Première partie

1. L’environnement de la recherche

Le mot *communication,* qui a toujours gardé la même signification à travers les siècles, a radicalement changé dans sa pratique. Cette interaction entre êtres humains, aussi bien sur le plan intellectuel que sur le plan social, a pu récemment évoluer grâce au développement des moyens techniques de communication. Bien évidemment, l'évolution de la technologie incarnée par l'Internet au cours de ces dernières années a joué un rôle majeur dans l'élargissement de l'horizon de la communication dans le monde entier quelles que soient leur langue et leur lieu de résidence. Les nouvelles technologies ont facilité les communications et l'accès à tous les aspects de la vie, à la fois quotidienne, scientifique, intellectuelle ou concernant les relations humaines, etc.

« L'électronisation», c'est à dire la création de données sur un format électronique, a commencé à prendre de plus en plus d'importance par rapport au support papier.

Des milliers de documents électroniques sont ainsi distribués chaque année dans tous les domaines, cependant ils présentent un grand problème d'accès à leur contenu. Or, le développement rapide et intensif des moyens de diffusion de l'information requiert un système permettant de définir les composants de ces documents, ceci afin que l'utilisateur trouve facilement l'information recherchée. Donc ces documents auront besoin d'un nouveau traitement.

Plusieurs normes ont commencé à voir le jour pour mettre de l'ordre dans cette évolution rapide des documents électroniques. La diversité exige toujours la normalisation : une tendance assez forte a été initialisée par des institutions et/ou des chercheurs pour établir une norme standard qui puisse englober facilement le plus possible de normes déjà créées par différents chercheurs quel que soit leur lieu de résidence dans le monde.

Pour pouvoir établir une norme propre aux documents électroniques qu'ils soient d'origine purement informatique ou numérisée à partir du support papier ou du microfilm, ancienne ou récente, la structuration des ces documents sera toujours la première étape nécessaire pour faciliter à la fois l'accès à ces documents et la diffusion des informations qu'ils contiennent. Contrairement aux documents électroniques, les documents sur support papier manuscrits ou imprimés exigent un processus de conversion vers un format électronique. Ce processus qui s'appelle la *numérisation* (en

anglais *digization*) prend deux formes : la numérisation en mode image et la numérisation en mode texte. La structuration de documents numériques quelle que soit leur origine (électronique, papier, microfilm, etc.) exige des règles spécifiques qui, en quelque sorte, équivalent à une norme de catalogage. Cette norme s'exprime en terme de métadonnées (*metadata*). Ces métadonnées (les données sur les données) seront développés de manière exhaustive dans la section (1.4.2 page 37). L'objectif de ce mémoire de thèse est de mettre en oeuvre les techniques et les procédures d'analyse les plus récents sur des manuscrits arabes anciens. En particulier, ceci nécessitera la création des métadonnées nécessaires. En d'autres termes, on essayera de mettre la « nouveauté » au service de « l'ancienneté », tout ceci afin de mettre en valeur ce fond de patrimoine rare et particulier.

1.1. La numérisation : aspects généraux

Numériser un document donné, c'est simplement transférer ce document d'un support analogique (parchemin, métal, cire, bande magnétique, etc.) sur un support informatique. Le processus de la numérisation signifie notamment la création d'une représentation informatisée (numérique) d'une imprimerie analogique. Selon Lee Stuart la numérisation est « *the conversion of an analog signal or code into a digital signal or code* »[4].

Pour Yannick Maignien, la numérisation est le passage d'un état à un autre, du physique au virtuel, du matériel à l'immatériel. C'est donc « la migration des collections physiques vers l'immatériel, celle du réel vers le virtuel, celle de l'analogique vers le numérique, celle du texte vers l'hypertexte et l'on pourrait relever longtemps les dualités qui sont souvent employées dans des réflexions sur ce sujet »[5].

