurai pas pu parvenir à cestade.*

*A mes sœurs et frères.*

*A tous mes amis (es).*

*CuZna*

# *Dédicaces*

*Je dédie ce modeste travail :*

*A ma très chère mère, sans laquelle je n'aurais pas  
pu parvenir à ce stade.*

*A mon très cher père.*

*A mes sœurs et frères.*

*A tous mes amis (es).*

*Cuzna.*

## **Résumé**

Les archivistes ont tendance de passer vers l'ère numérique, en faisant des versions numérisées des manuscrits. Mais le travail ne s'arrête pas là, car même numérisés sous forme d'images, les manuscrits sont inaccessibles en mode plein texte dans une bibliothèque ou sur internet à l'aide d'un moteur de recherche et c'est à cet instant qu'intervient le catalogage.

Le catalogage est la description des caractéristiques internes et externes spécifiques d'un document. La description se fait en utilisant un format d'encodage. Plusieurs formats d'encodage ont vu le jour. Nous pouvons citer entre autre, l'EAD et TEI. Devant la multiplicité des standards d'encodages, il deviendrait difficile de mener une recherche exhaustive.

L'objectif de notre travail est de concevoir un système interopérable offrant la possibilité d'accès à tout catalogue, quel que soit son format d'encodage.

# Table des matières

<b>Introduction général</b> .....	001
<b>Position du problème et solution proposée</b> .....	003
<b><u>Première partie : État de l'art</u></b>	
<b>Chapitre 1: Notions générales</b>	
<b>Sous-chapitre 1 : Notions sur les manuscrits arabes</b> .....	005
Introduction.....	005
I. Les anciens manuscrits arabes .....	005
1. Qu'est-ce qu'un manuscrit ? .....	005
2. Caractéristiques générales des manuscrits arabes anciens.....	005
II. Caractéristiques du manuscrit arabe.....	006
1 – La notice bibliographique.....	006
2 – Rédaction de la notice bibliographique de manuscrit.....	007
III. Conclusion .....	011
<b>Sous-chapitre 2 : Numérisation des manuscrits</b> .....	012
Introduction.....	012
I. Définition de la numérisation .....	012
II. Objectifs de la numérisation des manuscrits.....	013
III. Les modes de numérisation.....	013
IV. Conclusion .....	014
<b>Sous-chapitre 3 : Catalogage des manuscrits arabes</b> .....	015
Introduction.....	015
I. Notions sur les catalogues .....	015
1. Qu'est-ce qu'un catalogue ?.....	015
2. Qu'est-ce le catalogage ?.....	015
3. Les avantages du catalogage .....	015
4. Types des catalogues.....	016
5. Le fonctionnement du catalogue .....	016
II. Le catalogage des manuscrits arabes .....	016
1. Les catalogues traditionnels des manuscrits arabes .....	016
2. Le catalogage informatisé des manuscrits arabes.....	017
3. Qu'est-ce qu'une métadonnée ? .....	018
4. Utilités des métadonnées.....	018
5. Les métadonnées et la description des manuscrits arabes.....	018

6. Les éléments descriptifs des manuscrits arabes.....	019
7. Choix des métadonnées descriptives des manuscrits arabes numérisés.....	020
III. Indexation des données .....	022
1. Modes d'indexation.....	024
V. Conclusion .....	025

#### **Sous-chapitre 4 :Concepts d'interopérabilité .....**

Introduction.....	026
I. Qu'est que c'est l'interopérabilité.....	026
II. Les principes d'interopérabilité .....	027
III. Les approches d'interopérabilité.....	027
IV. Les niveaux d'interopérabilité.....	028
1. L'interopérabilité technique.....	028
2. L'interopérabilité sémantique .....	029
3. L'interopérabilité organisationnelle.....	030
V. Conclusion .....	031

### **Chapitre 2:Encodage du catalogue**

#### **Sous-chapitre 1 :La description de la DTD EAD**

Introduction.....	032
I. Historique .....	032
II. Définition de l'EAD.....	032
III. Caractéristique de l'EAD.....	033
IV. Avantage du format EAD .....	033
V. Structure de l'EAD.....	033
VI. Description des manuscrits avec la DTD EAD.....	048
VII. Conclusion.....	048

#### **Sous-chapitre 2 : La description de DTD TEI**

Introduction.....	049
I. Historique.....	049
II. Définition de la TEI.....	049
III. Caractéristiques de la TEI.....	050
IV. Avantages de la TEI.....	050
V. Principes de balisage .....	050
VI. La structure générale de la TEI.....	051
VII. La TEI-Manuscript Description (TEI-ms) .....	053
VIII. Encodage de manuscrits arabes avec la TEI-ms.....	061
IX. Conclusion .....	062

#### **Sous-chapitre 3 : La description du DC**

Introduction.....	063
I. Historique.....	063
II. Objectifs .....	063

III.	<i>Le Dublin Core simple</i> .....	063
IV.	<i>Le Dublin Core qualifié</i> .....	072
	1. <i>Les éléments supplémentaires</i> .....	073
	2. <i>Les éléments de raffinements ou qualifiants</i> .....	073
V.	<i>Forces du Dublin Core</i> .....	075
VI.	<i>Dublin Core et les manuscrits arabes</i> .....	075
VII.	<i>Conclusion</i> .....	075

## **Deuxième partie : Mise en œuvre de la solution d'interopérabilité**

### **Chapitre 1 : Enrichissement des formats d'encodage**

<i>Introduction</i> .....	076
I. <i>Pourquoi employer Dublin Core comme format de médiateur ?</i> .....	078
1. <i>Le protocole OAI-PMH et le Dublin Core</i> .....	077
II. <i>Étude de l'accès aux manuscrits numérisés</i> .....	077
III. <i>Enrichissement de la TEI-<i>ms</i> et son adaptabilité à la description des manuscrits arabes</i> .....	079
IV. <i>Enrichissement de l'EAD et son adaptabilité à la description des manuscrits arabes</i> .....	079
V. <i>Enrichissement de Dublin Core et son adaptabilité à la description des manuscrits arabes</i> .....	086
VI. <i>Correspondance entre EAD, TEI-<i>ms</i> enrichie &amp; Dublin Core</i> .....	105
VII. <i>Conclusion</i> .....	108

### **Chapitre 2 : Conception et implémentation**

<i>Introduction</i> .....	109
I. <i>Le processus de recherche dans les catalogues de manuscrits arabes</i> .....	109
II. <i>La recherche d'information dans les documents XML</i> .....	109
1. <i>Manipulation des documents XML</i> .....	109
2. <i>Indexation</i> .....	110
3. <i>Techniques de stockages et d'interrogation des documents XML</i> .....	116
III. <i>L'implémentation du moteur de recherche</i> .....	116
1. <i>Schéma de stockage</i> .....	117
2. <i>L'architecture générale du processus de recherche</i> .....	123
IV. <i>Présentation des outils de développement</i> .....	129
V. <i>Présentation des interfaces graphique</i> .....	131
VI. <i>Conclusion</i> .....	136
<i>Conclusion générale</i> .....	137
<i>Annexe A</i> .....	138
<i>Annexe B</i> .....	144
<i>Bibliographie</i> .....	149

# Liste des figures

<b>Figure 01</b> :Exemple d'un manuscrit arabe.....	006
<b>Figure 02</b> :Exemple d'un colophon.....	009
<b>Figure 03</b> :Exemple de décors de manuscrit .....	009
<b>Figure 04</b> :Procédure de numérisation des manuscrits .....	012
<b>Figure 05</b> :Schéma du processus de recherche d'information.....	023
<b>Figure 06</b> :Principes d'interopérabilité sémantique.....	030
<b>Figure 07</b> : Les quatre niveaux d'interopérabilité.....	030
<b>Figure 08</b> : Schéma d'un instrument de recherche EAD.....	035
<b>Figure 09</b> :Les éléments obligatoires de l'EAD.....	036
<b>Figure 10</b> :Les sous-éléments de l'élément <did>.....	037
<b>Figure 11</b> :Description générale d'un document XML TEI.....	052
<b>Figure 12</b> :L'encodage global d'un document encodé avec la TEI.....	052
<b>Figure 13</b> :Les sous-éléments de l'élément <msDesc>.....	054
<b>Figure 14</b> :Interopérabilité basés sur le petit dénominateur commun .....	077
<b>Figure 15</b> :Structure de la solution de l'interopérabilité des catalogues en utilisant Dublin Core .....	078
<b>Figure 16</b> :Structure de l'élément <origination> .....	080
<b>Figure 17</b> :Structure de l'élément <altformavail>.....	080
<b>Figure 18</b> :Structure de l'élément <bioghist>.....	082
<b>Figure 19</b> :Structure de l'élément <colophon>.....	082
<b>Figure 20</b> :Structure de l'élément <scopecontent> .....	083
<b>Figure 21</b> : Structure de l'élément <physdesc>.....	083
<b>Figure 22</b> : Structure de l'élément <originalsloc> .....	084
<b>Figure 23</b> : Structure de l'élément <imprint>.....	084
<b>Figure 24</b> : Structure de l'élément <archdesc> .....	085
<b>Figure 25</b> : Structure de l'élément <odd>.....	085
<b>Figure 26</b> : Structure de l'élément <summary> .....	086
<b>Figure 27</b> : Structure de l'élément <dc :name>.....	087
<b>Figure 28</b> : Structure de l'élément <dc :WasAlive> .....	090

<b>Figure 29</b> : Structure de l'élément <dc:terms:extent>.....	092
<b>Figure 30</b> : Structure de l'élément <dc:provenance> .....	094
<b>Figure 31</b> : Structure de l'élément <dc:publisher>.....	094
<b>Figure 32</b> : Structure de l'élément <dc:terms:tableOfContents> .....	095
<b>Figure 33</b> : Structure de l'élément <dc:volNumber> .....	096
<b>Figure 34</b> : Structure de l'élément <dc:foliation> .....	096
<b>Figure 35</b> : Structure de l'élément <dc:summary>.....	097
<b>Figure 36</b> : Structure de l'élément <dc:signature>.....	098
<b>Figure 37</b> : Structure de l'élément <dc:condition>.....	099
<b>Figure 38</b> : Structure de l'élément <dc:handNote> .....	100
<b>Figure 39</b> : Structure de l'élément <dc:decoDesc> .....	100
<b>Figure 40</b> : Structure de l'élément <dc:buinding>.....	101
<b>Figure 41</b> : Structure de l'élément <dc:separatedMaterial>.....	101
<b>Figure 42</b> : Structure de l'élément <dc:imprint>.....	102
<b>Figure 43</b> : Structure de l'élément <dc:translation> .....	103
<b>Figure 44</b> : Structure de l'élément <dc:originalsloc> .....	104
<b>Figure 45</b> : Structure de l'élément <dc:odd>.....	105
<b>Figure 46</b> : Principe de migration.....	105
<b>Figure 47</b> : Parsing d'un document XML.....	110
<b>Figure 48</b> : Indexation basée sur des champs.....	113
<b>Figure 49</b> : Indexation basée sur les chemins.....	113
<b>Figure 50</b> : Indexation basée sur les arbres.....	114
<b>Figure 51</b> : Présentation d'un fichier inverse.....	115
<b>Figure 52</b> : Schéma du processus d'accès aux images des manuscrits.....	117
<b>Figure 53</b> : Le digramme de classes.....	119
<b>Figure 54</b> : L'architecture générale du processus de recherche.....	124
<b>Figure 55</b> : L'interface principale de NetBeans .....	130
<b>Figure 56</b> : L'interface d'accueil.....	131
<b>Figure 57</b> : L'interface de la réindexation.....	132
<b>Figure 58</b> : L'interface de fin de réindexation .....	132
<b>Figure 59</b> : L'interface de recherche.....	133



# ***Introduction générale***

# Introduction générale

Les manuscrits sont document témoins de l'histoire. Les civilisations anciennes ont utilisé les manuscrits pour consigner les événements importants de l'histoire, ou même simplement enregistrer les transactions au jour le jour de la gouvernance locale. Notre compréhension de notre passé dépend en grande partie de ces documents.

De toute évidence l'importance des manuscrits ne peut être estimée. Toutefois les manuscrits ont tendance à se détériorer rapidement après manipulations répétées en raison de leur fragilité.

Les archivistes ont tendance de passer vers l'ère numérique, en faisant des versions numérisées des manuscrits. Mais le travail ne s'arrête pas là, car même numérisés sous forme d'images, les manuscrits sont inaccessibles en mode plein texte dans une bibliothèque ou sur internet à l'aide d'un moteur de recherche et c'est à cet instant qu'intervient le catalogage.

Le catalogage est la description des caractéristiques internes et externes spécifiques d'un document. La description se fait en utilisant un format d'encodage. Plusieurs formats d'encodage ont vu le jour. Nous pouvons citer entre autre, l'EAD et TEI. Devant la multiplicité des standards d'encodages, il deviendrait difficile de mener une recherche exhaustive.

L'objectif de notre travail est de concevoir un système interopérable offrant la possibilité d'accès à tout catalogue, quel que soit son format d'encodage.

Afin d'atteindre notre objectif, nous présentons au premier lieu, la problématique de notre thème et la solution que nous proposons, puis nous avons subdivisé notre travail en deux parties :

● **Première partie** : État de l'art, elle se compose de deux chapitres :

➤ **Chapitre 1** : Dans ce chapitre nous parlerons des généralités pour introduire le lecteur dans le domaine des manuscrits arabes anciens. En suite nous introduirons le concept de la numérisation. Juste après nous verrons les différents concepts du catalogage, ainsi les protocoles décrivant les manuscrits arabes, et nous définirons à la fin l'interopérabilité, et ses différents principes.

➤ **Chapitre 2** : Divisé en trois parties, nous présentons en premier l'EAD, en second la TEI et nous terminerons par le Dublin Core.

# Introduction générale

● **Deuxième partie** : Mise en œuvre de la solution d'interopérabilité, elle est composée de deux chapitres :

- **Chapitre 1** : Enrichissement des formats d'encodage, dans ce chapitre nous enrichirons le format EAD, puis nous adapterons le format Dublin Core.
- **Chapitre 2** : Conception et implémentation de notre plate-forme d'interopérabilité, dont nous décrivons le processus de recherche dans le catalogue des manuscrits arabes, et les différents outils utilisés pour le d'enveloppement, l'architecture du système de recherche, ainsi que les différents modules qui le composent suivis des captures d'écrans expliquant le fonctionnement de notre système de recherche d'information.

Nous terminerons ce mémoire par une conclusion générale de notre projet et ses perspectives.

# ***Problématique***

## *Problématique*

Les manuscrits arabes constituent un trésor universel pour l'humanité. Ils se retrouvent constamment menacés par l'effritement à cause de la précarité et du manque de moyens des lieux de conservation. La manipulation de ces œuvres constitue un danger supplémentaire de détérioration. Par conséquent, la transformation de ces derniers en documents électroniques par la technique de numérisation a permis à la fois leurs préservations dans de meilleurs états et a rendu leur consultation possibles à plusieurs chercheurs simultanément grâce à l'internet, mais une question mérite d'être soulevée :

## *Comment se fera l'accès aux manuscrits digitalisés ?*

Avant la mise en ligne d'un manuscrit numérisé, le documentaliste procède à la description du manuscrit. Au fait, il s'agit de créer un ensemble de notices bibliographiques qui constitueront le catalogue des manuscrits numérisés.

Le catalogue constitue un bon outil d'accès aux manuscrits numérisés. Mais, au vue de la multiplicité des formats de catalogage, le résultat n'est pas garanti pour accéder à tous les manuscrits répondant favorablement à la même requête utilisateur : Un taux de silence important. Il serait donc intéressant de réduire ce taux. Mais comment faire ? C'est dans ce contexte la que notre travail se situe, la question qui se pose alors est rendre ces différents systèmes interopérables.

## *Solution proposée*

Notre objectif est de rendre disponible toutes les images des anciens manuscrits arabes numérisés qui satisfont une requête utilisateur spécifique, indépendamment du format d'encodage.

Bien que de nombreuses méthodes de catalogage aient été mise au point, une norme de catalogage standard pour les manuscrits arabes n'existe pas. Chaque bibliothèque utilise ces propres outils de catalogage.

# Position du problème et solution proposée

La mise en place d'une plateforme d'interopérabilité est basée sur l'utilisation du format Dublin Core. Un format universel d'interopérabilité, simple et très utilisé sur Internet. Le système se base sur le moissonnage des données sources issues des formats les plus utilisés pour le catalogage des manuscrits arabes anciens : l'EAD et la TEI-xml.

Notre méthodologie d'approche est décrite comme suit :

- ✓ Étudier les deux normes d'encodages l'EAD et la TEI connu par leur description des manuscrits.
- ✓ Déceler les insuffisances de l'un et de l'autre des formats (les deux normes ne décrivent pas les manuscrits arabes d'une façon exhaustive.
- ✓ Proposer des modifications et ajouts d'éléments (métadonnées) tout en respectant leurs structures.
- ✓ Adapter un nouveau format d'encodage: Le Dublin Core, protocole utilisé par OAI-PMH dans le but d'assurer l'interopérabilité entre les différentes sources numériques.

# **Première partie : État de l'art**

**Chapitre 1 : Notions générales**

**Chapitre 2 : Encodage du catalogue**

# **Chapitre 1 : Notions générales**

**Sous-chapitre 1: Notions sur les manuscrits arabes**

**Sous-chapitre 2: Numérisation des manuscrits**

**Sous-chapitre 3: Catalogage des manuscrits arabes**

**Sous-chapitre 4: Concept d'interopérabilité**

## ***Introduction***

Témoin d'une époque glorieuse du monde musulman, le manuscrit arabe est recherché pour divers aspects. Premièrement, un intérêt particulier est porté par les chercheurs à son contenu, qui constitue un fonds juste partiellement exploré et deuxièmement, son aspect matériel constitue un intérêt non des moindres par rapport au premier.

L'objectif de notre étude dans cette partie est d'étudier la description générale des manuscrits arabes et leurs caractéristiques afin d'élaborer un mode d'accès électronique à ces manuscrits numérisés en mode image.

### ***I. Les anciens manuscrits arabes :***

#### ***1. Qu'est-ce qu'un manuscrit ?***

Un manuscrit (du latin scriptus) est un texte écrit à la main [wiki 01 2012]. Le manuscrit est le véritable témoin de la présence d'un texte à une époque donnée. Il véhicule les connaissances de cette ère. Il serait donc intéressant de pouvoir accéder au contenu de ces manuscrits. Pour ce faire, il faudrait mettre en place un système de descriptions adéquates des manuscrits.

Notons par ailleurs, qu'un manuscrit est tout d'abord une œuvre produite d'une manière artisanale, utilisant des matériaux rares et coûteux. Il est possible de ce fait un document archéologique qui peut être approché à travers ses caractéristiques matérielles et son histoire.

#### ***2. Caractéristiques générales des manuscrits arabes anciens :***

Les documents anciens sont d'une valeur inestimable, ils montrent souvent des dégradations plus ou moins importantes à cause des conditions de préservation et risquent ainsi de devenir inexploitable [Mokdad 1992]. La structure des manuscrits arabes peut être résumée comme suit [Toumazet 1987] :

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



Figure 01 : Exemple d'un manuscrit arabe.

## II. Caractéristiques du manuscrit arabe

Un manuscrit est un document unique, qu'il faut examiner depuis sa fabrication, jusqu'à l'intervention des lecteurs et des possesseurs successifs. Par conséquent, les rédacteurs de notices descriptives de manuscrits doivent relever tous les aspects caractérisant les manuscrits, relatifs à sa composition et à son contenu.

### 1 – La notice bibliographique

L'objectif principal de la notice bibliographique est de reporter d'une manière ordonnée et sous une forme claire et lisible les résultats de l'analyse du manuscrit. Le contenu de la notice est présenté selon une terminologie bien précise et codifiée [IRHT 2006].

La notice bibliographique permet au lecteur d'avoir une idée générale sur le manuscrit avant de le consulter. Elle doit donc signaler les richesses et les particularités renfermées par ce dernier.

La notice bibliographique devra s'abstenir, par contre, de discussions et de développements critiques du manuscrit. Elle doit se limiter aux éléments descriptifs qu'ils soient matériels, historiques ou de contenu du manuscrit[soualah 2008].

## 2 – Rédaction de la notice bibliographique de manuscrit

La réussite de tout projet de numérisation de manuscrits repose essentiellement, sur la facilité et l'efficacité d'accès aux manuscrits. Ce qui signifie l'élaboration d'une notice bibliographique décrivant succinctement le manuscrit.

La rédaction d'une notice est une opération délicate et précise, qui demande un savoir-faire dans le domaine de catalogage.

### 2.1 – Les champs bibliographiques

Le champ bibliographique résume l'ensemble des éléments descriptifs d'un manuscrit. Le caractère d'unicité des manuscrits leur donne la particularité d'être différents les uns des autres : Il s'agit d'un ensemble d'éléments définissant d'une manière unique les manuscrits [soualah2008]. L'identité de ces éléments sera très utile pour la création d'un format de description électronique, propre aux manuscrits arabes [keilah 2004].

Parmi les éléments d'identification nous retrouvons principalement :

- a) **La mention de responsabilité** : L'écriture manuelle du manuscrit fait apparaître une nouvelle mention de responsabilité qui est celle du "copiste", pouvant être mentionnée à la suite du nom de l'auteur [keilah2004].
- b) **Nom des propriétaires** : Le nom de(s) propriétaire(s) ou possesseur(s) du manuscrit est également un indice important pour les chercheurs désirant suivre le développement historique du manuscrit. La structure des noms arabes anciens présente une grande variation, due à la complexité de la structure du nom arabe. En effet, ce dernier se retrouve en général, défini selon quatre éléments distincts à savoir "Ism", "Laqab", "Kunya" et "Nisba"
  - ✚ Ism : Le "Ism" est un nom donné dans l'ordre direct.
  - ✚ Laqab : Le "Laqab" désigne le nom de célébrité de la personne.
  - ✚ Kunya : La "Kunya" est une marque de distinction appliquée à des personnages de premier plan pour les honorer.
  - ✚ Nisba: La "Nisba" est un adjectif formé à l'aide du suffixe *i* afin d'indiquer l'origine, le lieu de naissance ou de résidence d'une personne. Exemple : le mathématicien "alKhawarizmi" se réfère à Khwarizm, nom d'un lieu situé en Asie centrale.
- c) **Titre** : Le titre de l'ouvrage ou le titre de la série est l'élément principal d'identification qui dans la majorité des cas, est présenté en page de titre.

## 2.2 – Éléments descriptifs des manuscrits arabes

Les champs bibliographiques cités plus haut ne sont pas suffisants pour l'identification d'un manuscrit, car il peut en exister des œuvres qui, à travers les âges ont perdu ces informations [soualah 2008]. Pour décrire les manuscrits arabes on a divisé les éléments descriptifs en trois volets principaux, l'unité codicologique et le volume, étude du texte et enfin les notes.

### 2.2.1 – L'unité codicologique et le volume :

Un manuscrit peut être formé de plusieurs volumes, un volume peut être formé à son tour de plusieurs cahiers. La description de l'unité codicologique se résume à la description des volumes et éventuellement des cahiers, du papier, de l'écriture et aussi de la copie.

- a) **Le volume :** Un manuscrit peut être composé de plusieurs cahiers de parchemin ou de papier. Il est donc important d'observer l'organisation du volume. Cette opération permet de déterminer si le manuscrit est homogène ou composite, complet ou mutilé, si les cahiers sont reliés en ordre ou dans le désordre.

Pour ce faire, il faut considérer les signes de numérotation et de renvoi. Ils permettent de se repérer dans le manuscrit.

- **La cote :** Elle définit le numéro d'ordre séquentiel
  - **La foliotation :** C'est la numérotation de chaque feuillet formé par l'ensemble recto verso.
  - **La signature :** C'est la numérotation des cahiers. Elle se trouve soit sur le premier feuillet du cahier, au recto, soit sur le dernier au verso ou soit, sur les deux.
  - **La réclame :** C'est le premier mot du cahier suivant, inscrit dans la marge inférieure de la dernière page du cahier précédent. La réclame permet au couseur et au relieur de vérifier la bonne succession des feuillets et des cahiers.
  - **Libellé général :** Le libellé d'un manuscrit est défini par :
    - **La matière :** Décrit la matière de base de fabrication d'un manuscrit.
    - **La dimension des pages :** La mesure des pages
    - **Nombre total de folios :** Le papier étant un produit rare, les auteurs transcrivent sur les deux faces d'une feuille, on parle alors de folio.
- b) **Le colophon :** Il s'agit de la formule finale dans laquelle le copiste mentionne son nom, la date selon les calendriers hégire et grégorien et le lieu de la copie [IRHT2006]. L'image qui ce suit illustre un exemple d'un colophon :



Figure 02 :Exemple d'un colophon

- c) **La reliure :** C'est l'assemblage des cahiers formant un codex. De ce fait, elle doit être décrite au même titre que les autres éléments du codex. La reliure intéresse d'une part l'histoire du livre et des techniques artisanales, d'autre part l'histoire intellectuelle [IRHT 2006].
- d) **Le décor :** Le décor peut se présenter sous une forme peinte ou dessinée. Il apporte des indices de datation et de localisation du manuscrit. Il joue plusieurs rôles : symbolique, historique, et parfois, purement esthétique. Par conséquent, il permet à la fois d'articuler le texte et au lecteur de se repérer.



Figure 03 : Exemple de décors de manuscrit.

- e) **Type de support :** Plusieurs types de supports ont été utilisés pour rédiger les manuscrits arabes : le papier et le parchemin (le parchemin est la peau d'animal épilée et effleurée ayant subi un traitement non tannant puis un séchage sous tension la rendant propre à recevoir l'écriture sur ses deux faces [Muzelle 1985]).
- f) **Cahiers :** Les cahiers sont décrits par les éléments suivants :
  - Collation des cahiers (nombre de cahiers, somme des folios).
  - Anomalie de composition (feuillet manquant, feuillet ajouté, ordre de lecture).
  - Numéros de cahier.

**g) Préparation de la page avant l'écriture :** Elle décrit les éléments suivants :

- La surface écrite.
- Le nombre de lignes par page.
- La description des piqûres (série de petits trous ou fentes alignés sur le feuillet, destinés à guider le traçage de la réglure).
- Palimpseste : Représente le support d'écriture réutilisé dont le contenu pourrait disparaître à cause d'un lavage ou grattage.

**h) Écriture :** L'écriture se trouve décrite par les éléments suivants :

- Correction de copiste : Raturage, grattage, omissions, rectifications, collation, additions marginales et surlinéaires.
- Organisation du manuscrit : Description de l'articulation des textes, des chapitres entre eux et de la hiérarchie des subdivisions. La structure du manuscrit diffère selon les sujets traités. Les manuscrits arabes se retrouvent structurés en "kytab", "djuz", "bab", "fasl" [soualah 2008].
- Système d'écriture : Le degré de cursivité, la couleur d'encre, la taille, la particularité d'écriture, la proportion des lettres, l'inclinaison.

### 2.2.2 – Étude du texte :

Porte sur une description du contenu textuel du manuscrit sans pour autant s'intéresser au contenu proprement dit. La description du texte se fait en général en cinq éléments, que nous décrivons dans ce qui suit :

- Matière scientifique : Appelée encore genre, il s'agit de déterminer le thème sur lequel se porte le texte selon la division traditionnelle des sciences arabo-musulmanes (*fiqh, tafsir, adab, kalam, medecine, astronomie, etc.*).
- Résumé : le résumé contiendra une description contextuelle du volume, tout en notant les mots clés.
- Références : Il s'agit de reporter les références des différentes éditions récentes et antérieures du texte.
- Incipit : Les trois séquences, *basmala, hamdala* et entrée en matière.
- Explicit : Dernier mot du texte avant le colophon.

### 2.2.3 – Notes :

- La transmission de savoir : Attestations de lecture, attestations de collation, certificats d'audition (*samã*), certificats de transmission (*igãza*).
- Notes diverses : Marques de possession, inscription de waqf, table des matières, vers, invocations, exercices d'écriture, essais de plume, prix de vente.

### ***III – Conclusion***

Le catalogue des manuscrits Arabes anciens est un outil pertinent pour l'accès à ces ressources numérisées. Une bonne opération de catalogage repose sur la détermination de métadonnées représentative du manuscrit. Ainsi, la description du manuscrit devra se faire d'une manière minutieuse. Elle repose sur la description des caractéristiques codicologiques, paléographiques et l'histoire du manuscrit.

Dans ce chapitre, nous avons présentés les différentes métadonnées que devrait contenir une notice bibliographique, certes délicates et multiples mais d'une utilité essentielle pour l'accès aux manuscrits arabes numérisés.

## Introduction

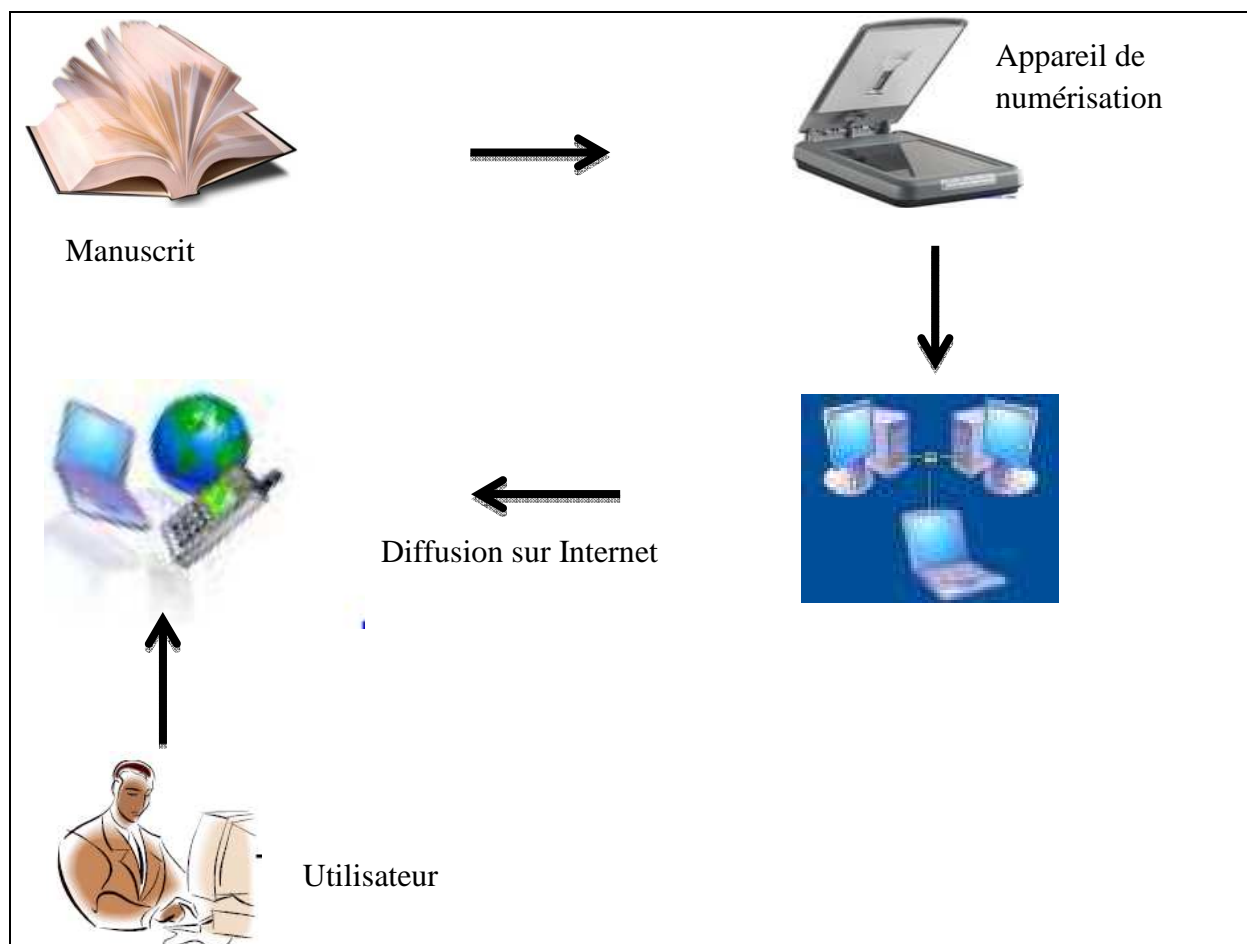
Dans cette partie, nous allons présenter l'aspect de la numérisation des manuscrits, tout en mentionnant ses objectifs et ses enjeux.

### I. Définition de la numérisation

La numérisation est une procédure qui consiste en la dématérialisation du support physique. Il s'agit d'une reproduction de la forme physique du document sous une forme électronique.

La numérisation (ou digitalisation), consiste à créer une image du document à l'aide d'une caméra numérique ou d'un scanner. Une haute résolution est souvent nécessaire pour restituer les éléments les plus fins de l'écriture et des graphismes.

La procédure de numérisation est décrite par le schéma suivant :



**Figure 04** : Procédure de numérisation des manuscrits.

## II. Objectifs de la numérisation des manuscrits :

La numérisation des manuscrits arabes anciens est motivée par plusieurs objectifs parmi lesquels nous citons [Cedelle-Joubert et al 2002]:

- **Objectif de préservation** : Il est reconnu qu'aujourd'hui la numérisation est un moyen de préservation plutôt que de conservation. En effet, un document en format numérique est de consultation très aisée et de manipulation facile. La numérisation permet de ce fait, d'augmenter le taux de consultation du document sans un impact physique sur l'original.

- **Objectif de diffusion** : L'accès physique au manuscrit représente un danger potentiel qui risque de réduire considérablement la durée de vie du manuscrit. Il est donc important de trouver un mode d'accès au manuscrit sans pour autant le manipuler physiquement. La diffusion du format numérique du manuscrit en Intranet de la bibliothèque, voire sur Internet constitue une solution salubre qui permet d'une part, d'accéder à l'œuvre et d'autre part, de limiter sa sortie des lieux de rangement.

Par ailleurs, il est important de rappeler la grande souplesse qu'offre la diffusion numérique d'un document.

- **Objectif de valorisation** : Tout document non consulté peut être considéré comme un document mort. En effet, une information invisible, combien même pertinente, ne peut guère servir l'humanité. Sa mise en ligne joue un double rôle : Le premier consiste à permettre son utilisation par le lecteur alors que le second permet sa sortie de l'anonymat. Cet aspect de valorisation de l'œuvre ne peut qu'être bénéfique pour la bibliothèque, lieu de conservation du manuscrit.

- **Objectif de l'aide à la recherche** : Comme la diffusion ou la valorisation des collections auprès d'un public élargi, l'amélioration des conditions de la recherche est l'un des objectifs que les bibliothèques fixent souvent à leurs programmes de numérisation. Dans ce domaine, la numérisation peut en effet rendre des services sans comparaison avec ce qui existe déjà par ailleurs (reproductions, éditions ou index sur supports papiers par exemple).

Encore faut-il, une fois de plus, que soient mises en œuvre les conditions nécessaires à cet usage optimal du document numérique : cohérence des corpus choisis, réflexion sur le mode de numérisation à adopter, sur les index. Là intervient la responsabilité des bibliothèques qui ont, nécessairement, des choix à faire et des décisions à prendre en amont.

## III. Les modes de numérisation

### a) Numérisation en mode image

Ce mode de numérisation sert à produire une présentation photographique de chaque page de document à numériser, obtenant ainsi une copie numérisée en fac-similé électronique du document originale.

A ce stade, il nous semble important de préciser qu'il existe trois types de numérisation en modes images [Farag 2012]:

- ✚ Le mode bitonal : Image noir et blanc, fichiers non volumineux. Ce mode est essentiellement employé pour la numérisation des documents en bon état.
- ✚ Le mode niveaux de gris : C'est un mode qui permet de capturer les nuances du document, les densités de l'encre, mais traduit aussi les taches et ombres en niveaux de gris plus ou moins élevés, ce qui peut gêner la lecture.
- ✚ Le mode couleurs RVB (Rouge Vert Bleu) : Ce mode permet aussi de capturer les nuances du document, les densités de l'encre, mais traduit les taches et piqures du papier dans leur couleur d'origine, ce qui facilite la lecture.

#### ***b) Numérisation en mode de texte***

Cette technique de numérisation revient à coder le texte en tant que tel. Le mode de numérisation de texte doit contenir compte trois caractéristiques élémentaires, communes à tout document :

- Le codage des systèmes d'écriture.
- La structure physique du document.
- La structure logique du document.

Ces deux modes de numérisation ont également des inconvénients. Tel que, le mode texte prend trop de temps, ainsi le mode image présente des difficultés pour la recherche à texte intégral. Une solution intelligente de numérisation n'est pas de choisir seulement un des deux modes, plutôt de les combiner pour obtenir les meilleurs résultats.

## ***IV. Conclusion***

La numérisation des manuscrits arabes est aujourd'hui le meilleur moyen pour diffuser auprès d'un public élargi des représentations aussi fidèles que possible de ces « trésors » patrimoniaux et de leurs contenus.

La numérisation n'est qu'un outil parmi d'autres pour atteindre ces objectifs. En particulier, ce qui intéresse l'utilisateur dans la numérisation c'est la valeur ajoutée qu'elle apporte par rapport aux outils traditionnels (microfilms,...etc) : un meilleur confort de lecture, la couleur, la possibilité de tourner les pages, une diffusion dans le temps et l'espace démultipliée, etc.

## ***Introduction***

Les bibliothèques et les centres de documentation possèdent des documents organisés en collections. Leurs missions sont : le stockage des documents pour permettre leur accès à l'utilisateur, et la conservation, mais aussi la création des outils qui permettent d'accéder aux caractéristiques intellectuelles et physiques des documents sans avoir à les consulter, ces outils sont les catalogues.

Dans cette partie, nous allons présenter des notions sur le catalogage d'une manière générale, tout en focalisant sur le catalogage des manuscrits arabes anciens.

### ***I. Notions sur les catalogues***

#### ***1. Qu'est-ce qu'un catalogue ?***

Le catalogue est un outil de recherche qui identifie et localise des documents présents dans un établissement documentaire (bibliothèque, musée...etc.) [Catalogue 2012].

#### ***2. Qu'est-ce le catalogage ?***

Le catalogage est le processus d'enregistrement des informations détaillées, et de la description des caractéristiques internes et externes spécifiques d'un document.

#### ***3. Les avantages du catalogage***

Parmi les différents avantages qu'offre le catalogage nous citons [Piérache 2008] :

- Description et identification d'un document : Le catalogage consiste en la description de la forme (auteur, titre, années d'apparitions...etc.).
- Indexation : C'est la description du contenu d'un document.
- Récupération des éléments d'identification d'un document qui diffère selon le type du document. Cela permet aux utilisateurs de connaître et de chercher l'existence d'un document, ou bien de sélectionner un document.
- L'élaboration des notices bibliographiques strictement normalisées.

## **4. Types des catalogues**

### **4.1. Le catalogue imprimé**

Ce sont des catalogues sur fiches, leur mise à jour est appelée *intercalation* [Piérache 2008]. Ce type de catalogage ne permet pas de trouver la totalité des fonds documentaires, et l'interrogation à distance.

### **4.2. Le catalogue informatisé**

Le catalogue informatisé est un outil de recherche consultable via un ordinateur [Catalogue 2012]. Il annonce la disponibilité d'un document pour un lecteur.

## **5. Le fonctionnement du catalogue**

Le catalogue étant un ensemble de notices bibliographiques, il faut pouvoir rechercher les notices pertinentes en fonction des critères de recherches qui sont les points d'accès à ces notices [Piérache 2008] :

### **1. Le catalogue imprimé**

Avec ce genre de catalogues les critères des recherches sont peu nombreux (auteurs, titre, collection, matière,...). Pour cela il faudra maintenir plusieurs fiches pour un seul ouvrage.

### **2. Le catalogue informatisé**

Vu l'existence d'une seule notice bibliographique dans ce type de catalogage, les critères de recherche peuvent être multipliés à loisir. Cela avec la minimisation des problèmes de classement de fiches, ce qui permet d'avoir un gain du temps.

## **II. Le catalogage des manuscrits arabes**

Les manuscrits arabes sont des documents d'une grande utilité et d'importance patrimoniale, pour cela, ils doivent être conservés, et aussi mis à la disposition de toute personne désirant les consulter. Ce qui est réalisable à l'aide du catalogue.

En effet, le processus de catalogage des manuscrits arabes a évolué d'une manière remarquable, en commençant par les catalogues imprimés jusqu'à l'arrivée d'internet et l'apparition des catalogues en ligne.

### **1. Les catalogues traditionnels des manuscrits arabes**

Avant l'apparition d'internet et l'informatisation des majorités des domaines y compris le domaine des manuscrits, le catalogage s'est effectué d'une manière traditionnelle.

Le catalogage traditionnel des manuscrits arabes n'avait pas de forme précise à suivre, le catalogage varie d'un catalogueur à un autre selon l'intérêt du manuscrit et aussi les spécialités des catalogueurs.

Les anciens catalogueurs réalisent leurs catalogages en se basant sur une description formelle des documents, et en fournissant par exemple des informations bibliographiques telles que : le titre, l'auteur, l'année de publication. Le catalogage des manuscrits arabes anciens est l'œuvre de deux écoles différentes [Soualah 2008] :

**1.1. L'école arabe :** les travaux de catalogage des manuscrits arabes datent de l'époque abbasside. Les catalogueurs de cette école utilisaient deux catégories d'éléments pour décrire les manuscrits :

- a) **Éléments fixes :** Ce sont des éléments qui sont utilisés dans la plupart des catalogues : le titre de manuscrit, l'auteur de manuscrit, le copiste, la date de réalisation du manuscrit, le style d'écriture, une description matérielle du manuscrit (la tomaison, la pagination, le nombre de lignes par page, le format).
- b) **Éléments variables :** Lieu de réalisation du manuscrit, l'incipit, l'encre utilisé, l'état du manuscrit.

**1.2. L'école orientaliste :** L'école orientaliste donne une importance majeure à la recherche et l'analyse de l'évolution littéraire arabes, et aussi aux catalogages des manuscrits. Pour le catalogage, les catalogueurs orientalistes ont utilisé les éléments de description suivants : le titre, l'auteur, la date de manuscrits, le type d'écriture, la pagination, le format, le copiste, le lieu de réalisation du manuscrit, l'incipit et le nombre de lignes par page.

## 2. Le catalogage informatisé des manuscrits arabes

Afin d'informatiser des anciens manuscrits, divers projets ont lieu. Les premiers catalogues apparus ont utilisé le format MARC<sup>1</sup> (Machine Readable Catalogue or Cataloguing), en absence de tout format spécifiquement destiné aux manuscrits à leur époque.

Avec l'émergence de l'internet et l'évolution du document électronique en texte intégral d'autres genres de formats sont apparus, citons à titre d'exemple : le format SGML (Standard Generalized Markup Language), HTML (HyperText Markup Language) et XML (eXtended Mark up Language) qui s'adapte particulièrement à la technologie internet. A ce titre signalons que plusieurs normes d'encodage sont apparues qui permettent dans leur majorité

---

<sup>1</sup> Les formats MARC sont des formats d'échange de données bibliographiques permettant d'informatiser les catalogues de bibliothèques.

de décrire les manuscrits, avec la détermination d'un ensemble d'éléments d'accès aux ressources numérisées, appelées les métadonnées.

### ***3. Qu'est-ce qu'une métadonnée ?***

Les métadonnées sont généralement définies comme « données sur les données » ou encore « information sur les données ». Elle se présente comme une liste structurée d'informations qui décrivent les données ou les services stockées dans les systèmes d'information. Ces métadonnées peuvent contenir une brève description sur le contenu, les objectifs, la qualité et la localisation de la donnée ainsi que les informations relatives à sa création [Ontario 2009].

### ***4. Utilités des métadonnées***

L'utilité des métadonnées se résume comme suit :

1. La description de tous les éléments associés à une ressource comme : les droits, formats, dates, etc.
2. Facilite l'organisation de l'information.
3. L'amélioration de la pertinence et de l'exhaustivité des recherches, le tri et le filtrage des résultats [Morel Pair 2005]
4. Edition, gestion, recherche, réutilisation, diffusion et publication de multiples contenus (textes, images, vidéos).
5. Facilite le repérage de l'information, recherche et localisation des documents.

### ***5. Les métadonnées et la description des manuscrits arabes***

Les projets de numérisation nécessitent la création des métadonnées comme moyen d'accès spécifique à ces manuscrits. Le but principal des métadonnées est de « matérialiser en quelque sorte le lien entre le document et la collection » [Soualah 2008].

Cette matérialisation dépend des besoins particuliers des lecteurs. Qui peuvent être des caractéristiques descriptives du manuscrit : titre, auteur, etc.

Les métadonnées doivent satisfaire quatre besoins : les catalogueurs, les utilisateurs, les experts techniques et les administrateurs. Pour cela, il faut qu'elles contiennent des informations créées par l'ensemble des acteurs mentionnés ci-dessous [Keilah 2004]:

- ✚ Premièrement, il faut donner aux catalogueurs la possibilité de saisir toutes les informations dont ils ont besoin pour l'intégration d'un nouveau document dans la collection. Les métadonnées doivent aussi concerner non seulement le document lui-même, mais aussi ses composantes graphiques ou textuelles en vue d'une recherche par le contenu.
- ✚ Deuxièmement, il faut donner aux utilisateurs de multiples points d'accès à la collection, afin de permettre non seulement l'identification des documents pertinents mais aussi une recherche fine sur le contenu.

- ✚ Troisièmement, les experts techniques doivent trouver toutes les informations dont ils ont besoin concernant le fichier numérique, ce qui les aidera dans la transformation et la préservation des informations dans le futur.
- ✚ Quatrièmement, les administrateurs ont besoin des métadonnées de gestion. Cela pour évaluer l'activité de la collection en termes de coût d'acquisition de documents, de saisie, d'accès, de satisfaction des usagers et pour permettre de faire des bilans périodiques.