Le Conseil Supérieur des Bibliothèques (en France) considère, dans son rapport annuel de 1995, que « la numérisation n'est qu'une forme moderne d'autres techniques comme le microfilmage, depuis longtemps utilisée pour permettre la communication sur des

[4] Lee, Stuart D. Digital imaging: a practical handbook. N.Y: Neal-Schuman Publishers, Inc, 2000. 194pages (P.3)

[5] Maignien, Yannik. *De l'imprimé au numérique, la migration du document*. (Actes du colloque Roanne mois du patrimoine écrit le patrimoine en mouvement. Migration de l'écrit au fil des siècles. 1-2 Oct, 1996a), P.160.

supports dits "de substitution" pour des documents très demandés, fragiles ou rares », Yannick Maignien n'est pas de cet avis. Pour lui, « la numérisation n'est pas que le support de substitution, la reproduction d'un objet matériel, textuel, sonore ou iconique, c'est une "autre" machine éditoriale, non encore développée, c'est un "autre" agencement des relations aux connaissances et aux autres, une autre communauté en gestation ».

Pour lui, la numérisation est ce « passage au virtuel, mixte inédit, sensible et intelligible, créant un espace de simulation, intégrant une dimension imaginaire, au sens à la fois d'illusion, mais aussi de création ». À son avis, la numérisation des documents entraîne avec elle une nouvelle conception de l'encyclopédisme »[6].

Jean Max Noyer souligne la même idée, en ce sens que la transformation, le stockage et l'archivage de textes, de sons et d'images viennent « susciter de nouvelles pratiques, faire apparaître de nouveaux objets à l'analyse, engager de nouveaux questionnements, créer des conditions favorables à l'établissement de nouvelles transversalités et ce, en raison donc des médiations émergentes, des technologies intellectuelles traitant la matière numérique »[7].

Dans le monde des bibliothèques, la numérisation est à l'origine d'un nouveau concept qui est celui de la bibliothèque virtuelle qui a complètement bouleversé le fonctionnement et l'agencement des tâches communément connues dans la bibliothèque dite classique. « La numérisation ne vient pas se rajouter aux fonctions d'acquisition, de conservation, de communication, de catalogage, de transmission de l'imprimé. La numérisation vient en modifier transversalement chacune des composantes ».[8]

1.2. La technique de la numérisation

Selon M. Lenart et D. Goldwaser, on peut distinguer deux techniques de gestion électronique de documents, en l'occurrance celles qui gèrent des documents en mode

---

[6] Maignien, Yannick. La Bibliothèque virtuelle : ou de l'ars memoria à xanadu. Bull. Bibl. France. T.40. N°.2, 1995a. p.10

[7] Noyer, Jean-Max. *Numérisation et image*. 1 page
http ://www.info.unicaen.fr/bnum/jelec/solaris/d01/1noyer2.html

[8] Maignien**,** Yannick. *La Bibliothèque virtuelle : ou de l'ars memoria à xanadu*. Bull. Bibl. France. T.40. N°.2, 1995a. page. 12.

image et celles qui gèrent des documents en mode texte. Le mode image résulte d'une « capture des images, à l'aide de moyens électroniques tels que le scanner (le numériseur) qui offre des documents (noir et blanc ou en couleur), la caméra vidéo qui permet le stockage provisoire des images analogiques sur support magnétique (tel que la vidéo disque) ou encore l'appareil photo- numérique»[9]

La différence entre les deux types de documents numérisés est l'utilisation ou non d'une reconnaissance optique de caractères (ROC ou OCR). « Le mode image ne donne qu'une image du texte, c'est à dire qu'il ne permet pas d'effectuer de traitement sur le contenu. Par contre, le mode texte permet de travailler sur le texte et donc, de procéder à des recherches sur le contenu »[10]

Depuis le début des années 90, de nombreux projets ont vu le jour aux Etats-Unis aussi bien qu'en Europe qui visent à numériser les fonds patrimoniaux existant au sein de leurs bibliothèques publiques et/ou privées, centres culturels etc. Ces initiatives ont pour but principal de conserver ces documents rares et précieux, de faciliter l'accès aux utilisateurs et de bien diffuser les informations se trouvant dans ces documents, tout en utilisant la nouvelle technologie.

1.2.1. Pourquoi numériser ?

Les manuscrits font partie de la mémoire collective de l'humanité. Numériser ce patrimoine aide à diminuer le risque de la disparition de ces manuscrits. Il est donc nécessaire aussi bien pour les manuscrits que pour les chercheurs d'avoir une copie électronique de ces manuscrits.