## 6. Les éléments descriptifs des manuscrits arabes

Plusieurs protocoles de description des manuscrits arabes ont été proposés, parmi lesquels on distingue [Soualah 2008] :

### 6.1. Protocoles défini par la fondation Al-Saoud

C'est un protocole de description des manuscrits arabes issu du congrès tenu au Maroc par la fondation El-Saoud du 05 au 07 juin 1989. Ce protocole propose un ensemble de sept éléments appelés champs, ainsi chaque champ est formé d'un groupe de sous-éléments le décrivant. Ces sept champs sont :

- ✚ Identification du manuscrit (حقل التعريف)
- ✚ Contenu (حقل المضمون)
- ✚ Description structurelle (حقل الوصف البادي للمخطوط)
- ✚ Description physique (حقل الوصف المادي للمخطوط)
- ✚ Additifs (حقل الإضافات)
- ✚ Notices supplémentaires et traduction (حقل البيانات الإضافية و بيانات الترجمة)
- ✚ Observations (حقل الملاحظات المتعلقة بالحقول السابقة)

### 6.2. Protocole défini par l'IRHT

C'est un protocole qui contient une série d'éléments descriptifs des manuscrits, structurées en plusieurs niveaux hiérarchiques et proposées par l'IRHT (Institut de Recherche en Histoire et Texte). Ce protocole est constitué des éléments suivant :

- ✚ Signalement du manuscrit.
- ✚ Support.
- ✚ Encre.
- ✚ Etat du manuscrit.
- ✚ Organisation du volume.
- ✚ Ecriture.
- ✚ Mise en page.
- ✚ Décoration.
- ✚ Reliure.
- ✚ Héraldique.
- ✚ Histoire.
- ✚ Contenu.

Les deux protocoles El-Saoud et IRHT traitent les manuscrits dans ces divers aspects codicologie, paléographie et histoire.

Par ailleurs, ils proposent deux approches différentes de catalogage tel que : l'IRHT propose le mode de catalogage par spécimen (à l'exemplaire), ou le manuscrit est considéré comme un travail indivisible tout en signalant s'il est hétérogène ou pas, puis procède à la description de chaque volume tandis que la fonction El-Saoud propose le mode de catalogage par volume.

En conclusion il est impératif de combiner les deux protocoles précédents afin d'en proposer un nouveau protocole, résultat de l'union des deux ensembles de métadonnées proposées par les deux institutions.

## ***7. Choix des métadonnées descriptives des manuscrits arabes numérisés***

L'accès aux manuscrits numérisés se fait grâce aux notices bibliographiques, Pour cela, cette dernière doit contenir une description complète du manuscrit à travers un ensemble d'éléments descriptifs nécessaire pour des recherches spécifiques selon des critères bien précis, ces métadonnées découlent du protocole enrichi obtenu à partir de la combinaison des deux protocoles précédents. Mais la bonne rédaction de la notice ne suffit pas seule pour avoir un accès aux manuscrits, ce qui implique l'intégration d'une adresse URI (Uniform Resource Identifier), et d'un autre ensemble de métadonnées qui ont comme intérêt l'enrichissement des méthodes d'accès à ces manuscrits.

A cet effet, les métadonnées descriptives des manuscrits peuvent être structurées en deux ensembles différents, ainsi deux catalogues différents, le premier sert à fournir une description des manuscrits alors que, le deuxième sert à augmenter les points d'accès aux images de manuscrit par les possibilités de navigation dans la base de données d'image :

### ***7.1. Les catalogues bibliographiques de manuscrit***

C'est le type de catalogues traditionnels, ou chaque catalogue est composé d'un ensemble de fiches classées par ordre alphabétique, connue sous le nom de notices bibliographiques. L'accès à ce catalogue se fait en recherchant le titre, l'auteur, le sujet ou tout autre champ descriptif. L'utilisation de ce type de catalogage traditionnel conduit l'utilisateur à s'intéresser beaucoup plus à la description du manuscrit selon un protocole définis.

Le meilleur protocole utilisé est celui issu des deux protocoles présentés respectivement par la fondation Al-Saoud ainsi que l'IRHT étudiés auparavant. Un protocole de ce genre a été déjà proposé par [Soualah & Hassoun 2011], ce dernier contient les métadonnées suivantes:

### 1. *Identification du manuscrit*

- ✚ Numéro d'identification.
- ✚ Numéro du volume : si le manuscrit se compose de volume.
- ✚ Nombre de copies microfilmées: si le manuscrit a été filmé.
- ✚ Foliation ou pagination : elle doit indiquer le type de numérotation utilisé.

### 2. *Table des matières du manuscrit*

- ✚ Le titre du manuscrit : un manuscrit peut avoir plusieurs titres :le titre des auteurs, le titre des catalogueurs, le titre des pointes à tracer, etc.).Parfois, cette information est manquée, ainsi elle doit être fournie par le catalogueur.
- ✚ Les noms des auteurs : Prénom, nom de famille, surnom, nom de célébrité qui sont équivalent à *ism*, *laqab*, *kunya*, *nisba*.
- ✚ Date de décès : elle doit être citée dans le calendrier d'hégire et dans son équivalence en calendrier grégorien.
- ✚ Date de naissance de l'auteur.
- ✚ Période d'existence.
- ✚ Sujet du manuscrit.
- ✚ Incipit : premières expressions de manuscrit.
- ✚ Explicit : dernières expressions de manuscrit.
- ✚ Colophon.
- ✚ Table des matières.
- ✚ Index.
- ✚ sommaire du manuscrit.

### 3. *Description physique du manuscrit*

Matériel de fabrication, nombre de folios, nombre de lignes par la page, Palimpseste ou non, signature, état du manuscrit (qualité), taille de la couverture, état de conservation, taille des folios, le modèle d'écriture, schémas, illumination, encre (couleur), relieur, folios détachées de la copie.

### 4. *Additifs*

- ✚ Certificat de lecture: قراءات
- ✚ Certificat d'audition : سماعات
- ✚ Certificat de transmission : اجازات
- ✚ Corrections : التصحيحات

### 5. *Références de copies*

- ✚ Les copies du manuscrit disponible dans d'autres bibliothèques.
- ✚ Copies imprimées.
- ✚ L'imprimeur.
- ✚ Références de publication.
- ✚ Références de traduction.
- ✚ Sources.




 Observation.

#### 6. *Histoire de manuscrit*

Possession de manuscrit (les noms des personnes) et type d'acquisition (achat, prêt et donation).

### 7.2. *Catalogue de manuscrit utilisé pour l'édition en ligne*

Des informations qui soutient la recherche documentaire des textes sont ajoutées aux notices bibliographiques, dans le but de les enrichir et aussi d'avoir un accès aux images :

-  La table des matières : chaque ligne contient un lien se dirigeant à l'image correspondante de manuscrit.
-  Annotations: L'information rapportée par le chercheur sur l'image peut être ajoutée au catalogue afin d'améliorer l'accès de manuscrit
-  L'index : Les index sont des ensembles de termes ou de mots clés, se rapportant à la description d'un document, et classés de façon ordonnée [Index 2012].

Les index sont destinés par le catalogueur et servent à localiser précisément des concepts, des lieux ou des noms particuliers qui ont été cités dans le manuscrit.

### *III. Indexation des données :*

L'indexation est une étape majeure dans la réalisation des systèmes de recherche d'information (SRI). Ce dernier, est un système informatique qui permet de retourner à partir d'un ensemble de documents, ceux dont le contenu correspond le mieux à un besoin en informations d'un utilisateur, exprimé à l'aide d'une requête [Zemerli 2004].

Le processus de recherche d'information se résume essentiellement en trois étapes, qui sont représentées schématiquement à travers le processus en "U" ci-dessous:

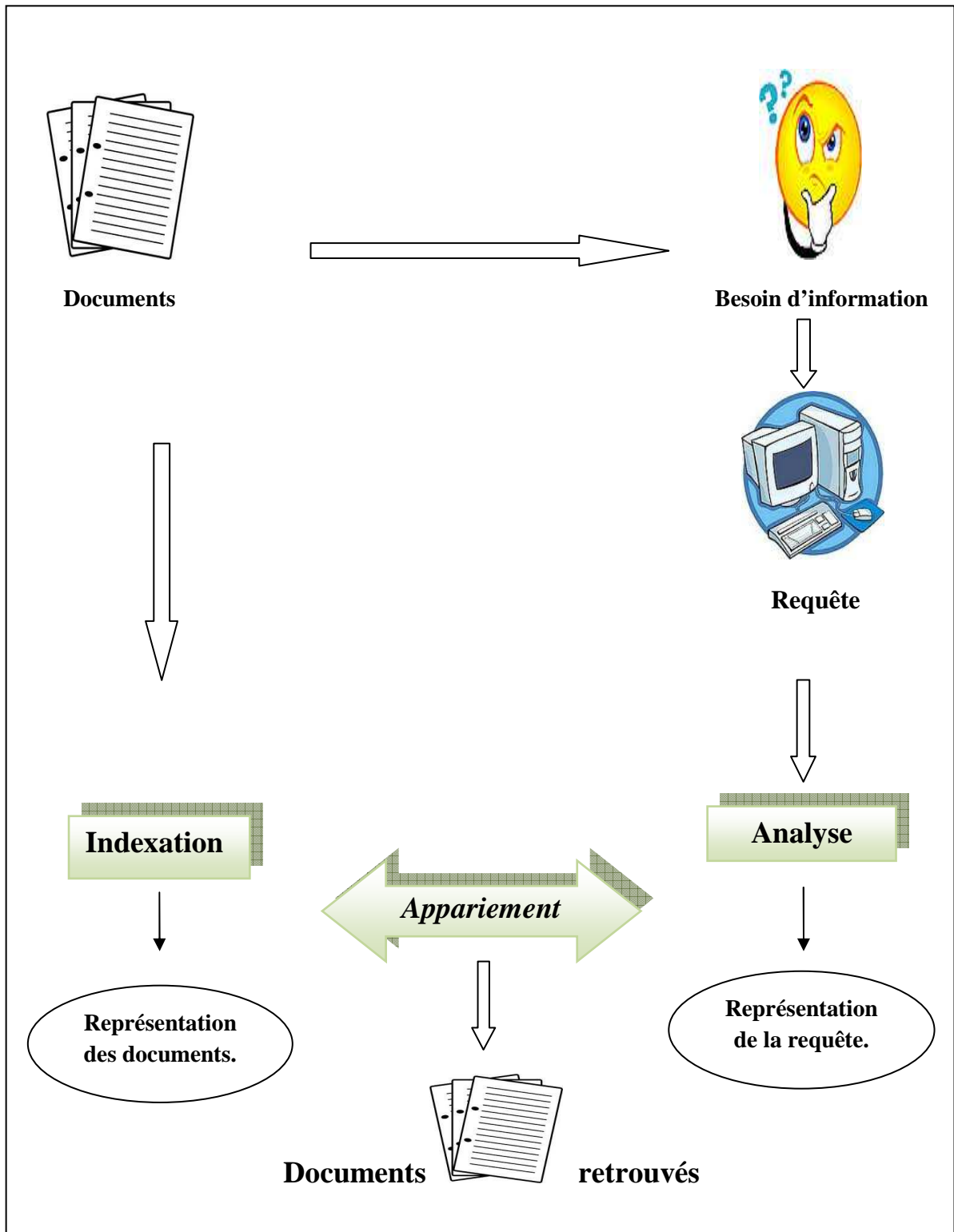


Figure 05 : Schéma du processus de recherche d'information.

A partir de ce schéma on distingue les étapes suivantes:

- a) **Analyse** : Cette étape a pour objectif la transformation de la requête formulée par l'utilisateur, en une représentation interne équivalente à celle des documents.
- b) **Appariement** : Elle consiste alors à mettre en correspondance et à calculer le degré d'appariement des représentations internes des documents et de la requête.
- c) **Indexation** : Elle consiste à déterminer et extraire les termes représentatifs du contenu d'un document ou d'une requête, qui couvrent au mieux leur contenu sémantique. La qualité de la recherche dépend en grande partie de la qualité de l'indexation.

**1. Modes d'indexation** : Il existe deux types d'indexation :

### **1.1. Indexation manuelle**

Elle est réalisée par un expert humain (documentaliste). Après la lecture et l'étude complète du corpus documentaire, le documentaliste, tout en se basant sur ses connaissances dans le domaine en question et en utilisant de méthodes et techniques bien connus, choisit les descripteurs qui couvrent au mieux la collection. Cette technique se caractérise par sa lenteur et la subjectivité de ses résultats. De plus, elle est inadaptée lorsque le volume ou le nombre de documents à indexer devient important.

### **1.2. Indexation automatique**

Elle vient pour combler les lacunes de l'indexation manuelle en termes de temps et d'efficacité des résultats, par l'automatisation des étapes de celle-ci. Plusieurs méthodes peuvent être utilisées pour la détermination des termes d'indexation dont la plus simple se base sur une analyse statistique du texte (elle considère que la présence ou l'absence d'un terme est significative pour le sujet d'un document).

En ce qui concerne l'indexation des manuscrits arabes le catalogage consiste à mettre en place trois instruments de recherche : les notices bibliographiques des manuscrits, la table des matières de chaque manuscrit et l'index associé, qui sont constituées en format XML.

En effet, les informations à indexer sont contenues dans les notices bibliographiques, les tables de matières et les tables d'index. L'indexation peut se faire par champs (Titre, auteurs, sujets,...etc.).

A partir de là, on déduit l'utilité du format XML dans le processus de catalogage des manuscrits arabes et tous les avantages qu'il fournit afin de mener un bon travail de catalogage.

#### ***IV. Conclusion***

Le catalogue est destiné à identifier les documents et à faciliter leur recherche ou leur localisation. Les manuscrits arabes anciens sont des documents certes composites, complexes et instables mais ne se différencient pas des autres ouvrages et ils doivent obéir au terme de catalogage.

Dans cette partie, nous avons présentés des notions sur le catalogage des manuscrits arabes, les différents types qui existaient avant et après l'apparition de l'outil informatique, sur les protocoles de métadonnées disponibles et destinés pour le catalogage de ces manuscrits, ainsi l'accès aux manuscrits arabes numérisés en tenant compte des processus de recherche d'information et d'indexation.

## ***Introduction***

De multiples systèmes d'encodages du catalogue manuscrits sont utilisés, à titre d'exemple nous pouvons citer la TEI (Text Encoding Initiative), l'EAD (Encoded Archival Description), l'UNIMARC (Universal Machine Readable Cataloging) et bien d'autres. Chaque système présente diverses particularités de structuration des données.

L'interrogation simultanée des différents catalogues sous divers formats d'encodage rend le système d'interrogation complexe, nécessitant la prise en compte de la structure logique de chaque format. Il serait donc intéressant de rendre invariants les requêtes, offrant une meilleure souplesse pour les utilisateurs désirant interroger tous les catalogues à la fois. Il s'agit de ce fait, de mettre en place une plateforme unique assurant l'interopérabilité entre les différents systèmes de catalogage.

Dans ce chapitre, nous allons définir le concept d'interopérabilité et ses différents principes.

### ***I. Qu'est que c'est l'interopérabilité***

L'interopérabilité est la capacité technique que plusieurs systèmes, qu'ils soient identiques ou radicalement différents, puissent communiquer et opérer ensemble sans ambiguïté, sans conflit de systèmes ou de contenu [Staccini 2007].

Dans [Interopérabilité 2012] l'interopérabilité en informatique est définie comme la capacité des systèmes informatiques (machines, systèmes, applications...) à échanger des informations et à travailler ensemble. Elle concerne:

- ✓ L'échange d'informations pour des applications spécifiques.
- ✓ L'échange d'informations à l'aide de protocoles adaptés (XML...).
- ✓ La communication entre systèmes hétérogènes à l'aide de protocoles communs (TCP/IP...).
- ✓ L'invocation de services (services Web...).

L'interopérabilité informatique consiste à s'assurer que des systèmes et leurs composants (des applications et des plateformes) peuvent se parler de façon à maximiser les opportunités d'échange et de réutilisation de l'information. Dans un monde de plus en plus numérique, c'est un objectif important, auquel les métadonnées participent en complémentarité avec d'autres processus.

L'interopérabilité est une notion absolument cruciale pour le réseau téléphonique mondial et Internet. Par essence, des matériels divers et variés sont mis en œuvre dans ces réseaux hétérogènes aux côtés d'une panoplie encore plus vaste de matériels informatiques et de logiciels. Elle est aussi cruciale pour l'ensemble de l'économie, car dans presque tous les domaines d'activité dans l'industrie, dans les services, et dans l'agriculture, on emploie des systèmes informatiques, qui aujourd'hui communiquent d'une entreprise à l'autre par l'intermédiaire de réseaux informatiques (Internet, extranet, messageries électroniques). L'interopérabilité est une nécessité dans le domaine du commerce électronique.

En effet, le demandeur de services et le fournisseur de services doivent pouvoir interfonctionner en employant des structures et des types de données communs.

## ***II. Les principes d'interopérabilité***

L'interopérabilité se base sur deux principes [Wiki 02 2012] :

### **◆ *Nécessité des normes***

L'interopérabilité nécessite que les communications obéissent à des normes<sup>2</sup>, clairement établies et univoques. Ces documents techniques définissent souvent des exigences, parfois accompagnées de recommandations plus ou moins optionnelles. Si la norme est correctement écrite, deux systèmes qui satisfont aux exigences doivent dialoguer ensemble sans souci particulier. Ils peuvent ainsi évoluer librement sans risque de casser cette possibilité de communication, tant qu'ils respectent la norme définissant leurs interfaces.

### **◆ *Les données véhiculées dans des interfaces***

En pratique, l'interopérabilité touche tous les domaines de l'informatique. Ce sont les règles de cohérence des données véhiculées qui gouvernent l'interopérabilité. Ainsi, Les données de référence<sup>3</sup> employées par plusieurs applications sont généralement celles qui pilotent l'interopérabilité. Dans des contextes où coexistent les données structurées (les bases de données) et les données non structurées (les documents, textes, images), on considère généralement aujourd'hui que les données communes sont constituées par des métadonnées .

Par conséquent, Les interfaces de programmation (API) sont à la base de l'interopérabilité informatique, Ces API peuvent s'appliquer à différents types de ressources informatiques (bases de données) ou applications.

## ***III. Les approches d'interopérabilité***

### ***1. Les normes ouvertes et standards communautaires*** [Wiki 02 2012]

Il existe des normes publiées, parfois d'abord à l'état de brouillon, dont les essais de mise en œuvre permettront d'en trouver les failles et d'en corriger les défauts, puis de candidat à la publication et enfin de recommandation officielle ou de document d'information.

Cette publication est ouverte, chacun à la possibilité d'étudier ces documents et de tenter de développer un système conforme à ces standards. De plus, le fait que la rédaction soit

---

<sup>2</sup>Une norme désigne un état habituellement répandu ou moyen considéré le plus souvent comme une règle à suivre.

<sup>3</sup> Les données de référence, aussi appelée données maîtres ou master data, sont des données non transactionnelles (autrement dit non directement liées aux opérations) évoluant peu avec le temps, critiques pour les processus opérationnels et utilisées pour structurer les données transactionnelles dans les systèmes informatiques.

relativement ouverte à la communauté évite de voir des normes publiées qui ne satisfont qu'une minorité qui détiendrait un pouvoir de décision sur leur contenu. Exemples de standards ouverts<sup>4</sup> : XML, XHTML...etc....

## **2. Les formats fermés et propriétaires [Wiki 02 2012]**

À l'opposé, le frein majeur à une interopérabilité optimale est l'utilisation dans des matériels et logiciels de formats dont seuls leurs concepteurs auraient les clefs. Cette fermeture est souvent volontaire car elle vise, dans le cas d'un format de fichier propriétaire, à s'assurer qu'un utilisateur n'utilisera pas un autre logiciel pour lire ses données. Exemples de formats fermés et propriétaires : Les messageries instantanées propriétaires comme Yahoo! Messenger ou MSN Messenger dont les protocoles ne sont pas compatibles et maintenus non interopérables.

## **IV. Les niveaux d'interopérabilité**

On distingue 3 niveaux d'interopérabilité :

### **1. L'interopérabilité technique**

L'interopérabilité technique désigne le recours à la définition et l'utilisation d'interfaces technologiques, des normes et des protocoles, en vue de créer des systèmes d'information collaboratifs fiables, efficaces et performants capables d'échanger l'information [Yahia 2011].

Sur un plan technique, l'interopérabilité se réalise à trois niveaux techniques complémentaires [Morel-pair 2005]:

1. Une description des ressources avec des sémantiques communes issues de différents jeux de métadonnées standardisés,
2. Un contexte générique d'implémentation de ces descriptions dans des langages structurés standardisés, interprétables par les machines,
3. Des protocoles informatiques d'échange de ces données normalisées.

Le tableau suivant résume le positionnement des niveaux techniques d'interopérabilité vu précédemment sous des différents standards traditionnels et récents :

---

<sup>4</sup>Un standard ouvert est tout protocole de communication, d'interconnexion ou d'échange et tout format de données interopérable et dont les spécifications techniques sont publiques et sans restriction d'accès ni de mise en œuvre.

	Standards traditionnels	Standards récents
Jeux de métadonnées	MARC	Dublin Core MARC-XML, MODS EAD LOM...
Cadre générique d'implémentation	ISO 2709 ISAD(G)	XML RDF espaces de nom URL
Protocoles	WAIS FTP Z39.50	http OAI-PMH SRU/SRW

## 2. L'interopérabilité sémantique

L'interopérabilité sémantique consiste à faire en sorte que la signification des informations échangées ne soit pas perdue dans la procédure d'interopération et qu'elle soit préservée et comprise par les personnes, les applications et les institutions concernées (les formats des messages, la structuration, la sémantique des constituants). Elle nécessite que l'interopérabilité technique soit effective [Yahia 2011].

### 2.1. Les principes d'interopérabilité sémantique

L'interopérabilité sémantique se base sur quatre principes nécessaires qui sont [Staccini 2007]:

#### a. Métadonnées

Les métadonnées comme on les a définies auparavant, se sont des informations descriptives sur les ressources. L'utilisation de métadonnées descriptives et standardisées améliore la recherche des informations pertinentes dans un réseau de ressources.

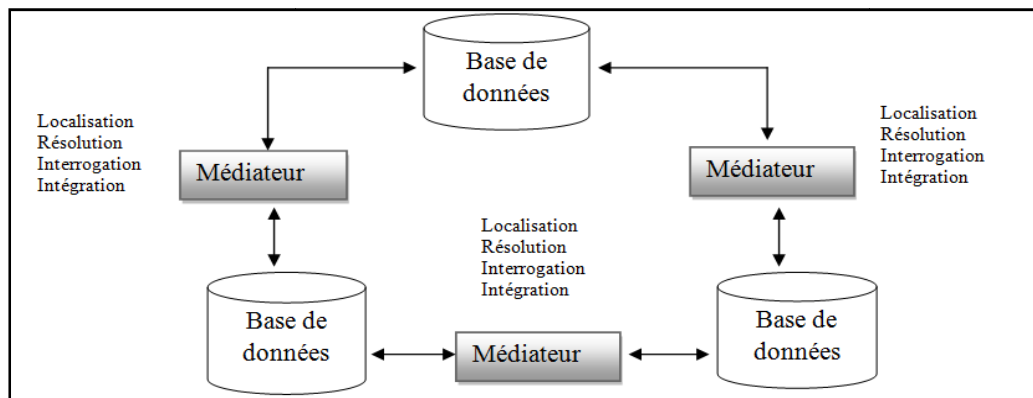
#### b. Ontologie

L'ontologie est une description formelle des concepts, des rôles et des relations qui existent pour un agent ou une communauté d'agents. Elles fournissent une compréhension commune d'un domaine qui peut être communiquée, tout en jouant un rôle majeur dans les échanges d'information.

#### c. Médiateurs

Un médiateur est un adaptateur de données situées sur un réseau entre un client et un serveur de données (le client peut être une autre base de données). Par ailleurs, un médiateur est un composant logiciel qui résout les conflits schématiques<sup>5</sup> et sémantiques<sup>6</sup>.

<sup>5</sup>Les conflits schématiques résultent d'une structuration et d'une classification différente des informations. Ils sont étroitement liés aux choix de conception.



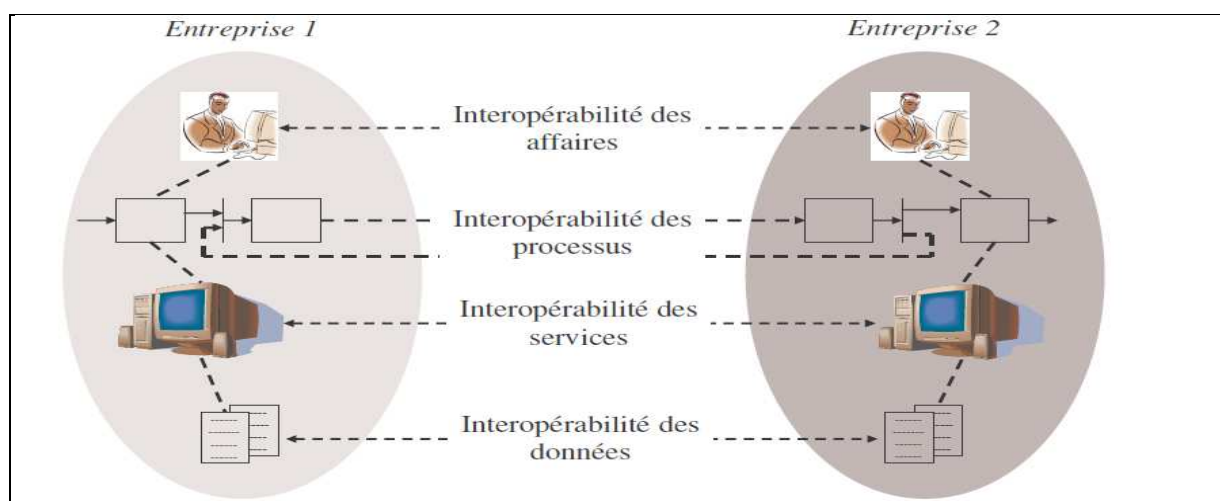
**Figure 06 :** Principes d'interopérabilité sémantique [Staccini 2007].

### 3. L'interopérabilité organisationnelle

L'interopérabilité organisationnelle est la capacité d'identifier les acteurs et les procédures organisationnelles intervenant dans la fourniture d'un service spécifique et de parvenir à un accord entre ces acteurs et à des procédures sur la manière de structurer leur interaction. En d'autres termes, il s'agit de définir les « interfaces d'entreprise » et de s'intéresser aux rôles des entités et des acteurs en interaction avec les systèmes d'information. Elle nécessite que l'interopérabilité sémantique soit effective [Yahia 2011].

#### 3.1. Les niveaux d'interopérabilité organisationnelle

L'interopérabilité dans l'entreprise peut se percevoir au travers de quatre niveaux : l'interopérabilité des données, l'interopérabilité des services, l'interopérabilité des processus et l'interopérabilité des affaires. Ces différents niveaux sont illustrés par la figure suivante [Daclin 2008] :



**Figure 07 :** Les quatre niveaux d'interopérabilité.

<sup>6</sup>Les conflits sémantiques proviennent des différences d'interprétation des informations partagées entre différents domaines d'application.

A partir de ce schéma on distingue :

- a. L'interopérabilité des données :** Elle se réfère aux différents modèles de données de travail collaboratif et à l'organisation de schémas conceptuels.
- b. L'interopérabilité des services :** Elle permet d'identifier, de composer et de créer des fonctions communes à diverses applications (conçues et mises en application indépendamment) en résolvant les différences syntaxiques et sémantiques ainsi qu'en trouvant les points de raccordements des diverses bases de données hétérogènes.
- c. L'interopérabilité des processus :** Elle vise à faire travailler ensemble divers processus selon un besoin spécifique de l'entreprise et notamment à relier des processus internes de deux entreprises pour créer un processus commun.
- d. L'interopérabilité des affaires :** Elle se rapporte à l'harmonisation de l'organisation globale des entreprises malgré les différents modes de prise de décision, méthodes de travail, culture d'entreprise, etc. de sorte qu'une activité conjointe puisse être développée.

## **V. Conclusion**

A travers ce chapitre nous avons présentés les concepts généraux d'interopérabilité y compris ses principes, ses approches, ainsi que ses différents niveaux, tel que l'interopérabilité technique, l'interopérabilité sémantique et l'interopérabilité organisationnelle. En effet, le choix d'un niveau d'interopérabilité dans notre cas, nécessite une étude des différents formats de codage des documents. Pour ce faire, le chapitre suivant traitera les principes et les caractéristiques des formats : EAD, TEI et Dublin Core.

## **Chapitre 2 : Encodage du catalogue**

**Sous-chapitre 1: La description de la DTD EAD**

**Sous-chapitre 2: La description de la DTD TEI**

**Sous-chapitre 3: La description de DC**

## ***Introduction***

L'internet a sans doute apporté à notre métier les évolutions indispensables permettant notamment la publication, la diffusion aux publics et l'échange entre professionnels des éléments de description de nos fonds d'archives. Dans ce domaine, l'impulsion est venue des États-Unis avec la mise en œuvre de standards d'encodage de des éléments de description, permettant leur gestion et leur lecture. Parmi ces standards nous allons présenter l'EAD.

### ***I. Historique*** [MICAELLI 2012]

Le développement de l'EAD a commencé en 1993 à l'Université de Californie à Berkeley. L'objectif du projet était de créer un standard d'encodage des descriptions de documents d'archives qui puisse être interprété par un ordinateur, permette une grande facilité d'interrogation et contienne des informations beaucoup plus riches que les notices traditionnelles.

En juillet 1995, un groupe d'experts américains a été créé pour continuer le travail commencé à Berkeley. Peu après, la Société des archivistes américains ont décidé de prendre part au développement de la DTD<sup>7</sup>, et en janvier 1996 la Bibliothèque du Congrès s'est engagée à assurer la maintenance informatique et la diffusion de l'information sur la DTD. Développée initialement en SGML, la DTD EAD a été rendue conforme à l'XML afin de permettre une meilleure diffusion sur le web. La version EAD 1.0 a été publiée à la fin d'août 1998 et une seconde version, EAD 2002, en décembre 2002.

### ***II. Définition de l'EAD***

L'EAD est un standard d'encodage des instruments de recherche<sup>8</sup> de type archivistique (inventaires, répertoires, catalogues de collection) qui permet de créer, de publier, et de conserver ces instruments de recherche en version électronique [Starck 2008]. Basé sur un langage de balisage numérique XML (eXtensible Markup Language) qui facilite l'échange automatisé de contenus entre systèmes d'informations.

#### **Pourquoi DTD EAD ?**

L'EAD est la DTD utilisée pour décrire des fonds d'archives, des collections de manuscrits et plus généralement tout ensemble hiérarchisé de documents.

---

<sup>7</sup>La DTD (Document Type Définition)est un standard qui définit l'ensemble des règles et des propriétés que doit suivre un document édité au format XML ou SGML pour qu'il soit valide.

<sup>8</sup>Un instrument de recherche est un ouvrage, un fichier ou une base de données qui décrit le contenu d'un ensemble de documents conservés par un service d'archives.

### **III. Caractéristique de l'EAD**

La DTD EAD a été créée pour structurer des inventaires d'archives, ou tout autre outil de recherche relatif aux archives.

On structure des données à l'aide de balises qui ont pour rôle :

- a) Faire apparaître les relations hiérarchiques entre les différentes parties de l'inventaire.
- b) Préciser le contenu de l'information.

L'opération qui consiste à placer les balises dans un texte est l'encodage. Chaque balise porte un nom (celui de son élément). A chaque élément, on peut joindre des attributs qui ont diverses fonctions.

L'EAD présente peu d'éléments obligatoires. L'instrument de recherche peut être encodé à volonté de l'utilisateur. En effet, on peut se contenter d'encoder la structure, cet aspect permet une lecture linéaire. Toutefois, on peut encoder les noms de personnes, de lieux, des dates, ...etc. cet autre aspect permet une lecture de recherche en utilisant un index particulier.

### **IV. Avantages du format EAD**

L'EAD est basée sur les principes suivants :

- a) Une hiérarchie permettant d'aller du général au particulier.
- b) La mise en évidence d'une description à plusieurs niveaux.
- c) Une information appropriée à chaque niveau de description.
- d) La liaison de chaque description à celle du niveau supérieur.
- e) Ce standard est indépendant des plates-formes logicielles puisqu'il est développé sur les langages SGML et XML

### **V. Structure de l'EAD**

#### **1. Les attributs de l'EAD**

Des attributs sont associés à la plupart des éléments contenus dans l'EAD pour les qualifier. Les DTD permettent de déclarer des attributs et leurs valeurs, et de les associer aux éléments. Les attributs peuvent prendre différents types de valeurs : ces valeurs peuvent être du texte ou des chaînes de caractères alphabétiques ou numériques.

Chaque attribut est constitué par :

- **Un nom :** Qui correspond à une propriété d'un élément.

- **Une valeur** : Selon le contexte dans lequel il apparaît.

### Exemple

```
< [balise] [attribut]="[valeur]">
```

La plupart des attributs sont facultatifs à l'exception de 2 qui sont obligatoires, l'attribut **LEVEL** et l'attribut **COLS**

- ✚ L'attribut **LEVEL** : De l'élément description archivistique <archdesc> et de l'élément groupe de description archivistique <archdescgrp>, il permet de définir le niveau de description.

Exemple : **LEVEL** = collection, série, class...etc.

- ✚ L'attribut **COLS** : De l'élément groupe d'éléments de tableau <tgroup>, il donne le nombre de colonnes dans un tableau.

Exemple : <tgroup **COLS** ="4">

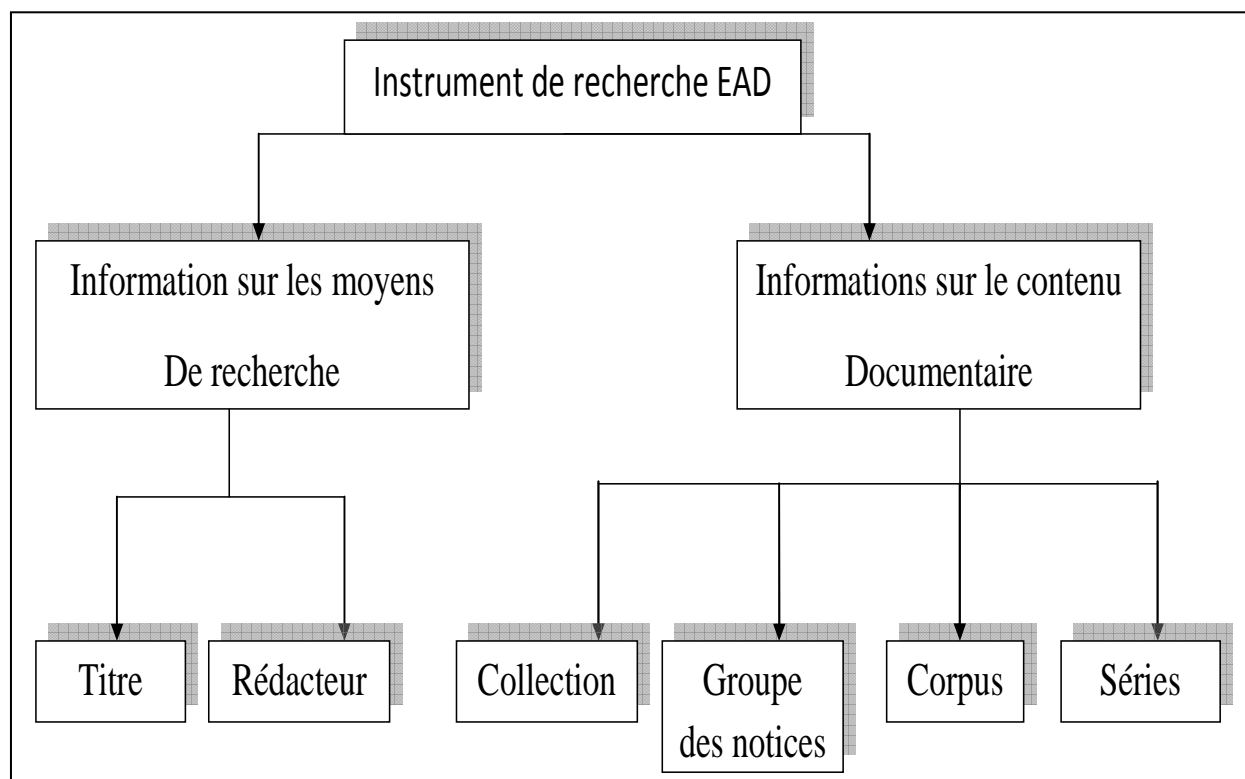
## 2. Les éléments de l'EAD

Un instrument de recherche encodé en EAD comprend deux segments importants ; un en-tête et une description archivistique.

Le *premier segment* fournit les informations sur l'instrument de recherche (exemple : titre, rédacteur, date de rédaction, etc.).

Le *deuxième segment* <archdesc> englobe toutes les informations décrivant un corpus documentaire dans son ensemble et recouvre les différents niveaux hiérarchiques des informations concernant le fonds.

La figure suivante récapitule la structure d'un instrument de recherche EAD.



**Figure 08 :** Schéma d'un instrument de recherche EAD

L'EAD comprend 146 éléments, dont 8 sont obligatoires pour avoir un document XML valide (conforme à la DTD).

Trois éléments principaux s'inscrivent sous un élément racine <ead> :

- **<eadheader>** : En-tête EAD, qui contient des informations bibliographiques et descriptives sur l'instrument de recherche.
- **<frontmatter>** : Préliminaires, qui sert à créer une première de couverture de l'instrument de recherche.
- **<archdesc>** : Description de l'unité documentaire.

La figure suivante représente les huit (08) éléments obligatoires et leurs hiérarchies dans l'EAD.

```
<ead>
<eadheader>En-tête EAD
<eadid>Identifiant de l'IR</eadid>
<filedesc>Description de l'IR
<titlestmt>Mention de titre de l'IR
<titleproper>Titre propre de l'IR</titleproper>
</titlestmt>
</filedesc>
</eadheader>
<archdesc>Description archivistique de l'Unité de description
<did>Identification et description de l'Unité de description</did>
</archdesc>
</ead>
```

Figure09 : Les éléments obligatoires de l'EAD.

Notre intérêt d'étudier de la DTD EAD est d'apprécier à quel point cette dernière apporterait une réponse exhaustive à la description des manuscrits arabes eux-mêmes et non à la description des instruments de recherches, pour cela on s'intéresse seulement au troisième et dernier segment <archdesc>.

### L'élément <archdesc>

Un élément *obligatoire*, englobe toutes les informations décrivant un ensemble documentaire et recouvre les différents niveaux hiérarchiques de cette description. Il contient les 29 sous-éléments suivants:

- |                     |                   |                 |
|---------------------|-------------------|-----------------|
| 1.did               | 11.altformavail   | 21.odd          |
| 2.scopecontent      | 12.accruals       | 22.originalsloc |
| 3.bioghist          | 13.accessrestrict | 23.phystech     |
| 4.custodhist        | 14.userrestrict   | 24.prefercite   |
| 5.bibliography      | 15.controlaccess  | 25.processinfo  |
| 6.separatedmaterial | 16.appraisal      | 26.runner       |
| 7.relatedmaterial   | 17.arrangement    | 27.note         |
| 8.otherfindaid      | 18.descgrp        | 28.dao          |
| 9.dsc               | 19.fileplan       | 29.daogrp       |
| 10.acqinfo          | 20.Index          |                 |

1. **<did>** : Identification et description, élément *obligatoire* qui contient au moins un élément (il est disponible a chaque niveau). Cet élément doit permettre à l'utilisateur de déterminer si les documents concernés sont utiles pour sa recherche. Il constitué de:

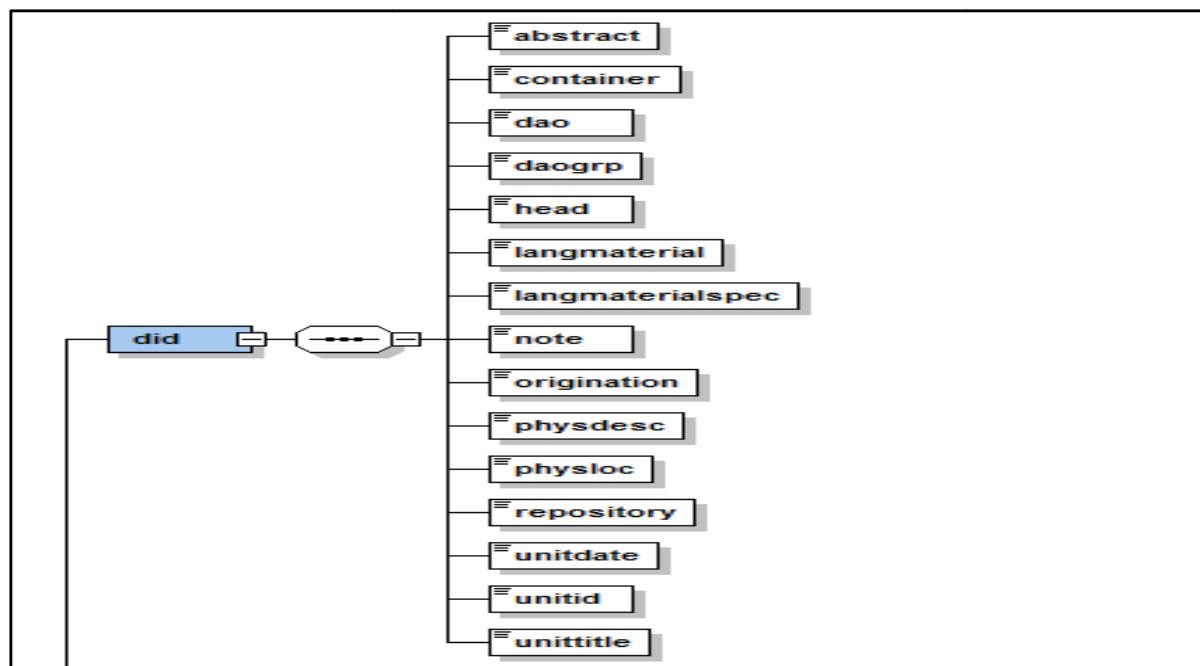


Figure10 : Les sous-éléments de l'élément <did>.

1.1 **<abstract>** : Résumé (brèves description).

1.2 **<container>** : Contenant les informations qui contribuent à localiser les unités documentaires décrites en indiquant le type de conditionnement des documents et en identifiant tout numéro séquentiel assigné à ces conditionnements.

1.3 **<dao>** : Élément de lien qui utilise les attributs ENTITYREF ou HREF pour lier les informations contenues dans l'instrument de recherche aux représentations électroniques des unités documentaires décrites.

1.4 **<daogrp>** : Élément englobant contenant au moins deux éléments Emplacement d'objet archivistique numérique <daoloc> qui peuvent être considérés comme un groupe.

1.5 **<head>** : Pour donner un titre ou l'intitulé d'une partie de texte, doit être placé en premier, avant les autres éléments.

**1.6** *<langmaterial>* : Langue des unités documentaires, pour donner les langues principales des documents décrits avec l'attribut LANGCODE.

**1.7** *<materialspec>* : Élément contenant des données spécifiques à un type ou à une forme de documents, pour lesquelles aucun autre élément de description ne convient.

**1.8** *<note>* : Compléter des informations dans <did> et non des indications sur le contenu. Il peut utiliser l'élément <p> pour les paragraphes.

**1.9** *<origination>* : Désigne la personne morale ou physique qui a produit, rassemblé ou constitué l'ensemble documentaire. Il peut inclure les sous-éléments suivants :

- *<persname>* : Nom de personne.
- *<famname>* : Nom de famille.
- *<corpname>* : Nom propre identifiant un organisme ou un groupe de personnes agissant comme une entité : par exemple, les noms d'associations, d'institutions...etc.
- *<name>* : Élément contenant un nom propre ou une expression nominale identifiant une entité qu'il est difficile de baliser avec les éléments présentés précédemment.

**1.10** *<physdesc>* : Description physique de l'ensemble. Il peut contenir les sous-éléments suivants:

- *<extent>* : Élément fournissant des informations sur le nombre de documents décrits ou sur l'espace qu'ils occupent. Il inclut les mesures telles que cubage ou métrage linéaire, nombre de feuillets, de volumes, de bobines de microfilms, de photographies ou d'autres supports spécifiques.
- *<dimensions>* : Élément fournissant des informations sur la dimension des documents en cours de description.
- *<physfacet>* : Particularité physique, c'est un élément répétable, on utilise l'attribut "type" pour distinguer les différentes particularités physiques, techniques, support, reliure.
- *<genreform>* : Élément qui précise le style ou la technique de création du contenu intellectuel (genre) du document, et la disposition des informations ou la fonction des objets (forme).

**1.11** *<physloc>* : Localisation physique, lorsque les documents sont physiquement conservés par une autre institution que celle qui assure l'accès intellectuel, ou bien sur un autre site.

**1.12** *<repository>* : Pour désigner l'organisme responsable de l'accès intellectuel aux documents.

**1.13** *<unitdate>* : Élément indiquant la date l'unité documentaire décrite. Il utilise les

attributs suivants :

- TYPE : Pour indiquer si l'élément <unitdate> représente des dates extrêmes ou des dates prépondérantes.
- CERTAINTY : Pour indiquer si la date a été restituée ou estimée approximativement.
- DATECHAR : Pour une expression caractérisant la nature des dates, telle que "création", "modification" ou "collecte".
- CALENDAR : Spécifie le calendrier dont découle la date tel que le calendrier grégorien ou le calendrier julien.
- ERA : Indique la période au sein de laquelle les années sont numérotées (ère commune ou ère chrétienne)

**1.14** <unitid> : Élément contenant une chaîne de caractères alphanumériques servant de référence univoque ou de numéro de contrôle à l'unité documentaire décrite.

**1.15** <unittitle> : Élément contenant l'intitulé, d'origine ou forgé, des unités documentaires décrites. Il peut consister en un mot, une phrase, un caractère ou un groupe de caractères.

Voici un exemple qui résume les éléments présentés précédemment [EAD 2004] :

```
<did>
<head>Orientation bibliographies</head>
<unitdate>1950-1961</unitdate>
<unitid>1422 W</unitid>
<unittitle>Fonds du haras national de Pompadour</unittitle>
<physdesc>469 cartons et registres, 96 mètres linéaires</physdesc>
<repository>Archives départementales de l'Ain </repository>
<container>Portefeuille</container>
<origination><corpname>France. Bureau de l'enregistrement et des domaines de
Bligny-sur-Ouche </corpname>
</origination>
<abstract>La sous-série T199 appartient à l'ensemble des papiers privés tombés dans le
domaine public qui forme la série T des Archives nationales...etc.</abstract>
<langmaterial>Toutes les lettres reçues par Persigny sont en<languagecode="fre">
français</language>
</langmaterial>
<physloc>Fonds conservé au Musée Picasso</physloc>
<materialspeg type="durée">L'enregistrement dure 92 minutes 18 secondes 22 trames
</materialspeg>
<daohref="406AP57.jpg" </dao>
<daogrp>
<daolochref="image_marches.jpg" title="plan des marches du maître-autel de l'église,
XVIIIe siècle"></daoloc> 91<daolochref="image_chaire.jpg" title="Plan de la chaire
de l'église de l'hôpital, 1731"></daoloc>
</daogrp>
</did>
```

2. **<scopecontent>** : Élément donnant l'essentiel des informations sur la portée des unités documentaires décrites et sur les sujets qu'elles couvrent.

**Exemple [EAD 2004]**

```
<scopecontent><p> Le fonds Roger Planchon est issu de l'activité de Roger Planchon dans les différentes troupes théâtrales qu'il a dirigées ....</p></scopecontent>.
```

3. **<bioghist>** : Élément contenant un texte rédigé ou une chronologie, qui place de façon concise les documents d'archives dans leur contexte, en fournissant des informations à propos de leur(s) producteur(s)on utilisant les éléments suivants :

- **<chronlist>** : Liste chronologique.
- **<chronitem>** : Élément d'une liste chronologique.
- **<date>** : Pour encoder toute date qui mérite de l'être, sauf les dates de production des unités documentaires décrites
- **<event>** : Décrit ou nomme un événement

**Exemple [EAD 2004]**

```
<bioghist>  
<chronlist>  
<chronitem>  
<date>6 novembre 1872</date>  
<event>Naissance à Niort de Gaston Chérau</event>  
</chronitem>  
<chronitem><date>1926</date>  
<event>Gaston Chérau succède à Elémir Bourges à l'Académie Goncourt </event>  
</chronitem>  
</chronlist>  
</bioghist>
```

4. **<custodhist>** : Élément qui énumère les changements successifs de propriété, de responsabilité et de conservation des unités documentaires en cours de description.

**Exemple [EAD 2004]**

```
<custodhist>  
<p>Le second inventaire a fait partie de la collection Boucher de Molandon </p>  
</custodhist>
```

**5. <bibliography>** : Élément permettant de citer les œuvres qui ont pour source ou sujet les documents en cours de description. On peut réaliser une bibliographie simple ou méthodique.

**a) Bibliographie simple**

On utilise les éléments suivants :

- **<head>** : Contient le titre de l'ouvrage.
- **<bibref>** : Élément fournissant une référence bibliographique et/ou un lien hypertexte vers une œuvre.

**Exemple [EAD 2004]**

```
<bibliography>  
<head>Ouvrages généalogiques et biographiques</head>  
<bibref>Mignon, Paul-Louis. - Charles Dullin... - Lyon : la Manufacture, 1990.</bibref>  
</bibliography>
```

**b) Bibliographie méthodique**

On ouvre un élément **<head>**, puis autant d'éléments **<bibliography>** et **<head>** qu'il y a de rubriques. On entre ensuite les références dans **<bibref>**.

**Exemple [EAD 2004]**

```
<bibliography>  
<head>Bibliographie</head>  
<bibliography>  
<head>Sources imprimées</head>  
<bibref>Rueff, Jacques Paris, 1977</bibref>  
</bibliography>  
</bibliography>
```

**6. <separatedmaterial>** : Éléments donnant des informations sur les documents qui ont été séparés du fond.

**Exemple [EAD 2004]**

```
<separatedmaterial><p>D'autres papiers provenant de Paul Reynaud sont conservés au  
ministère des Affaires étrangères de 1939-1942.</p></separatedmaterial>
```

**7. <relatedmaterial>** : Élément donnant des informations sur des sources complémentaires.

**Exemple [EAD 2004]**

```
<relatedmaterial>  
<head>Sources externes</head>  
<p>Dossiers concernant les rapports entre le Conseil général du Finistère et le Conservatoire  
de l’Affiche en Bretagne de la Direction de l’Enseignement</p>  
</relatedmaterial>
```

**8. <otherfindaid>** : Élément qui donne des informations sur des instruments de recherche supplémentaires ou complémentaires relatifs aux unités documentaires décrites.

**Exemple [EAD 2004]**

```
<otherfindaid><p>Répertoire numérique dactylographié, par Gisèle Drigeard et Véronique  
de Lavarde, sous la direction de Hélène Say, directeur, 2003, 68pages </p></otherfindaid>
```

**9. <dsc>** : Élément englobant qui rassemble des informations sur les regroupements hiérarchiques des documents en cours de description, il est immédiatement suivie d'un certain nombre d'éléments Composant <c> qui s'imbriquent éventuellement les uns dans les autres.

**Exemple [EAD 2004]**

```
<dsc>  
<c LEVEL="series">  
<c LEVEL="file">  
<c LEVEL="item"></c>  
</c>  
</c>  
</dsc>
```

**10. <acqinfo>** : Élément sur la provenance directe des unités documentaires en cours de description et les circonstances de leur entrée. Il inclut les dons, legs, transferts, achats...

**Exemple** [EAD 2004]

```
<acqinfo><p>Type d'entrée : versement. </p></acqinfo>
```

**11. <altformavail>** : Élément qui fournit des informations sur les reproductions des unités documentaires en cours de description.

**Exemple** [EAD 2004]

```
<altformavail><p>Les cahiers de doléances ont été microfilmés </p></altformavail>
```

**12. <accruals>** : Élément qui fournit des informations sur les accroissements prévisibles de l'ensemble documentaire.

**Exemple** [EAD 2004]

```
<accruals><p>Dépôt, 1983 ; sera complété ultérieurement.</p></accruals>
```

**13. <accessrestrict>** Élément qui fournit des informations sur les conditions de mise à disposition des unités documentaires en cours de description.

**Exemple** [EAD 2004]

```
<accessrestrict>  
<p>Il n'y a pas de restriction juridique à la consultation de  
ces Documents</p></accessrestrict>
```

**14. <userrestrict>** : Élément qui fournit des informations sur les conditions d'utilisation des unités documentaires décrites, une fois la communication autorisée.

**Exemple [EAD 2004]**

```
<userrestrict>  
<p>Reproduction et publication avec l'accord du déposant. </p>  
</userrestrict>
```

**15. <controlaccess>** : Élément englobant qui désigne des points d'accès pour les unités documentaires décrites et permet la recherche par listes d'autorité. Il utilise les sous éléments suivants :

- **<subject>** : Élément contenant un terme qui identifie un sujet associé aux unités documentaires décrites ou traité dans celles-ci.
- **<corpname>**
- **<persname>**

**Exemple [EAD 2004]**

```
<controlaccess>  
<subject>Parti politique</subject>  
<corpname>Mouvement républicain populaire (MRP)</corpname>  
<persname>Abelin, Pierre</persname>  
</controlaccess>
```

**16. <appraisal>** : Élément qui fournit des informations sur la manière dont a été déterminée la valeur archivistique des documents, selon leur utilité administrative, juridique et fiscale du moment.

**Exemple [EAD 2004]**

```
<appraisal><p>Aucune élimination n'a été pratiquée dans les dossiers de travail du  
commissaire à l'aménagement du Marché d'intérêt national de Paris-Rungis, où sont mêlés  
documents de synthèse et notes manuscrites prises au cours de lectures, de réunions,  
d'entretiens téléphoniques</p></appraisal>
```

**17. <arrangement>** : Élément qui contient des informations sur la façon dont les documents décrits ont été divisés en unités plus petites.

**Exemple [EAD 2004]**

```
<arrangement><p>Classement par commune.</p></arrangement>
```

**18. <descgrp>** : Élément permettant d'opérer des regroupements à l'intérieur de l'élément description archivistique<archdesc>.

**Exemple [EAD 2004]**

```
<descgrp>  
<head>Conditions d'accès et d'utilisation</head>  
<accessrestrict>  
<head>Restrictions d'accès:</head><p>Aucune</p>  
</accessrestrict>  
<userrestrict>  
<head>Conditions de publication et de reproduction:</head>  
<p>La reproduction est soumise à l'autorisation de l'Académie </p>  
</userrestrict>  
</descgrp>
```

**19. <fileplan>** : Élément fournissant des informations sur tout type de classement utilisé par les personnes à l'origine de la création ou de la recension des unités documentaires décrites, pour les classer, les stocker et les retrouver.

**Exemple [EAD 2004]**

```
<fileplan><p>Classement par ordre alphabétique</p></fileplan>
```

**20. <index>** : Élément donnant une liste de mots clés et de pointeurs de référence regroupés dans l'éléments <indexentry>, pour améliorer l'accès aux unités documentaires en cours de description.

**Exemple [EAD 2004]**

```
<index>
<head>Index des correspondants de Persigny</head>
<indexentry>
<persname>Bonaparte, Pierre-Napoléon</persname>
<ref target="c44AP11">44AP/11</ref>
</indexentry>
</index>
```

**21. <odd>** : Élément qui donne, sur les documents décrits, des informations qu'on ne peut facilement placer à l'intérieur des éléments description archivistique <archdesc> et composant <c>.

**Exemple [EAD 2004]**

```
<oddtype="note-historique">
<p>Cette propriété fut acquise par Jean Péan en 1634. Il était le père de Marie Péan.</p>
</odd>
```

**22. <originalsloc>** : Élément qui donne des informations sur l'existence, la localisation, l'accessibilité et/ou la destruction des documents originaux lorsque les unités documentaires décrites sont des reproductions.

**Exemple [EAD 2004]**

```
<originalsloc>
<p>Microfilm des papiers de Jacques-François Begouën, député du bailliage de Caux à l'Assemblée constituante (1743-1831) et de ses descendants </p></originalsloc>
```

**23. <phystech>** : Élément englobant permettant de regrouper des informations concernant l'aspect ou la structure matérielle des unités documentaires décrites.

**Exemple** [EAD 2004]

```
<phystech><p>Lisible uniquement avec lampe de Wood</p></phystech>
```

**24. <prefercite>** : Mention conseillé, est utilisé pour donner la formule selon laquelle les utilisateurs devront mentionner les documents lorsqu'ils y font référence dans une publication et en particulièrement en cas de reproduction.

**Exemple** [EAD 2004]

```
<prefercite>  
<head>Mention recommandée</head>  
<p>Bibliothèque de l'Institut de France (Paris), ms. ..., fol. ...</p>  
</prefercite>
```

**25. <processinfo>** : Élément contenant des informations sur la gestion de l'entrée, le classement, la description, la préservation, le stockage et tous les traitements préalables à l'exploitation des unités documentaires décrites.

**Exemple** [EAD 2004]

```
<processinfo><p> Notice rédigée par Madeleine Cottin</p></processinfo>
```

**26. <runner>** : Élément facultatif de mise en forme qui permet de faire figurer un en-tête, un bas de page ou un filigrane.

**Exemple** [EAD 2004]

```
<runner placement="header">Guide des archives du Jura</runner>
```

- 27. **<note>** : *Expliqué au paravent*
- 28. **<dao>** : *Expliqué au paravent*
- 29. **<daogrp>** : *Expliqué au paravent*

## VII. Description des manuscrits avec la DTD EAD :

Notre intérêt d'étudier de la DTD EAD est de voir à quel point ce standard apporterait à la description des manuscrits arabes. Malgré son titre, l'encodage EAD qui est destiné en particulier à la description d'archives, permet de décrire des manuscrits mais de manière marginale [Kaileh 2004]. Après avoir étudié l'EAD on constate que celle-ci n'offre aucune description implicite ou intuitive des manuscrits arabes. Un effort d'analyse et de structuration doit être effectué au préalable par l'encodeur, ce qui ne rend pas la tâche facile au vu de l'aspect œuvre archéologique que représentent les manuscrits arabes.

## VIII. Conclusion

La DTD EAD permet une description à plusieurs niveaux (fonds, sous fonds, série, sous série, pièce,...etc.) elle permet aussi d'insérer des balises permettant d'identifier la nature et les niveaux des parties de texte ainsi balisées et de développer à partir de ces balises des outils d'interrogation et de recherche.

L'EAD qui est un modèle de description adapté aux archives, initialement conçu avec SGML puis adapté à XML, cependant le partage d'un même vocabulaire permet facilement d'échanger et d'exploiter des fichiers XML. Il offre une grande souplesse (146 éléments non obligatoires dont l'ordre peut varier) tout en permettant une description très fine et complexe.

## ***Introduction***

Dans cette partie, nous allons décrire un autre standard utilisé pour la description de manuscrits qui est la TEI (Text Encoding Initiative). La TEI joue un rôle important dans le domaine de catalogage des manuscrits médiévaux. Dans cette partie nous allons décrire les principes et les caractéristiques de la TEI et les outils d'encodage que ce standard utilise.

### ***I. Historique***

La *Text Encoding Initiative* (TEI) est un projet international qui vise à la mise au point d'un ensemble de normes pour la préparation et l'échange de textes électroniques. La naissance de la TEI a eut lieu lors d'une réunion organisée en novembre 1987 à VassarCollege (PoughKeepsie, New York) [Ide & Véronis 1996].

Cette norme a été créée officiellement en 1988, surnommée la TEI P1, avec la participation d'un grand nombre de chercheurs à travers le monde sous forme de comités. Depuis la TEI a connu l'apparition de différentes versions (chaque version apporte des améliorations par rapport à sa précédente). La TEI P2 vit le jour en 1992, ensuite la TEIP3 en 1994 sous forme de *Guidelines for Electronic Text Encoding and Interchange* (Recommandations pour le codage et l'échange des textes informatisés) en SGML.

En Juin 2003 la TEI P4, a convertit les recommandations de la TEI P3 en XML. En final, la TEI P5 est apparue en Novembre 2007, qui est une révision majeure des recommandations de la TEI P4. Cette version fournit d'importantes nouvelles fonctionnalités de codage [TEI P5 2012].

### ***II. Définition de la TEI***

TEI (Text Encoding Initiative) est un projet universitaire pluridisciplinaire visant à uniformiser autant que possible le codage de documents en vue de leur échange et de leur publication en ligne ou hors ligne [Difournaud 2008]. La TEI permet d'encoder des textes en particulier les textes littéraires et linguistiques. La TEI procède à l'encodage suivant une structure arborescente hiérarchique (divisions, chapitres, sous-chapitres, sections, et jusqu'à ses parties les plus complexes telles que citations, vers, noms propres mentionnés dans le texte, soulignement et autres mises en évidence).

### III. *Caractéristiques de la TEI*

La TEI est une norme de balisage [Telma 2007], de notation et d'échange des documents électroniques, fondée sur le SGML et s'appuyant sur le XML. La TEI se présente sous forme d'un ensemble de recommandations rassemblées en modules distincts [Crassous 2009] :

- ✓ Un ensemble d'éléments organisés en modules différents, les « **tag sets** » que l'on sélectionne en fonction des besoins spécifiques de l'utilisateur pour former un Document Type Définition (DTD).
- ✓ Un ensemble de recommandations, les « *Guidelines* » qui expliquent comment utiliser la DTD.

Une DTD conforme à la TEI est composée de quatre fragments de DTD [Beaudry 2007] :

- Un ensemble de balises obligatoires (Core tag sets)
- Un ensemble de balises de base (base tag set).
- Un ensemble de balises additionnelles (additional tag set).
- Un ensemble de balises auxiliaires (auxiliary tag set).

Les balises du noyau sont toujours présentes, mais les jeux de balises de base, additionnelles ou auxiliaires peuvent «changer» dans la DTD en fonction des besoins des utilisateurs [Ide & Véronis 1996]. A cet effet, l'appellation de TEI Pizza Chef a été donnée à cette norme, d'abord prendre le noyau, choisir une base et enfin, ajouter les garnitures de choix.

### IV. *Avantages de la TEI*

- ✓ Le codage de la TEI indépendant des «plateformes», c'est-à-dire indépendant de marque d'ordinateur, de système commercial de manipulation de texte, mais aussi de réseau [André 1996].
- ✓ L'architecture modulaire proposée par la TEI permet de choisir des ensembles d'éléments répondant aux besoins d'encodage d'un type particulier de texte : poésie, pièces de théâtre, dictionnaires, corpus linguistiques, manuscrits, critiques textuelles, transcriptions de discours oraux.

### V. *Principes de balisage*

Les éléments, qui sont les structures de base des documents XML, sont délimités par des balises de début et des balises de fin qui portent le contenu. TEI repose sur XML pour sa syntaxe [Telma 2007]. Par conséquent, la TEI est conforme au principe de balisage utilisé par XML, tel que les éléments et les attributs. Tout document TEI doit avoir une liste d'éléments valides et un ensemble d'attributs.

## 1. Les attributs

On appelle « attribut » un paramètre associé à un élément, Un attribut est lui-même composé d'une valeur. Il existe [Difournaud 2008] :

- Des attributs globaux (utilisés par tous les éléments). Par exemple, l'attribut `n` pour un changement de ligne avec son numéro.

**Exemple** : `<lb n="36">`

- Des attributs particuliers spécifiques à quelques éléments, tel que l'attribut `"when"` qui est un attribut réservé à l'élément `<date>` pour l'encodage de la date au format ISO (aaaa-mm-jj).

**Exemple** : `<date when="1515-04-28">28 avril 1515</date>`

## 2. Les éléments

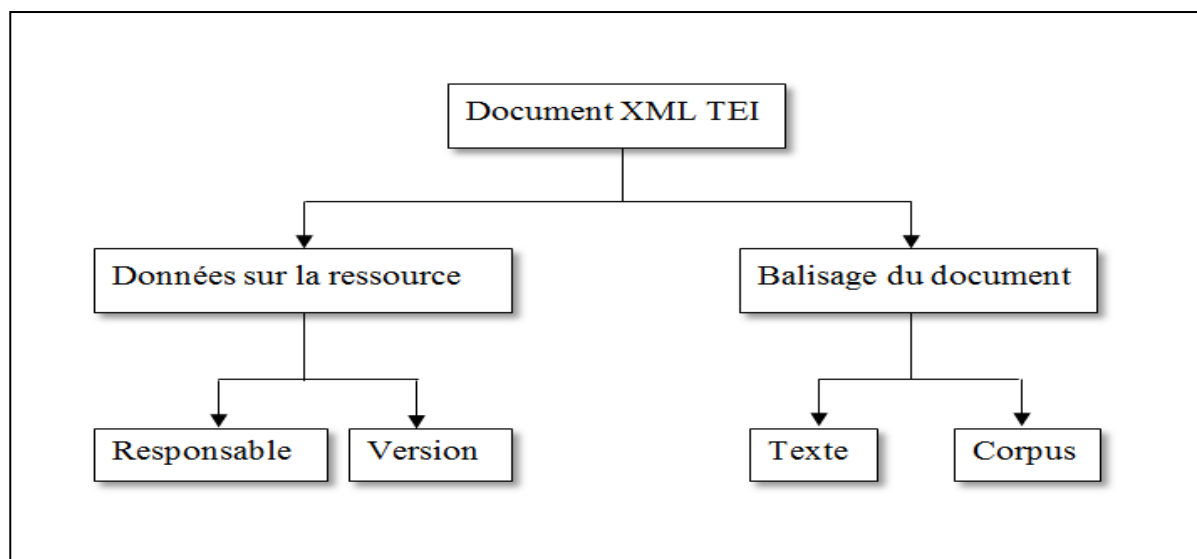
Un texte encodé avec la TEI est un arbre d'éléments, permettant de le décomposer en unités d'informations définies par des balises de début et de fin de contenu. La TEI se présente avec plusieurs éléments constituant les métadonnées évalués 521 éléments entre obligatoires et facultatifs, regroupés en 146 classes, et 22 modules, parmi lesquelles on trouve la TEI-`ms`.

## VI. La structure générale de la TEI

La structure générale du document XMLTEI (TEI Lite ou la TEI simplifiée), à travers son balisage, est décomposée en deux jeux de métadonnées complémentaires :

- L'en-tête (`teiHeader`) qui intègre à la fois des données sur la source et des méta-métadonnées sur la transcription (responsables, version, mode de disponibilité,...) en d'autre terme tous les éléments de catalogage d'un ouvrage sont précisés à ce niveau.
- Le balisage proprement dit du document (`text`). Il peut être assez léger, mais toujours formel. Il permet essentiellement, l'échange de références ou de corpus.

Le schéma suivant illustre la structure d'un document XML TEI :



**Figure 11:** Description générale d'un document XML TEI.

Parmi les 521 éléments de la TEI on peut distinguer 5 éléments principaux qui sont précédé par l'élément racine <tei>.

```

<tei>
<teiHeader>
...
</teiHeader>
<text>
<front>...</front>
<body>
<div>
...
</div>
</body>
<back>...</back>
</text>
</tei>
  
```

**Figure 12:** L'encodage global d'un document encodé avec la TEI.

### a) *L'en-tête TEI <teiHeader>*

Élément donnant des informations descriptives et déclaratives qui constituent une page de titre au début de tout texte conforme à la TEI.

### **b) Texte de l'œuvre <text>**

Élément englobant l'ensemble du texte. Le texte est structuré au maximum en cinq composantes ; deux sont obligatoires et trois optionnelles :

- 1.<text> : Élément obligatoire, il peut être unitaire ou composite comme dans un recueil;
- 2.<front> : Élément optionnel, il regroupe les pièces situées avant le début du texte lui-même comme la page de titre de l'ouvrage et les pièces liminaires ;
3. <group> : Élément optionnel, il regroupe plusieurs textes ;
- 4.<body> : Élément obligatoire, il regroupe le corps entier d'un texte unitaire ;
- 5.<back> : Élément optionnel, il regroupe les annexes qui suivent le corps du texte.

*Le principal objectif* d'étudier la TEI est la possibilité d'encoder les manuscrits arabes avec cette norme d'une manière fine et cohérente. Pour cela, nous nous intéresserons à l'étude de module TEI-ms (Text Encoding Initiative Manuscript Description), qui est destiné à la description des manuscrits.

## **VII. La TEI-Manuscript Description (TEI-ms)**

La TEI propose un module spécifique pour la description des manuscrits nommé TEI-ms. La TEI Manuscript Description dans sa version P5 est une révision majeure des lignes directrices qui offre de nombreuses fonctionnalités nouvelles et améliorées. Elle a été officiellement publiée en Novembre 2007 [TEI P5 2012].

### **1. Caractéristiques de la TEI-ms**

Pour la description des manuscrits la TEI-ms propose deux méthodes différentes :

#### **i. Méthode simple**

Le manuscrit est décrit à l'aide d'une série de paragraphes imbriqués dans des identificateurs définissant un aspect particulier de ce dernier.

#### **ii. Méthode complexe**

Le manuscrit est décrit d'une manière précise et complète à l'aide d'une notice dotée d'une structure de balisage plus détaillée.

La première méthode est intuitive et simple à utiliser. Elle offre aux catalogueurs le cadre familier pour la description des manuscrits, mais elle présente également de grandes difficultés pour une indexation conceptuelle. La seconde méthode nécessite le respect de la structure TEI-ms, mais il facilite l'étape de l'indexation et la recherche d'information.

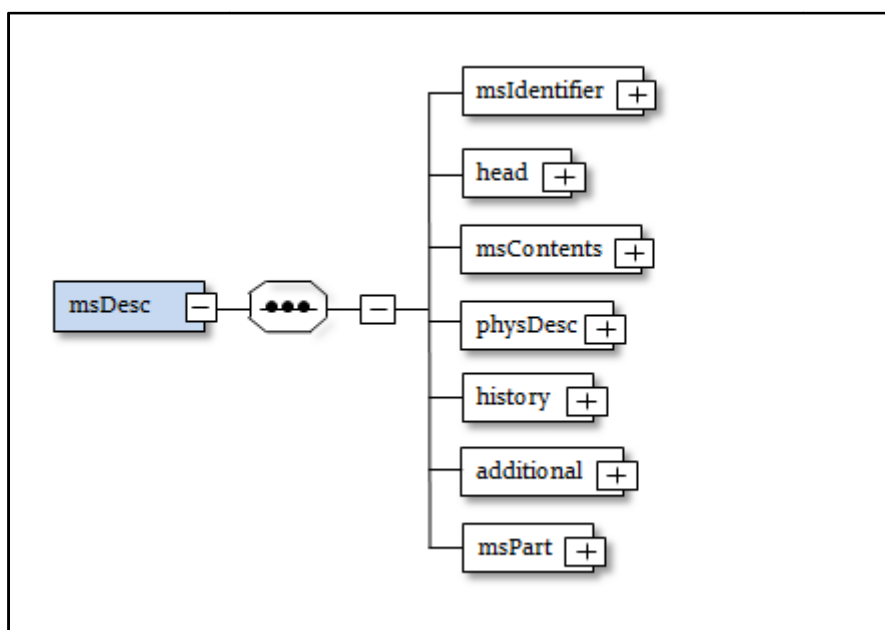
## 2. Éléments de la TEI-ms

Parmi les nouveautés apportées par la TEI-ms nous citons :

- ✚ La TEI-ms possède des éléments descriptifs standards ce qu'on appelle *les éléments de niveau de phrase* utilisés pour décrire au niveau d'un paragraphe <p> des caractéristiques d'un élément approprié du manuscrit.
- ✚ Utilisé pour la description des manuscrits la TEI-ms propose l'élément <msDesc>, qui est l'élément racine de tous les éléments descriptifs des manuscrits. Cet élément doit être placé juste après l'élément <sourceDesc>. Il est composé de 7 éléments qui ont à leur tour des ensembles de sous-éléments.

### L'élément <msDesc>

<msDesc> ; Manuscript Description, c'est l'élément *réservé à l'encodage des manuscrits*. Il décrit un seul manuscrit à la fois, il est formé de sept éléments comme le montre la figure suivante :



**Figure13** : Les sous-éléments de l'élément <msDesc>.

**1. <msIdentifier>** : Manuscript Identifier, élément obligatoire permet d'identifier un manuscrit. Les éléments suivants sont disponibles au niveau de cet élément :

- <country> : Pays, contient le nom d'une unité géopolitique, comme une nation, pays, colonie ou une communauté.
- <region> : Contient le nom d'une unité administrative comme un état.

- **<settlement>** : Contient le nom d'un règlement.
- **<institution>** : Contient le nom d'une institution telle qu'une bibliothèque dans laquelle se trouve le lieu de conservation du manuscrit.
- **<repository>** : Contient le nom d'une organisation qui est titulaire du manuscrit.
- **<collection>** : Contient le nom d'une collection de manuscrits.
- **<idno>** : Identifiant, qui fournit toute sorte d'identifiant utilisé pour identifier un objet, tel qu'un élément de données bibliographiques, une personne, un titre, une organisation, d'une manière standardisée.
- **<altIdentifier>** : Autre identifiant, contient un identificateur de remplacement ou un ancien employé d'un manuscrit, comme un numéro de catalogue ancien.
- **<msName>** : Nom du manuscrit, contient toute forme de nom alternatif structuré utilisé pour un manuscrit.

### Exemple[TEI P5 2012]

```

<msIdentifier>
  <pays> États-Unis </pays>
  <region> Californie </region>
  <settlement> San Marino </règlement>
  <repository> Huntington Library </repository>
  <collection> El </collection>
  <idno> 26 C 9 </idno>
  <msName>Le chaucer Ellesmere</msName>
  <collection>Oriental Collection</collection>
  <altIdentifier><idno>2MS BS1145 I8</idno></altIdentifier>
</MsIdentifier>

```

2. **<head>** : Contient tout type d'en-tête, par exemple le titre d'une section, ou l'intitulé d'une liste, d'un glossaire, d'une description de manuscrits, il fournit aussi son titre et son nom.

### Exemple [TEI P5 2012]

```

<head>Marsilius de Inghen, AbbreuiataphisicorumAristotelis; Italy, 1463.</head>

```

3. **<msContents>** : Élément utilisé pour décrire le contenu intellectuel du manuscrit ou d'une partie de manuscrit, il est constitué des éléments suivants :

- a) **<msItem>** : Contient les éléments relatifs à l'information bibliographique d'un manuscrit. Sous cet élément on peut retrouver les éléments suivants :
- **<author>** : Contient le nom de la personne responsable du contenu du manuscrit.
  - **<repStmt>** : Enoncé de responsabilité, fournit une déclaration de responsabilité pour le contenu intellectuel d'un texte, édition, un enregistrement ou une série, lorsque les éléments spécialisés pour les auteurs, éditeurs...etc, ne suffisent pas ou ne s'appliquent pas.
  - **<title>** : Contient un titre pour tout type de travail.
  - **<rubric>** : Contient les mots qui indiquent le début ou la fin d'une division, généralement séparé du texte par l'encre rouge, par le type d'écriture ou une autre taille de l'écriture.
  - **<incipit>** : Contient les premiers mots du texte du manuscrit.
  - **<quote>** : Citation, contient une phrase ou un passage attribué par le narrateur ou l'auteur à un organisme externe pour le texte.
  - **<explicit>** : Contient les dernières phrases du manuscrit par lesquelles l'auteur termine son œuvre.
  - **<finalRubric>** : Contient la chaîne de mots qui désigne la fin d'une division de texte.
  - **<colophon>** : Contient les informations fournies par le copiste sur l'élaboration du manuscrit.
  - **<decoNote>** : Élément de base pour décrire un aspect spécifique de la décoration d'un manuscrit.
  - **<listBibl>** : Liste de citations, contient une liste de citations bibliographiques de toute nature.
  - **<bibl>** : Référence bibliographique, contient une citation bibliographique.
  - **<note>** : Contient une note ou une annotation.
  - **<textLang>** : Langue du texte, encode les noms des langues cirées dans le manuscrit.

### Exemple [TEI P5 2012]

```

<msContents>
  <msItem>
    <locus>f. 1-223</locus>
    <author>RadulphusFlaviacensis</author>
    <title>Expositio super Leviticum</title>
    <incipit>
      <locus>f. 1r</locus>
      Forte Herveimonachi</incipit>
    <explicit>
      <locus>f. 223v</locus>
      Benedictio salis et aquae</explicit>
    </msItem>
  </msContents>

```

- b) **<msItemStruct>** : Contient la description structurée d'un item ou d'une œuvre, dans le contenu intellectuel d'un manuscrit ou d'une partie d'un manuscrit.
- c) **<summary>** : Sert à encoder un aperçu du contenu du manuscrit.

4. **<physDesc>** : La description physique du manuscrit se fait en utilisant trois grands aspects du manuscrit :

- Description de l'objet
- Description de l'écriture, de décoration et autres annotations
- Reliure et matière

#### 4.1. Description de l'objet

Elle est assurée par l'élément **<objectDesc>**, qui est utilisé pour regrouper les parties de la description physique qui se rapportent spécifiquement à l'objet de texte portant son format, la constitution de mise en page. Cet élément comporte deux parties ; une description du support et une description de la mise en page.

##### a. Description du support

Elle est étiquetée en utilisant les éléments suivants :

- **<supportDesc>** : Description de soutien, regroupe des éléments décrivant le support physique pour la partie écrite d'un manuscrit.
- **<support>** : Contient une description des matériaux qui constituent le support physique pour la partie écrite d'un manuscrit. Il contient l'élément **<material>** décrivant les matériaux utilisés pour fabriquer un manuscrit
- **<extent>** : Décrit la taille approximative d'un texte tel qu'il est stocké sur un certain type de support. Il contient l'élément **<dimensions>** :<dimensions> Contient la spécification du document. Il peut contenir
  - **<height>**: Encode la hauteur d'un manuscrit.
  - **<width>**: Encode la largeur d'un manuscrit.
  - **<depth>** : Encode l'épaisseur d'un manuscrit.
  - **<dim>** : Contient une seule mesure faisant partie d'une spécification dimensionnelle d'une certaine sorte.
- **<collation>** : Contient une description de la façon dont les feuilles ou bifolios sont physiquement disposés.
- **<foliation>** : Décrit le système de numérotation ou des systèmes utilisés pour compter les feuilles ou les pages dans un codex.
- **<condition>** : Contient une description de l'état physique du manuscrit.

**Exemple [TEI P5 2012]**

```

<objectDesc>
  <supportDesc>
    <material>paper</material>
    <foliation><p>NeuereFolierung, die auch das Vorsatzblattmitgezählt
    hat.</p></foliation>
    <extent>ii + 321 leaves<dimensions >
    <height>35</height>
    <width>27</width>
    </dimensions>
    </extent>
    <collation></collation>
    <condition><p>The manuscript shows signs of damage from water and mould on its
    outermost leaves.</p></condition>
  </supportDesc>
</objectDesc>

```

**b. Description de la mise en page**

Cette partie sert à décrire et à documenter la mise en page du manuscrit (la manière dont le texte et l'éclairage sont disposés sur la page). Elle contient les deux éléments suivants :

1. **<layout>** : Décrit comment le texte est disposé sur la page, y compris des informations sur toute décision.
2. **<layoutDesc>** : Description de mise en page, recueille l'ensemble des descriptions de mise en page d'un manuscrit.

**4.2. Description de l'écriture, de décoration et autres annotations**

Description des styles d'écriture comme la façon dont il est disposé sur la page, les éléments décoratifs, toute notation musicale utilisée et toutes les annotations marginales. Ce deuxième aspect de la description physique du manuscrit est décrit par les éléments :

1. **<handDesc>** : Description des mains, contient une description de tous les différents types d'écriture utilisés dans un manuscrit.
2. **<handNote>** : Note sur la main, décrit un style particulier ou la main distingués dans un manuscrit.
3. **<scriptDesc>** : Contient une description des scripts utilisés dans un manuscrit ou d'une source similaire.
4. **<scriptNote>** : Décrit un script particulier distingué dans la description d'un manuscrit ou d'une ressource.

5. **<typeDesc>** : Contient une description des polices de caractères ou d'autres aspects de l'impression d'une source ou d'un autre incunable imprimé.
6. **<typeNote>** : Décrit une police particulière ou un autre signe typographique important distingué au sein de la description d'une ressource imprimée.
7. **<decoDesc>** : Contient une description de la décoration d'un manuscrit, soit en une suite de paragraphes, ou comme une séquence de l'élément **<decoNote>**.
8. **<decoNote>** : Note sur la décoration, contient une note décrivant soit un élément décoratif d'un manuscrit, ou d'une catégorie assez homogène de tels éléments.
9. **<musicNotation>** : Contient une description du type de la notation musicale.
10. **<addition>** : Contient une description des ajouts significatifs trouvés dans un manuscrit, tels que gloses marginales ou autres annotations.

### 4.3. Reliure et matière

Le troisième volet de la description physique du manuscrit est relatif à la reliure et la matière de fabrication du manuscrit. Il est défini par les éléments :

- **<bindingDesc>** : Description de liaison, décrit les liaisons actuelles et anciennes du manuscrit, soit comme une série de paragraphes ou comme une série de différents éléments, un pour chaque reliure du manuscrit.
- **<binding>** : Contient une description d'un liant, à savoir le type de revêtement, conseils, appliquées à un manuscrit.
- **<sealDesc>** : Description joint, décrit les sceaux ou autres objets attachés à un manuscrit.
- **<seal>** : Contient une description d'un sceau similaire, attaché à un manuscrit.
- **<accMat>** : Matériel d'accompagnement, donne des détails sur tout matériel d'accompagnement étroitement associé au manuscrit, tel que documents non contemporains ou fragments reliés avec le manuscrit à une époque antérieure

## 5. **<history>**

Cet élément définit l'historique du manuscrit. Il contient les éléments suivants :

- **<origin>** : Contient des informations concernant le lieu, la date et toute autre information relative à la production d'un manuscrit: cet élément peut être utilisé pour fournir les preuves sur le contenu de **<origDate>** et de **<origPlace>**.
- **<provenance>** : Contient des informations concernant n'importe quel aspect de l'histoire du manuscrit (y compris des informations sur sa production et / ou son acquisition par le présent ou le dernier propriétaire connu, lorsque ceux-ci n'apparaissent pas dans **<origin>** ou **<acquisition>**).

- **<acquisition>** : Contient les informations relatives à l'acquisition du manuscrit par le possesseur actuel.

### Exemple [TEI P5 2012]

```

<history>
  <origin>
    <p>Written in Durham during the mid-twelfth century </p>
  </origin>
  <provenance>
    <p>Recorded in two medieval catalogues of the books belonging to Durham Priory,
    made in 1391 and 1405.</p>
  </provenance>
  <provenance>
    <p>Given to W. Ollefy by William Ebchester, Prior (1446-56) and later belonged to
    Henry Dalton, Prior of Holy Island (Lindisfarne) according to inscriptions on ff. 4v and
    5.</p>
  </provenance>
  <acquisition>
    <p>Presented to Trinity College in 1738 by Thomas Gale and his son Roger.</p>
  </acquisition>
</history>

```

## 6. <additional>

Définit l'aspect administratif et bibliographique du manuscrit. Il est défini par trois éléments de base :

- **<adminInfo>** : Renferme les informations relatives à l'institution de conservation, à la disponibilité et la conservation du manuscrit.
- **<surrogates>** : contient des informations sur toute reproduction numérique ou photographique du manuscrit en cours de description.
- **<listbibl>** : Contient une liste de citations bibliographiques de toute nature.

## 7. <msPart>

Cet élément décrit d'autres manuscrits réunis en un seul. Dans le cas où un manuscrit serait formé par l'assemblage de plusieurs manuscrits, <msPart> permettrait de décrire chaque manuscrit comme un objet à part et totalement indépendant des autres volumes du manuscrit.

**Exemple [TEI P5 2012]**

```

<msDesc>
<msIdentifier>
<repository>Bibliothèque Municipale</repository>
<idno>MS 3</idno>
<msName>Maurdrampus Bible</msName>
</msIdentifier><!-- other elements here -->
<msPart>
<msIdentifier>
<idno>MS 6</idno>
</msIdentifier><!-- other information specific to this part here -->
</msPart>
<msPart>
<msIdentifier>
<idno>MS 7</idno>
</msIdentifier><!-- other information specific to this part here -->
</msPart>
<!-- other msParts here -->
</msDesc>

```

**VIII. Encodage de manuscrits arabes avec la TEI-ms**

En termes d'encodage de manuscrits dans leur ensemble, le module TEI-ms se considère comme étant le plus proche à fournir une réponse adéquate aux problèmes rencontrés par les catalogueurs, tel qu'il offre un ensemble concis des éléments descriptifs pour le codage des métadonnées sur les manuscrits.

Ce module a été développé à l'origine pour répondre aux besoins des catalogueurs et des chercheurs qui travaillent avec les manuscrits médiévaux.

Mais malgré tous ces points forts le module TEI-ms a encore des faiblesses qui le rendent pas particulièrement adapté pour décrire des manuscrits arabes.

Sachant que, le module TEI P5 est un outil adéquat pour le catalogage des manuscrits, mais il lui manque certaines fonctionnalités permettant une description exhaustive des manuscrits arabes anciens [Soualah & Hassoun 2011].

## ***IX. Conclusion***

La TEI est un système extensible, modulaire constituant un modèle abstrait. La modularité de la TEI constitue une de ses grandes forces, lui permettant d'être un modèle universel et interdisciplinaire.

Toutefois la TEI comporte certaines difficultés. Paradoxalement, les atouts que sa richesse et sa généralité la rendent complexe à utiliser.

Mais malgré ces faiblesses le module TEI-ms d'après des études qui ont été faites, il peut être utilisé pour les manuscrits médiévaux occidentaux et les manuscrits arabes et sa après quelques modifications et ajouts.

## ***Introduction***

Après avoir présenté deux normes d'encodage des catalogues dans les deux premières parties de ce chapitre, présent, nous allons étudier un autre format de représentation de métadonnées qui se nomme le Dublin Core.

### ***I. Historique***

Contrairement à ce que son nom laisse présager, le Dublin Core ne possède aucune racine en Irlande ! C'est plutôt une ville de l'Ohio, aux États-Unis, qui voit naître cette norme très répandue aujourd'hui. En 1995, le National Center for Supercomputing Applications et Online Computer Library Center (OCLC) réunissent leurs forces pour former un groupe de travail. Il s'agit de la Dublin Core Metadata Initiative (DCMI), une organisation dont la mission consiste à promouvoir l'adoption à grande échelle de métadonnées standards et interopérables dans la description des ressources électroniques. À travers ce mandat, le comité vise à permettre un meilleur repérage des documents sur Internet. Depuis la création de la DCMI, plusieurs équipes de travail provenant de différents domaines collaborent activement au développement du projet afin d'améliorer à la fois sa simplicité et son exhaustivité. À l'heure actuelle, les métadonnées du Dublin Core sont les plus connues et leur succès est sans doute redevable à leur grande simplicité [aifbd 2012].

### ***II. Objectifs***

L'objectif du Dublin Core au travers de la définition d'un ensemble de champs est de fournir un socle commun d'éléments descriptifs est de faciliter l'utilisation de ces informations dans des environnements applicatifs, afin améliorer le signalement et la recherche de ressources au-delà des diverses communautés et des nombreux formats descriptifs propres à chaque spécialité, tout en restant suffisamment structuré.

### ***III. Le Dublin Core simple***

Le Dublin Core compte 15 éléments de base répartis en 3 grandes catégories [guideDC2008] :

### 1. Contenu

- Couverture
- Description
- Type
- Relation
- Source
- Sujet
- Titre

### 2. Propriété intellectuelle

- Collaborateur
- Créateur
- Éditeur
- Droits

### 3. Instanciation

- Date
- Format
- Identifiant
- Langue

Cet ensemble de base répond à 5 principes fondamentaux :

- ✓ Tous les éléments sont optionnels.
- ✓ Tous les éléments sont répétables.
- ✓ Le Dublin Core est extensible. Il est possible de raffiner la description d'un élément d'une façon normalisée au moyen de sous-éléments. Cela restreint sa signification sans la changer fondamentalement.
- ✓ Le Dublin Core est international (on l'utilise dans plus de 20 langues).
- ✓ Chaque élément possède l'un des quatre statuts suivants :
  - **Obligatoire** : L'élément doit impérativement figurer dans la notice
  - **Obligatoire si applicable** : Si l'information est connue, elle doit figurer dans la notice
  - **Recommandé** : Il est souhaitable que cet élément figure dans la notice
  - **Facultatif** : L'élément peut figurer ou non dans la notice

<dc:coverage> : Couverture	
Définition	La portée ou la couverture spatio-temporelle de la ressource.
Commentaires	Généralement, l'élément couverture inclut une position géographique (le nom d'un lieu où ses coordonnées géographiques), une période de temps (un nom de période, une date ou un intervalle de temps). Ne pas employer pour indiquer la période ou le lieu de création de la ressource [guideDC 2008].
Exemple	<dc:coverage>Savoie, Duché de (Royaume de Sardaigne). 1720-1861</dc:coverage>
Statut	Recommandé si applicable

<dc:description> :Description	
Définition	Une présentation du contenu intellectuel de la ressource.
Commentaires	L'information porte sur la description intellectuelle des documents et non à la description matérielle de celui-ci.  La description peut inclure un résumé, une table des matières, une référence à une représentation graphique du contenu ou un compte-rendu libre [siteDC 2007].
Exemple	<dc:description>Collection : Collection Hetzel ; 11-12</dc:description> [guideDC 2008]
Statut	Facultatif



<code>&lt;dc:source&gt;</code> :Source	
Définition	Référence à une ressource dont la ressource décrite est dérivée.
Commentaires	<ul style="list-style-type: none"> <li>- L'indication des pages ou feuillets fait partie de la source si seule une partie de l'originale a été numérisée,</li> <li>- Si l'originale numérisée est elle-même une reproduction, la notice décrit les données bibliographiques de la source, afin d'en faciliter l'accès aux utilisateurs [guideDC 2008].</li> </ul>
Exemple	<code>&lt;dc:source&gt;</code> Bibliothèque nationale de France, département Littérature et art, 8-Y2-21488 (1) <code>&lt;/dc:source&gt;</code> .
Statut	Recommandé

<code>&lt;dc:subject&gt;</code> : Sujet	
Définition	Thème du contenu de la ressource.
Commentaires	Cet élément est utilisé pour décrire le contenu intellectuel de la ressource et il ne sert pas à préciser la forme ni le genre de la ressource.
Exemple	<code>&lt;dc:subjectxml:lang= "fr"&gt;</code> Homme -- Effets du climat <code>&lt;/dc:subject&gt;</code>
Statut	Obligatoire si applicable

<code>&lt;dc:title&gt;</code> : Titre	
Définition	Nom donné à la ressource.
Commentaires	S'il existe un ou plusieurs titre(s) parallèle(s) dans une autre langue, on crée pour chacun d'eux un <code>&lt;dc:title&gt;</code> avec un attribut « <code>xml:lang</code> ».
Exemple	<pre>&lt;dc:title xml:lang="fr"&gt;Bibliographie internationale des sciences historiques&lt;/dc:title&gt;  &lt;dc:title xml:lang="ita"&gt;Bibliografia internazionale delle scienze storiche&lt;/dc:title&gt;</pre>
Statut	Obligatoire et répétable

<code>&lt;dc:contributor&gt;</code> : Contributeur / Collaborateur	
Définition	Entité responsable de contributions au contenu de la ressource
Commentaires	<ul style="list-style-type: none"> <li>- Les exemples de contributeur comprennent une personne, une organisation ou un service.</li> <li>- En plus du nom du contributeur, cet élément peut contenir ses dates biographiques.</li> <li>- on crée autant d'éléments <code>&lt;dc:contributor&gt;</code> que d'entités responsables de contributions au contenu de la ressource</li> </ul>
Exemple	<pre>&lt;dc:contributor&gt;Browne, Hablot Knight (1815-1882). Illustrateur&lt;/dc:contributor&gt;</pre>
Statut	Recommandé si applicable

<code>&lt;dc:creator&gt;</code> : Créateur	
Définition	L'entité principalement responsable de la création du contenu de la ressource.
Commentaires	<ul style="list-style-type: none"> <li>- On crée autant d'éléments <code>&lt;dc:creator&gt;</code> que de co-auteurs.</li> <li>- Outre le nom et le prénom, cet élément contient, si elles sont disponibles, les dates de naissance et le cas échéant de décès de la personne.</li> </ul>
Exemple	<code>&lt;dc:creator&gt;Dickens, Charles (1812-1870)&lt;/dc:creator&gt;</code>
Statut	Obligatoire si applicable

<code>&lt;dc:publisher&gt;</code> : Éditeur	
Définition	L'entité responsable de la mise à disposition ou diffusion de la ressource.
Commentaires	<ul style="list-style-type: none"> <li>- Les exemples d'éditeur comprennent une personne, une organisation ou un service.</li> <li>- Le nom de l'éditeur peut être complété par la mention du lieu d'édition entre parenthèses.</li> <li>- Le lieu de l'édition est limité à la ville d'édition, les adresses plus précises ne sont pas retenues.</li> </ul>
Exemple	<code>&lt;dc:publisher&gt;Crépin-Leblond (Moulins)&lt;/dc:publisher&gt;</code>
Statut	Obligatoire si applicable

<code>&lt;dc:rights&gt;</code> : Gestion des droits.	
Définition	Informations sur les droits associés à la ressource. L'information sur les droits englobe souvent les droits de propriété intellectuelle.
Commentaires	<p>Cet élément peut indiquer ou bien :</p> <ul style="list-style-type: none"> <li>- Les restrictions portant sur l'utilisation de la ressource.</li> <li>- Le titulaire des droits auxquels il est nécessaire de s'adresser pour connaître les conditions d'utilisation de la ressource.</li> <li>- Un URL pointant vers un document décrivant les droits applicables à la ressource.</li> </ul>
Exemple	<code>&lt;dc:rightsxml:lang="fre"&gt;domaine public&lt;/dc:rights&gt;</code>
Statut	Obligatoire

<code>&lt;dc:date&gt;</code> : Date	
Définition	Date de la création ou de mise à disposition de la ressource.
Commentaires	<p>Cet élément doit être utilisé pour indiquer la date de création ou publication de la ressource. Les dates qui ne sont pas celles de la production du document ne sont pas à intégrer dans l'élément.</p> <p>Il est recommandé d'encoder la valeur de la date conformément au profil défini dans la norme ISO 8601, qui comprend (notamment) des dates suivant la forme AAAA-MM-JJ.</p>
Exemple	<code>&lt;dc:date&gt;1883&lt;/dc:date&gt;</code>
Statut	Obligatoire si applicable

<code>&lt;dc:format&gt;</code> : Format	
Définition	La manifestation physique ou numérique de la ressource.
Commentaires	<ul style="list-style-type: none"> <li>- L'élément format peut comprendre les dimensions de la ressource</li> <li>- L'élément format peut être utilisé pour décrire le format sous lequel les documents entiers sont accessibles, même si, dans l'interface de consultation, d'autres formats peuvent intervenir (txt, png...).</li> </ul>
Exemple	<code>&lt;dc:format&gt;text/xml&lt;/dc:format&gt;</code>
Statut	Obligatoire

<code>&lt;dc:identifiant&gt;</code> : Identifiant de la ressource	
Définition	Référence univoque à la ressource dans un contexte donné.
Commentaires	Il est recommandé d'identifier la ressource au moyen d'une chaîne de caractères ou d'un numéro conforme à un système formel d'identification.
Exemple	<code>&lt;dc:identifiant&gt;ISBN 9783447054676&lt;/dc:identifiant&gt;</code>
Statut	Obligatoire et répétable

<code>&lt;dc:language&gt;</code> : Langue	
Définition	La langue du contenu intellectuel de la ressource.
Commentaires	<p>Cet élément s'applique aux ressources textuelles (langue dans laquelle elle a été écrite) ainsi qu'aux ressources audio et vidéo qui comportent des passages parlés.</p> <p>Pour les œuvres multilingues, on crée autant de <code>&lt;dc:language&gt;</code> que de langues utilisées [guideDC 2008].</p> <p>Il est recommandé d'utiliser une des valeurs définies dans la norme ISO 639 [ISO639]<sup>9</sup>, qui définit des codes de langues primaires à deux et à trois caractères. Exemples : " en " ou " eng " pour l'anglais, " fr " ou " fre " pour le français.</p>
Exemple	<code>&lt;dc:language&gt;ger&lt;/dc:language&gt;</code> [le document est écrit en allemand]
Statut	Obligatoire si applicable

#### IV. Le Dublin Core qualifié :

Les quinze éléments de base de Dublin Core sont considérés comme un dénominateur commun et pour la plupart des cas sont insuffisamment précis. Les éléments de base ont été étendus (ou précisés) par un ensemble d'autres termes, parfois appelés *qualifiants*.