Les objectifs de numérisation sont nombreux : ils aident à la sauvegarde, à la préservation, à l'accès et à la valorisation de documents rares et précieux.

Numériser un tel document et le mettre sur l'Internet permet aux chercheurs de l'étudier, de le récupérer sur son ordinateur, de l'annoter, de lui associer des commentaires et des liens avec d'autres documents. On voit, ici, apparaître la notion de

---

[9] Lenart, M et D. Goldwaser. *Applications documentaires de la GED dans les bibliothèques et centres de documentation*. - Paris : Ajoure Editeur, 1993. 101 pages

[10] Béquet, Gaëlle. La numérisation des documents patrimoniaux, 12 pages http: //www.culture/conservatio/fr/preventi/documents/c13.pdf

poste de lecture assistée par ordinateur et les possibilités de travail coopératif entre les chercheurs. Rendre les manuscrits sur l'Internet accessibles à un travail collectif permet à plusieurs chercheurs dans des laboratoires différents plus ou moins dispersés et éloignés, de participer à un projet de recherche commun sur une collection de manuscrits située dans une bibliothèque donnée ou dans un ensemble de bibliothèques réparti dans le monde entier.

1.2.2. Que faut-il numériser ? Constitution de la collection

Le processus de numérisation peut concerner plusieurs médias et mettre en œuvre différentes méthodes. La chaîne de numérisation passe par plusieurs étapes (les documents doivent être sélectionnés, préparés, mis en mémoire, diffusés et conservés), et toutes ces étapes doivent être organisées en fonction de l'usage final qui sera fait de la collection numérisée, usage qui doit, bien sûr, avoir été défini au préalable.

Il est nécessaire de définir la collection concernée avant de commencer le processus de numérisation. Dans le choix de la collection à numériser, il faut prendre en considération les critères suivants:

- Le contexte général et les objectifs du projet.
- Le besoin des utilisateurs. (Voir le résultat de l'enquête auprès d'experts des manuscrits arabes la quatrième partie de thèse page 158)
- Rassembler les documents.

1.2.3. Le traitement des documents avant numérisation.

Avant de numériser, le document doit passer par toute une succession de processus qui correspond à une série des questions : toutes ces opérations ont un coût qu'il faut avoir prévu dans l'élaboration du budget.

1.2.3.1. Dans quel état sont-ils ?

Les conditions de conservation du document ont un effet important sur l'état général des manuscrits. Les imperfections de la conservation peuvent être la conséquence (surtout en ce qui concerne les manuscrits arabes) de la situation économique de la bibliothèque ou de l'institution qui possède les documents et à l'état général des

bâtiments où se trouve le document. L'humidité et la chaleur abîment le document et laissent des traces sur le papier.

La restauration est donc essentielle suivant deux points de vue :

- Le document fragile est consolidé et risque moins d'être abîmé pendant le processus de numérisation.
- Elle permet d'obtenir un document numérisé de la meilleure qualité possible

1.2.3.2. Le document est-il complet ?

C'est une question qui se pose pendant l'étude de document avant la numérisation. Une étude sur l'importance de ce document est indispensable pour arriver à la décision de numériser ou pas. On trouve quelque fois nécessaire de numériser le document, malgré le manque de page ou d'une partie du document. Numériser un tel document et le diffuser sur l'Internet peut permettre aux chercheurs de l'étudier quand même et peut-être de participer à la localisation des parties manquantes. On voit ici le rôle qui peut être joué par un travail collaboratif.

1.2.3.3. Où et sur quel matériel doit-on les numériser ?

Une des premières questions porte sur la qualité du résultat de la numérisation. Un matériel très performant est-il nécessaire ? Plusieurs cas de figure peuvent se présenter :

- L'institution dispose d'un matériel performant et elle numérise sur place.
- L'institution dispose d'un matériel moins performant mais préfère éviter le déplacement des documents et numérise sur place.
- L'institution ne dispose pas du matériel nécessaire. Il se pose alors la question du déplacement de la collection ou de l'utilisation d'un support intermédiaire. Le recours à une sous-traitance proche permet de tolérer un déplacement contrôlé des documents ; les transferts des documents sur un support, comme le microfilm par exemple, est une autre solution.