Le DCMI a défini trois éléments supplémentaires étendant l'ensemble des quinze éléments de base et une trentaine de raffinements d'éléments.

<sup>9</sup>ISO 639 code pour la représentation des noms de langues, codage sur trois caractères

**1. Les éléments supplémentaires :**

<dcterms: audience>	
Définition	Une classe d'entités à qui la ressource est destinée ou est utile.
Exemple	<dcterms: audience>Enseignants du primaire</dcterms: audience>

<dcterms: provenance>	
Définition	Une déclaration des changements significatifs de propriété et de garde de la ressource depuis sa création pour son authenticité, son intégrité et son interprétation.
Exemple	<dcterms :provenance>Cette copie a appartenu à Benjamin Spock</dcterms :provenance>

<dcterms: rightsHolder>	
Définition	La personne physique ou morale possédant ou gérant les droits sur la ressource.
Exemple	<dcterms: rightsHolder>University of Bath</dcterms: rightsHolder>

**2. Les éléments de raffinement ou qualifiants :**

Aux éléments du Dublin Core viennent s'ajouter des qualifiants qui précisent encore l'usage des éléments. Par exemple pour dire que la date mentionnée est une date de modification du document sa représentation est comme suit :

<dcterms: modified>1992-03-13 <dcterms: modified>

Le tableau suivant résume tous les qualifiants et leurs significations [Pleade 2005]:

Qualifié	Qualifiants	Signification
<dc :title>	<dcterms: alternative>	Comprend toute autre forme de titre
<dc :description>	<dcterms: tableOfContents>	Table des matières
	<dcterms: abstract>	Résumé
<dc : rights>	<dcterms: accessRights>	Restrictions de l'accès à la ressource
<dc : relation>	<dcterms: isVersionOf>	Est une version de (édition, adaptation, traduction): changement dans le contenu
	<dcterms: hasVersion>	Contient d'autres versions
	<dcterms: isReplacedBy>	Est remplacé par
	<dcterms: replaces>	Remplace
	<dcterms: isRequiredBy>	Est requis par
	<dcterms: requires>	Requiert
	<dcterms: isPartOf>	Est une partie de
	<dcterms: hasPart>	A comme partie
	<dcterms: isReferencedBy>	Est référencé par
	<dcterms: references>	Référence
	<dcterms: isFormatOf>	Est un autre format de (changement sur le format par sur le contenu)
	<dcterms: hasFormat>	A pour autre format
<dc : coverage>	<dcterms: spatial>	Couverture spatiale
	<dcterms: temporal>	Couverture temporelle
<dc :audience>	<dcterms: mediator>	
	<dcterms: educationLevel>	
<dc :date>	<dcterms: created>	Date de création
	<dcterms: valid>	Date de validité
	<dcterms: available>	Date de disponibilité
	<dcterms: issued>	Date de parution
	<dcterms: modified>	Date de modification
	<dcterms: dateAccepted>	Date d'acceptation (thèse, article...etc).
	<dcterms: dateCopyrighted>	Date du copyright
	<dcterms: dateSubmitted>	Date de soumission
<dc : identifier>	<dcterms: bibliographicCitation>	Référence bibliographique de la ressource
<dc : format>	<dcterms: extent>	Étendue de la ressource (taille, durée)
	<dcterms: medium>	Support

## **V. Forces de Dublin Core :**

Les éléments de base du Dublin Core répondent à des critères de simplicité et de concision. Ainsi, tous les types de créateurs peuvent les utiliser.

Chacun des éléments possède un ensemble de qualificatifs et d'attributs normalisés que l'on peut employer pour restreindre davantage leurs sens.

L'identification des champs est conçue à partir d'une sémantique interdisciplinaire. Par exemple, dans la catégorie de la propriété intellectuelle, on privilégie le terme « créateur », ce qui permet de faire autant allusion à un auteur de texte qu'à un photographe ou un illustrateur.

Son envergure est internationale grâce à ses nombreux représentants qui s'impliquent sur presque tous les continents. Il existe au moins une vingtaine de versions linguistiques du Dublin Core à ce jour [aifbd 2012].

## **VI. Dublin Core et les manuscrits arabes :**

La norme Dublin Core offre un encodage simple des documents. La description se fait de manière marginale, de sorte qu'une notice DC contient les champs nécessaires afin de repérer les documents sans prendre en considération leurs structures et leurs caractéristiques.

Mais ce qu'on vient de dire ne se considère pas comme étant un obstacle devant l'utilisation de cette norme pour l'encodage des manuscrits arabes, il peut avoir recours à des champs supplémentaires pour décrire ces documents avec un minimum de précision dans le but de l'enrichir afin qu'elle s'adapte à la description de ces manuscrits et toutes leurs particularités.

## **VII. Conclusion**

La description des métadonnées selon le Dublin Core se présente donc sous la forme d'un ensemble de couples « propriété – valeur ». Ces valeurs peuvent être données dans un format libre (chaînes de caractères interprétables par des humains, mais sans signification particulières pour les machines),

Les éléments du Dublin Core fournissent un vocabulaire de concepts avec des définitions en langage courant.

Le Dublin Core permet une interopérabilité minimale entre les systèmes conçus indépendamment les uns des autres.

**Deuxième partie : Mise en œuvre de  
la solution d'interopérabilité**

**Chapitre 1 : Enrichissement des formats  
d'encodages**

**Chapitre 2 : Conception et implémentation**



**Chapitre 1 : Enrichissement des  
formats d'encodages**

## ***Introduction***

L'accès aux catalogues encodés sous différents formats d'encodage (EAD, TEI, Dublin Core ...etc.) se retrouvant sur de différents serveurs, oblige l'utilisateur à interroger l'un après l'autre. Chacun serveur est susceptible de contenir les informations recherchées. Devant la multiplication des standards d'encodages, nous serions amenés à concevoir un langage d'interrogation spécifique à chaque format. Cette solution est complexe et fastidieuse à la fois, rendant le système d'interrogation instable et fortement conditionné à la base à interroger. La recherche d'un outil d'interrogation standard et invariant par rapport à la mise à jour de la base de données.

La mise en œuvre de ce type de solution nécessite la conception d'un middleware dont le rôle principal est de convertir chaque copie des notices à partir d'un format d'encodage source vers un format d'encodage cible. Le format source peut varier selon les choix des institutions. Par contre, le format cible constitue un dénominateur commun à tous les acteurs.

Dans notre travail, le middleware est le format Dublin Core, un format hautement utilisé sur le Web et très représentatif des demandes utilisateur, alors que les formats d'encodage source sont limités à la TEI et à l'EAD, qui sont les formats les plus utilisés pour l'encodage des manuscrits arabes anciens.

Par ailleurs, l'intégration d'un nouveau format d'encodage source ne modifie en rien la solution d'interrogation de la base de données. Ainsi, notre objectif est de mettre en place une base documentaire, formée par les notices descriptives des manuscrits Arabes anciens, réalisant l'interopérabilité des différents formats d'encodage des manuscrits Arabes anciens.

### ***I. Pourquoi employer Dublin Core comme format de médiateur ?***

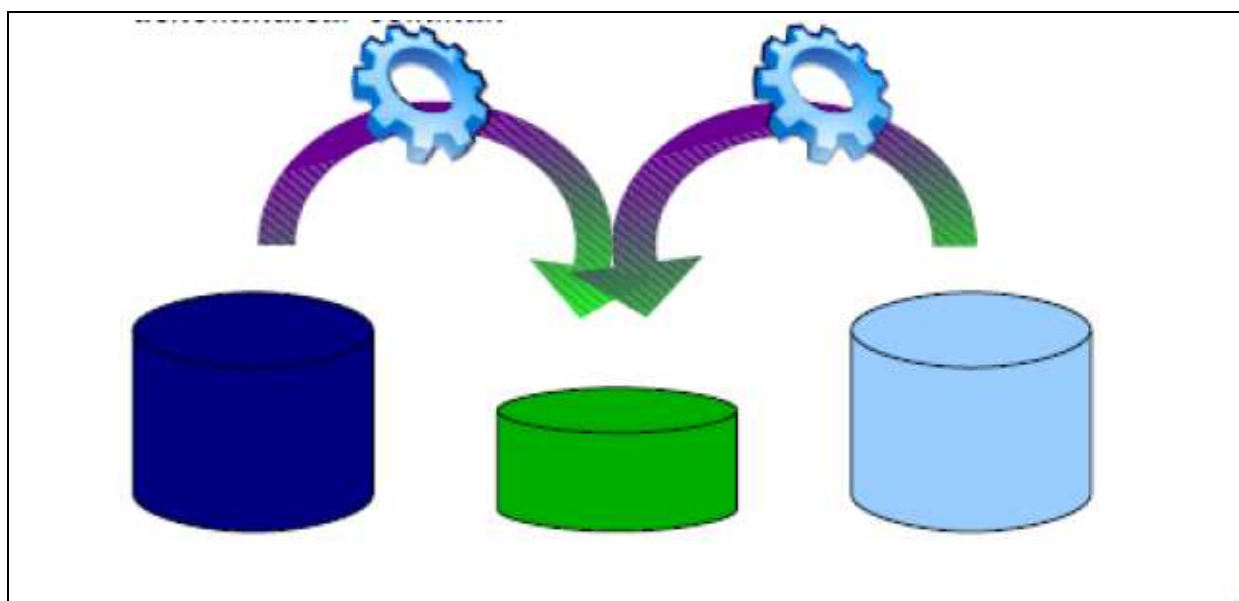
Dublin Core fournit un socle commun d'éléments descriptifs pour améliorer le signalement et la recherche de ressources au-delà des diverses communautés et des nombreux formats descriptifs propres à chaque spécialité, tout en restant suffisamment structuré [Solari 2008].

Le principal atout du Dublin Core, c'est l'interopérabilité, et c'est pour cela que l'OAI-PMH impose le Dublin Core, partant du principe qu'il permet de décrire tout type de ressources.

C'est vrai, le Dublin Core est un vrai atout pour l'interopérabilité car il permet à une multitude de gens d'exprimer la même chose de la même manière, à un niveau de granularité plus fin. Au lieu que chacun réinvente la roue et crée sa propre propriété "titre", ou "auteur", on utilise le Dublin Core, qui permet au moins, pour un large ensemble de ressources décrites, de repérer assez facilement les "titres" et les "auteurs". Ensuite, on complète avec d'autres propriétés créées exprès ou prises dans d'autres vocabulaires.

## 1. Le protocole OAI-PMH et le Dublin Core

Le protocole OAI-PMH, est mis en place dans les années 1990. Il propose une approche qui implique dès sa conception une préoccupation de convergence entre les données de la recherche. Afin d'assurer l'interopérabilité entre les différentes sources de données, il exige l'utilisation d'un format de données minimal commun, le Dublin Core. Les données ainsi récoltées, c'est-à-dire récupérées dans les bases réparties, pour être versées dans une base commune qui servira de support à la création de nouveaux services. On parlera alors d'interopérabilité basée sur le plus petit dénominateur commun (*smallest common denominator interoperability*) [Bermes 2011].



**Figure 14:** Interopérabilité basée sur le plus petit dénominateur commun [Bermes 2011].

Nous remarquons que cette approche présente des limites, vue que les différentes sources sont obligées de simplifier leurs données pour les faire entrer dans ce format commun, ce qui causera soit la suppression de nombreuses informations (perte d'information), soit leur concaténation dans des champs de métadonnées généralistes qui seront difficiles à exploiter.

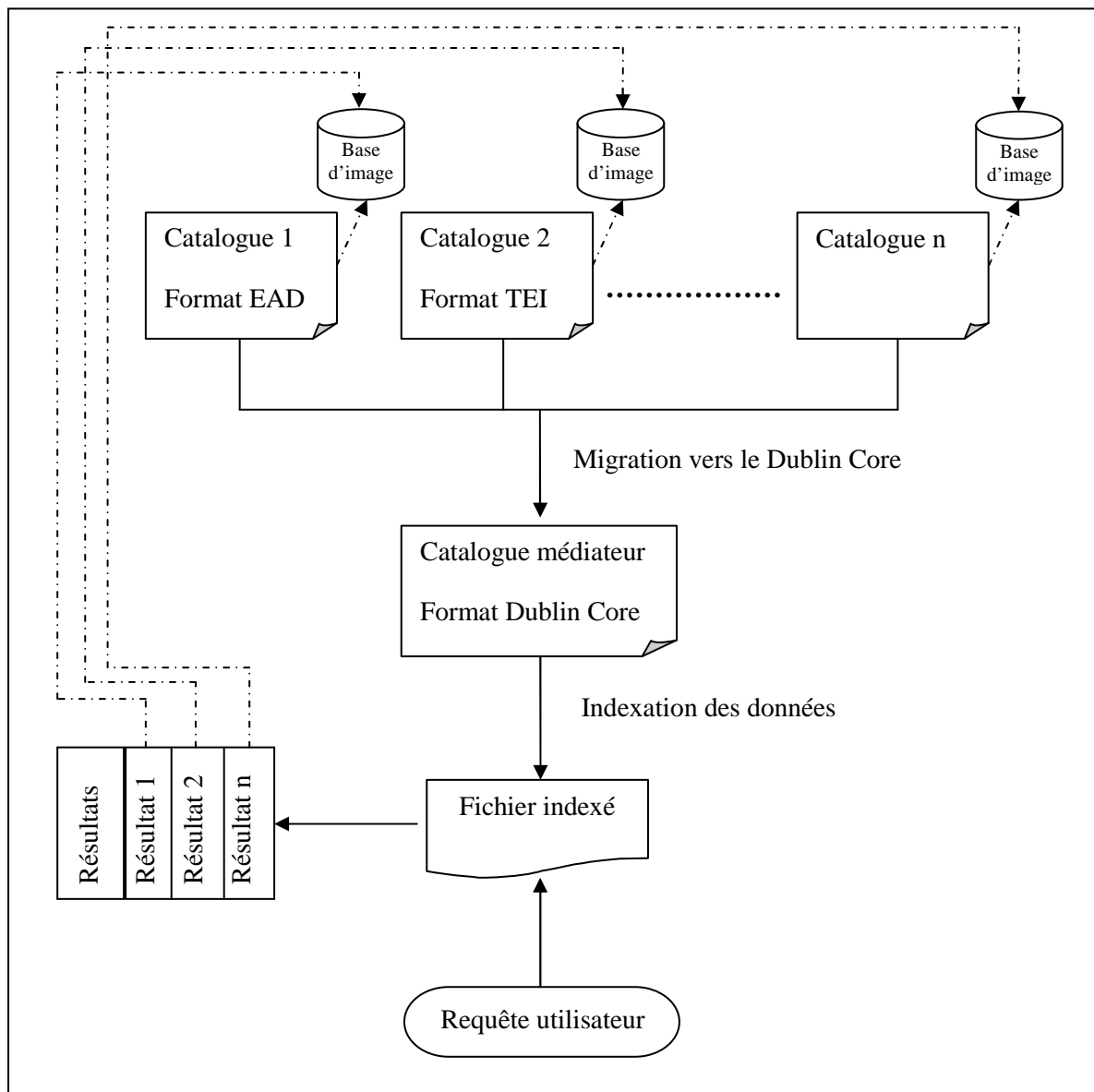
## II. Étude de l'accès aux manuscrits numérisés

Nous nous sommes inspirés du modèle de protocole OAI-PMH et le rôle du noyau de Dublin Core en fournissant l'interopérabilité pour ce protocole. L'idée est que le noyau de Dublin Core prend le rôle d'un ensemble de minimum, qui est employé comme un pont entre les différents catalogues dans divers formats.

La stratégie est de produire des notices bibliographiques qui doivent être décrites selon le format de métadonnées de Dublin Core à partir des catalogues de base existants.

L'accès à ces notices se fait selon une *recherche libre*. Afin de rendre ces descriptions des documents accessibles et récupérables sur les différents serveurs, chaque description de document doit inclure une URL pointant vers le document décrit, dans sa version numérique.

Le schéma ci-dessus décrit la structure de l'utilisation du noyau de Dublin Core pour soutenir les besoins d'interopérabilité des catalogues des manuscrits arabes digitalisés :



**Figure 15 :** Structure de la solution de l'interopérabilité des catalogues en utilisant Dublin Core.

Par ailleurs, la production des notices bibliographiques décrites selon le format de métadonnées de Dublin Core à partir des catalogues de base existants, n'offre pas une

description complète des manuscrits arabes vue que les catalogues de base déjà ne décrivent pas les manuscrits arabes d'une façon exhaustive comme nous avons mentionné dans le deuxième chapitre de la première partie du mémoire. Pour cela un enrichissement des normes EAD et TEI au préalable est inévitable afin de faire migrer les données à partir de ces deux standards vers le Dublin Core, pour que cette dernière présentera les manuscrits arabes d'une façon exhaustive. L'enrichissement des ces différentes normes constituera le contenu des sections qui viennent. Cette procédure d'enrichissement va se baser également sur l'ensemble des métadonnées du protocole proposé par [Soualah&Hassoun 2011] décrit dans le chapitre 2 de la première partie de ce mémoire. Ce dernier, relève les différentes particularités et caractéristiques des manuscrits arabes anciens.

### ***III. Enrichissement de la TEI-ms et son adaptabilité à la description des manuscrits arabes***

Nous avons déjà dit que la TEI-ms malgré tous les outils qu'elle a apporté à la description des manuscrits, elle souffre encore de quelques imperfections en ce qui concerne la description et le catalogage des manuscrits arabes anciens.

Par conséquent, la solution d'enrichir cette norme et la rendre capable de décrire toutes les particularités des manuscrits arabes anciens, est une nécessité à mettre en œuvre. En effet, le sujet d'enrichissement de la TEI-ms a été déjà traité par monsieur [Soualah 2008]. Ce dernier a proposé des modifications et des ajouts au contenu de cette norme en lui permettant de décrire amplement les manuscrits arabes.

### ***IV. Enrichissement de l'EAD et son adaptabilité à la description des manuscrits arabes***

Du fait que le but de sa création a été le catalogage des documents d'archives, la norme EAD présente des insuffisances par rapport à la description des manuscrits arabes anciens et leurs différentes caractéristiques.

Par conséquent, il serait intéressant d'apporter des enrichissements à l'EAD dans le but de l'adapter aux différents traits des manuscrits arabes anciens.

C'est ce point là que nous allons traiter dans ce qui suit, de façon à proposer un enrichissement de la structure de l'EAD, on apportant des modifications et des nouveaux éléments essentiels pour la description des manuscrits arabes, en se basant sur le protocole décrit précédemment.

#### ***1. Structure des noms arabes anciens***

La complexité de la structure du nom arabe présente une grande variation, à savoir "Ism""Laqab""Kunya""Nisba", l'EAD ne prend pas en considération tous ces détails. Afin que l'accès au manuscrit numérisé se fasse immédiatement avec cette norme nous proposons ce qui suit :

L'élément <persname> correspondra à "Ism" et l'élément <famname> correspondra à son tour à "Laqab", et enfin l'élément <name> contiendra l'attribut TYPE avec les valeurs "Kunya" et "Nisba" comme suit :

<name TYPE="Kunya">

<name TYPE="Nisba">

## 2. Le copiste

Les manuscrits arabes possèdent une nouvelle mention de responsabilité qui est celle du "copiste". Le nom de l'auteur figure au début du texte et/ou sur la page de titre, celui du copiste au colophon [kaileh 2004] que l'EAD ne prend pas en considération.

Dans l'élément <altformavail> qui fournit des informations sur les reproductions des unités documentaires, nous rajoutons un nouvel élément <colophon> qui va contenir les informations sur le copiste (nom, lieu, date...etc).

**Remarque :** Notons que les éléments qui identifient le nom seront rajoutés au nom de l'auteur ainsi qu'au nom du copiste comme suit :

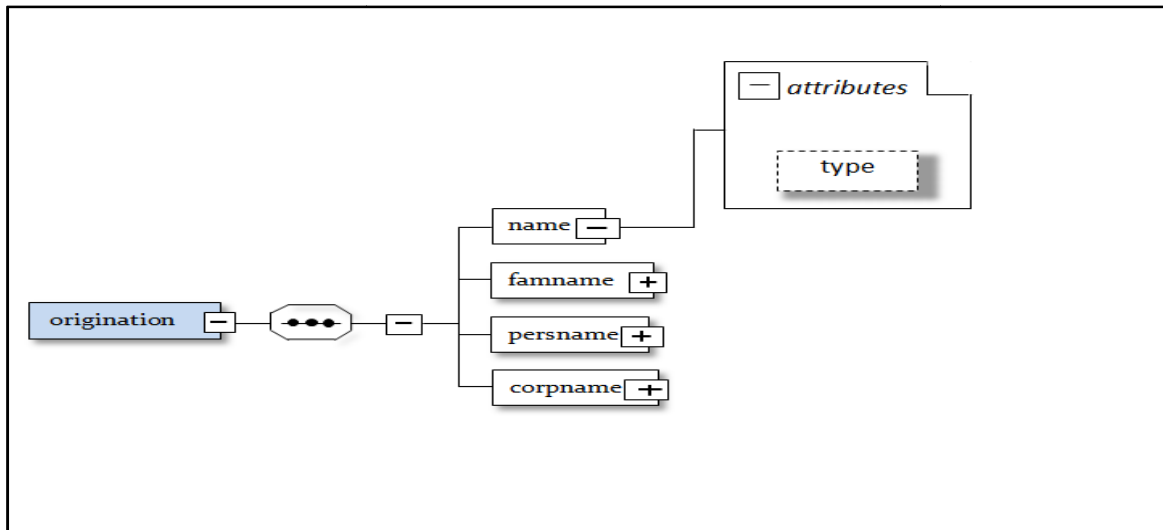


Figure 16 : Structure de l'élément <origination>.

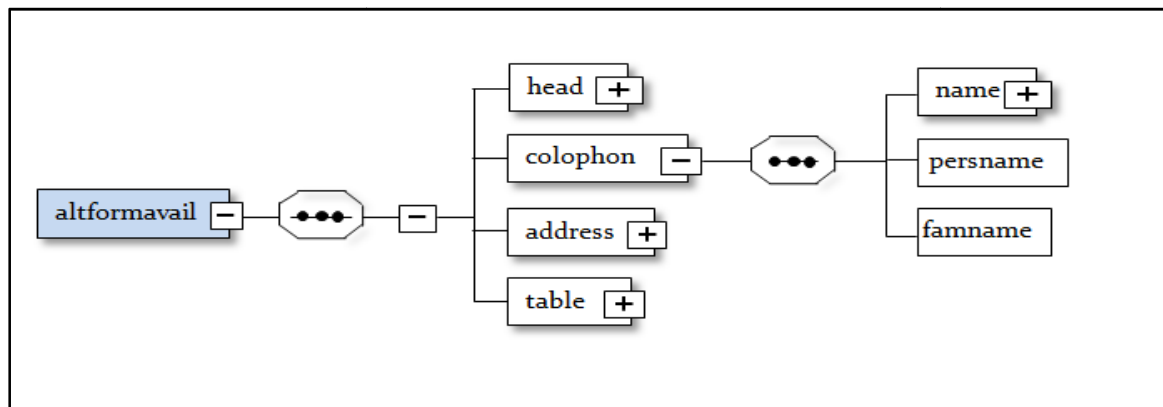


Figure 17 : Structure de l'élément <altformavail>.

### 3. Système de datation

Le système de datation le plus souvent utilisé dans les manuscrits arabes utilise le jour de la semaine, le jour du mois, l'année (julienne, hégirienne ou grégorienne) et parfois même le siècle. L'EAD prend tous ces aspects en considération sauf pour le jour de la semaine pour ce faire on rajoute l'élément <day>.

Toutefois, il arrive que seule le jour ou le mois ou l'année qui est connus et même des fois seulement le siècle soit connu. Dans ce cas nous introduisons un ensemble d'éléments <month>, <year> et <century>, qui indiqueront respectivement juste le mois, l'année et le siècle.

**Exemple :** Pour coder les deux dates de création d'un manuscrit :

08 شعبان 1004 هـ / 1595 م  
ق 12 هـ

```
<unitdate CALENDAR="hegyre" CERTAINLY="effective">
<day>08</day>
<month>شعبان</month>
<year>1004 </year>
</unitdate>
<unitdate CALENDAR="grégorien" CERTAINLY="effective">
<year>1595</year></unitdate>

<unitdate>CALENDAR="hegyre" CERTAINLY="approximative" >
<century>12</century>
```

**Remarque :** Notons que les éléments qui identifient la date seront rajoutés à date de naissance et de décès de l'auteur dans l'élément <bioghist> avec les deux éléments <dateBirth> et <dateDeath>, ces deux derniers éléments seront rajoutés aussi dans l'élément <colophon> afin d'identifier la date de naissance et de décès du copiste comme l'illustre les schémas suivants :

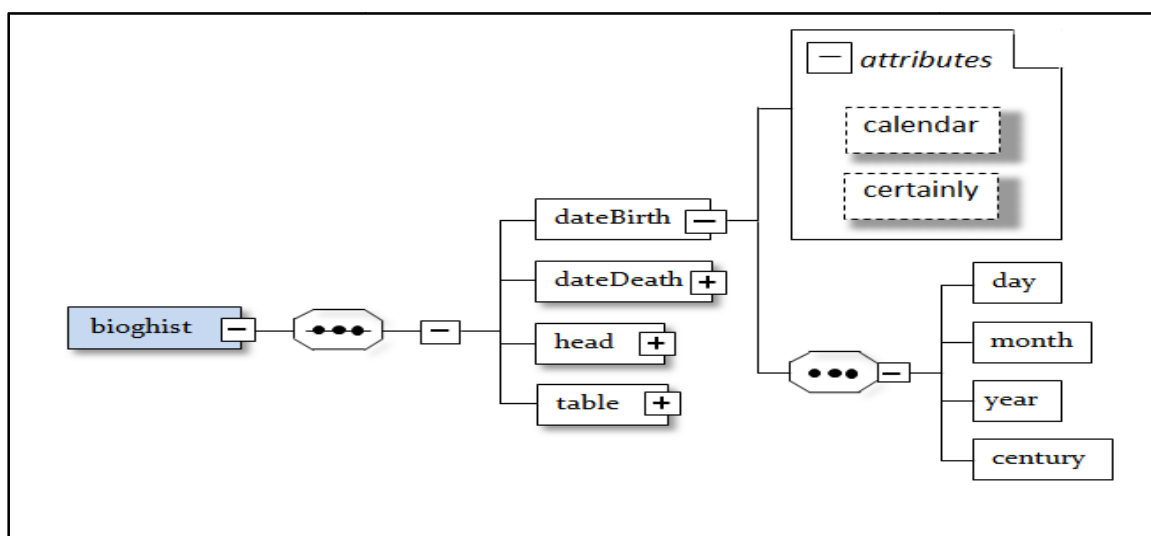


Figure 18 : Structure de l'élément <bioghist>.

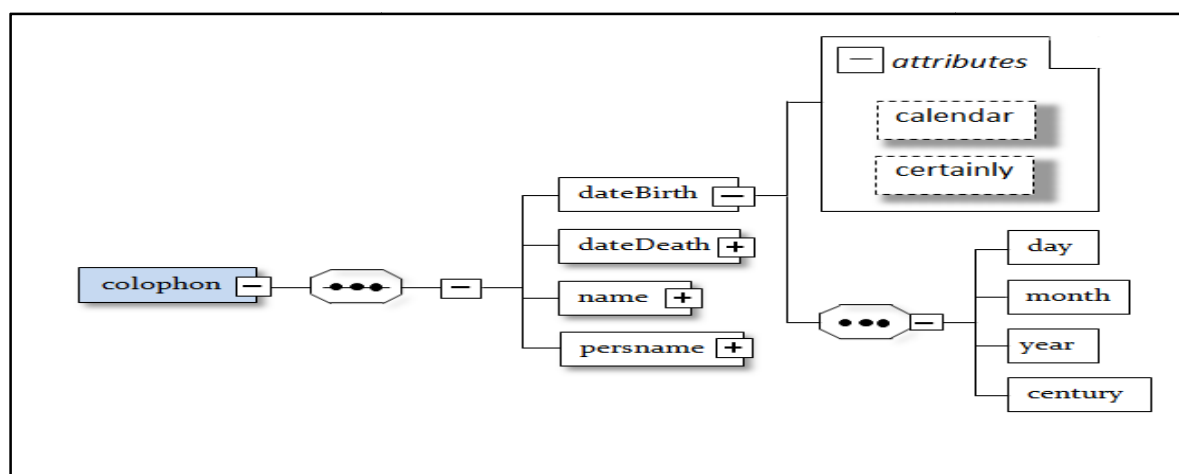
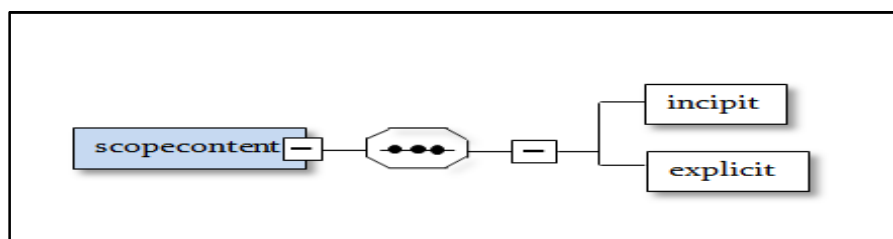


Figure 19 : Structure de l'élément <colophon>.

#### 4. Les aspects additifs relatifs au contenu

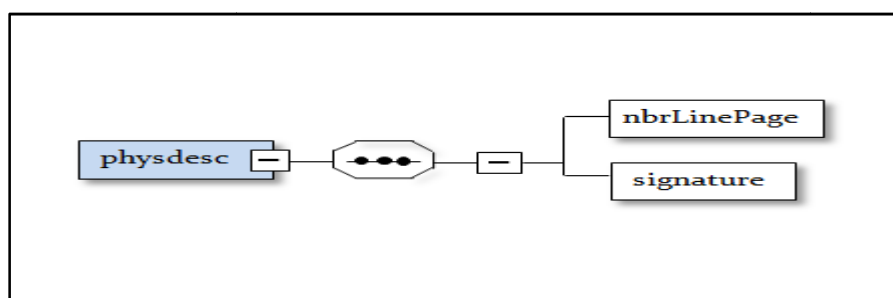
Parmi les caractéristiques majeures des anciens manuscrits arabes, on trouve "l'incipit" (c'est à dire la première phrase du texte) ainsi que "l'explicit" (la dernière phrase du texte manuscrit), sont des éléments de recherche et d'identification très importants, mais que l'EAD ne prend pas en considération. Pour ceci nous proposons de rajouter dans l'élément <scopecontent>(présentation du contenu) les deux éléments <incipit> et <explicit> qui représentent la première et la dernière phrase du texte manuscrit.



**Figure 20:** Structure de l'élément <scopecontent>.

Les manuscrits arabes sont caractérisés par le nombre de lignes par page et leurs signatures, ces aspects que l'EAD ne traite pas.

Pour ce faire nous rajoutons deux éléments <nbrLinePage> et <signature> pour décrire respectivement le nombre de lignes par page et la signature du manuscrit, sous l'élément <physdesc> qui permet de regrouper des informations concernant l'aspect ou la structure matérielle des unités documentaires décrites.



**Figure 21:** Structure de l'élément <physdesc>.

## 5. La transmission de savoir

L'EAD ne prend pas en considération la transmission de savoir des manuscrits arabes, définie dans la première partie du premier chapitre.

Nous avons créé sous l'élément <originalsloc> un élément <certify>, contenant l'attribut "type", qui peut prendre les valeurs "listner" (écoute=samâ), "transmission" (transmission=ijaza), "reader" (lecteur=qarî) ou "comparison" (collation=muqabalat).

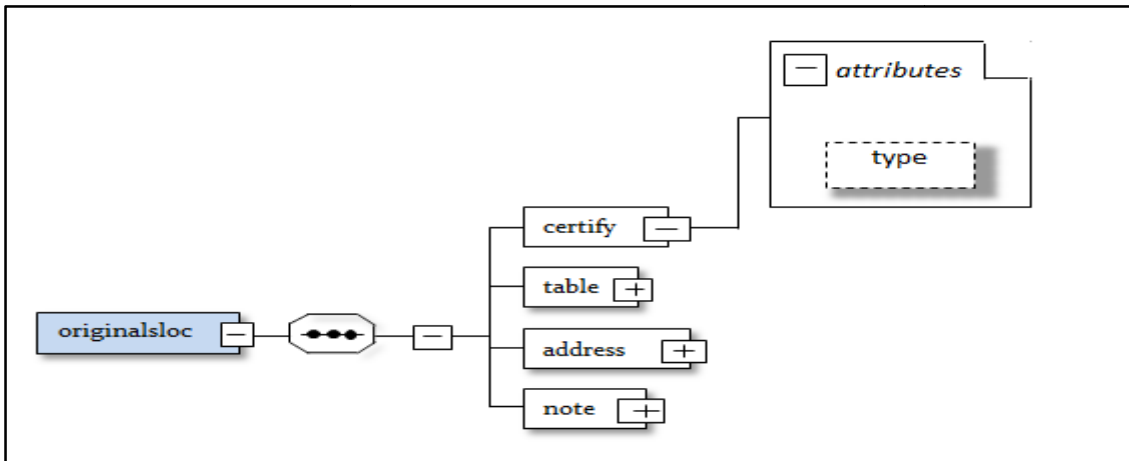


Figure 22 : Structure de l'élément <originalsloc>.

## 6. Autres éléments descriptifs

### 6.1. Référence de la copie du manuscrit dans d'autres bibliothèques :

Sous l'élément <imprint> nous avons rajouté l'élément <copie> qui contient l'attribut type qui prend la valeur, "true" si la copie est imprimée, "false" sinon, et par conséquent <namePrinter> pour le nom de l'imprimeur.

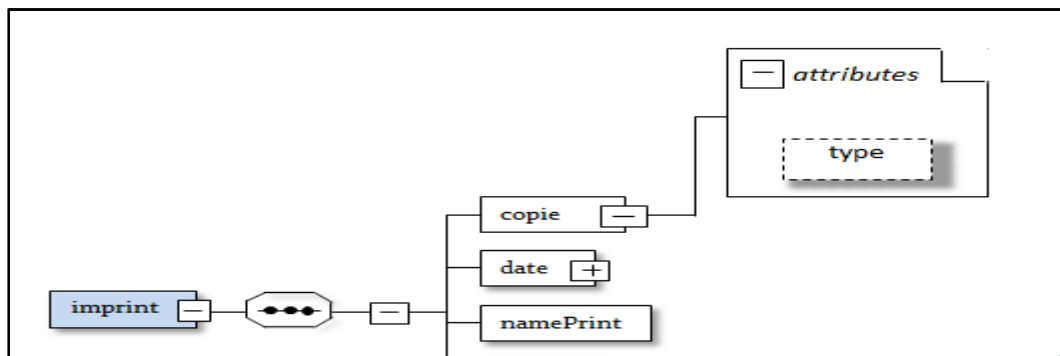


Figure 23 : Structure de l'élément <imprint>.

### 6.2. Informations sur les copies et leurs références

Sous l'élément <altformavail> on rajoute l'élément <refCopie> pour désigner la référence de la copie et sous l'élément <extent> qui fournit des informations sur le nombre de documents décrits et leurs mesures, donc la structure actuelle de <extent> telle que proposée par EAD, prête à confusion et rend la fonction de recherche d'information plus ou moins délicate. C'est pour cela que nous proposons d'introduire dans cet élément deux sous-éléments <nbrCopie> pour préciser juste le nombre de copies qui existe, et <volNum> qui décrit le numéro du volume.

### 6.3. Références de publication et de traductions

Sous l'élément <publicationStm< (désigne la mention de publication décrit sous l'élément <fildesc> qui a son tour se trouve sous l'élément <eadheader>) nous rajoutons l'élément <extref> pour référencer la mention de publication, ainsi nous rajoutons sous l'élément <archdesc> les deux éléments <translations> et <refTranslat> pour décrire les traductions et leurs référence.

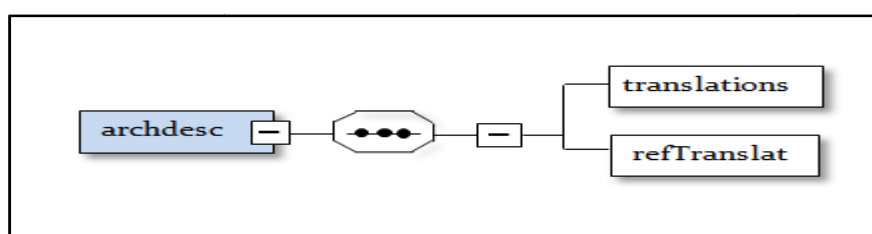


Figure 24 : Structure de l'élément <archdesc>.

### 6.4. Autre données descriptives

Sous l'élément <odd> (autre données descriptives) nous rajoutons l'élément <observation> pour désigner les observations, et l'élément <palimpsest> qui définira le palimpseste.

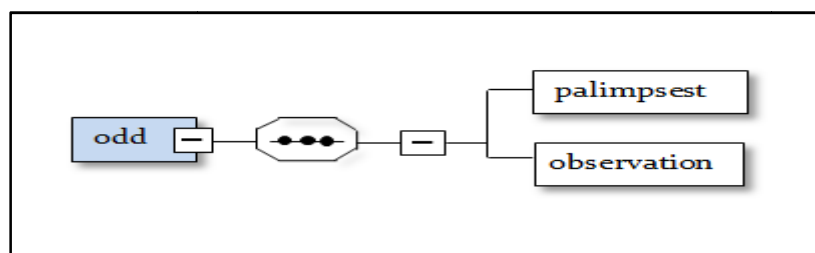


Figure 25 : Structure de l'élément <odd>.

### 6.5. Sommaire et table des matières

Rappelons que l'EAD réserve l'élément <scopecontent> l'essentiel des informations sur la portée des unités documentaires décrites et sur les sujets qu'elles couvrent. A cet effet, nous proposons de reprendre cet élément pour décrire sommairement le manuscrit, qui sera contenu sous le nouveau élément <summary>, sous cet élément on rajoute l'élément <fileplan> déjà existant sous <archdesc> fils <ead> et qui fournit des informations sur tout type de classement, nous proposons de rajouter sous <fileplan> l'ensemble d'éléments à savoir (<djuz>, <kytab>, <bab>, <fasl>), ces derniers seront à leur tour dotés d'un attribut nommé "value" qui prend ses valeurs dans l'ensemble {true, false}, pour signifier l'existence ou l'inexistence de l'élément dans la structure du manuscrit, dotés aussi des deux éléments <title> et <ptr> indiquerons le titre et sa référence. Le schéma suivant résume tous ces éléments.

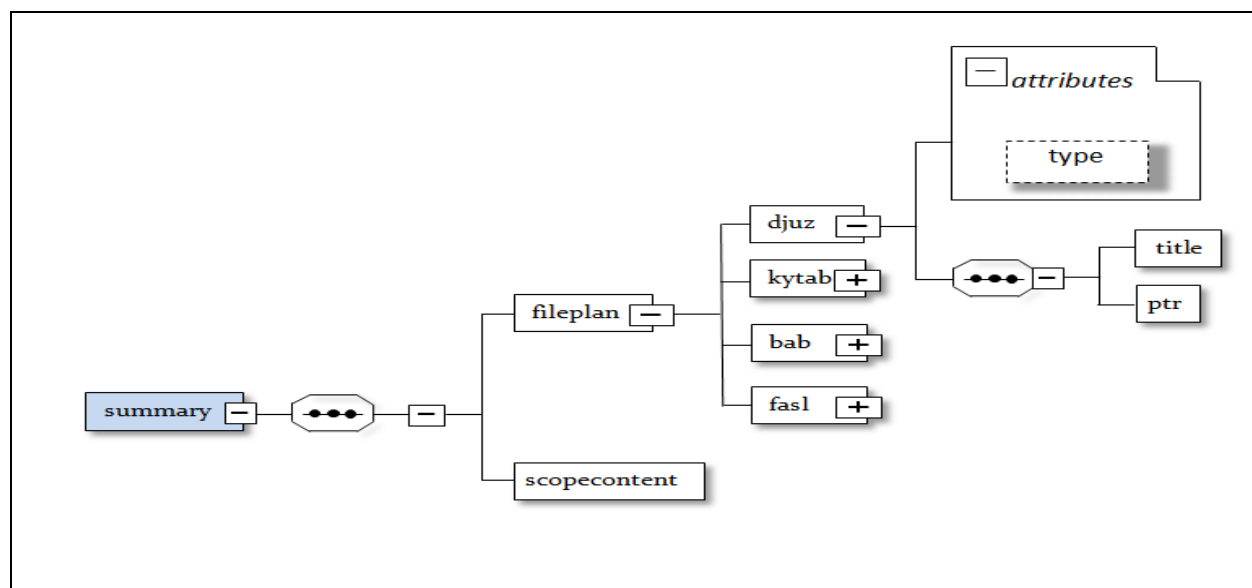


Figure 26 : Structure de l'élément <summary>.

## V. Enrichissement de Dublin Core et son adaptabilité à la description des manuscrits arabes

Le format d'encodage Dublin Core offre peu de marge de manœuvres pour le catalogage des manuscrits arabes anciens, vu le grand nombre de champs descriptifs sur les quelles se base le processus de description de ces manuscrits. Par contre, comme cité précédemment l'adaptation de ce format peut se faire selon un enrichissement de son contenu et de sa structure.

A cet effet, dans cette section nous allons recenser l'ensemble des aspects des manuscrits arabes anciens non décrits par le Dublin Core et nous proposerons de les intégrer dans celle-ci, tout en respectant la structure générale de Dublin Core. Ce qui constituera un enrichissement de cette dernière. Durant le processus d'enrichissement nous nous sommes basés sur les différentes métadonnées du protocole enrichi vu auparavant.

### 1. L'auteur

La description du nom de l'auteur tel que Dublin Core le définit avec l'élément <dc :créator> est très simple, elle ne s'applique pas sur les manuscrits arabes anciens, qui présentent une complexité accrue dans la description du nom d'auteur.

En effet, l'auteur peut se présenter comme l'un des meilleurs candidats afin d'assurer un accès uniforme et non ambigu aux manuscrits arabes numérisés. Malheureusement, l'accès à ces manuscrits via les noms des auteurs se voit un peu complexe vu la présentation des noms arabes anciens selon une grande variation et une structure différente. A cet effet, ce dernier se présente en générale sous forme de quatre éléments distincts : Ism, Laqab, Kunya et Nisba. Lors de la recherche d'un manuscrit, il est imprévisible de savoir à quel élément l'utilisateur

en fera appel, pour cela et afin d'assurer un accès aux manuscrits avec la prise en considération de la structure des noms arabes et par l'intermédiaire de n'importe quel éléments des quatre qu'on a cités auparavant. La réalisation de cet objectif passe impérativement par l'ajout des éléments : <name>, <ism>, <laqab>, <kunya> et <nisba> qui sont utilisés dans la TEI-ms enrichiet aussi dans EAD enrichi, leur présentation en Dublin Corese fait comme suit:

✚ <dc :name> : c'est un élément englobant, qui va être constitué des éléments :

1. <dc : ism> : Définit le « ism » de l'auteur.
2. <dc : laqab> : Définit le « laqab » de l'auteur.
3. <dc : kunya> : Définit la « kunya » de l'auteur.
4. <dc : nisba> : Définit la <nisba> de l'auteur.

La structure de <dc :creator> peut être présentée ainsi :

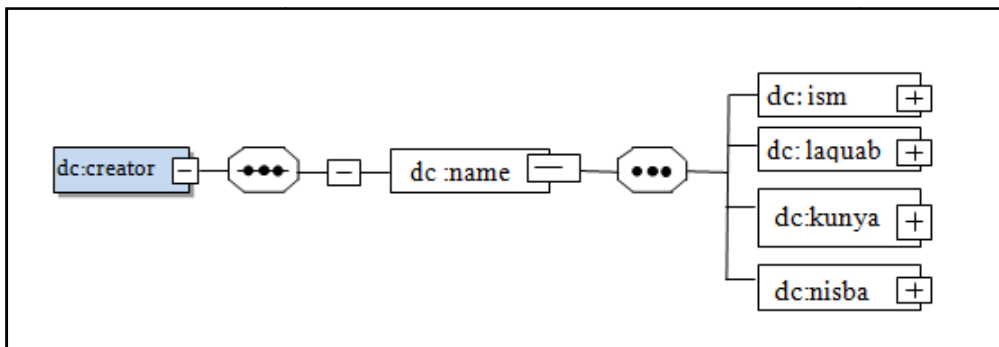


Figure 27: Structure de l'élément<dc :name>.

### Exemple

L'encodage du nom d'auteur « محمود بن عمر بن محمد بن احمد الخوارزمي الزمخشري , جار الله أبو القاسم »

se fait comme suit :

```

<dc :créator>
<dc :name>
<dc :isme> محمود</dc : isme>
<dc :laqab> بن عمر بن محمد بن احمد</dc : laqab>
<dc :kunya> الخوارزمي الزمخشري</dc : kunya>
<dc :nisba> جار الله أبو القاسم</dc : nisba>
</dc :name>
</dc :créator>
    
```

## 2. Système de datation

Le Dublin Core propose l'élément `<date>` et ces différents affinements pour définir les dates citées dans les documents encodés, cette date est présentée sous la forme AAAA-MM-JJ selon le calendrier grégorien.

Par ailleurs, en ce qui concerne les manuscrits arabes anciens, ils utilisent dans la plus part de temps un système de datation selon le calendrier hégire. Du fait que le Dublin Core ne propose aucun élément pour la prise en compte de ce genre des systèmes de datation, nous allons introduire l'attribut "*value*" utiliser dans la TEI-ms. Cet attribut pourra prendre les valeurs "*hégire*" et "*grégorien*" représentant respectivement le système de datation hégire et grégorien.

Les dates citées dans un manuscrit arabes peuvent être sous la forme JJ-MM-AAAA spécifiant le jour, le mois et l'année d'un événement précis, comme elle peut avoir juste le jour ou le mois ou bien l'année. Dans le but de présenter ces cas, Nous proposons aussi d'introduire ces nouveaux éléments issus de l'EAD enrichi `<dc : day>`, `<dc : month>` et `<dc : year>` qui servent à :

- ✚ `<dc : day>` : définir le jour.
- ✚ `<dc : month>` : définir le mois.
- ✚ `<dc : year>` : définir l'année.

### Exemple

L'encodage de la date 14 شعبان عام 1002 هـ se fait comme suit :

```

<dc :date value= " hegyre">
<dc :day>14</dc :day>
<dc :month>شعبان</dc :month>
<dc :year>1002</dc :year>
</date>
```

## 3. Ere de l'auteur

Les auteurs des manuscrits arabes anciens sont aussi décrits à côté de leurs noms par leurs dates de décès et la date de naissance selon le calendrier hégire ou grégorien. Par contre cet aspect n'est pas pris en considération par le Dublin Core qui utilise l'élément `<dc :date>` ainsi que ces affinements pour l'encodage des dates qui concerne le manuscrit lui même y compris : sa date de création, mais pas son auteur c'est-à-dire que ce format n'offre aucune précision sur les dates de naissance ou de décès des auteurs.

Afin d'améliorer la description fournie par le Dublin Core pour l'auteur de manuscrit nous proposons d'ajouter deux éléments de la TEI-ms qui sont `<dc :birth>` et `<dc :death>` pour reporter respectivement la naissance et le décès de l'auteur.

### <dc :birth>

<dc :birth>est un élément qui sert à décrire la date de naissance de l'auteur de manuscrit. Également avec l'insertion de l'attribut *value* ajouter précédemment pour le type de date hégire ou grégorien, cet élément peut contenir l'ensemble des éléments cités auparavant dans l'élément <dc :date> tel que :<dc :day>, <dc :month>,<dc :year>.

Parfois, les dates de naissances ou de décès des auteurs cités sur les manuscrits peuvent être vraie c'est-à-dire la date est effective et exacte, comme elles peuvent être approximatives et estimatives. Pour cette raison on propose d'introduire un attribut « *type* » qui aura la valeur « *effective* » pour signifier l'exactitude de la date saisie ou la valeur « *approximative* » pour une date approximative.

#### Exemple

```
<birth type = "effective" value = "gregore">
<year>981</year>
</birth>
```

### <dc :death>

Cet élément décrit la date de décès de l'auteur, avec la prise en compte de tous les attributs et les sous-éléments ajoutés pour la description de la date de naissance.

#### Exemple

```
<death type = "approximative" value = "hegyre">
<year>1573</year>
</death>
```

Toutefois il arrive que la date de naissance ou de décès d'un auteur puisse ne pas être connue, seulement le siècle de naissance ou de décès qui est motionné sur le manuscrit.

Pour cela, il est impératif d'ajouter un autre élément qu'on peut appeler <dc : century> pour l'indication de siècle.

#### Exemple

```
<death type = "approximative" value = "hegyre">
<century>•11 ق</century>
</death>
```

#### 4. Période d'existence de l'auteur

Notons qu'il arrive parfois de ne pas connaître la date de naissance de l'auteur, seule la période de son vécu est connue. Cette période peut être donnée sous forme de date, de siècle ou de période. Cet aspect fait défaut à Dublin Core, comme nous l'avons dit auparavant, ne le prend pas en considération.

Pour ce faire nous proposons d'introduire `<dc:wasAlive>` dans l'élément `<dc:créateur>` qui découle de la norme TEI-ms enrichi, pour signifier : "était vivant". Cet élément contiendra la période d'existence de l'auteur, il peut contenir les attributs suivants :

- + **"notBefore"** et **"notAfter"**, si l'élément `<dc:wasAlive>` définit un intervalle de date.
- + **"type"** qui prend les valeurs **"effective"** pour signifier que la date saisie est vraie ou **"approximative"** pour une date approximative.
- + **"evidence"** pour mentionner l'origine de la date qui peut être interne « **internal** » au manuscrit, externe « **external** » ou attribuée « **attributed** » par le catalogueur

De plus, l'élément `<wasAlive>` peut renfermer les éléments `<dc:century>`, `<dc:year>` et

`<dc:p>` pour définir respectivement un siècle, une année ou une date sous une forme quelconque.

Cette structure peut être représentée comme suit :

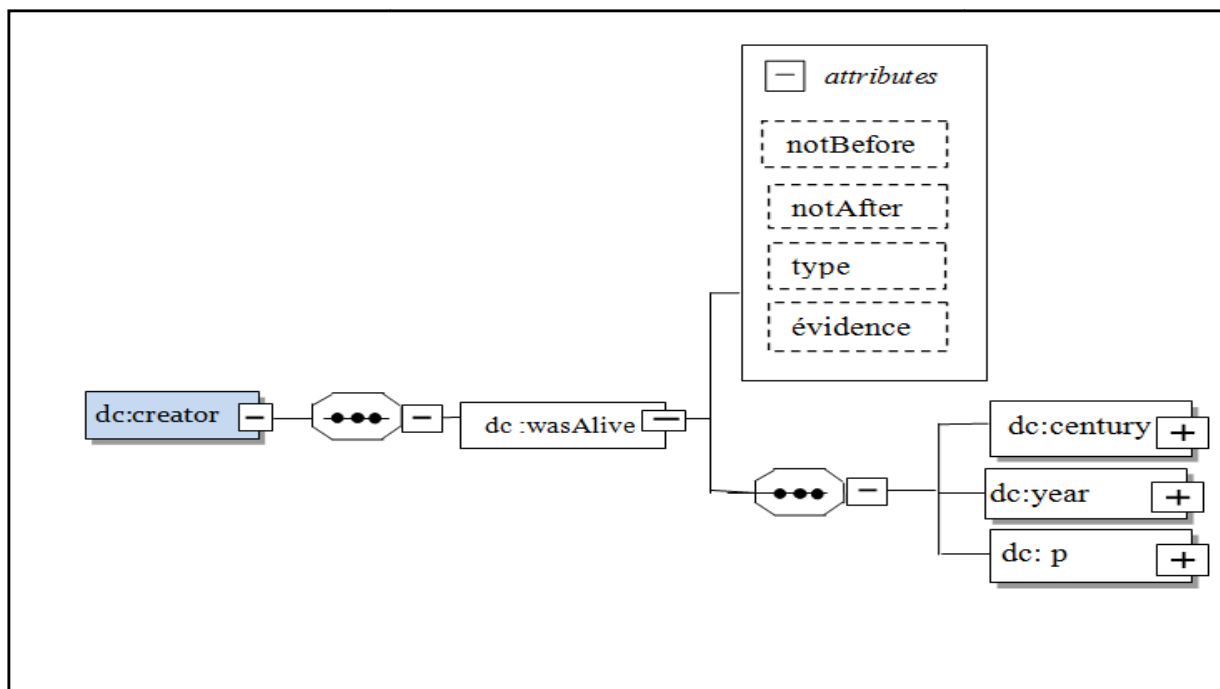


Figure 28 : structure de l'élément `<dc:wasAlive>`.

## 5. Nombres de copies de microfilm, nombres de folios, le nombre de lignes par page et taille de folios

Le Dublin Core propose un affinement dans l'élément <dc :format> appelé <dcterms :extent> qui sert à décrire la dimension ou la durée d'une ressource. Du fait de ses utilisations variées, cet affinement comprend généralement une valeur numérique et une légende nécessaire à l'interprétation de la valeur numérique. A cet effet, l'affinement <dcterms :extent> peut inclure les éléments suivants :

### 5.1. Nombres de copies de microfilm

Un manuscrit arabe ancien peut être conservé sous forme d'un microfilm qui est une photographie de très petit format qui permet de reproduire sur film l'image de ce manuscrit. Un lieu de conservation d'un manuscrit, une bibliothèque par exemple, peut avoir plusieurs copies de microfilm d'un manuscrit pour le mettre à disposition de plusieurs utilisateurs à la fois. Dans ce cas, l'ajout d'un élément pour l'identification de chaque copie est nécessaire. A cet effet, nous allons introduire à l'intérieur de <dcterms :extent> l'élément <dc :nbrCopie> issu de l'élément <nbrCopie> de l'EAD enrichis. Cet élément définit le nombre des copies microfilmées du manuscrit.

### 5.2. Nombre de folios

Le mot folios signifie les feuilles d'un ouvrage ou d'un manuscrit comprenant le recto et le verso, donc nombre de folios veut dire le nombre de feuillets de manuscrit. Le format Dublin Core n'offre pas un élément précis pour l'encodage de cet aspect. Pour cela, l'ajout d'un élément <dc :nbrFolios> à l'affinement <dcterms :extent> suffit afin d'indiquer le nombre de folios de manuscrit.

### Exemple

```
<dcterms :extent>
<dc :nbrFolios> ٦٦٧ <dc :nbrFolios>
</dcterms :extent>
```

### 5.3. Le nombre de lignes par page

Dans le but de fournir à l'utilisateur le nombre de ligne par page d'un manuscrit, l'ajout d'un nouvel élément est jugé important. Cet élément issu de la norme EAD enrichis va indiquer le nombre de ligne par page du manuscrit décrit, qui portera comme nom <dc :nbrLinePage> et il sera inclus également dans l'affinement <dcterms :extent>.

#### 5.4. Taille de folios

La taille de folios représente les dimensions du manuscrit. Le Dublin Core permet de définir toutes les dimensions d'un manuscrit mais d'une manière non précise c'est-à-dire qu'il ne réserve pas un élément spécifique pour les dimensions. A cet effet, nous proposons d'ajouter un nouvel élément `<dc:dimensions>` issu de l'élément `<dimensions>` de la TEI-ms enrichie ainsi que ces sous éléments : `<height>`, `<width>`, `<depth>` et `<dim>`. Ces éléments auront la forme suivante sous le Dublin Core : `<dc:height>`, `<dc:width>`, `<dc:depth>` et `<dc:dim>`. Par conséquent ces éléments introduits dans l'affinement `<dcterms:extent>` servent à définir toutes les dimensions du manuscrit.

Après tous ces ajouts la structure de l'affinement `<dcterms:extent>` sera la suivante :

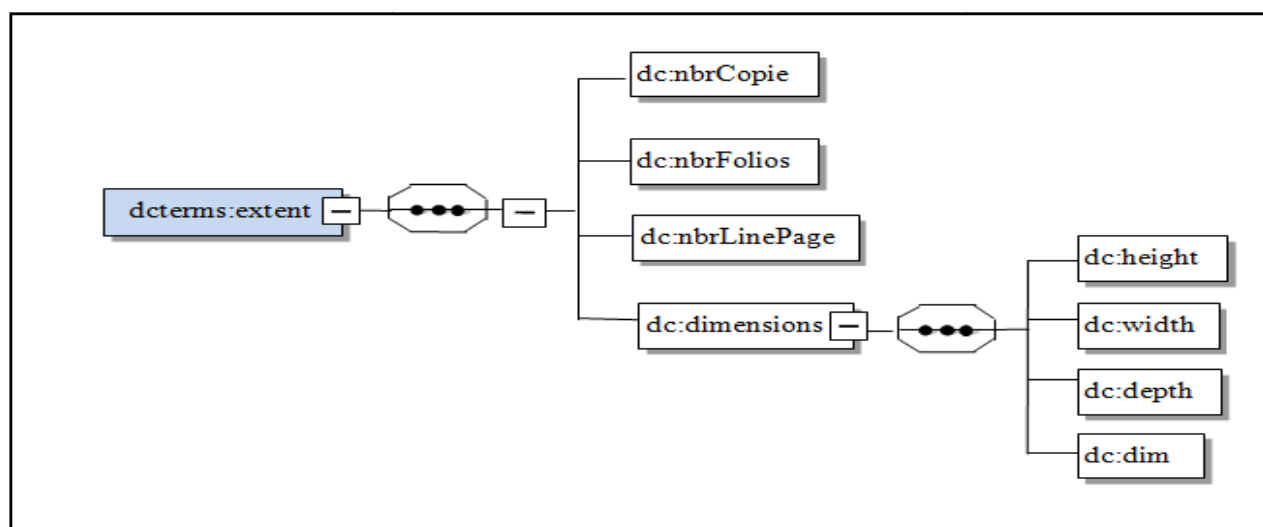


Figure 29: structure de l'affinement `<dcterms:extent>`.

#### 6. Histoire du manuscrit

Les aspects historiques des documents sont pris en considération par le Dublin Core d'une manière rigoureuse, du fait qu'il propose l'élément `<dc:provenance>` afin de décrire seulement les différents possesseurs du manuscrit avant son acquisition par l'institution de conservation.

Par ailleurs, l'histoire du manuscrit arabe peut inclure aussi l'origine du manuscrit et les modes des différentes acquisitions de ce dernier.

Pour ce faire nous proposons d'ajouter à l'élément `<dc:provenance>` l'élément `<dc:p>` ainsi que les deux éléments `<dc:origin>` et `<dc:acquisition>` issus respectivement des éléments `<origin>` et `<acquisition>` de la TEI-ms enrichis.

- a) `<dc:origin>` : Cet élément peut contenir les sous-éléments : `<dcterms:created>`, `<dc:origPlace>` et `<p>` :

- **<dc:terms:created>**: C'est un affinement de l'élément <dc:date> qui sert à définir la date de création de manuscrit, il peut contenir effectivement tous les attributs et les sous-éléments ajoutés pour la description des dates, tel que :
  - 1) L'attribut « value » avec la valeur « hégire » ou « grégorien » pour décrire une date hégire ou une date grégorienne.
  - 2) L'attribut « type » pour exprimer l'exactitude de la date donnée avec la valeur « effective » ou l'incertitude de la date avec la valeur « approximative ».
  - 3) Le sous-élément <dc:day>, pour coder le jour
  - 4) Le sous-élément <dc:month> pour coder le mois
  - 5) Le sous-élément <dc:year> pour coder l'année.
- **<dc:origPlace>** : Un élément offrant la possibilité de définir le lieu de création de manuscrit
- **<p>** : Élément paragraphe qui peut contenir n'importe quelle information sur la production du manuscrit.

**2. <dc:acquisition>** : cet élément permet de décrire le mode d'acquisition du manuscrit par l'institution de conservation actuelle. Sachant que l'acquisition du manuscrit peut se faire selon plusieurs procédures, parmi lesquelles nous citons : l'achat, le don, le prêt et le Waqf.

L'achat et le don font de l'institution le propriétaire légitime du manuscrit, alors que le prêt est généralement réalisé pour des objectifs de numérisation ou d'étude particulière sur le manuscrit. Dans ce cas, il serait toujours intéressant d'identifier le possesseur. Parfois, la présence du manuscrit dans l'institution est le résultat d'un Waqf : c'est le synonyme de habs en arabe [soualah 2008].

Dans le but d'arriver à décrire tous ces types d'acquisition, on propose d'introduire l'attribut « type » dans l'élément <dc:acquisition>. Cet attribut peut avoir comme valeur : « purchase », « gift », « prêt » ou « wakf ».

- ✚ « purchase » : pour dire que le type d'acquisition était un achat.
- ✚ « gift » : pour dire que le type d'acquisition était un don.
- ✚ « prêt » : pour dire que le type d'acquisition était un prêt.
- ✚ « wakf » : pour dire que le type d'acquisition était un wakf.

Le schéma suivant représente toute cette structure :

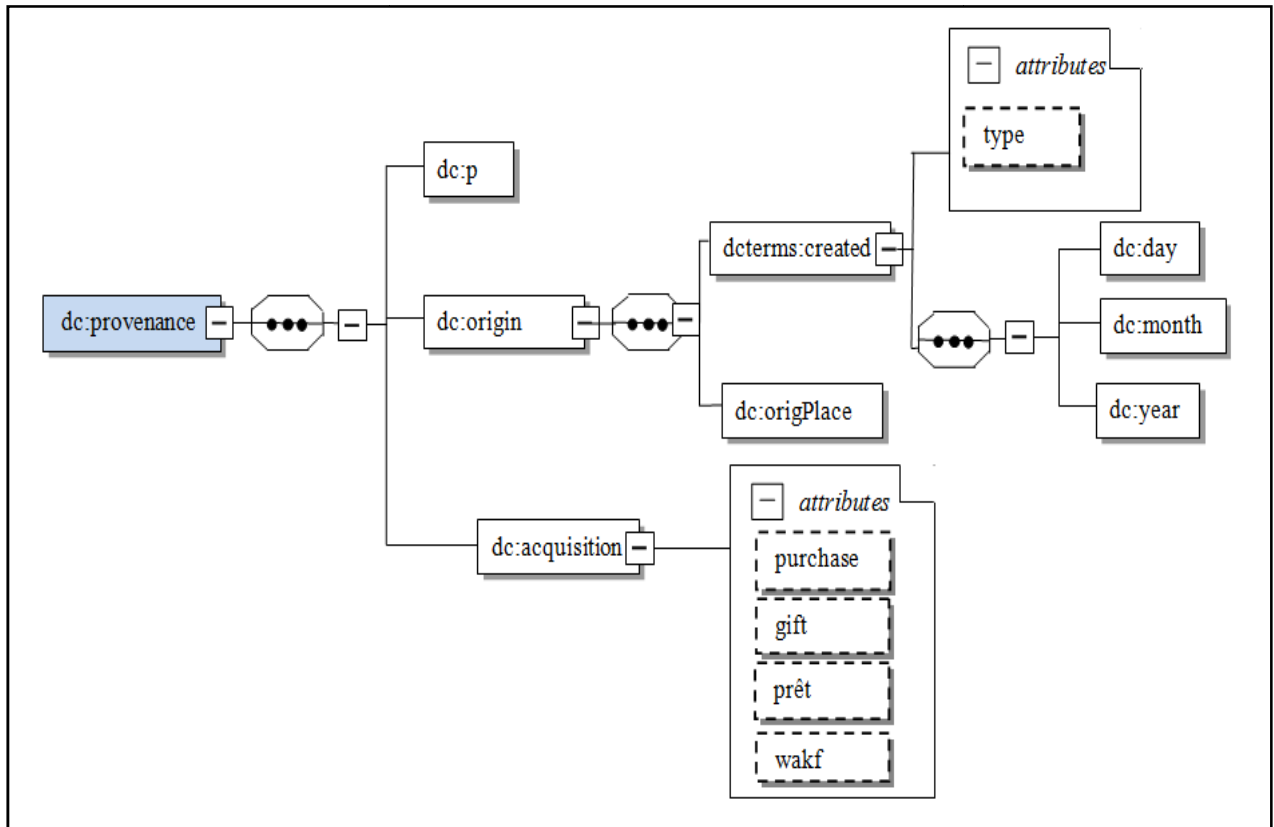


Figure 30 : structure de l'élément <dc :provenance>.

## 7. Références de publication

Dans le cas où le manuscrit a été déjà publié, il est nécessaire d'informer l'utilisateur en lui donnant l'URL ou bien la référence du site dans lequel le manuscrit a été publié. Le Dublin Core permet de décrire l'entité responsable de la mise à disposition de la ressource grâce à l'élément <dc :publisher>. Afin d'améliorer la description fournie par le Dublin Core. Nous allons introduire à l'intérieur de l'élément <dc :publisher> ces éléments <dc :entity> et <dc:ref>.

- ✚ <dc :entity> : pour la description de l'entité responsable de la publication.
- ✚ <dc :ref> : pour décrire la référence de publication.
- ✚

La structure de l'élément <dc :publisher> sera représentée comme suit :

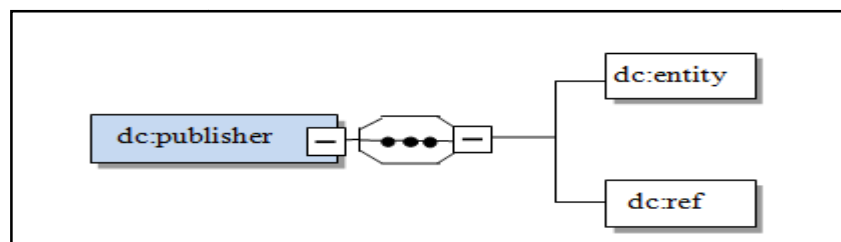


Figure 31 : Structure de l'élément <dc :publisher>.

Pour ce qui est des autres métadonnées du protocole enrichis sur lequel nous nous sommes basées et qui n'existe pas parmi les éléments standards du Dublin Core, nous proposons de les insérer à l'intérieur de l'élément `<dc:description>`. Cet élément est conçu pour porter le contenu intellectuel en fournissant un compte rendu de contenu de la ressource numérique. La description peut inclure un résumé, une table des matières, une référence à une représentation graphique du contenu.

En effet, cet élément peut contenir une liste d'éléments non limitée, et fournit le meilleur candidat pour l'exécution du système d'indexation des textes libres. Pour cela, nous l'avons choisi pour qu'il soit un élément conteneur d'un ensemble de sous-éléments descriptifs des métadonnées restantes du protocole.

## 8. Les éléments descriptifs

### 8.1. Table des matières

Les tables des matières des manuscrits arabes décrivent en générale la structure logique du manuscrit, les sous titres du manuscrit ainsi que les numéros de page de ces sous-titres. La structure des manuscrits arabes se décompose en "kytab", "djuz", "bab", "fasl". Le Dublin Core décrit la table des matières avec l'affinement `<dcterms:tableOfContents>` sous l'élément `<dc:description>`. Dans le but d'adapter la description des tables des matières offerte par le Dublin Core selon la structure logique des manuscrits arabes, nous ajoutons l'éléments suivant `<dc:repartition>` issu de l'élément `<repartition>` de la TEI-ms. Cet élément aura comme sous éléments :

- ◆ `<dc:kytab>` : Pour décrire le classement en kytab.
- ◆ `<dc:djuz>` : Pour décrire le classement en djuz.
- ◆ `<dc:bab>` : Pour décrire le classement en bab.
- ◆ `<dc:fasl>` : Pour décrire le classement en fasl.

Tous ces sous éléments seront dotés d'un attribut nommé "value" qui prend ses valeurs dans l'ensemble {true, false}, pour signifier l'existence ou l'inexistence de l'élément dans la structure du manuscrit. Le schéma ci-dessus résume cette structure :

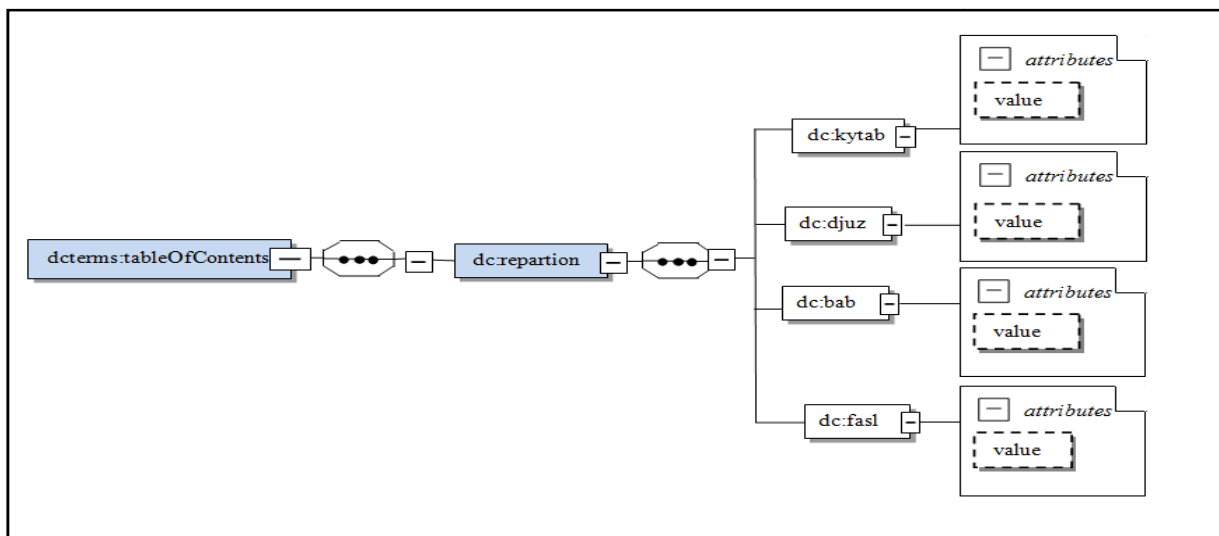
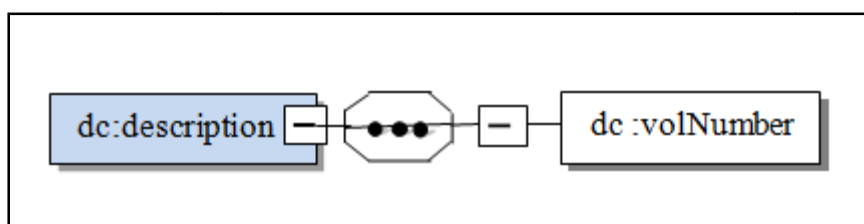


Figure 32: Structure de l'affinement `<dcterms:tableOfContents>`.

### 8.2. Numéro du volume

Les manuscrits arabes anciens se caractérisent par leur structure divisionnaire, dans un autre sens, un manuscrit peut être constitué de plusieurs volumes. Afin d'identifier chaque volume nous pouvons utiliser l'élément <dc : identifier>, mais cette description peut engendrer une certaine ambiguïté due à la réutilisation de ce dernier.

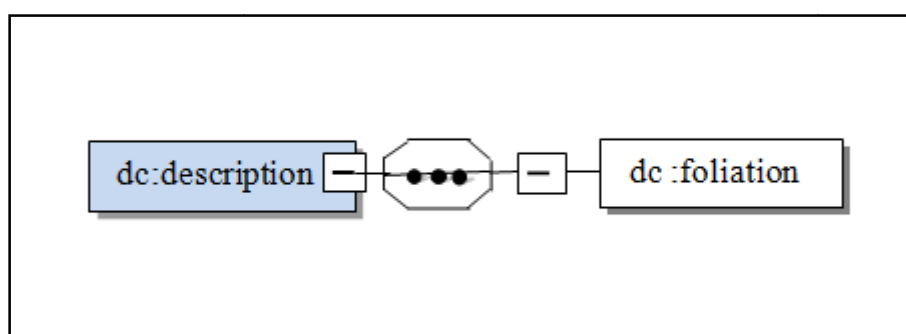
Dans le but d'éviter ce genre de situation, se pose l'obligation d'ajouter un nouvel élément. Pour cela, nous allons introduire l'élément <dc :volNumber> de la TEI- ms enrichie qui signifie le numéro de volume, utilisé afin d'identifier chaque volume à part. Cet élément peut être représenté avec la figure suivant :



**Figure 33:** Structure de l'élément <dc :volNumber>.

### 8.3. Foliation ou pagination

Cette métadonnée doit indiquer le type de numérotation utilisé. Le format actuel du Dublin Core n'offre aucune possibilité de décrire cet aspect, à partir de là nous allons introduire l'élément <dc :foliation> issu de l'élément <foliation> de la TEI- ms enrichie, qui sert à décrire le système de numérotation ou des systèmes utilisés pour compter les feuilles ou les pages dans un codex. Cette structure est illustrée dans la figure ci-dessus :



**Figure 34:** Structure de l'élément <dc :foliation>.

### 8.4. Incipit et Explicit

Le mot incipit désigne les premiers mots d'un texte, pour les manuscrits arabes l'incipit peut être la première phrase de manuscrit. Ainsi, Le terme explicit veut dire les derniers mots de manuscrit, qui signifie son achèvement. Le Dublin Core n'offre aucun élément permettant de décrire le début ou la fin d'un document, alors nous proposons d'ajouter un élément issu de l'EAD enrichie <scopecontent> et ses sous élément <incipit> et <explicit>, qui représentés par

les éléments suivants :<dc :incipit> et <dc :explicit> sous Dublin Core. Ces éléments servent à contenir respectivement la première et la dernière phrase du manuscrit.

**Exemple**

```

<dc :incipit>
البِسْمِلة... " الحمد لله رب العالمين... أما بعد حمدا لله مستحق الحمد و ملهمه... فإن كتابالخلاصة
الألفية في علم العربية...
</dc :incipit>
<dc :explicit>
... الحمد لله العلي الأجلل الواسع الفضل الموهوب المجزل... و الحمد لله على إتمامه و مولانا محمد
خاتم النبيين... و حسبنا الله و نعم الوكيل... و هذا آخر الكلام و الحمد لله على التمام."
</dc :explicit>

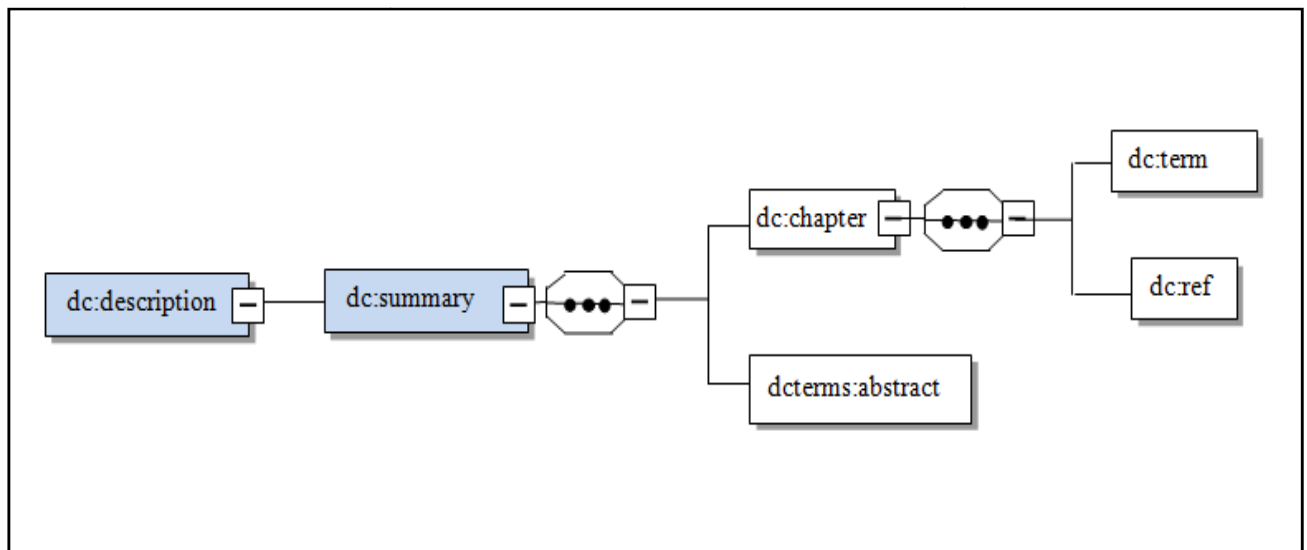
```

**8.5.Sommaire**

En général, les sommaires mentionnent les différents chapitres du manuscrit, et leurs numéros de pages. Le format Dublin Corene contient pas d'éléments pouvant décrire le sommaire du manuscrit. Dans ce cas, l'ajout d'un élément est essentiel.Ce dernier nommé <dc :summary> issu de l'élément<summary> de la TEI-ms enrichie sui sert à définir le sommaire. Cet élément aura également comme sous-éléments : <dc :chapter>,<dc :ref>, <dc :term> et <dcterms :abstract>.

- ◆ <dc :chapter> : cet élément sert à définir les chapitres du manuscrit.
- ◆ <dc :ref> : cet élément sert à fournir la référence du chapitre.
- ◆ <dc :term> :Cet élément présente le terme référencé.
- ◆ <dcterms :abstract> : cet affinement de l'élément <dc :description>, sera utilisé pour donner un résumé sur le sommaire du manuscrit.

Le schéma ci-dessus représente la structure de l'éléments <dc :summary> :



**Figure 35:**Structure de l'élément <dc :summary>.

## 8.6. Index

Les index comme nous les avons déjà définis sont des termes fournis par le catalogueur dans le but de faciliter la recherche. En effet, la recherche en utilisant cet index se fait d'une manière implicite, tel que le catalogueur choisit un terme quelconque à l'intérieur d'un manuscrit pour qu'il soit un indexé, alors la recherche de ce manuscrit sera facile en utilisant les termes de l'index. Le Dublin Core sous sa forme actuelle ne donne aucun élément permettant de définir les index utilisés par le catalogueur.

Par conséquent, l'ajout d'un tel élément est nécessaire. Pour ce faire, nous proposons d'utiliser la description des index fournie par la TEI-*ms* enrichie. A cet effet, nous allons introduire l'élément `<dc:index>` qui aura comme attribut « `indexName` » avec la valeur "keywords". L'élément `<dc:index>` va être constitué des deux sous-éléments `<dc:term>` et `<dc:ref>` :

- ✚ `<dc:term>` : sert à décrire les mots indexés.
- ✚ `<dc:ref>` : sert à donner les références des mots indexés.

### Exemple [Soualah 2008]

```
<index indexName = "keywords">
  <term> الحضارة
  <ref>45,57</ref>
</term>
</index>
```

## 8.7. Signature

Les manuscrits arabes anciens se caractérisent par l'existence des cahiers qui peuvent être un pliage d'une feuille ou la réunion de plusieurs feuilles pliées. La signature représente la numérotation de cahier. Cette numérotation se trouve sur le premier feuillet du cahier, au recto, ou sur le dernier au verso ou bien, sur les deux.

Le format Dublin Core ne permet pas de décrire ce concept, pour cela nous proposons d'ajouter l'élément `<dc:signature>` issu de l'élément `<signature>` de l'EAD enrichie. Cet élément peut être représenté comme suit :

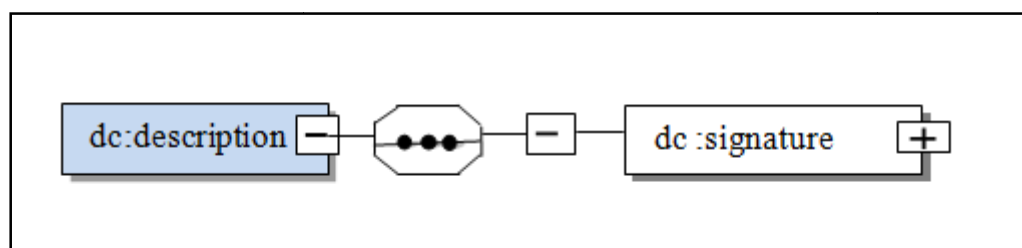


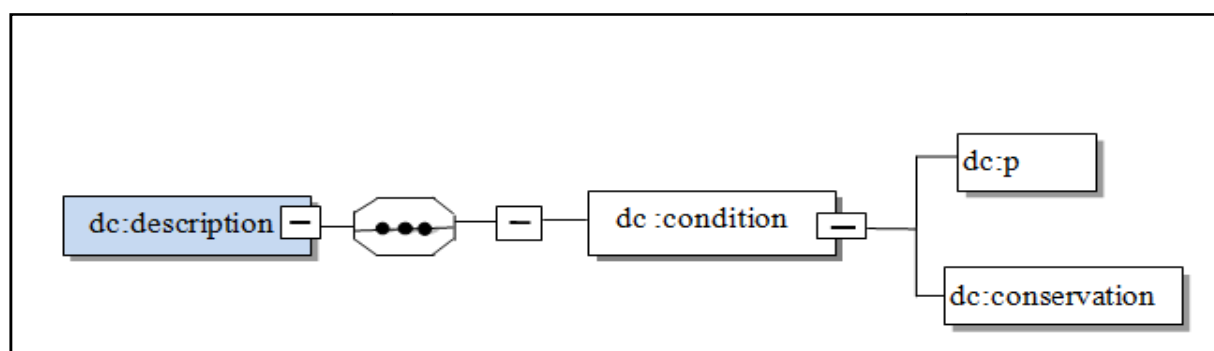
Figure 36: structure de l'élément `<dc:signature>`.

## 8.8. Etat du manuscrit et leur état de conservation dans des bibliothèques

Un manuscrit peut être complet ou incomplet en présentant des parties manquantes. Dans ce cas, il faut informer l'utilisateur sur l'état du manuscrit qu'il veut consulter s'il contient tous ces folios, ainsi que toutes ces parties. Pour cela, nous allons introduire l'élément `<condition>` de la TEI-*ms* enrichie, sous la forme `<dc:condition>` afin de décrire l'état physique du manuscrit y compris sa complétude. Cet élément peut avoir un ou plusieurs éléments `<dc:p>` pour la description de l'état sous forme de paragraphes.

Par ailleurs, Le manuscrit arabe peut exister au sein de différentes bibliothèques et cela sous forme de copie. Une bibliothèque peut conserver un manuscrit selon des divers types de copies, citons par exemple : une copie sous forme d'un microfilm, une copie imprimée etc...

A cet effet, nous allons introduire l'élément `<dc:conservation>` comme un sous-élément de `<dc:condition>` afin de décrire l'état de conservation d'un manuscrit, selon la structure suivante :



**Figure 37:** Structure de l'élément `<dc:condition>`.

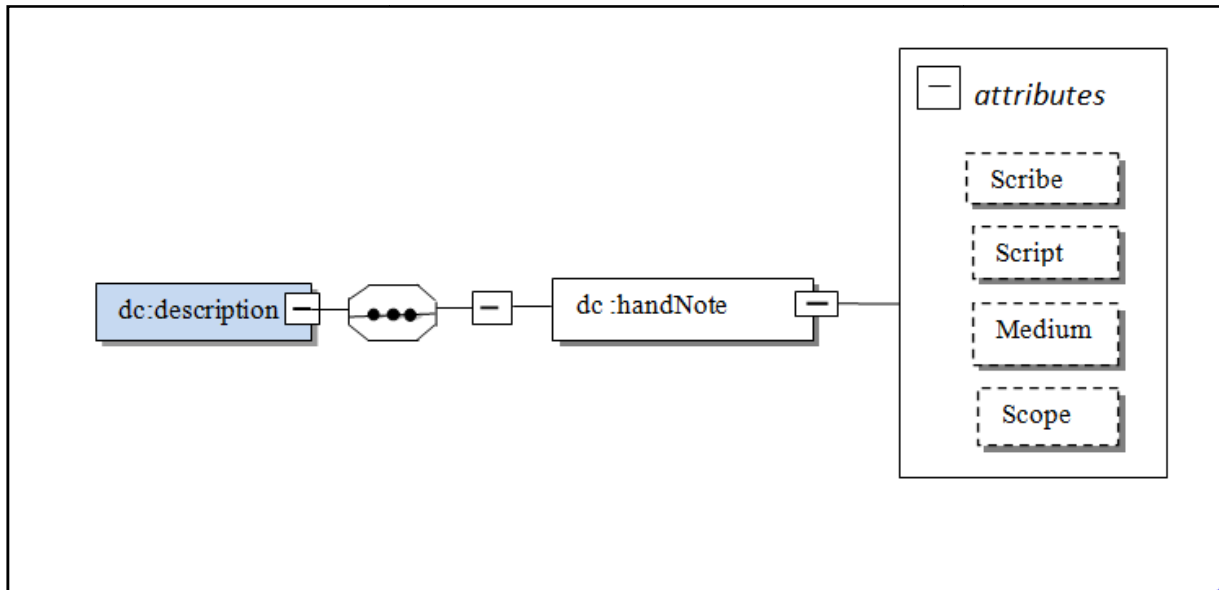
### 8.9. Le modèle d'écriture et Encre

Les styles calligraphiques arabes se sont développés d'une manière remarquable. En effet, il existe plusieurs styles des écritures arabes tel que : le style koufi, le style naskhi, ...etc.

La calligraphie utilisée dans le manuscrit doit être décrite, du fait que chaque manuscrit peut être caractérisé par les styles d'écriture employés et la couleur de l'encre utilisé. A cet effet, l'ajout d'un élément pour la description des styles d'écritures et les couleurs de l'encre de manuscrit est essentiel, alors nous proposons l'élément `<dc:handNote>` qui est issu de l'élément `<handNote>` de la TEI-*ms* enrichis. Cet élément va contenir également les attributs suivants :

- ❖ **Scribe** : Contient le nom ou les noms des personnes ayant participé à la rédaction du manuscrit.
- ❖ **Script** : Décrit le style d'écriture, par exemple coufi, maghribi, diwani, farisi, ...etc.
- ❖ **Medium** : Décrit la teinte ou le type d'encre et sa couleur.
- ❖ **Scope** : Spécifie le taux de participation du calligraphe à l'écriture du manuscrit.