Bien entendu, les considérations budgétaires interviennent fortement dans le choix de la solution.

1.2.3.4. Qu'est ce que numériser ?

Le processus de numérisation soit pour les images fixes, soit pour les textes, passe par un scanner mais on peut numériser directement ou par l'intermédiaire du microfilm par exemple. La numérisation serait réalisée soit par un scanner normal, soit par l'intermédiaire d'une camera, surtout pour les livres délicats comme les manuscrits par exemple. Selon l'expérience du projet Celtic et Medieval Manuscript Project qui a été réalisé à l'université d'Oxford, Julius Smith[11], le photographe du projet, dit que le manuscrit a été numérisé selon un angle de 45 degrés. Ceci donne la possibilité à un manuscrit donné d'être ouvert avec sécurité à un angle de 90 et 100 degrés. Pour faciliter la prise d'image numérisée d'un manuscrit avec un relieur, un léger système porteur, orientable suivant différents axes, a été conçu pour permettre à chaque manuscrit d'être confortablement posé pendant le processus de numérisation "*In order to facilitate the digital imaging of bound manuscripts, a lightweight yet sturdy cradle has been designed to allow each manuscript to sit comfortably whilst being subjected to periods of digital scanning*"[12]

La chaleur et la lumière sont deux facteurs négatifs qui affectent les manuscrits. Donc le choix d'une caméra avec un filtrage est important pour avoir le moins de lumière possible pendant le processus de numérisation. Un éclairage froid est donc nécessaire.

1.2.3.5. Le poids d'une image et les techniques de compression.

L'obtention d'une photographie numérique de chaque page, c'est le mode image. Une image numérique est constituée d'un ensemble de points élémentaires appelés pixels. En fonction des caractéristiques des pixels, il existe trois modes d'images qui chacun sert pour une tâche bien spécifique:

- Les pixels du premier mode « image binarisée » sont représentés par 1 bit (deux valeurs possibles : noir ou blanc)

[11] Lee, Stuart D. Digital imaging: a practical handbook. New York: Neal-Schumen, Inc. 2000. 194pages (P. 79)
[12] ibid.

- Les pixels du deuxième mode sont représentés par un octet et permettent d'avoir des images avec 256 niveaux de gris. Il faut alors un octet pour représenter chaque pixel.
- Généralement, les pixels du troisième mode sont représentés par un ensemble de 24 bits, ce qui permet d'obtenir 16,8 millions de couleurs. Pour représenter chaque pixel, il faut alors 3 octets.

1.2.3.5.1. La définition (ou résolution) d'une image

L'image numérisée est d'une qualité d'autant meilleure que les pixels (ou points de l'image) sont plus resserrés. Cette caractéristique est appelée définition (ou résolution) de l'image. Elle s'exprime en « points par pouce » ou dpi (dot per inches).

Une bonne image demande donc une bonne résolution. Mais il faut considérer les conséquences du choix de cette définition en termes de durée de numérisation, de volume de stockage et de durée de transmission sur l'Internet.

Chaque page numérisée correspond, dans l'ordinateur, à un fichier dont la taille se mesure de la façon suivante :

Soit une image de 4 pouces par 6 pouces. Cette image est représentée dans l'ordinateur par un ensemble de lignes et de colonnes.

- Une définition de 300 dpi donne :Une ligne est composée de 1200 pixels (4 pouce avec 300 pixels par pouce)
- Il y a 1800 lignes (6 pouces avec 300 pixels par pouces)
- L'image complète comprend donc 2 160 000 pixels (1800 lignes de 1200 pixels chacunes)

S'il s'agit de pixels de couleurs à 16 millions de nuances, chaque pixel est représenté par 3 octets. Cette image occupe donc un volume (pèse) 6 480 000 octets (6,48 mégaoctets – pour mémoire, une disquette ne peut contenir que1,4 mégaoctets).