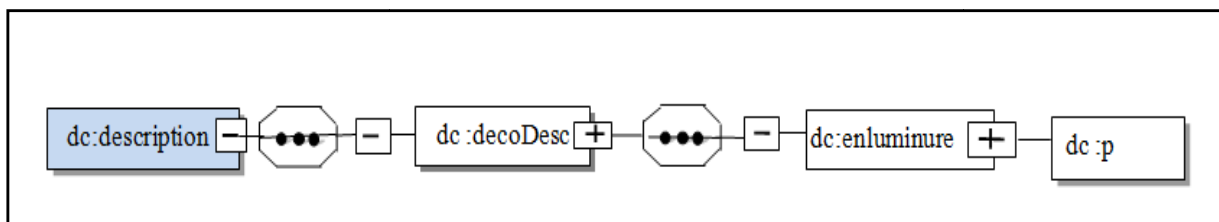
Le schéma ci-dessus représente la structure de l'élément `<dc:handNote>` :



**Figure 38:** structure de l'élément <dc:handNote>.

### 8.10. Enluminure

L'enluminure est une peinture ou un dessin exécuté à la main qui décore ou illustre un texte, généralement un manuscrit. Les enluminures font partie de la décoration des manuscrits. Le Dublin Core avec son format actuel n'offre pas la possibilité au catalogueur pour décrire cet aspect. Pour cela nous allons ajouter l'élément <dc:decoDesc> issu de l'élément <decoDesc> de la TEI-ms enrichis. Cet élément contient l'élément <dc:enluminure>, qui à son tour va être constitué de l'élément paragraphe <dc:p> pour décrire toutes les enluminures existantes dans le manuscrits sous formes d'un ou de plusieurs paragraphes. Le schéma suivant résume cette structure :



**Figure 39:** structure de l'élément <dc:decoDesc>.

### 8.11. Reliure

Parmi les caractéristiques des manuscrits arabes on trouve la reliure, qui peut être fabriqué selon des manières différentes. Elle peut être aussi changée par les propriétaires du manuscrit si l'ancienne reliure était dans un mauvais état. Dans le but de décrire cette reliure et tous ces changements y compris la description de la reliure actuelle et l'ancienne on va ajouter l'élément <dc:binding> issu de la TEI-ms enrichie. Cet élément sera constitué de l'élément <dc:p> pour décrire ces reliure et leur changements. Cette structure est décrite comme suit :

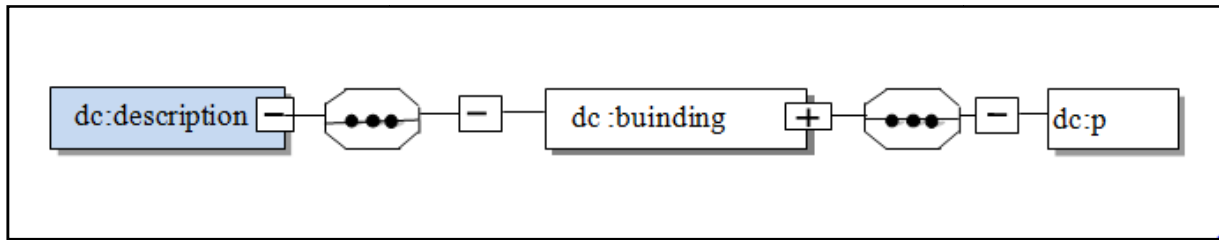


Figure 40: Structure de l'élément <dc:binding>.

### 8.12. Feuilles détachées

Vu l'ancienneté des manuscrits arabes, ils se considèrent comme étant des documents fragiles à cause de leur état physique détruit. Par conséquent, un manuscrit qui est dans une bibliothèque, peut ne pas être complet à cause de mauvaise reliure par exemple. Ces feuilles détachées ou séparées du manuscrit doivent être identifiées.

À cet effet, l'existence d'un élément qui décrit ces feuilles est essentiel, alors nous avons ajouté au format Dublin Core un nouvel élément <dc:separatedmaterial> qui est issu de l'élément <separatedmaterial> de l'EAD enrichi. L'élément ajouté va avoir comme sous-éléments <dc:p> et <dc:ref>.

- ✚ <dc:p> : pour décrire ces feuilles détachées.
- ✚ <dc:ref> : pour donner la référence de ces feuilles.

Le schéma suivant décrit la structure de cet élément :

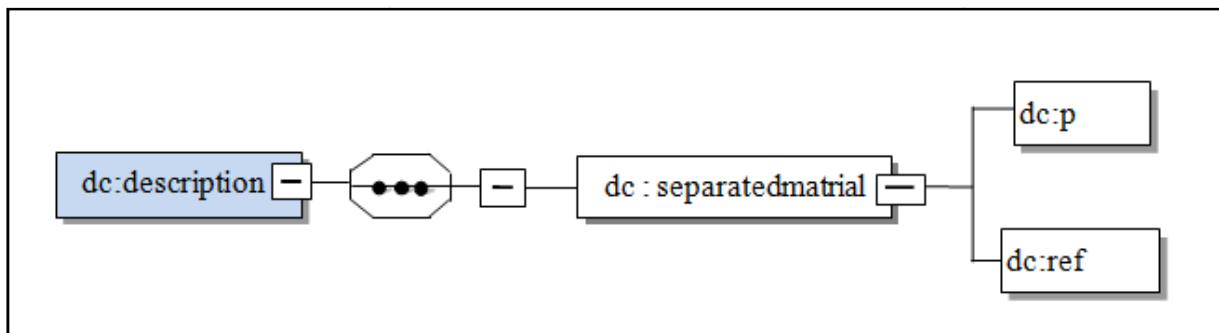


Figure 41: structure de l'élément <dc:separatedMaterial>.

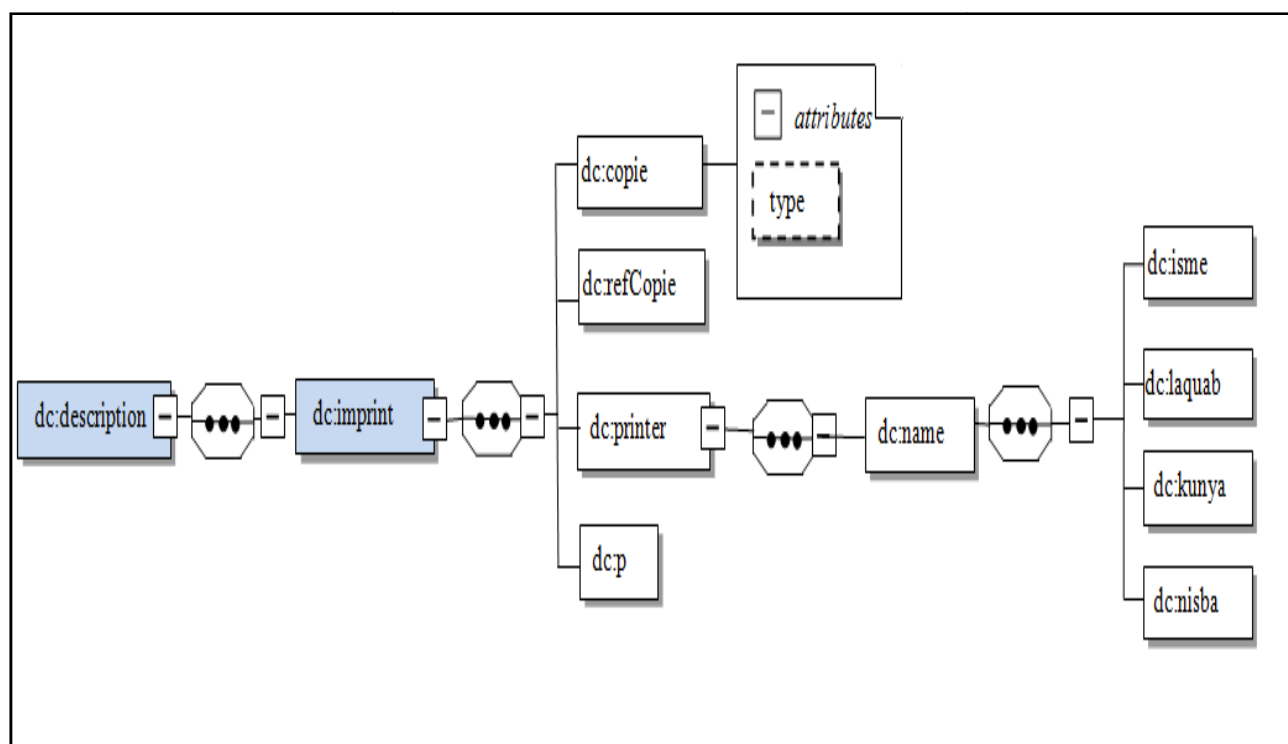
### 8.13. Copies imprimées, imprimeur et références des copies du manuscrit disponible dans d'autres bibliothèques

Les manuscrits arabes anciens peuvent avoir plusieurs copies imprimées dans divers lieux, du fait que le copiste ou l'imprimeur, peut copier ou imprimer plusieurs copies et les déposer dans des instituts ou des bibliothèques différentes. Afin de décrire tous les aspects qui sont en relation avec ces copies y compris leurs références et leurs imprimeurs, nous ajoutons

l'élément <dc :imprint> de l'EAD enrichie qui peut inclure toutes les informations concernant les copies imprimées. Cet élément va avoir également les éléments suivants:<dc :copie>, <dc :refCopie>, <dc :printer> et<dc :p>.

- ✚ <dc :copie> : élément constitué de l'attribut « type » qui prend la valeur "true" si la copie est imprimée, "false" sinon.
- ✚ <dc :RefCopie> : Afin d'identifier la copie.
- ✚ <dc :printer> : Afin de définir l'imprimeur c'est-à-dire le nom de la personne responsable sur l'impression de la copie. Pour décrire le nom nous avons besoin de l'élément <dc :name> et de tous ces éléments qui permettent de définir le nom d'une manière complète.
- ✚ <dc :p> : un paragraphe pour décrire la copie imprimée, donner toute sorte d'information sur elle, et aussi pour citer bibliothèques ayant d'autres copies.

Le schéma suivant représente la structure de l'élément <dc :imprint> :



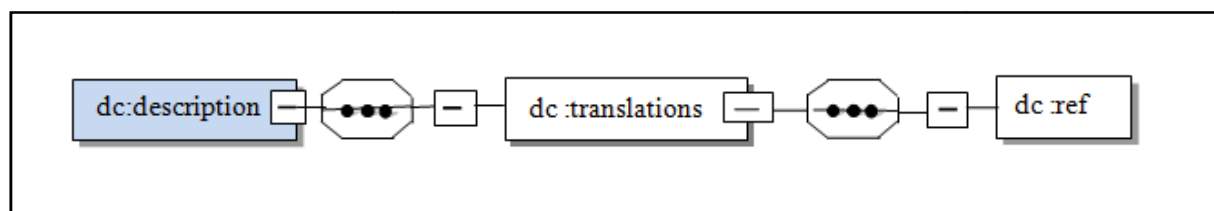
**Figure 42:** Structure de l'élément <dc :imprint>.

#### 8.14. Références de traduction

Nous avons dit auparavant que les manuscrits arabes peuvent avoir des annotations marginales pour les corrections. Mais elles peuvent concerner aussi d'autres aspects comme par exemple les traductions (الترجمة) vers d'autres langues, citées dans des références bibliographiques.

La structure actuelle du Dublin Core ne permet pas de décrire ce cas, pour cela nous allons ajouter l'élément <dc:translations> issu de l'élément <translations> de la TEI-ms qui peut être constitué de un ou de plusieurs éléments <dc:ref>, afin de fournir la référence de document dans lequel existe la traduction du manuscrit.

La figure suivante résume la structure de l'élément <dc:translations> :



**Figure 43:** Structure de l'élément <dc:translations>.

### 8.15. Additifs et Corrections

Les manuscrits arabes sont réalisés par des adeptes qui après avoir reçu un certificat de transmission (igaza) de leur maître seront habilités à rédiger le discours écouté après avoir obtenu un certificat d'audition (samâ) du maître. Une fois finie, on procède à la lecture de l'œuvre par un lecteur (qarî).

Pour permettre au catalogueur d'encoder tous ces aspects, nous allons ajouter l'élément <originalsloc> de l'EAD enrichie et son sous-élément <certify>, sous la forme <dc:originalsloc> et <dc:certify>. L'élément <dc:certify> représente l'adepte, il est défini par l'attribut « type » qui peut prendre les valeurs suivantes : « listner », « transmission » ou « reader ».

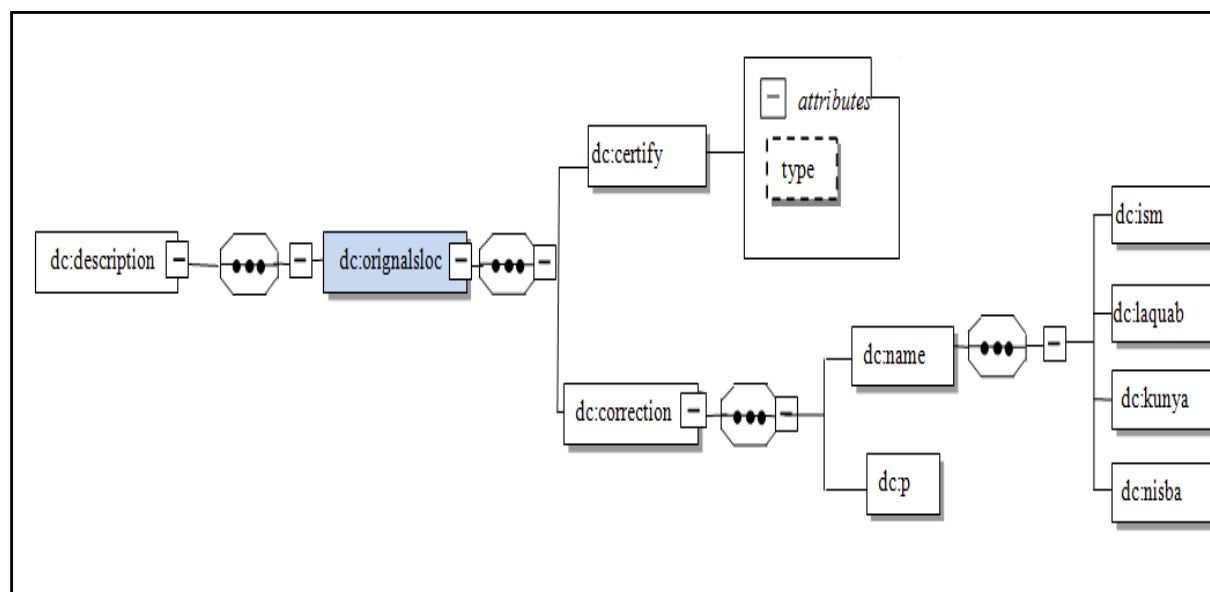
- ✚ « listner » : pour définir le certificat d'audition « سماعات ».
- ✚ « transmission » : pour définir le certificat de transmission « اجازات ».
- ✚ « reader » : pour définir le certificat de lecture « قراءات ».

Des annotations marginales sont souvent portées sur les manuscrits arabes. Elles peuvent concerner les différentes corrections (التصحیحات) portées par l'auteur ou un lecteur. Par ailleurs, la structure du Dublin Core ne permet pas de décrire ces particularités, à cet effet et afin de mentionner ces corrections, nous introduisons dans l'élément <dc:originalsloc> un nouvel élément issu de l'élément <correction> de la TEI enrichie qui sera nommé <dc:correction>.

Cet élément peut être composé des éléments suivants <dc:name> et <dc:p> :

1. <dc:name> : pour décrire les noms des personnes responsables sur ces corrections, qui peut renfermer tous les sous-éléments ajoutés précédemment pour décrire les noms.
2. <dc:p> : élément paragraphe qui sert à donner une description des corrections portées.

Le schéma ci-dessus résume cette structure :



**Figure 44:** structure de l'élément <dc:originalsloc>.

### 8.16. Palimpsesteetobservations

Les manuscrits arabes anciens peuvent être écrits sur des supports déjà utilisés, c'est ce qu'on appelle le palimpseste, qui est une caractéristique dans la description d'un manuscrit. Aussi, le catalogage du manuscrit peut inclure des informations complémentaires sur n'importe quel aspect du manuscrit ou toutes sortes d'informations qui ne sont pas déjà citées à travers les autres éléments, et qui peuvent être nécessaires pour l'utilisateur.

A cet effet, pour décrire ces deux aspects il est impérativement essentiel d'ajouter un nouvel élément. Pour cela, nous allons utiliser l'élément <odd> qui découle de l'EAD enrichie et ces sous-éléments <palimpsest> et <observations>. Avec le Dublin Core ces éléments seront nommés respectivement : <dc:odd>, <dc:palimpsest> et <dc:observations>.

<dc:odd> : élément englobant qui va être constitué de :

- <dc:palimpsest> : pour décrire le support réutilisé.
- <dc:observations> : cet élément regroupera toutes les informations supplémentaires en utilisant l'élément <dc:p> tel que la description se fera selon un ou des ensembles de paragraphes.

Le schéma suivant résume la structure de l'élément <dc:odd> :

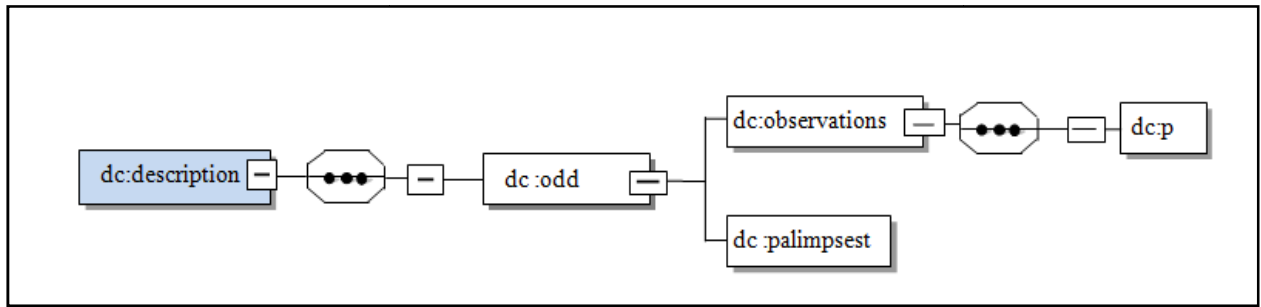


Figure 45: Structure de l'élément <dc:odd>.

## VI. Correspondance entre EAD, TEI-*ms* enrichie & Dublin Core

A travers la section précédente, nous avons effectués un enrichissement dans le contenu des trois normes d'encodage étudiées dans le but de les rendre capables de décrire les manuscrits arabes d'une façon explicite. Le travail fait sur le Dublin Core, il s'agit bien d'une migration du contenu des deux normes de base TEI-*ms* et EAD vers le format Dublin Core, sans perte d'informations.

Autrement dit, tous les champs descriptifs d'une notice bibliographique d'un manuscrit arabe encodée selon les normes TEI-*ms* et EAD retrouvent leurs correspondances dans une notice du même manuscrit encodée à son tour selon le format Dublin Core.

Le schéma ci-dessus représente graphiquement ce principe :

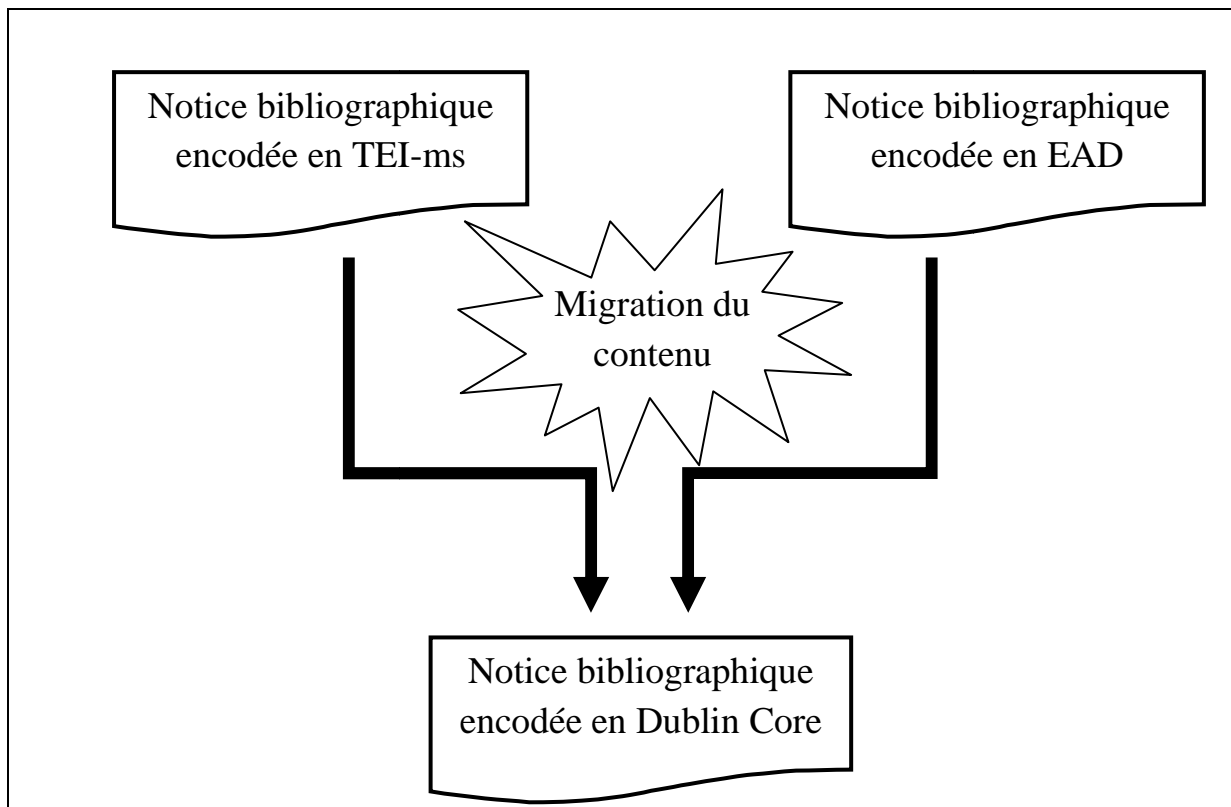


Figure 46: principe de migration.



-Explicit	<explicit>	<explicit>	<dc:explicit>
-Colophon	<scribe>	<colophon>	<dc:contributor>
-Table des matières	<tabContent>	<fileplan>	<dcterms : tableOfContents>
-Index	<index>	<index>	<dc:index>
-Sommaire du manuscrit	<summary>	<scopecontent>	<dc:summary>
<b>Description physique du manuscrit</b>			
-Matériel de fabrication	<support>	<physfacet type="support">	<dcterms: medium>
-Nombre de folios	<extent>	<extent>	<dc:nbrFolios>
-Nombre de lignes par page	<layout>	<nbrLinePage>	<dc:nbrlinePage>
-Palimpsest ou non	<palimpsest>	<palimpsest>	<dc:palimpsest>
-Signature (numérotation du cahier)	<signatures>	<signature>	<dc:signature>
-État du manuscrit (complet ou pas)	<condition>	<accruals>	<dc:condition>
-Dimension de la couverture	<dimensions>	<dimensions>	<dc :dimensions>
-État de conservation	<surrogates>	<container>	<dc:conservation>
-Taille de folios	<dimensions>	<dimensions>	<dc:dimensions>
-Style d'écriture	<handNote Script= "">	<physfacet style="">	<dc :handNote Script="" ">
-Schémas	<decoNotes schema="" ">	<genreform>	<dc:description>
-enluminure	<decoNotesenluminur e="" ">	<physfacet type="enluminure">	<dc:enluminure>
-Encre	<handNote medium="" ">	<physfacet type="encre">	<dc:handNote medium="" ">
-Reliure	<buidingDesc>	<physfacet type="reliure">	<dc:buiding>

-Folios détachées de la copie	<accMat>	<separatedmaterial>	<dc:separatedmaterial>
<b>Additifs</b>			
-Certificat de lecture	<scholar type="reader">	<certify type="reader">	<dc:certify type="reader">
-Certificats d'audition	<scholar type="listner">	<certify type="listner">	<dc:certify type="listner">
-Autorisation de transmission	<scholar type="transmission">	<certify type="transmission">	<dc:right>
-Corrections	<corrections>	<change>	<dc:correction>
<b>Références des copies</b>			
-Copie du manuscrit dans d'autres bibliothèques	<bibl>	<refCopie>	<dc :refCopie>
-Copie imprimée (Oui ou Non)	<docImprint>	<copie>	<dc :copie>
-Nom Imprimeur	<imprint>	<nameImprint>	<dc:printer>
-Références de publication	<ref>	<extref>	<dc :ref>
-Références de traductions	<translations>	<refTranslat>	<dc :translations>
-Sources	<source>	<resource>	<dc : source>
-Observations	<overview>	<observations>	<dc :observations>
<b>Histoire de manuscrit</b>			
-Possession du manuscrit	<provenance>	<custodhist>	<dc : provenance>
changeement de propriétaire	<acquisition>	<acqinfo>	<dc :acquisition>
-Acquisition du manuscrit			

## VII. Conclusion

A travers notre travail, nous avons proposés une adaptation du format Dublin Core selon les deux normes de bases. Ainsi, le format Dublin Core défini devient aisément capable de décrire les manuscrits arabes. De plus, une correspondance d'éléments entre Dublin Core et les deux normes de bases TEI-ms et EAD après les avoir enrichit, a été effectuée, ce qui permet d'effectuer une migration de données parfaite et sans perte d'informations.

L'adaptation du format Dublin Core permettra de générer des notices bibliographiques adaptées aux besoins et au profil des utilisateurs et facilitant ainsi, l'accès aux manuscrits numérisés en mode image grâce à un système d'hyperliens.

**Chapitre 2 : Mise en place de la  
solution d'interopérabilité**

## ***Introduction***

L'objectif de ce chapitre est de présenter les différentes étapes suivies afin de concevoir et d'implémenter un moteur de recherche qui donnera la possibilité d'effectuer des recherches dans des catalogues des manuscrits arabes numérisés et d'afficher les résultats de ces recherches. Au début de ce chapitre nous allons décrire le processus de recherche dans le catalogue des manuscrits arabes, dont les notices sont structurées selon le format XML. Ensuite, nous nous intéresserons à l'implémentation du moteur de recherche en décrivant les différents modules constituant le processus d'indexation. A la fin, nous allons faire une présentation des outils de développement, ainsi que les principales interfaces du moteur de recherche.

### ***I. Le processus de recherche dans les catalogues de manuscrits arabes***

L'objectif de recherche au sein d'un catalogue est de répondre à une requête spécifique d'un utilisateur, en affichant l'image correspondante du manuscrit. Le principe de la recherche est de réaliser une correspondance entre les termes de la requête et ceux des notices du catalogue.

Un catalogue contient un ensemble de notices bibliographiques. Les notices doivent être sauvegardées dans une base documentaire.

La recherche, se fait en deux étapes successives :

- L'étape d'indexation.
- L'étape d'interrogation.

### ***II. La recherche d'informations dans les documents XML***

Toute recherche d'information dans un document XML s'effectue selon trois étapes : manipulation, indexation et interrogation.

#### ***1. Manipulation des documents XML***

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

Le schéma suivant représente le principe de parsing d'un document XML :

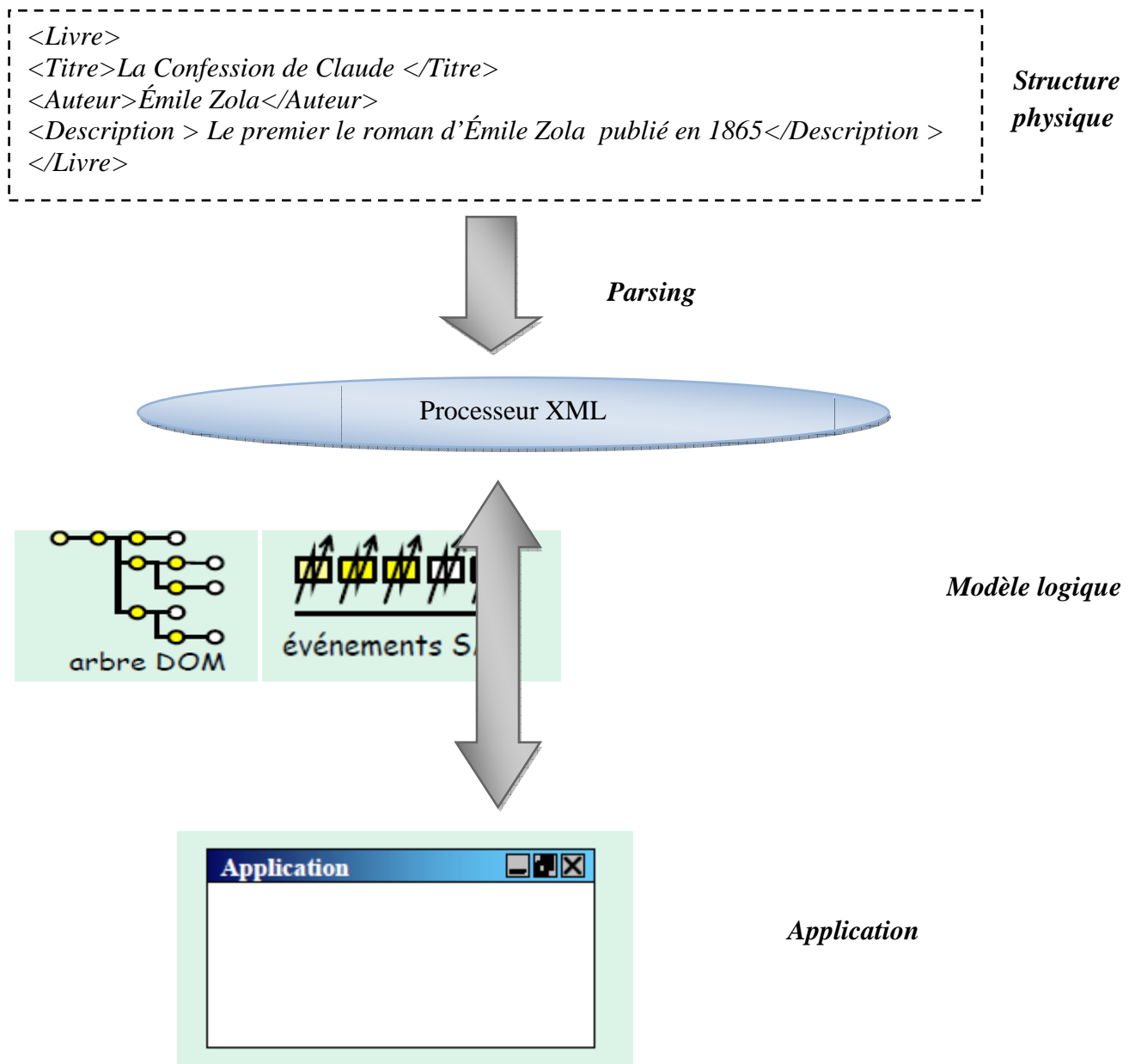


Figure 47: Parsing d'un document XML.

## 2. Indexation

L'indexation est une phase importante dans tout processus de recherche d'information. En effet, l'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à-dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse[Zargayouna2005].

La phase d'indexation est nécessaire dans le processus de recherche d'information, son efficacité est essentielle pour la pertinence de la recherche. En effet, un document mal indexé risque de devenir inaccessible pour les utilisateurs.

## **2.1. Le processus d'indexation des documents XML**

Le processus d'indexation est un processus de conversion des données citées sur des documents vers des représentations traitables par les systèmes informatiques. Ce processus se résume en générale en ces quatre étapes indispensables :

### **2.1.1. L'extraction des mots clés du document**

Tokenization en anglais, elle consiste à découper le document en unités lexicales. Chaque unité lexicale est une séquence de caractères entourée par des séparateurs d'unités [Harrathi2010].

### **2.1.2. La normalisation des mots clés du document**

Elle consiste à réduire les mots à leur forme canonique, à leur racine : toutes les formes d'un verbe par exemple sont regroupées à l'infinitif, tous les mots au pluriel sont ramenés au singulier ...etc [Torjmen 2009].

### **2.1.3. L'élimination des mots vides**

Dans cette étape les mots d'usage général et grammatical (les mots vides) sont éliminés. Du fait que ces mots apparaissent d'une manière uniforme dans les documents ils sont non utiles pour l'indexation et ils doivent être éliminés [Harrathi 2010]. On distingue deux techniques pour éliminer les mots vides :

- L'utilisation d'une liste de mots vides (aussi appelée anti-dictionnaire ou *stoplist*).
- L'élimination des mots dépassant un certain nombre d'occurrences dans la collection ou les mots rares de la collection.

### **2.1.4. La pondération des mots clés**

La pondération des termes permet de mesurer l'importance d'un terme dans un document. Cette importance est souvent calculée à partir de considérations et interprétations statistiques (ou parfois linguistiques). L'objectif est de trouver les termes qui représentent le mieux le contenu d'un document [Sauvaganat 2005].

## **2.2. Technique d'indexation des documents XML**

L'indexation de documents semi-structurés XML consiste à prendre en compte l'information structurelle dans la représentation du document, en plus de l'information textuelle [Torjmen 2009]. En effet, comme nous l'avons déjà défini un document XML est un document constitué par un ensemble de balises représentant la structure ainsi qu'un ensemble de données à l'intérieure de ces balises qui représentent le texte.

### 2.2.1. Indexation de l'information textuelle

L'indexation du contenu textuel d'un document XML consiste à indexer les éléments ou les nœuds de l'arbre du document. On distingue deux approches d'indexation [Harrathi 2010] :

#### a. Sous-arbres imbriqués

On considère que le contenu de chaque nœud de l'index est une unité atomique. Les termes des nœuds feuilles sont donc propagés dans l'arbre des documents. Comme les documents XML possèdent une structure hiérarchique, les nœuds de l'index sont imbriqués les uns dans les autres et par conséquent, l'index contient des informations redondantes.

#### b. Unité disjointes

Le document est décomposé en unités disjointes, de telle façon que le texte de chaque nœud de l'index est l'union d'une ou plusieurs parties disjointes. Ainsi, les techniques standards de l'indexation peuvent être appliquées sur ces nœuds.

### 2.2.2. Indexation de l'information structurelle

L'indexation structurelle peut se faire selon des granularités variées, c'est à dire que toute l'information structurelle n'est pas forcément utilisée dans le processus d'indexation [Torjmen 2009].

Parmi les approches proposées dans la littérature, on distingue trois types d'approches pour l'indexation de l'information structurelle : indexation basée sur des champs, indexation basée sur des chemins et enfin indexation basée sur des arbres.

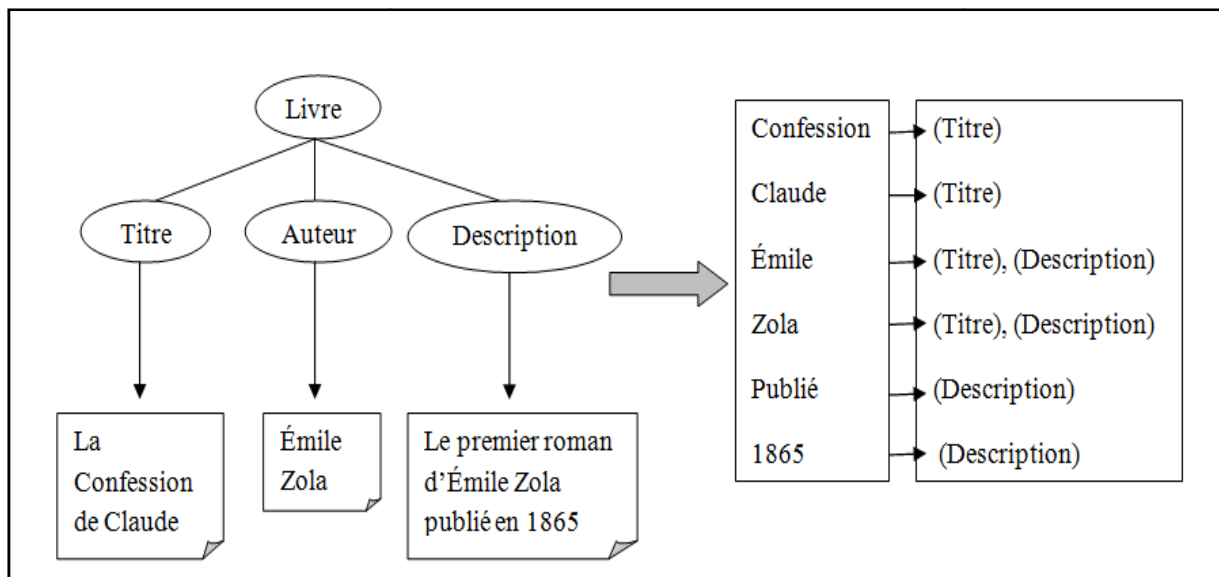
#### a. Indexation basée sur des champs

Dans cette indexation, chaque terme est associé au nom de la balise dans laquelle il apparaît. Cette méthode permet d'effectuer une recherche restreinte à certains champs, les termes de l'index sont construits en combinant le nom du champ avec les termes du contenu [Harrathi 2010]

#### Exemple

L'exemple ci-dessus présente l'indexation d'un extrait d'un document XML basée sur des champs :

```
<Livre>  
<Titre>La Confession de Claude </Titre>  
<Auteur>Émile Zola</Auteur>  
<Description > Le premier le roman d'Émile Zola publié en 1865</Description >  
</Livre>
```



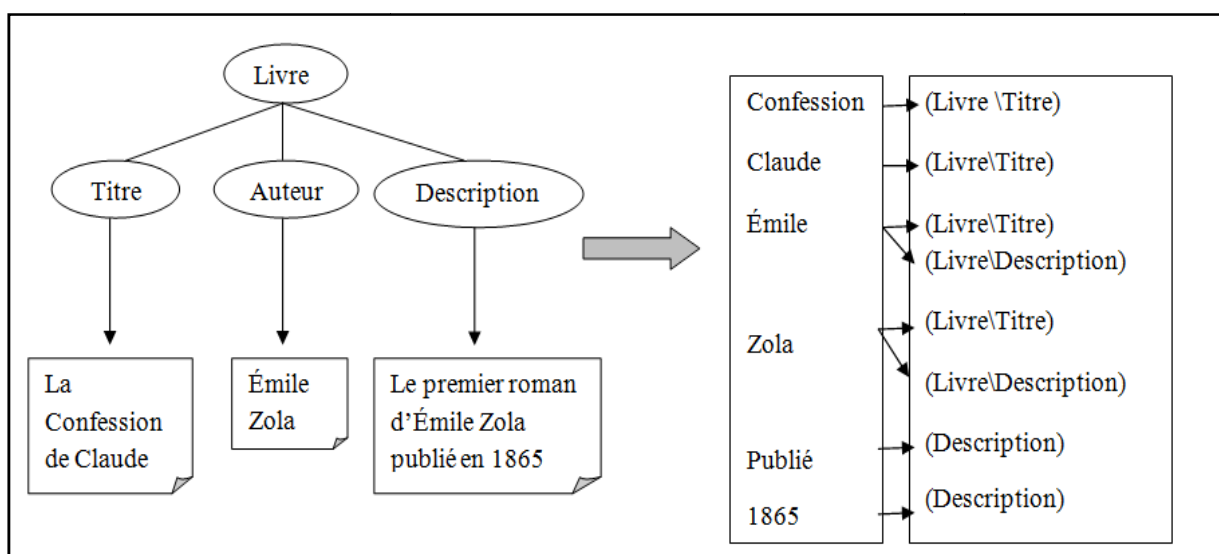
**Figure 48:** Indexation basée sur des champs.

**b. Indexation basée sur des chemins**

Cette technique facilite la navigation dans les documents en permettant la résolution des expressions de chemin XPath. Elle utilise un index plein texte sur le contenu, ce qui permet de retrouver rapidement des documents ayant des valeurs connues pour certains éléments ou attributs [Torjmen 2009].

**Exemple**

L'exemple ci-dessus présente l'indexation d'un document XML basée sur des chemins :



**Figure 49:** Indexation basée sur les chemins.

**c. Indexation basée sur des arbres**

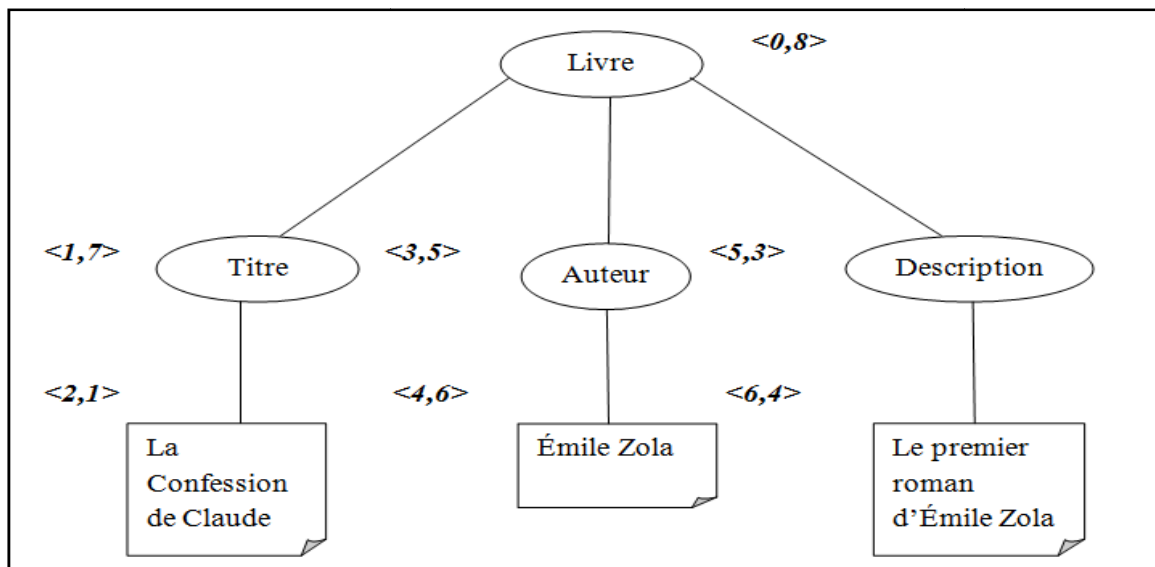
Dans cette technique, chaque nœud (élément) du graphe représentant le document XML est identifié par un identifiant unique. Les termes sont donc associés à cet identifiant, ce qui permet de localiser de façon précise l'endroit où ces termes sont apparus et de retrouver les relations hiérarchiques entre les éléments [Torjmen 2009].

Plusieurs méthodes d'identification structurales des nœuds ont été proposées dans la littérature, la plus utilisée est la méthode de numérotation Dietz.

**La technique de numérotation Dietz**

Cette technique de numérotation est la première qui a employé l'ordre de parcours d'arbre pour déterminer la relation *descendant/ascendant* entre n'importe quelle paire de nœuds d'arbre. Selon la technique de Dietz, chaque nœud de l'arbre XML possède deux valeurs : *pré-ordre et post-ordre* (pré et post) [Torjmen 2009].

La figure suivante présente un exemple d'un arbre XML dont les nœuds sont annotés par la numérotation de Dietz :



**Figure 50:** Indexation basée sur les arbres.

Cette figure présente un arbre XML dont les nœuds sont annotés par la numérotation de Dietz. Chaque nœud est marqué avec une paire d'entiers en pré-ordre et en post-ordre. A partir de l'arbre, nous pouvons constater par exemple que le nœud <1,7> est un ancêtre du nœud <2,1>, du fait que le nœud <1,7> vient avant le nœud <2,1>, selon l'ordre pré-ordre (C'est-à-dire, 1 < 2) et après le nœud <2,1> selon l'ordre post-ordre (c'est-à-dire 7 > 1).

L'approche d'index basée sur des arbres la plus conviviale, au vu des facilités qu'elle offre en particulier en ce qui concerne la détermination des relations ancêtres-descendants entre les différents nœuds du document.

### 2.3. Structure d'index [Sauvaganat 2005]

Afin de répondre plus rapidement à une requête, des structures de stockage particulières sont nécessaires pour mémoriser les informations sélectionnées lors du processus d'indexation. Les moyens de stockage les plus répandus sont les suivants : les fichiers inverses (« inverted files »), les tableaux de suffixes (« suffixArrays ») et les fichiers de signatures (« signature files »). Les fichiers inverses sont actuellement le meilleur choix possible pour la plupart des applications.

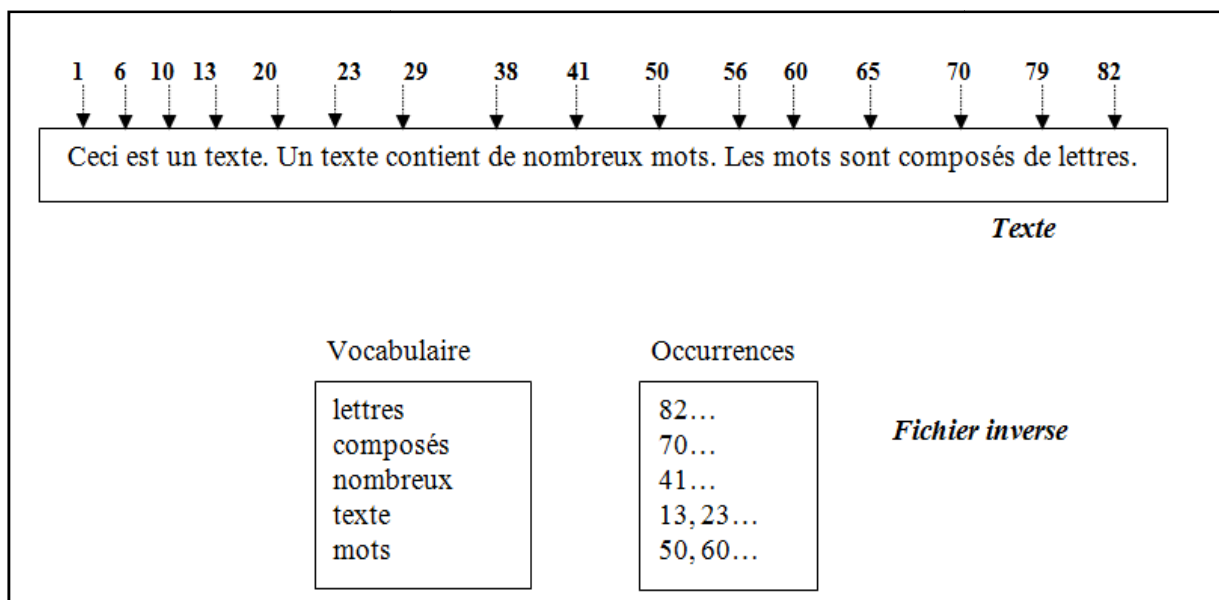
#### ✚ Le fichier inverse

Un fichier inverse est la structure la plus commune pour la recherche d'information. Chaque fichier est composé de deux éléments principaux :

1. **Le vocabulaire** : Qui est l'ensemble de tous les mots différents du texte.
2. **Les occurrences (posting) pour chaque mot** : Il s'agit de la liste de toutes les positions dans le texte pour lesquelles le mot apparaît.

#### Exemple

La figure suivante illustre une représentation d'un fichier inverse :



**Figure 51:**Présentation d'un fichier inverse.

### 3. Technique de stockage et d'interrogation des documents XML

#### 3.1. Le stockage

Dans le but d'exploiter les documents XML une solution de stockage devient indispensable. A cet effet, il existe deux solutions possibles : des SGBD relationnels et des SGBD natifs XML [Varandat 2002] :

- ✚ **SGBD relationnel** : Afin de stocker les documents XML, le modèle relationnel s'appuie sur des tables, des colonnes et des champs. Les informations sont rangées dans des cases. Une fois créée, la structure de donnée est fixe et très complexes à modifier.
- ✚ **SGBD natif XML** : Conçues pour le XML, les bases de données natives XML exploitent pleinement sa richesse. Elles stockent et manipulent directement des arbres XML au lieu de passer par une structure intermédiaire (table relationnelle). Elles possèdent des index spécialisés permettant d'accéder aux composants d'un arbre de documents XML : éléments, attributs et texte.

#### 3.2. L'interrogation

Plusieurs langages d'interrogation des documents XML ont été proposés, afin d'exprimer les requêtes utilisateurs portant à la fois sur le contenu et la structure des documents. Parmi ces langages nous citons :

- ✚ **XQuery** : Est un langage permettant la manipulation et la génération de documents XML à partir des sources XML, bases de données relationnelles ou Objets. Il permet de faire des requêtes sur des sources XML et de créer un résultat en format XML.
- ✚ **XPath** : Est un langage d'interrogation des documents XML. Il permet de sélectionner certaines parties d'un document XML : des sous-arbres, des nœuds, des attributsetc...

### III. L'implémentation du moteur de recherche

Le but de notre travail est d'avoir un accès vers les bases d'images des manuscrits arabes, à partir des catalogues, qui contient des notices codés selon le format Dublin Core adapté.

Le principe consiste à effectuer une recherche dans le catalogue. Cette recherche passera par les étapes de stockage et d'indexation. Nous pourrions résumer le processus d'accès aux images des manuscrits dans le schéma suivant :

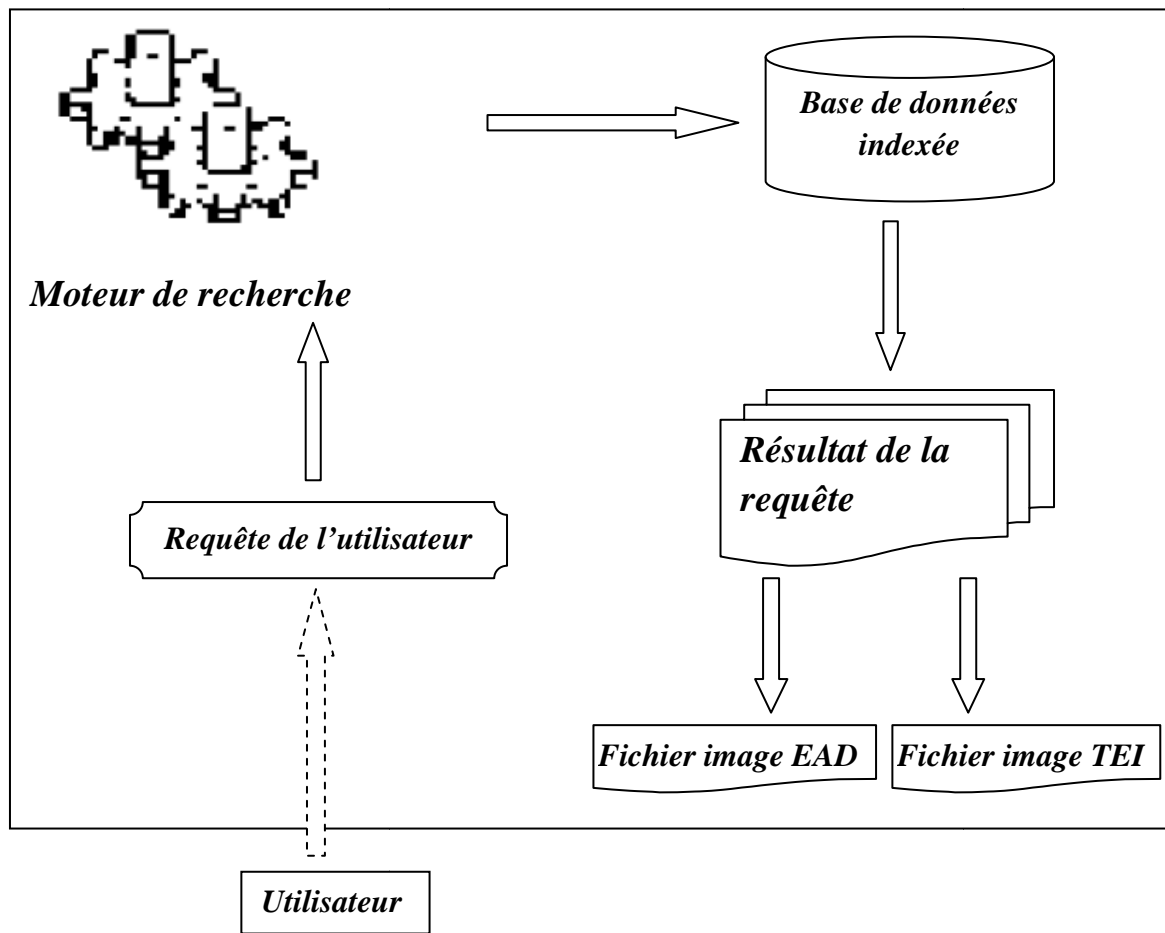


Figure 52: Schéma du processus d'accès aux images des manuscrits.

## 1. Schéma de stockage

### 1.1. Indexation

#### ✚ Structure d'index

Afin de réaliser notre solution, nous avons utilisé la technique du fichier inverse. Il s'agit d'une base de données indexée contenant la structure d'une notice, c'est-à-dire un document XML. Le modèle d'indexation devra chercher la notice dont les termes correspondent à ceux de la requête de l'utilisateur. En effet, chaque terme de catalogue doit contenir le numéro de la notice où il apparaît. C'est à partir des termes du catalogue que nous affichons la notice. La notice, affichée à son tour donne l'emplacement d'image du manuscrit.

#### ✚ La technique d'indexation des notices

En ce qui concerne l'indexation des notices bibliographiques, nous avons utilisé l'indexation basée sur les arbres. Ce type d'indexation des documents XML permet non seulement d'identifier les éléments du document mais aussi son contenu : les balises, les attributs et les données textuelles.

L'indexation permet l'identification des nœuds feuilles, contenant les termes et la détermination rapide des relations ancêtres-descendants.

## **1.2. Structure de la base de données**

### **1.2.1. La représentation des données**

Rappelons que la représentation des notices utilisée est la forme arborescente. Cette forme permet à la fois de représenter le contenu et la structure. En effet, l'arbre est composé d'un ensemble de nœuds qui peuvent être soit des éléments, des attributs, des données ou bien des commentaires. Dans le but de faciliter la navigation dans l'arbre, la détermination rapide des relations ancêtres-descendants et d'avoir un accès rapide à un nœud nous utiliserons la numérotation de Dietz vue plus haut.

Afin, de parcourir l'arbre du document XML, il faut préciser les valeurs pré-ordre et post-ordre. Ces valeurs seront assignées durant le parcours de l'arbre, pour assurer la gestion de la structure et de contenu. Ce mode de numérotation permet d'effectuer à la fois un parcours pré-ordre et un parcours post-ordre. Par contre, dans notre cas grâce à un tri dans l'ordre croissant de la numérotation pré-ordre, un simple parcours pré-ordre suffira pour déterminer les différents chemins ainsi que les nœuds et leurs parents.

### **1.2.2. La représentation conceptuelle des données**

Les notices bibliographiques seront stockées en utilisant un SGBD relationnel (exemple: PostgreSQL). Chaque notice bibliographique sera composée d'un ensemble : de balises, de termes indexés, d'attributs et de valeurs. Elle est représentée avec des tables : colonnes et champs. L'élaboration du diagramme de classe, se fait à partir de la définition de différentes règles de gestion.

### **1.2.3. Les règles de gestion**

Les règles de gestion sont définies à partir de la réalité avec la prise en considération de tous les principes et les aspects sur lesquels se base le domaine d'étude concerné. Dans notre cas et à partir des concepts étudiés, nous pouvons établir les règles suivantes :

1. Un catalogue peut contenir une ou plusieurs notices.
2. Un catalogue peut être composé d'un ou plusieurs catalogues (Cas de multilinguisme où un catalogue est formé de la version originale et des catalogues équivalents dans d'autres langues).
3. Une notice est définie à travers son identificateur et son nom.
4. Une notice peut être composée d'une ou plusieurs notices.
5. Une notice se décompose en balises et feuilles.
6. Les balises et les feuilles contiennent un ou plusieurs termes.
7. Un terme est décrit par son nom et son poids.
8. Une balise se caractérise par son identificateur, une valeur préordre, une valeur post ordre et une valeur parent.
9. Une notice peut avoir plusieurs balises qui ont les mêmes noms.
10. Une balise peut avoir un ou plusieurs attributs.
11. Une balise peut avoir un ou plusieurs valeurs d'un attribut.
12. Un attribut est décrit par son identificateur, son nom.
13. Une feuille est définie par un identificateur, une valeur pré-ordre, une valeur post-ordre et une valeur parent.

### 1.2.4. Le diagramme de classe

Partant des règles de gestion définies précédemment, le diagramme de classe utilisant le formalisme UML sera représenté comme suit :

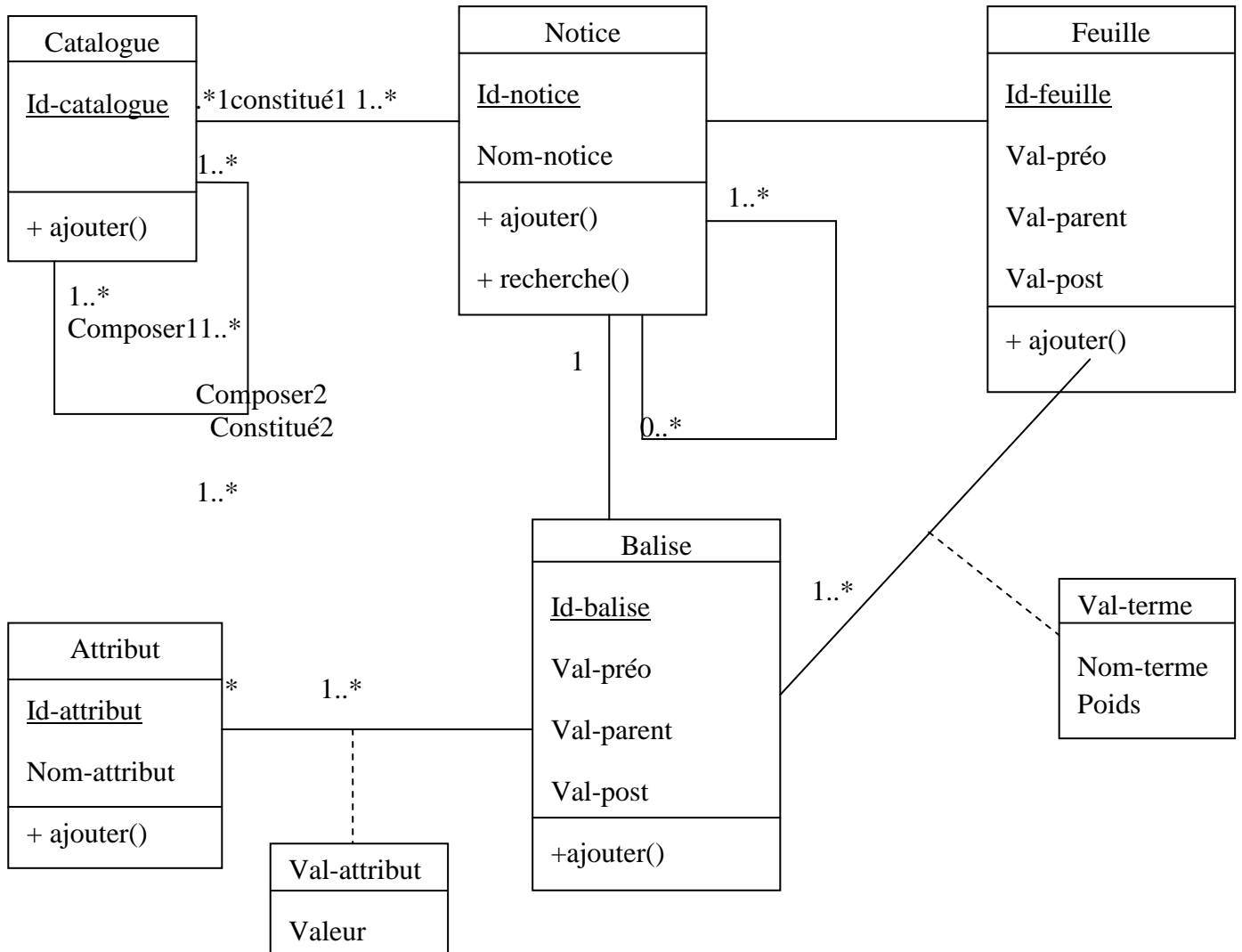


Figure 53: Le diagramme de classes.

### 1.2.5. La représentation relationnelle des données

Le diagramme de classe nous sert de référence pour l'implémentation de la base de données, car il met en évidence toutes les classes entités et leurs attributs. En effet, dans cette section nous allons traduire le diagramme de classe en schéma relationnel, tout en représentant l'implémentation de la base de données.

### **a. Transformation de diagramme en MLD**

Pour transformer le diagramme de classe UML en MLD (Modèle Logique de Données) on doit passer par les règles de transformation suivantes [Gnahou 2008]:

#### **1. Règles de transformation des classes d'objets**

- ✚ La classe se transforme en une table (relation).
- ✚ Les attributs de la classe deviennent des attributs de la table.
- ✚ Choisir un attribut (ou groupe d'attribut) de la classe pouvant jouer le rôle de clé primaire; si aucun attribut ne convient, il faut en ajouter un à la table.

#### **2. Transformation des associations**

Les règles de transformation des associations dépendent des cardinalités/multiplicité des associations, et nous pouvons distinguer trois grandes familles d'associations :

➤ **Association un à plusieurs [(1..1) ou (0..1)] à [(0..\*) ou (1..\*)]**

Dans ce cas il est nécessaire d'ajouter un attribut du type clé étrangère dans la relation fils de l'association. L'attribut porte le nom de la clé primaire de la relation père de l'association, c'est-à-dire la clé de la relation père migre dans la relation fils.

➤ **Association plusieurs à plusieurs [(0..\*) ou (1..\*)] à [(0..\*) ou (1..\*)]**

L'association ou la classe-association devient une relation. La clé primaire est composée des clés primaires des relations obtenues. Les éventuels attributs de la classe-association deviennent des attributs de la nouvelle relation.

➤ **Association binaires [(0..1) ou (1..1)] à [(0..1) ou (1..1)]**

Il faut ajouter un attribut clé étrangère dans la relation de la classe ayant la multiplicité minimale, l'attribut porte le nom de la clé primaire de la relation dérivée de l'entité (classe) connecté à l'association.

### **b. La représentation relationnelle des données**

A partir de diagramme de classes et les règles de transformation cités précédemment, nous obtiendrons le schéma relationnel suivant :

Catalogue (Id-catalogue).

Composer1 (Id-catalogue, Id-catalogue-composer1).

Notice (Id-notice, Nom-notice, Id-catalogue\*).

Composer2 (Id-notice, Id-notice-composer2).

Contient (Id-catalogue, Id-notice).

Feuille (Id-feuille, Val-préo, Val-parent, Val-post, Id-notice\*).

Constitué1 (Id-notice, Id-feuille).

Balise (Id-balise, Val-préo, Val-parent, Val-post, Id-notice\*).

Constitue2 (Id-notice, Id-balise).

Val-terme (Id-balise, Id-feuille, nom-terme, Poids).

Concerne (Id-feuille, Id-terme).

Attribut (Id-attribut, Nom-attribut).

Val-attribut (Id-balise, Id-attribut, Valeur).

### 1.2.6. La représentation physique de données

Après avoir élaboré le schéma relationnel de la base de données, nous allons passer à un autre niveau. Ce niveau concerne les structures de données à mettre en œuvre dans la base de données, ainsi nous aboutissons au modèle physique de données suivant :

#### 1. Table Catalogue

<i>Nom de champs</i>	<i>Type de données</i>	<i>Description</i>	<i>Clef</i>
Id-catalogue	Serial(5)	Identificateur de catalogue	Primaire

#### 2. Table Composer1

<i>Nom de champs</i>	<i>Type de données</i>	<i>Description</i>	<i>Clef</i>
Id-catalogue, Id-catalogue-composer1	Serial(10)	Identificateur	Primaire

#### 3. Table Notice

<i>Nom de champs</i>	<i>Type de données</i>	<i>Description</i>	<i>Clef</i>
Id-notice	Serial(5)	Identificateur de notice	Primaire
Nom-notice	Varchar(30)	Nom physique de la notice	
Id-catalogue	Serial(5)	Identifiant du catalogue contenant la notice.	Etrangère

#### 4. Table Composer2

<i>Nom de champs</i>	<i>Type de données</i>	<i>Description</i>	<i>Clef</i>
Id-catalogue, Id-notice-composer2	Serial(10)	Identificateur	Primaire

5. Table Feuille

<i>Nom de champs</i>	<i>Type de données</i>	<i>Description</i>	<i>Clef</i>
Id-feuille	Serial(5)	Identificateur d'une feuille	Primaire
Val-préordre	Serial(5)	Valeur pré-ordre	
Val-parent	Serial(5)	Valeur parent (Val-préo du neoud parent)	
Val-postordre	Serial(5)	Valeur post-ordre	
Id-notice	Serial(5)	Identificateur de la notice contenant la feuille	Etrangère

6. Table Constitué1

<i>Nom de champs</i>	<i>Type de données</i>	<i>Description</i>	<i>Clef</i>
Id-notice, Id-feuille	Serial(10)	Identificateur	Primaire

7. Table Contient

<i>Nom de champs</i>	<i>Type de données</i>	<i>Description</i>	<i>Clef</i>
Id-catalogue, Id-notice	Serial(5)	Identificateur de catalogue	Primaire

8. Table Balise

<i>Nom de champs</i>	<i>Type de données</i>	<i>Description</i>	<i>Clef</i>
Id-balise	Serial(5)	Identificateur de balise	Primaire
Val-préordre	Serial(5)	Valeur pré-ordre	
Val-parent	Serial(5)	Valeur parent( Val-préo du nœud parent)	
Val-postordre	Serial(5)	Valeur post-ordre	
Id-notice	Serial(5)	Identificateur de la notice contenant la balise	Etrangère

9. Table Constitue2

<i>Nom de champs</i>	<i>Type de données</i>	<i>Description</i>	<i>Clef</i>
Id-notice, Id-balise	Serial(10)	Identificateur	Primaire

**10. Table Val-terme**

<i><b>Nom de champs</b></i>	<i><b>Type de données</b></i>	<i><b>Description</b></i>	<i><b>Clef</b></i>
Id-balise, Id-feuille	Serial(5)	Identificateur du terme (identifiant de la balise et de la feuille contenant le nœud)	Primaire
nom-terme	Varchar(30)	Le terme indexé	
Poids-terme	Serial(10)	poids du terme qui est mis à jour chaque fois qu'on indexe de nouveaux documents	

**11. Table Attribut**

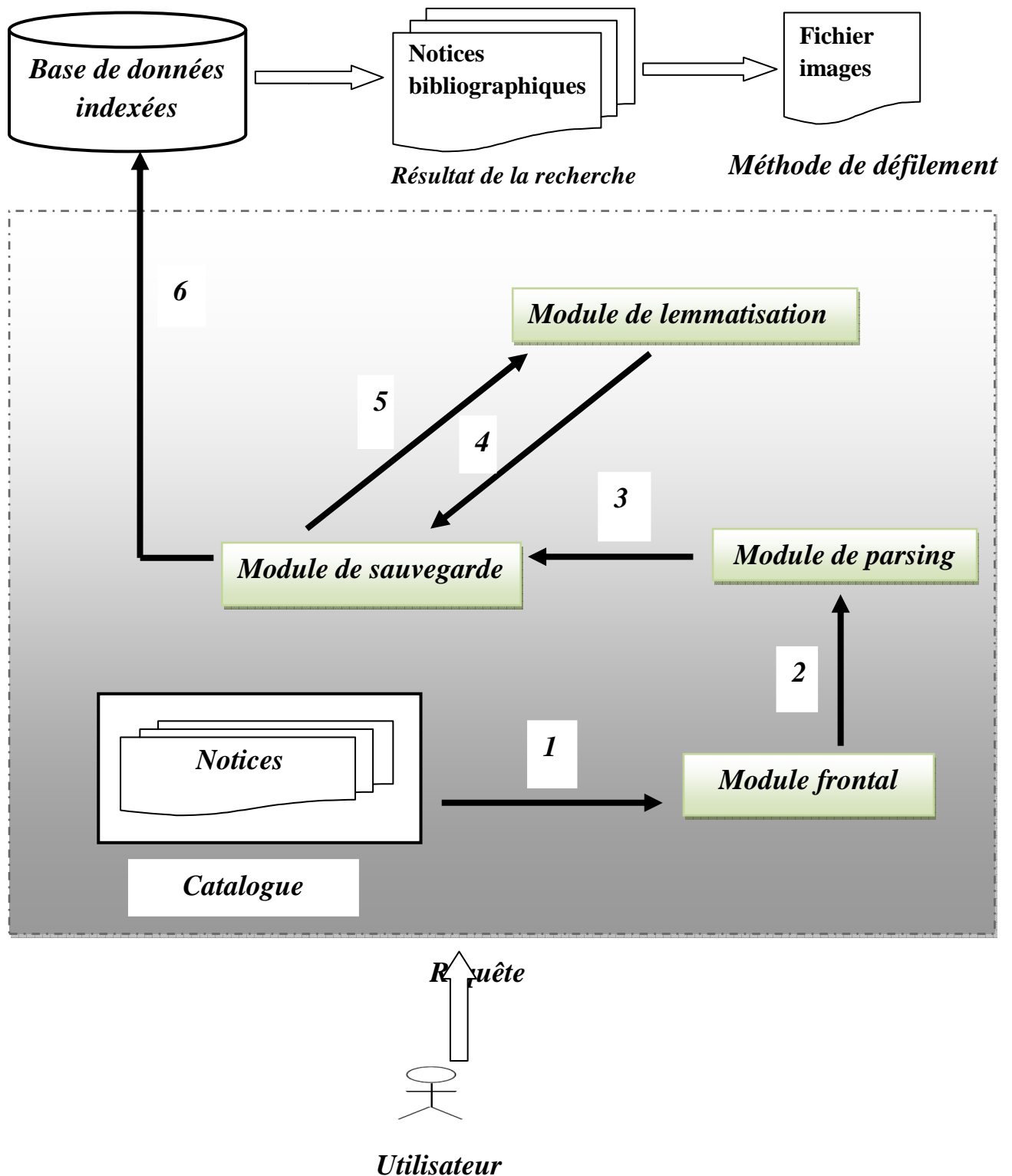
<i><b>Nom de champs</b></i>	<i><b>Type de données</b></i>	<i><b>Description</b></i>	<i><b>Clef</b></i>
Id-attribut	Serial(5)	Identificateur de l'attribut	Primaire
Nom-attribut	Varchar(30)	Nom de l'attribut	

**12. Table Val-attribut**

<i><b>Nom de champs</b></i>	<i><b>Type de données</b></i>	<i><b>Description</b></i>	<i><b>Clef</b></i>
Id-balise, Id-attribut	Serial(10)	Identificateur de la valeur (identifiant de la balise et de l'attribut contenant la valeur)	Primaire
Valeur	Varchar(30)	La valeur de l'attribut	

**2. *L'architecture générale du processus de recherche***

Le schéma ci-dessous représente l'architecture générale du processus de recherche, et en particulier les différents modules constituant le processus d'indexation et leur organisation :



**Figure 54:** L'architecture générale du processus de recherche.

À partir de ce schéma nous distinguons quatre modules qui constituent notre processus d'indexation : module frontal, module de parsing, module de lemmatisation, module de sauvegarde.

Dans ce qui suit nous allons développer ces différents modules, en précisant leur fonctionnement et leur utilité.

## 1. Module frontal

C'est un module essentiel qui a pour rôle de :

- Lister tout les chemins des notices du catalogue.
- Tester l'existence d'un chemin dans la base de données indexée.
- Retourner la liste des notices non indexées, afin de les indexer.

### Algorithme

#### Variable

- `_Vect_Liste_Catalogue` : vecteur de chaine de caractères destiné à contenir l'emplacementPhysique des fichiers issu du catalogue ;
  - `_Vect_Liste_BDD` : vecteur de chaine de caractères destiné à contenir l'emplacementPhysique des fichiers préalablement indexés et sauvegardés dans la BDD.
  - `Vect_Liste_Resultat` : vecteur de chaine de caractères destiné à contenir l'emplacementPhysique des fichiers candidats à une nouvelle indexation.
- `i, j` : entier ;

Trouvai : booléen

#### Début

Récupéré la liste des chemins des notices du catalogue : `Vect_Liste_Catalogue`.

Récupéré la liste des chemins figurant dans la base de donnée indexé dans :

`Vect_Liste_BDD`.

Trouvai = faux ;

**Pour** (`i=0` à taille `Vect_Liste_Catalogue`) **faire**

#### Début

**Pour** (`j=0` à taille `Vect_Liste_BDD`) **faire**

#### Début

**Si** (`Vect_Liste_Catalogue [i] = Vect_Liste_BDD [j]`) **alors**

{ Trouvai = vrai ; Sortir ; }

#### Fsi

**Si** trouvai = faux **alors**

Ajouter à `vect_list_Resultat` `vect_liste_Catalogue [i]`

#### Fsi

Trouvai := faux ;

#### Fin

#### Fin

#### Fin

## 2. Module de parsing

Le module de parsing a pour rôle de séparer le contenu d'une notice encodé en XML de la structure. La numérotation est basée sur le principe deDietz et le parcours du document en format XML (Notice bibliographique) se fait selon le mode de parcours pré-ordre.

Le module de parsing s'appuie sur L'API SAX (Simple API for XML). Notre but étant d'identifier les différents éléments de la notice afin de permettre une meilleure manipulation des différentes informations contenues dans ces éléments.

### Algorithme

**Variable :**

    Booléen : trouvai ;

    Chaîne de caractère : id notice ;

Pre=0 ; par=-1 ; niveau=0 ;

**Début**

    Pour chaque événement rencontré

**Début**

**Si** (événement= balise ouvrante) **alors**

**Si** (nom\_balise=dc :identifier) **alors**

                    Trouvai=vrai ;

**Fsi**

        niveau := niveau+1 ;

        vectpre[niveau] := pre ;

            Mettre dans vecteurBalise l'enregistrement (nom\_balise, Pre, par)

            //nom de la balise, valeur de pre-ordre, valeur de pré ordre du nœud père

**Si** (balise contient Attribut) **alors**

            Pour chaque Attribut

                Mettre dans vecteurAttribut l'enregistrement (nom\_attribut, Val-attribut, Pre)

**Fsi**

            Par := pre ;

        Pre := pre+1 ;

**Fsi**

**Si** (événement=Texte) **alors**

**Si** (trouvai=vrai) **alors**

                    Id\_notice :=texte

**Fsi**

                    Trouvai=faux ;

**Si** (texte est non vide) **alors**

                    Mettre dans vecteur Texte l'enregistrement (texte, pre, par)

                Pre := pre+1 ;

**Fsi**

**Fsi**

**Si** (événement=balise fermante) **alors**

                        niveau := niveau-1 ;

                        par := vect\_pre [niveau];

**Fsi**

**Fin**

**Fin**

### 3. Module de lemmatisation

Le module de lemmatisation reçoit en entrée un texte. Ce texte est issu des nœuds feuilles d'un document, ainsi il élimine les mots vides qui sont contenus dans la liste de mot vide correspondante, et retourne un vecteur contenant la liste des mots clés.

#### Algorithme

```

Variable
    Vecteur : special_Car ;(vecteur contient les caractères spéciaux).
    Vecteur : list_mot ;(Liste des mots de la balise feuille).
    Booléen : Mot_encour;
    Entier : i ;
Début
    Charger dans special_Car la liste de caractères spéciaux.
    Tant que (i<taille texte) faire
        Début
        Mot_encour := faux ;
        Si (texte[i] appartient à special_Car) alors
            Si (Mot_encour =vrai)
                Début
                Ajouté mot a list_mot ;
                Mot = chaine vide;
            Fin
        Mot_encour = faux;
        Fin
    Fsi
        Sinon
            Si (! Mot_encour) alors
                Début
                Mot := mot+texte[i] ;
            Fin
        Fsi
    Fsino
        i++;
    Fin tant que
    Ajouté mot a la liste des list_mot ;
Fin

```

#### 4. Module de sauvegarde

Ce module reçoit en entrée les résultats issus des trois modules précédents tels que : la liste des chemins, le résultat de parsing et le résultat de lemmatisation. Egalement, en sortie il actionne la sauvegarde dans la base de données de toutes les notices indexées.

#### Algorithme

Récupérer Vect\_Liste\_Resultat (vecteur de chaîne de caractères résultants de module frontal qui contient l'emplacement Physique des fichiers prêt à être indexé).

**Début**

**Tant que** ( $i < \text{taille Vect\_list\_resultat}$ ) **faire**

**Début**

Passer Vect\_list\_resultat [i] au module de parsing ; /\*Récupérer la notice\*/

Récupérer identifiant de la notice ;

**Sauvegarder**, nom\_notice et id\_notice dans la table notice.

**Tant que** ( $j < \text{taille vecteur\_texte}$ ) **faire**

**Début**

Pour chaque vecteur\_texte [j] sauvegarder dans la table feuille les valeurs pré ordres et parent, id\_notice.

Passer le texte de vecteur\_texte [j] comme paramètre au module de lemmatisation.

**Tant que** ( $k < \text{taille vect\_terme}$ ) **faire**

**Début**

**Sauvegarder** dans la table terme id\_feuille et le terme ;

$K := k + 1$  ;

**Fin**

**Fin Tant que**

$J := j + 1$  ;

**Fin tant que**

**Tant que** ( $p < \text{taille vect\_balise}$ ) **faire**

**Début**

Sauvegarder pour chaque vect\_balise [p] la valeur de preordre, nom de la balise et, id\_notice dans la table balise ;

$P := p + 1$  ;

**Fin**

**Fin tant que**

**Tant que** ( $m < \text{taille vect\_attribut}$ ) **faire**

**Début**

Sauvegarder pour chaque vect\_attribut [m] le nom de l'attribut, sa valeur et identifiant de la balise.

$M := m + 1$  ;

**Fin**

**Fin tant que.**

**Fin.**

**Fin.**

## ***IV. Présentation des outils de développement***

Avant de représenter les différentes interfaces graphiques du système de recherche, nous allons tout d'abord spécifier les outils utilisés pour nos développements et les avantages qu'ils offrent.

### ***1. Le langage de programmation Java***

Afin de réaliser l'interface permettant aux utilisateurs de manipuler notre système, nous avons choisi le langage Java.

Le langage java est un langage de programmation orienté objet mis au point par Sun Microsystems. Il s'inspire du C++ et des techniques approuvées en Small talk et autres langages (organisation en classe, utilisation d'un ramasse-miettes (garbagecollector), exécution à l'aide d'une machine virtuelle, gestion d'exception. . .), tout en évitant les inconvénients de la plupart de ces langages. Sa caractéristique principale est qu'il est indépendant de toute plate-forme, En effet la source est compilée en pseudo code ou bytecode puis exécutée par un interpréteur Java appelé Java Virtual Machine (JVM). Ce concept est à la base du slogan de Sun pour java : WORA (Write Once, RunAnywhere : écrire une fois, exécuter partout). En effet, le byte code, s'il ne contient pas de code spécifique à une plate-forme particulière peut être exécuté et obtenir quasiment les même résultats sur toutes les machines disposant d'une JVM. Le code reste indépendant de la machine sur laquelle il s'exécute. Cette indépendance est assurée au niveau du code source grâce à Unicode et au niveau du byte code.

Java est caractérisée aussi pour sa sécurité, un programme Java plante ne menace pas le système d'exploitation. Il ne peut pas y avoir d'accès direct à la mémoire, ainsi que sa simplicité de mise en œuvre.

### ***2. NetBeans***

Le JDK (Java Development Kit), fournit gratuitement par Sun, regroupe l'ensemble des éléments permettant le développement, la mise au point et l'exécution des programmes Java. Ainsinous avons utilisé JDK (Java Development Kit) pour l'exécution de codes java sous NetBeans7.0.1 qui est un IDE (IntegratedDevelopmentEnvironment) créé par Sun Microsytems. Ce dernier contient toutes les fonctionnalités d'un bon environnement de développement intégré (un debugger, un outil de modélisation UML, une interface graphique et fonctionnelle, etc.).

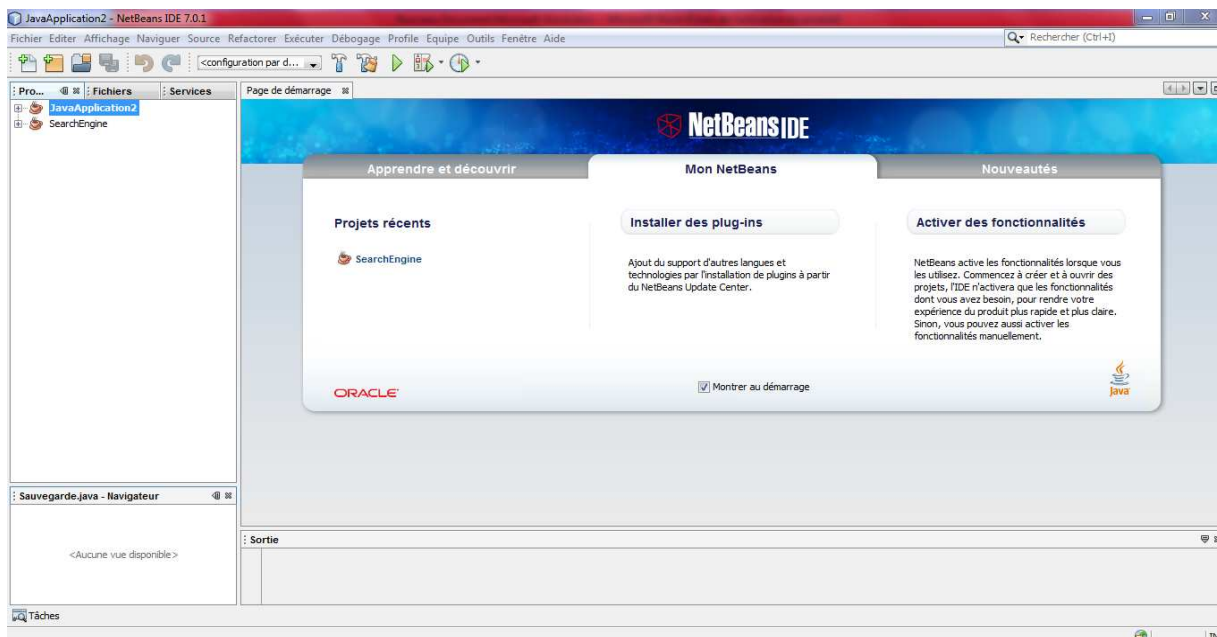


Figure 55: L'interface principale de NetBeans.

### 3. PostgreSQL

PostgreSQL est un SGBDR (Système de Gestion de Base de Données Relationnelles) développé à l'université de Californie au département des sciences informatiques de Berkeley, fonctionnant sur diverses plates-formes matérielles sous différents systèmes d'exploitation. PostgreSQL est un SGBDR robuste et puissant :

- Il reconnaît la plupart des spécifications SQL tout en offrant de nombreuses fonctionnalités modernes (requêtes complexes, clés étrangères, triggers, vues, intégrité transactionnelle)
- Il inclut l'API ODBC permettant à n'importe quelle application supportant ce type d'interface d'accéder à des bases de données de type PostgreSQL.
- Il peut être étendu de plusieurs façons par l'utilisateur en ajoutant par exemple de nouveaux types de données ou de nouvelles fonctions.

De plus, PostgreSQL est d'être un logiciel libre, c'est-à-dire gratuit et open source (les sources du logiciel sont disponibles dans [http : //www.postgresql.org](http://www.postgresql.org)).

### 4. L'interface Java/JDBC

Pour connecter l'application à la base de données, nous avons utilisé l'interface JDBC. JDBC est l'acronyme de Java Data Base Connectivity et désigne l'API définie par SUN Microsystems pour permettre un accès aux bases de données avec Java. Ensemble de classes Java regroupées dans le package `java.sql` et sont inclus dans la JDK depuis la version 1.1 de cette API qui permettent de se connecter à une base de données et de l'interroger afin d'en extraire des données. Pour pouvoir utiliser JDBC, il faut un pilote qui est spécifique au SGBD

qui contient la base avec laquelle on veut se connecter. Pour ce qui est de PostgreSQL on utilise " JDBC3 PostgreSQL pilote " disponible sur le site de Sun Microsystems.

Nous citons ci-dessous les 4 classes principales utilisées pour accéder à la base de données.

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

## V. Présentation des interfaces graphiques

### 1. Interface d'accueil

L'interface illustrée par la figure 56 ci-dessous représente l'interface principale de notre système de recherche.

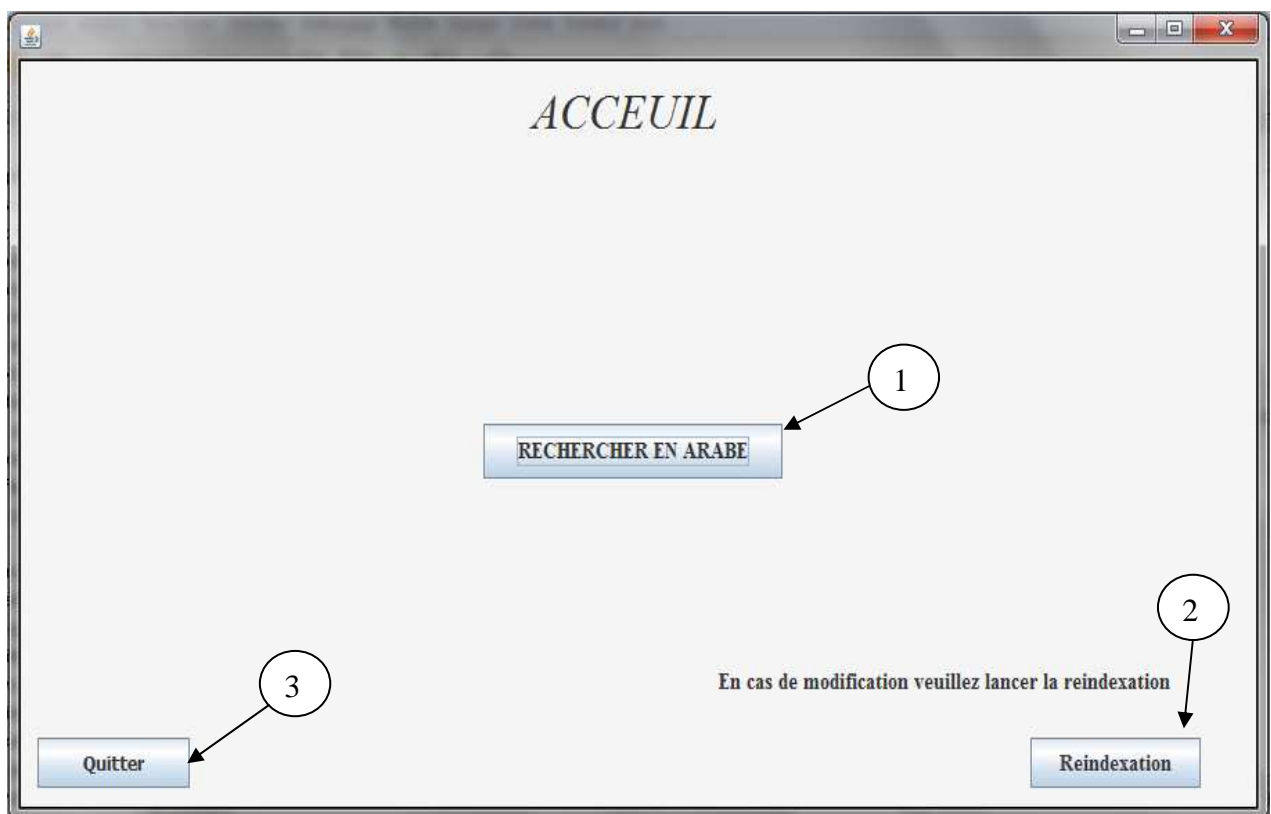


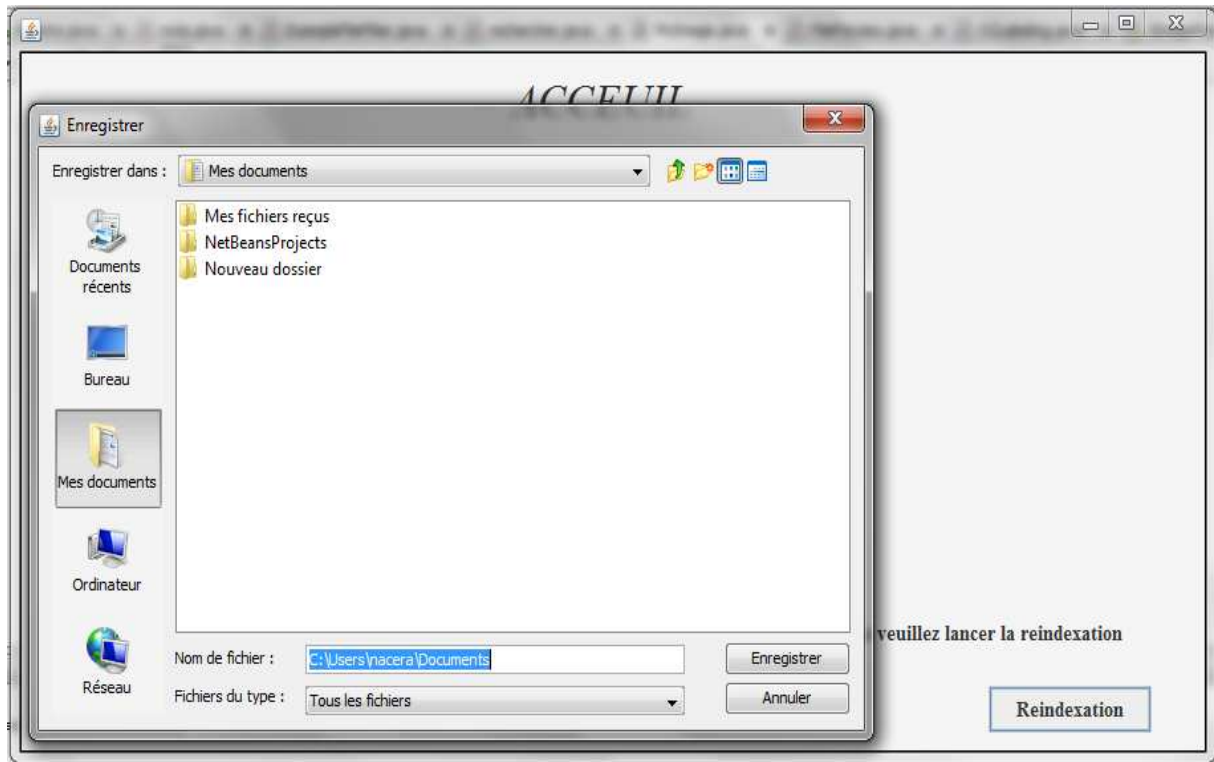
Figure 56: Interface d'accueil.

▪ **Fonctionnement**

1. Permet d'afficher l'interface de recherche.
2. Permet d'effectuer une réindexation.
3. Permet de quitter l'application.

**2. Interface de réindexation**

En cas de modification du catalogue, un clic sur le bouton Reindexation fait apparaître l'interface ci-après :



**Figure 57:** Interface de réindexation.

Cette interface permet de choisir un catalogue et de lancer la réindexation avec le bouton enregistrer.

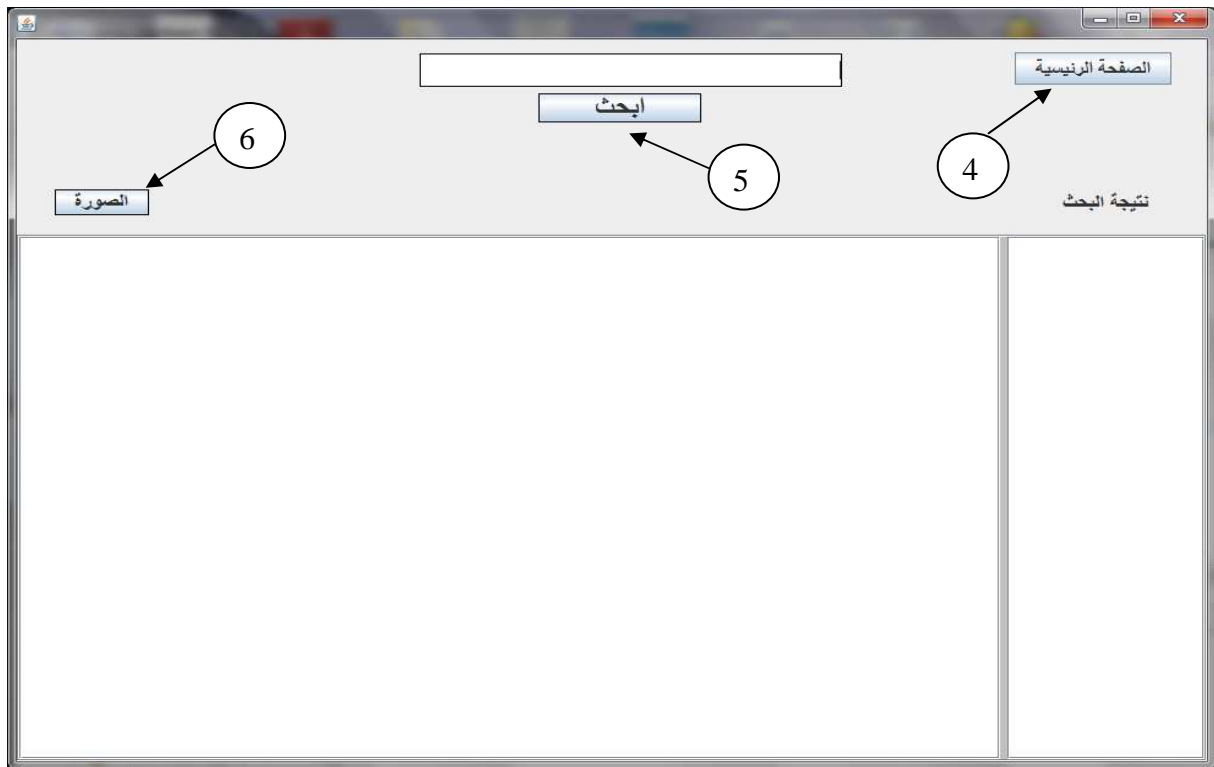
A la fin de réindexation l'interface ci-dessous s'affiche :



**Figure 85:**Interface de fin de réindexation.

### 3. Interface de recherche

Un clic sur le bouton RECHERCHER EN ARABE de l'interface d'accueil permet d'afficher l'interface suivante :

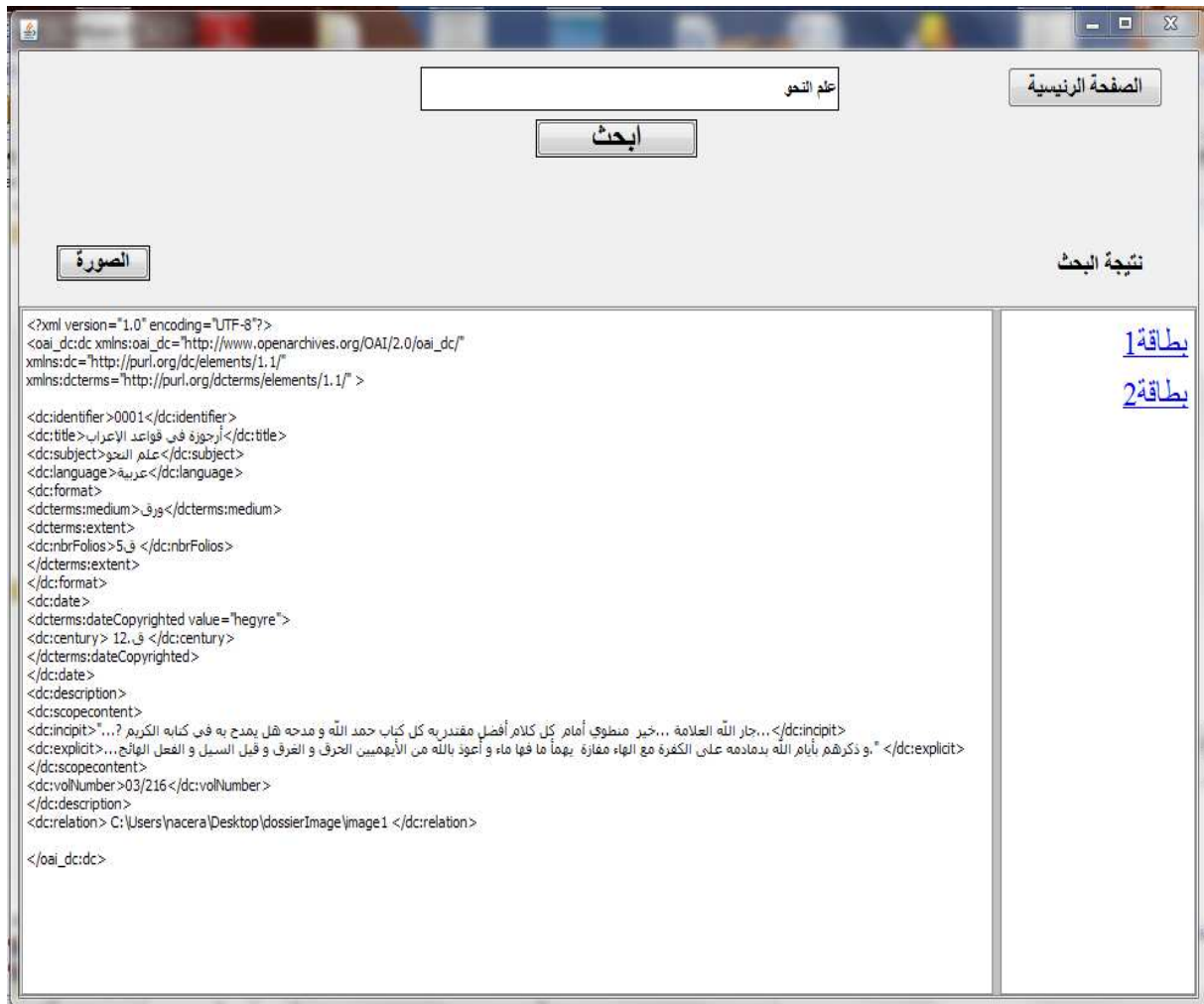


**Figure 59:** Interface de recherche.

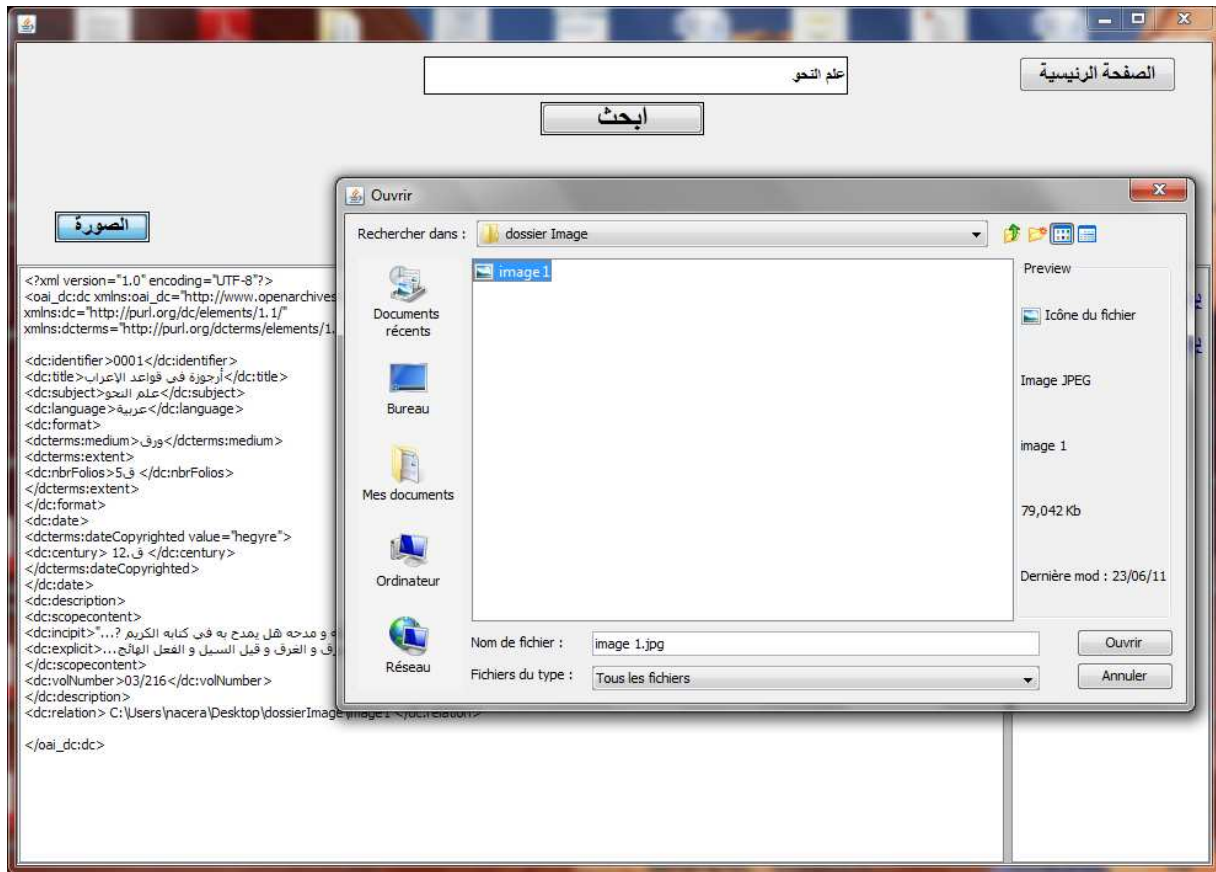
#### ▪ Fonctionnement

4. Permet d'afficher l'interface d'accueil.

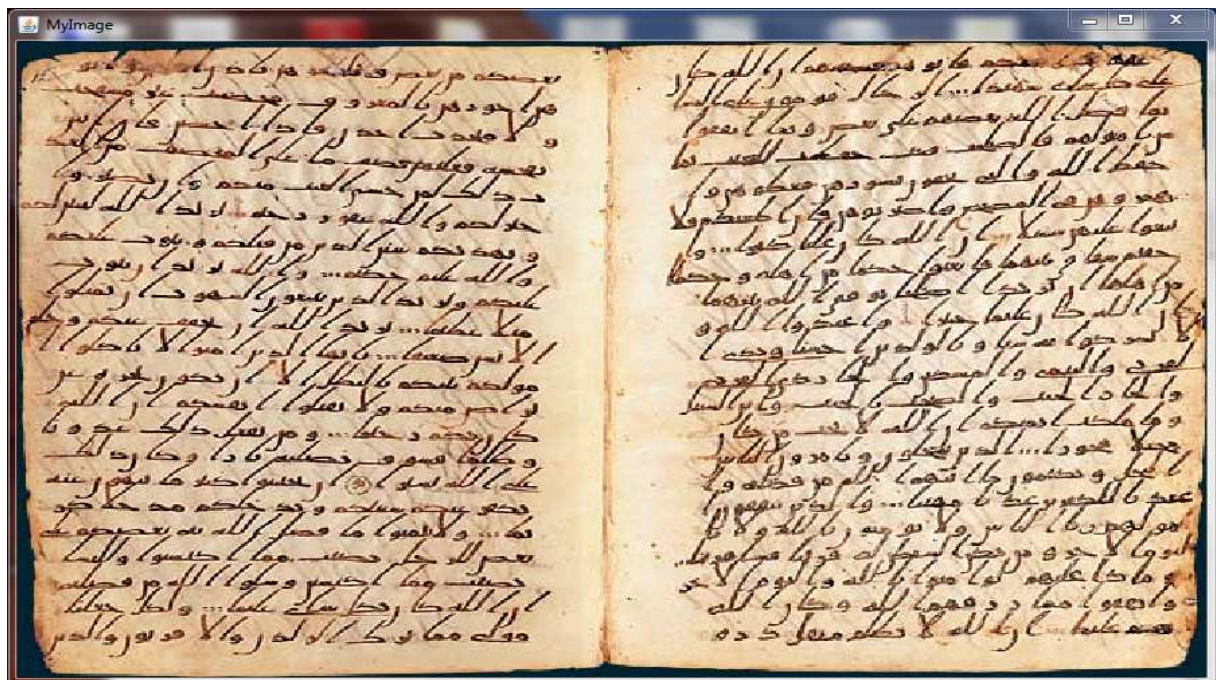
5. Permet de lancer une recherche après la saisie de la requête dans la zone d'édition.  
L'interface ci-dessous illustre un exemple de recherche.



6. Donne l'accès vers l'emplacement de l'image correspondante au résultat retourné. En cliquant sur ce bouton l'interface ci-dessous s'affiche :



Un clic sur le bouton ouvrir de cette interface, après la sélection de l'image dont l'emplacement est cité dans la notice bibliographique affichée, fait apparaître cette image. Cela est illustré par la figure ci-dessous :



## **VI. Conclusion**

Ce chapitre a été consacré à l'implémentation du moteur de recherche. Dans un premier temps, nous avons présentés les techniques d'indexation et de stockage utilisés, en se basant sur les principes de recherche d'informations dans les documents XML. Cet aspect a été suivi par les représentations conceptuelles, logiques et physiques de la base de données, ainsi que l'implémentation des différents modules d'indexation et de sauvegarde. A la fin, nous avons présentés outils de développement et les différentes interfaces graphiques.

# ***Conclusion générale***

# Conclusion générale

Le catalogage des manuscrits arabes est un aspect stratégique qu'il faut maîtriser pour une meilleure gestion des manuscrits. Il est la clé qui assure une bonne visibilité des manuscrits numérisés. En effet, la mise en ligne de ces derniers ne pourrait avoir d'intérêt sans l'existence d'un instrument de recherche efficace : Rôle de prédilection du catalogue.

Le problème de catalogage des manuscrits arabes numérisés réside dans la détermination de métadonnées portées par un protocole bien défini. Plusieurs travaux ont proposés leur propre démarche. Notre travail se place dans une perspective de fédérateur, en n'excluant aucune des propositions émises. Bien au contraire, nous poussons la réflexion plus loin en proposant d'intégrer les métadonnées de structure des manuscrits à l'instar de la table de matière, de l'index et des annotations. C'est en fait, un véritable travail de fourmi que devrait fournir le catalogueur afin d'accroître l'accessibilité aux manuscrits numérisés.

Notre travail ne se limite pas à l'aspect de catalogage, mais nous proposons une solution d'interopérabilité entre les catalogues quelque soit leur format d'encodage, ce qui constituera un catalogue collectif virtuel des manuscrits arabes numérisés, cet aspect sera mis en œuvre en faisant appel au format Dublin Core. Le catalogue collectif virtuel permettra d'atteindre l'ensemble des manuscrits arabes numérisés répondant à une requête donnée et ce, quelque soient le format d'encodage du catalogue de ces œuvres et de leurs lieux de stockage.

Notre travail trouve son utilité dans la mise en place d'une plateforme virtuelle d'accès aux manuscrits arabes numérisés. Notre système effectuera des passages sur les différents systèmes qui auraient mis en lignes leurs catalogues. La procédure de moissonnage se chargera pour faire migrer les données des divers systèmes vers notre plateforme. Ainsi, toute interrogation de notre système se verra offrir les réponses les plus exhaustives sur l'existence de tout manuscrit, du fait qu'un système fédérateur a été mis en place.

Notons toutefois, qu'une approche ontologie pourrait offrir de meilleurs résultats du fait de l'intégration de l'aspect sémantique, qui fait défaut à notre système de recherche d'information, basé sur le modèle booléen.

# ***Annexe A***

## *XML*

### *I. Définition*

Le XML, acronyme d'eXtensibleMarkupLanguage (langage de balisage extensible), est un langage informatique qui sert à enregistrer des données textuelles.

XML est donc un langage de balises, qui permet de codifier de manière générale toutes sortes de données. Les balises n'étant pas prédéfinies, on peut les définir librement via un fichier qui contiendrait la définition des balises (par exemple avec la DTD).

### *II. Enjeux de XML*

1. XML inclut les informations de structure en plus du contenu en données, ce qui simplifie les traitements automatiques (parsing, génération de rapports, etc).
2. Format proche du texte, donc lisible par les humains dans certaines limites.
3. Format qui tend à s'imposer partout : la plupart des langages de programmation offrent des outils pour lire/écrire en XML, ainsi que la plupart des SGBD permettent d'exporter en XML, ce qui facilite le portage de leurs données à un autre système.

### *III. Le document XML*

#### *1. Définition*

« Un document est un ensemble d'informations sélectionnées, assemblées et organisées pour permettre à un utilisateur donné de remplir une mission donnée »<sup>1</sup>. Cette définition de 1991 est plus que jamais d'actualité. Initialement destinée aux seuls documents sur papier, elle convient aux fichiers XML, qu'ils représentent des documents papier, électroniques, ou des données.

En particulier, on comprend que cette définition convienne parfaitement bien aux fragments XML échangés entre systèmes : seul le mot « utilisateur » devrait maintenant être remplacé par l'expression « utilisateur ou système ». Désormais, tout fichier XML bien formé est appelé document, plus précisément document XML.

#### *2. Structure d'un document XML*

En réalité un document XML est structuré en trois parties :

**2.1. Prologue :** C'est une sorte d'entête comportant des informations sur la version de la norme XML utilisée, les instructions de traitements ainsi une déclaration de type de document (DTD), il est constitué de :

---

<sup>1</sup>Guide pour la réalisation de spécifications d'élaboration de documentation structurée » – juin 1991 – Direction Générale de l'Armement.

### a. Déclaration de XML

Cette partie est facultative, elle fournit des informations globale sur la norme XML tel que :

```
<? Xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
```

On distingue trois informations fournies dans cette déclaration :

- ✚ **Version** : La version XML utilisée dans le document.
- ✚ **Encoding** : Représente le jeu de codage de caractères utilisé. Par défaut, l'attribut *encoding* a la valeur UTF-8. Cela permet à l'ordinateur de savoir quel caractère il doit afficher en réponse aux combinaisons de 1 et de 0 que contient le fichier sur le disque dur.
- ✚ **Standalone** : Dépendance du document par rapport à une déclaration de type de document. *Standalone* a la valeur *yes*, le processeur de l'application n'attend aucune déclaration de type de document extérieure au document. Sinon, le processeur attend une référence de déclaration de type de document. La valeur par défaut est *yes*.

### b. Instruction de traitement

Une instruction de traitement est une instruction interprétée par l'application servant à traiter le document XML. Elle ne fait pas totalement partie du document. Les instructions de traitement qui servent le plus souvent sont la déclaration XML ainsi que la déclaration de feuille de style. Exemple d'instruction de traitement :

```
<?xml-stylesheet type="text/css" href="biblio.xsl"?>
```

Dans cet exemple :

- L'application est *xml-stylesheet*, le processeur de feuille de style de l'application traitant le document XML.
- Deux feuilles de style différentes peuvent être utilisées, les XSL (propres au XML) ainsi que les CSS (feuilles de style apparues avec le HTML).
- L'attribut *type* indique de quel type de fichier il s'agit (*text/css* pour les feuilles de style CSS, par exemple).
- L'attribut *href* indique l'URL du fichier. Cette instruction de traitement est notamment utilisée par les navigateurs Internet pour la mise en forme du document.

### c. Déclaration de type de document (DTD)

Cette déclaration facultative, permet de définir la structure de document. Voici un exemple de déclaration de type de document :

```
<!DOCTYPE biblio SYSTEM "biblio.dtd">
```

Ce type de déclaration est celui d'une déclaration de type de document externe. Elle définit l'ensemble des éléments utilisables dans le document, y compris l'élément-racine ainsi que l'emplacement où se trouve le fichier biblio.dtd dans lequel se trouve définie la structure du document.

### 1) *Les commentaires*

En XML, les commentaires commencent par `<!--` et se terminent par `-->`

### 2) *L'arbre d'éléments*

Un document XML peut se représenter sous la forme d'une arborescence d'*éléments*. Cette arborescence comporte :

- i. ***L'élément racine*** : L'élément-racine, est la base de document XML. Il est unique et englobe tous les autres éléments. Il s'ouvre juste après le prologue, et se ferme à la toute fin du document.
- ii. ***Les éléments*** : Les éléments forment la structure même du document : ce sont les branches et les feuilles de l'arborescence. Ils peuvent contenir du texte, ou bien d'autres éléments, qui sont alors appelés « éléments enfants », l'élément contenant étant quant à lui appelé logiquement « élément parent ».
- iii. ***Les attributs*** : Tous les éléments peuvent contenir un ou plusieurs attributs. Chaque élément ne peut contenir qu'une fois le même attribut. Un attribut est composé d'un nom et d'une valeur.

### *Exemple*

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<biblio>
<livre><!-- Élément enfant titre -->
<titre>Les Misérables</titre>
<auteur>Victor Hugo</auteur>
<nb_tomes>3</nb_tomes>
</livre>
<livre>
<titre>L'Assommoir</titre>
<auteur>Émile Zola</auteur>
<couverture couleur="rouge" />
</livre>
</biblio>
```

- iv. ***Les entités*** : Il existe des entités définissables et définies. Elles peuvent être analysables ou non, internes ou externes. La déclaration des entités s'effectue au sein de la DTD. Elles peuvent être utilisées aussi bien dans la DTD que dans le document XML.

**Exemple**

Caractère	Entité
&	&amp;
<	&lt;
>	&gt;

- v. **Les sections CDATA** : Une section CDATA est une section pouvant contenir toutes sortes de chaîne de caractères. Une section CDATA permet de définir un bloc de caractères ne devant pas être analysés par le processeur XML. Cela permet entre autres de garder dans un bloc de texte un exemple de code à afficher tel quel.

Exemple d'utilisation de CDATA :

```
<![CDATA [Une balise commence par un < et se termine par un >.]]>
```

- vi. **Règles de composition** : Un certain nombre de règles de base doivent être respectées :

1. Un nom d'élément ne peut commencer par un chiffre. Si le nom n'est composé que d'un seul caractère, ce doit être une lettre comprise entre « a » et « z » pour les minuscules, « A » et « Z » pour les majuscules. S'il est composé d'au moins deux caractères, le premier peut être « \_ » ou « : ». Le nom peut ensuite être composé de lettres, chiffres, tirets, tirets bas et deux points. La syntaxe XML est sensible à la casse (le format distingue majuscules et minuscules).
2. Toutes les balises portant un contenu non vide doivent être fermées. La balise de début, la balise de terminaison et le contenu entre les deux sont globalement appelés *élément* ;
3. Les balises n'ayant pas de contenu doivent se terminer par />.
4. Les valeurs d'attributs doivent être entre guillemets.

**Remarque** : Un document respectant tous ces critères est dit un document bien formé (wellformed).

### 3. Manipulation des docs XML

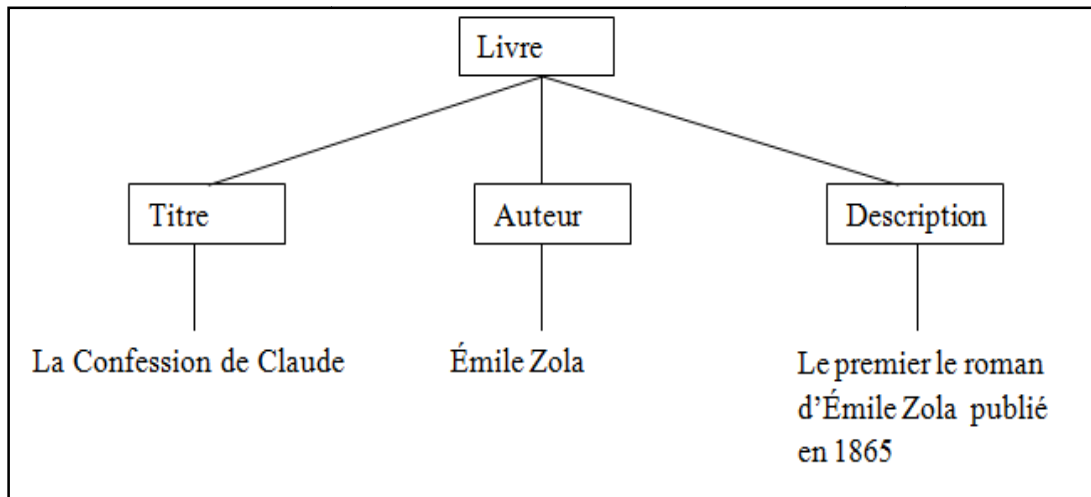
#### 3.1. DOM (Document Object Model)

DOM est l'acronyme de document Object model. C'est une spécification du W3C pour proposer une API qui permet de modéliser, de parcourir et de manipuler un document XML. Le principale rôle de DOM est de fournir une représentation mémoire d'un document XML sous la forme d'un arbre d'objets et d'en permettre la manipulation (parcours, recherche et mise à jour).

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

**Exemple :** Document XML et l'arbre DOM associé

```
<Livre>
<Titre> La Confession de Claude </Titre>
<Auteur>Émile Zola</Auteur>
<Description > Le premier le roman d'Émile Zola publié en 1865</Description >
</Livre>
```



**Figure :** Arbre DOM correspondant

### 3.2.SAX (Simple API for XML)

SAX est une API (Application Programming Interface) basée sur un modèle événementiel. SAX permet de déclencher des événements au cours de l'analyse du document XML. Une application utilisant SAX implémente généralement des gestionnaires d'événements, lui permettant d'effectuer des opérations selon le type d'élément rencontré.

Soit le document XML suivant

```
<Livre>
<Titre> La Confession de Claude </Titre>
<Auteur>Émile Zola</Auteur>
<Description > Le premier le roman d'Émile Zola publié en 1865</Description >
</Livre>
```

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

- start document.
- start élément: Livre.
- start élément: Titre.
- characters: La Confession de Claude.

- end élément: Titre.
- start élément: Auteur.
- characters: Émile Zola.
- end élément: Auteur.
- start élément: Description.
- characters: Le premier le roman d'Émile Zola publié en 1865.
- end élément: Description.
- end élément: Livre.
- end document.

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

L'API SAX définit les quatre interfaces suivantes :

- ✚ **DocumentHandler** : Possédant des méthodes renvoyant des événements relatifs au document :
  - startDocument() renvoyant un événement lié à l'ouverture du document.
  - startElement() renvoyant un événement lié à la rencontre d'un nouvel élément.
  - characters() renvoyant les caractères rencontrés.
  - endElement() renvoyant un événement lié à la fin d'un élément.
  - endDocument() renvoyant un événement lié à la fermeture du document.
- ✚ **ErrorHandler** : Possédant des méthodes renvoyant des événements relatifs aux erreurs ou aux avertissements.
- ✚ **DTDHandler** : Renvoie des événements relatifs à la lecture de la DTD du document XML.
- ✚ **EntityResolver** : Permet de renvoyer une URL (Uniform Resource Locator) lorsqu'une URI est rencontrée.

# *Annexe B*

## I. Protocole de description des manuscrits défini par la fondation El Saoud

Champ "Identification"	حقول التعريف
<ul style="list-style-type: none"> <li>- Numéro du manuscrit (Cote)</li> <li>- Numéro séquentiel du volume</li> <li>- Numéro de la copie microfilmée</li> <li>- Numérotation des feuilles</li> </ul>	<ul style="list-style-type: none"> <li>-رقم المخطوط (رقم الطلب)</li> <li>-الرقم التسلسلي للمخطوط</li> <li>-رقم النسخة المصورة</li> <li>-رقم الأوراق</li> </ul>
Champ "Contenu"	حقول المضمون
<ul style="list-style-type: none"> <li>- Titre du manuscrit</li> <li>- Auteur</li> <li>- Date de décès (Hégire)</li> <li>- Date de décès (Grégorien)</li> <li>- Thème</li> <li>- Langue</li> <li>- Incipit</li> <li>- Colophon</li> </ul>	<ul style="list-style-type: none"> <li>-العنوان الذي أتى بذكره بعد عبارة "سميته"</li> <li>-المؤلف</li> <li>-تاريخ الوفاة بالهجري</li> <li>-تاريخ الوفاة بالميلادي</li> <li>-الموضوع</li> <li>-اللغة</li> <li>-البداية بقسم الهالرحمن الرحيمهما كانتا البداية</li> <li>-النهاية</li> </ul>
Champ "Description première"	حقول الوصف البادي للمخطوط
<ul style="list-style-type: none"> <li>- Nom du copiste</li> <li>- Lieu de la copie</li> <li>- Date de la copie</li> </ul>	<ul style="list-style-type: none"> <li>-اسم الناسخ</li> <li>-مكان النسخ</li> <li>-تاريخ النسخ</li> </ul>
Champ "Description Physique du manuscrit"	حقول الوصف المادي للمخطوط
<ul style="list-style-type: none"> <li>- Numéro du volume</li> <li>- Numéro du cahier</li> <li>- Matière du manuscrit</li> <li>- Nombre de feuilles</li> <li>- Nombre de lignes par page</li> <li>- Etat du manuscrit (complet ou pas)</li> <li>- Dimension de la couverture (en mm)</li> <li>- Dimension des folios (en mm)</li> <li>- Style d'écriture</li> <li>- Dessins explicatif (Dessins, Schémas, Tableaux, ... etc.)</li> <li>- Enluminures</li> <li>- Encre (Couleur)</li> <li>- Type de reliure</li> <li>- Languette de fermeture</li> <li>- Résumé (Présent ou pas)</li> </ul>	<ul style="list-style-type: none"> <li>-رقم المجلد</li> <li>-رقم الجزء</li> <li>-مادة المخطوط</li> <li>-عدد الأوراق</li> <li>-عدد الأسطر في الورقة</li> <li>-حالة المخطوط (كاملا ناقص)</li> <li>-المقاس بالمليمتر للغلاف الخارجي (الطول والعرض)</li> <li>-المقاس بالمليمتر للورقة</li> <li>-نوع الخط</li> <li>-الرسوم والتوضيحية (الجدول، الرسوم، المخططات، الخرائط، ... الخ)</li> <li>-التزيين (زخارف، تذهيب... الخ)</li> <li>-المداد (لونه)</li> <li>-نوع التفسير</li> <li>-اللسان موجود (املا)</li> <li>-التعقيبية (هل هي موجودة املا)</li> </ul>

Champ "Additifs"	حقل الإضافات
<ul style="list-style-type: none"> <li>- Contenu <ul style="list-style-type: none"> <li>o Lecture</li> <li>o Collation</li> <li>o Correction</li> <li>o Noms correcteurs</li> </ul> </li> <li>- Autres <ul style="list-style-type: none"> <li>o Autorisation de transmission</li> <li>o Certificat de lecture</li> <li>o Passation de possession</li> <li>o Folios détachées de la copie</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>- المتعلقة بالمضمون <ul style="list-style-type: none"> <li>o القراءة</li> <li>o المقابلات</li> <li>o التصحيحات</li> <li>o اسم المصححون</li> </ul> </li> <li>- الغير متعلقة بالمضمون <ul style="list-style-type: none"> <li>o الاجازات</li> <li>o السماعات</li> <li>o الوقفات الملكية</li> <li>o الاوراق المنفصلة عن النسخة</li> </ul> </li> </ul>
Champ "Notice supplémentaire et traductions"	حقل البيانات الإضافية و بيانات الترجمة
<ul style="list-style-type: none"> <li>- Copie du manuscrit dans d'autres Bibliothèques</li> <li>- Copie imprimée (Oui ou Non)</li> <li>- Nom Imprimeur</li> <li>- Références de publication</li> <li>- Références de traductions</li> <li>- Sources</li> </ul>	<ul style="list-style-type: none"> <li>- النسخ الموجودة من المخطوط في المكاتب الأخرى</li> <li>- اذ كانت النسخة مطبوعة (ام لا)</li> <li>- اسم الطابع</li> <li>- بيانات النشر</li> <li>- بيانات الترجمة</li> <li>- المصادر</li> </ul>
Champ "Observations"	حقل الملاحظات المتعلقة بالحقول السابقة

## II. Protocole de description des manuscrits défini par l'IRHT

Signalement rapide du manuscrit
<ul style="list-style-type: none"> <li>- Cote actuelle : Elle est composée du nom de la ville où est conservé le manuscrit, du dépôt, puis éventuellement du fonds, suivi du numéro et/ou de la lettre (ou du groupe de lettres) désignant le manuscrit.</li> <li>- Dimension en mm</li> <li>- Nombre de folios</li> <li>- Palimpseste ou non</li> <li>- Homogène ou hétérogène</li> <li>- Date donnée ou restituée (année si le manuscrit est daté, sinon siècle), origine.</li> </ul>
Support
<ul style="list-style-type: none"> <li>- Parchemin/papier</li> <li>- Si papier : vergeures, fils de chaîne, filigranes,...</li> </ul>

<b>Encre</b>
<ul style="list-style-type: none"> <li>- Encre à écrire</li> <li>- Encres de couleur (pour la décoration)</li> </ul>
<b>Organisation du volume</b>
<ul style="list-style-type: none"> <li>- Système de numération : foliotation, pagination</li> <li>- Signatures</li> <li>- Réclames</li> <li>- Cahiers : composition, succession</li> </ul>
<b>Préparation de la page avant l'écriture</b>
<ul style="list-style-type: none"> <li>- Piqûres</li> <li>- Réglure : instrument utilisé, schéma</li> </ul>
<b>Ecriture</b>
<ul style="list-style-type: none"> <li>- Distinction des différentes mains (ne pas oublier les gloses ou les commentaires)</li> <li>- Caractérisation des écritures</li> </ul>
<b>Mise en page</b>
<ul style="list-style-type: none"> <li>- Nombre de colonnes, de lignes</li> <li>- Surface écrite (en mm)</li> <li>- Titre courant, manchettes</li> <li>- Dispositions spéciales (tableaux...)</li> <li>- Gloses, commentaires...</li> </ul>
<b>Décoration</b>
<ul style="list-style-type: none"> <li>- Hiérarchie du décor</li> <li>- Description : décor du texte, de la page, illustration</li> <li>- Mentions écrites pour l'enlumineur</li> <li>- Notation musicale</li> </ul> <p>Remarque : A défaut de savoir décrire et interpréter ces éléments, signaler les folios où ils se trouvent.</p>
<b>Colophon, souscription ou mention similaire</b>
<ul style="list-style-type: none"> <li>- Type de mentions</li> <li>- Transcription</li> <li>- Critique : Interprétation des dates, identification des personnes et des lieux.</li> </ul>

Reliure
<ul style="list-style-type: none"> <li>- Technique (plat, couverture, dos, titre)</li> <li>- Décor (emplacement, motifs)</li> <li>- Eléments annexes (signets, boulons, fermoirs, chaîne,...)</li> <li>- Datation, localisation, histoire, remaniements, restaurations.</li> </ul>
Possesseurs et lecteurs
<ul style="list-style-type: none"> <li>- Marques de possesseurs</li> <li>- Interventions des éditeurs et des lecteurs</li> </ul>
Héraldique
<ul style="list-style-type: none"> <li>- Armoiries (écu, timbres, ornements, extérieurs)</li> <li>- Devises, emblèmes, chiffres</li> </ul>
Contenu
<p>Liste des textes dans l'ordre du manuscrit</p> <p>Folios extrêmes</p> <p>Etat matériel : lacune/fragment</p> <p>Identification</p> <ul style="list-style-type: none"> <li>- Auteur</li> <li>- Titre</li> <li>- Date de composition</li> <li>- Prose</li> <li>- Langue, dialecte</li> </ul> <p>Analyse du texte</p> <ul style="list-style-type: none"> <li>- Formule initiale (dédicace)</li> <li>- Pièces liminaires (préface, prologue, table)</li> <li>- Intitulé de début et de la fin (s'il est différent du titre)</li> <li>- Incipit (au moins cinq mots)</li> <li>- Particularités (divisions, lacunes, additions, gloses)</li> <li>- Explicit</li> <li>- Pièces terminales (postface, épilogue, envoi)</li> </ul> <p>Caractère</p> <ul style="list-style-type: none"> <li>- Texte intégral/extrait</li> <li>- Remaniement (source)</li> <li>- Compilation, florilège</li> <li>- Autres exemplaires manuscrits connus (pour les textes rares)</li> </ul> <p>Bibliographie</p> <ul style="list-style-type: none"> <li>- Edition(s) du texte</li> <li>- Sigle, appellation dans une édition, un répertoire, un catalogue moderne</li> <li>- Etudes dans lesquelles un manuscrit est décrit, cité ou reproduit en fac-similé</li> </ul>

**Histoire****Origine**

- Date de copie
- Lieu de copie (colophon ou formule de souscription du copiste)
- Artisan(s)

**Destinée**

- Personnes physiques ou morales pour qui et/ou sur l'ordre de qui le manuscrit a été exécuté; possesseurs successifs)
- Ex-libris ou ex-dono à transcrire en indiquant le folio où ils se trouvent
- Relevé de l'incipit et de l'explicit-repères, pour confrontation avec les catalogues anciens, avec renvoi éventuel à des catalogues médiévaux où le manuscrit est

mentionné.



# ***Bibliographie***

## Bibliographie

[TEI P5 2012]	TextEncoding Initiative. 10 Manuscript Description. Disponible sur le web à l'adresse <a href="http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html">http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html</a> . Dernière mise à jour le 17 Juin 2012. Consulté le 13/08/2012
[EAD 2004]	Description Archivistique Encodée, Dictionnaire des balises, Society of American Archivists octobre 2004.
[Aifbd 2012]	Association internationale francophone des bibliothécaires et documentaliste. Disponible sur le web sur le web à l'adresse : <a href="http://aifbd.org/index.php?option=com_content&amp;view=article&amp;id=134&amp;Itemid=96">http://aifbd.org/index.php?option=com_content&amp;view=article&amp;id=134&amp;Itemid=96</a> . Mise à jour le Dimanche, 19 Septembre 2010. Consulté le 12/09/2012.
[guideDC 2008]	Guide d'utilisation du Dublin Core (DC) à la BnF : Dublin Core simple et Dublin Core qualifié, avec indications pour utiliser le profil d'application de TEL, 2008
[Muzelle 1985]	Muzelle, Danis. Vocabulaire codicologique : répertoire méthodique des termes français relatifs aux manuscrits. Paris : Editions CEMI, 1985, 348p (P.39).
[Ide & Véronis 1996]	Nancy IDE et Jean VÉRONIS. Présentation de la TEI : TextEncoding Initiative, Department of Computer Science VassarCollege, Université de Provence et CNRS, juin 1996.
[Farag 2012]	Farag Ahmed. Techniques de numérisation : choix, problématiques et perspectives. Université d'Assiout Égypte. Disponible sur le web à l'adresse : <a href="http://www.membres.multimania.fr/ahmedfarag/documents/.../Techniques.pdf">www.membres.multimania.fr/ahmedfarag/documents/.../Techniques.pdf</a> . Consulté le 15/08/2012.
[Catalogue 2012]	Qu'est-ce qu'un catalogue. Disponible sur le web à l'adresse : <a href="http://www.bib.fsagx.ac.be/edudoc/defist/agroA/module">www.bib.fsagx.ac.be/edudoc/defist/agroA/module</a> . Consulté le 29/08/2012.
[Ontario 2009]	Définition et utilité des métadonnées. Disponible sur le web à l'adresse <a href="http://www.mnr.gov.on.ca/stdprodconsume/groups/lr/@mnr/.../266887.pdf">www.mnr.gov.on.ca/stdprodconsume/groups/lr/@mnr/.../266887.pdf</a> . Consulté le 12/08/2012.
[Morel Pair 2005]	Morel-pair Catherine. Panorama : des métadonnées pour les ressources électroniques. Ateliers des réseaux de la documentation scientifique, Arcachon. 2005.
[Index 2012]	Qu'est-ce qu'un index ?. Disponible sur le web à l'adresse : <a href="http://bibliotheque.u-pec.fr/besoin.../qu-est-ce-qu-un-index--411999.kjsp?">bibliotheque.u-pec.fr/besoin.../qu-est-ce-qu-un-index--411999.kjsp?</a> . Consulté le 02/09/2012.
[Zemerli 2004]	Zemirli W. Nesrine. Vers le développement d'un système de recherche d'information personnalisé intégrant le profil utilisateur. Université Paul Sabatier – Toulouse III. 2004.
[Interopérabilité 2012]	Interopérabilité. Disponible sur le web à l'adresse <a href="http://www.guideinformatique.com/definition-interoperabilite">www.guideinformatique.com/definition-interoperabilite</a> . Consulté le 15/09/2012.
[Varandat200	Stockage de contenus XML : base "native" ou relationnelle ? Par Marie

## Bibliographie

2]	Varandat publié le 06/11/2002. Disponible sur le web sur l'adresse : <a href="http://www.indexel.net/article/stockage-de-contenus-xml-base-native-ou-relationnelle">www.indexel.net/article/stockage-de-contenus-xml-base-native-ou-relationnelle</a> . consulté le 01/09/2012
[Solari 2008]	J.J.Solari. Niveaux d'interopérabilité des métadonnées Dublin Core. Disponible sur le web à l'adresse : <a href="http://www.yoyodesign.org/doc/dcmi/interoperability-levels/index.html">www.yoyodesign.org/doc/dcmi/interoperability-levels/index.html</a> . Publié le 20/12/2008. consulté le 12/09/2012.
[Zargayouna, 2005]	Zargayouna Haïfa. Indexation sémantique de documents XML. Thèse de doctorat de l'université Paris XI Orsay. 2005.
[Gnahou 2008]	Gnahou Christophe. Démarche de conception d'une BD relationnelle. Disponible sur le web à l'adresse : <a href="http://www.math-info.univ-paris5.fr/~gnahou/uml/coursUMLBD.pdf">www.math-info.univ-paris5.fr/~gnahou/uml/coursUMLBD.pdf</a> . consulté le: 10/09/2012.
[Soualah&Hassoun 2011]	Soualah Mohammed Ourabah et Mohamed Hassoun. « Which metadata for Ancient Arabic Manuscripts Cataloguing? ». 2011.
[siteDC 2007]	Guide d'utilisation du Dublin Core. Disponible sur le web sur à l'adresse : <a href="http://yoyodesign.org/doc/dcmi/usageguide/elements.html">http://yoyodesign.org/doc/dcmi/usageguide/elements.html</a> . Consulté le 15/09/2012
[Wiki 01 2012]	Encyclopédie universelle Wikipedia – la manuscrit: Notion generals sur les manuscrits. Disponible sur le web à l'adresse <a href="http://fr.wikipedia.org/wiki/manuscrit">http://fr.wikipedia.org/wiki/manuscrit</a> . Consulté le 30/07/2012
[Wiki 02 2012]	Encyclopédie universelle Wikipedia-interopérabilité en informatique. Disponible sur le web à l'adresse : <a href="http://fr.wikipedia.org/wiki/Interopérabilité_en_informatique">fr.wikipedia.org/wiki/Interopérabilité_en_informatique</a> . Consulté le 09/08/2012.
[Mokdad 1992]	L.MOKDAD. «Compression d'image pour la visiophonie». PFE INI 1992.
[Toumazet 1987]	J.J.TOUMAZET. «Traitement de l'image sur micro-ordinateur». Sybex 1987.
[IRHT 2006]	Livret du stage d'initiation au manuscrit médiéval, réalisé par l'Institut de Recherche en Histoire et du Texte (IRHT), Lyon – France. Département des manuscrits arabes - disponible en ligne sur <a href="http://aedilis.irht.cnrs.fr/stage/descriptionmanuscrit.htm">http://aedilis.irht.cnrs.fr/stage/descriptionmanuscrit.htm</a> .
[Soualah 2008]	Soualah Mohammed Ourabah. « Numérisation des manuscrits arabes : Catalogage et accès multilingue ». Mémoire de Magister en Informatique. Institut National de formation en Informatique I.N.I. 2008.
[keilah 2004]	Keilah Hala. «Accès à distance aux manuscrits arabes numérisés en mode image » Thèse pour obtenir le grade de Docteur de l'université de Lumière Lyon II. 2004.
[Cedelle-Joubert et al 2002]	Cedelle-joubert Laure et Charlette Buressi – Cadrer un projet. Article parue dans la revue "Conduire un projet de numérisation" Editions TEC&DOC – Juin 2002.
[Piérache 2008]	Piérache Nathalie. Le catalogage Description des documents : formats et normes. Ecole de documentation Hammamet, 2008.

## Bibliographie

[Staccini 2007]	Staccini Pascal. Normes et technologies d'interopérabilité. UFR Médecine Nice, université Nice-Sophia antipolis, 2007.
[Yahia 2011]	Yahia Esmâ. Contribution à l'évaluation de l'interopérabilité sémantique entre systèmes d'information d'entreprises : application aux systèmes d'information de pilotage de la production. Thèse de doctorat de l'université Henri Poincaré, Nancy 1. 2011.
[Daclin2008]	Daclin Nicolas, Vincent Chapurlat. Évaluation de l'interopérabilité organisationnelle et managériale des systèmes industriels : le projet carioner. Disponible sur le web à l'adresse : <a href="http://www.supdeco-montpellier.com/.../Daclin_Chapurlat_MTO_2008.pdf">www.supdeco-montpellier.com/.../Daclin_Chapurlat_MTO_2008.pdf</a> . consulté le 31/08/2012.
[Berms 2011]	Berms Emmanuelle. Convergence et interopérabilité : l'apport du Web de données. Bibliothèque nationale de France (BnF) Paris, France. 2011
[Difournaud 2008]	DufournaudNicole. Manuel d'encodage XML-TEI. Centre d'études supérieures de la renaissance. 2008.
[Crassous 2009]	Crassous Béatrice. Introduction à la TextEncodingInitiative. Ecole nationale supérieure des sciences de l'information et des bibliothèques. France. 2009.
[Beaudry 2007]	Beaudry Guylaine – La TextEncoding Initiative : Les moyens pour ajouter de la valeur à un texte numérisé – disponible sur le web à l'adresse : <a href="http://www.ebsi.umontreal.ca/cursus/index.html">http://www.ebsi.umontreal.ca/cursus/index.html</a> . Consulté le 17/09/2012.
[Harrathi 2010]	Harrathi Rami. Recherche d'information conceptuelle dans les documents semi-structurés. Thèse de doctorat spécialité informatique, de l'Institut Nationale des Sciences Appliquées de Lyon. 2010.
[Torjmen 2009]	TORJMEN Mouna. Approches de Recherche MultiMedia dans des Documents Semi-Structurés : Utilisation du contexte textuel et structurel pour la sélection d'objets multimédia. Thèse de doctorat de l'Université Paul Sabatier Spécialité: informatique. 2009.
[Sauvaganat 2005]	SauvagnatKaren. Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structurés. Thèse de doctorat spécialité informatique à l'université de Paul Sabatier de Toulouse. 2005
[Telma 2007]	Centre de Ressources Numériques TELMA – Présentation sur la TEI – disponible en ligne à l'adresse <a href="http://www.cntelma.fr/">http://www.cntelma.fr/</a> . Consulté le 07/08/2012.
[Starck 2008]	StarckJulien. EAD. École nationale supérieure des sciences de l'information et des bibliothèques. Créée le : 15 février 2008
[Micaelli 2012]	MICAELLI Eric. Réglementation et pratique quotidienne des archives. Publication - dimanche 28 février 2010. Mise à jour - lundi 7 mai 2012
[André 1996].	André Jacques. BALISES, STRUCTURES ET TEI. Cahiers GUTenberg. 1996
[Pleade 2005]	Les éléments du Dublin Core qualifié. Disponible sur le site : <a href="http://www.pleade.org/fr/documentation/descriptions/qdc.html">http://www.pleade.org/fr/documentation/descriptions/qdc.html</a> . Consulté le 22/08/2012.