Voici d'autres exemples de poids d'image :

Mode de numérisation	Résolution (dpi)			
	400	**300**	**200**	**100**
Noir et blanc (1bit)	0,48 Mo	0,27 Mo	0,12 Mo	0,03 Mo
Niveaux de gris (1 octet)	3,84 Mo	2,16 Mo	0,96 Mo	0,240 Mo
Couleur 24 bits (3 octets)	11,52 Mo	6,48 Mo	2,88 Mo	0,720 Mo

Table n°.1 : Exemples de poids d'image

1.2.4. Les formats d'image

Le tableau précédent montre que le résultat brut de la numérisation donne des images très pesantes. Une définition très fine et un grand nombre de nuances de couleurs conduit à des images très volumineuses. Dans le but de les alléger et de les rendre plus facilement manipulables et stockables, on applique à ces images des algorithmes de compression. Suivant l'algorithme utilisé, on obtient des formats d'image différents. Les trois formats les plus courants sont : le GIF (*Graphic Interchange Format*), le JPEG (*Joint Photographic Experts Group*) et le TIF (*Tagged Image File Format*).

1.2.4.1. Le GIF (Graphique Interchange Format)

Ce format, le plus répandu sur l'Internet comme format d'information graphique, a été crée en 1987 (*GIF87a format)* par *CompuServe Intercorporated* pour ses services en ligne. Mais le format a été amélioré en 1989 et une nouvelle édition est sortie sous le nom de *GIF89a*. Contrairement au JPEG, le format GIF est limité à 256 couleurs ou nuances de gris par pixel. Cette limitation dans la gamme des couleurs permet d'obtenir une taille de fichier relativement petite. Il est important de remarquer que le GIF comme le TIFF sont des techniques de compression sans perte d'information, c'est-à-dire qu'on peut revenir exactement à l'image initiale avant compression. Le format GIF peut prétendre, très valablement, à être utilisé dans l'art graphique et dans le dessin de lignes comprenant un nombre limité de couleurs. Être transparent est un autre avantage du GIF qui se conforme bien avec l'arrière plan de page web.

1.2.4.2. Le JPEG (Joint Photographic Expert Group)

Il a été créé à la fin des années 1980 par le Joint Photographic Expert Group (JPEG), un groupe d'experts nommés par des organismes nationaux de normalisation et par des industriels: ce format est connu comme étant le format le plus puissant utilisé pour diffuser sur le Web des images de grande qualité. Il est plus riche que le format GIF car

il permet l'utilisation de 16 millions de couleurs. Il est bien employé pour présenter des photographies. Egalement au format GIF, le JPEG est très répandu dans le Web (le Microsoft Internet Explorer et Netscape utilise bien JPEG pour diffuser les images en ligne) surtout grâce à son excellent algorithme de compression.

Depuis l'année 1992, cette méthode a été adoptée comme norme internationale. Le JPEG est libre de tout droit, donc gratuit, ce qui lui a permis d'être largement diffusé sur Internet, car il permet aussi bien de traiter les images en couleur qu'en niveaux de gris.

« Le JPEG est un format de compression des images numériques non entropique, c'est à dire qui ne conserve pas la qualité de l'image initiale »[13]

L'image sur JPEG est très fidèle à la copie originale «*JPEG retains the look – and feel of the original, thus layout features, graphics, and characters of any type are displayed*»[14]

Trois options de compression sont incluses dans le format JPEG :

1. Bonne qualité/ niveau de compression bas.
2. Bonne qualité / bon niveau de compression.
3. Qualité basse/ niveau de compression très élevé
4. Même avec l'option « bonne qualité/ niveau de compression bas » le fichier JPEG garde toujours une taille moins élevée que le fichier GIF.

1.2.4.3. Le format TIFF (Tagged Image File Format)

Il a été développé par Aldus et Microsoft. Grâce à son algorithme de compression sans aucune perte d'information, TIFF est le format le plus répandu dans le domaine de l'archivage de documents comme format principal pour la création et la conservation des images. Ce format est le plus utilisé en numérisation d'image, car il peut être lu par la plupart des plates-formes. Le seul inconvénient du format TIFF est dans sa taille d'image, qui est grande. Les fichiers TIFF sont lourds ; pour cela il est très difficile au navigateur Web de les lire. Le format TIFF ITU-T.6 avec son codage des pixels en 24-bits est beaucoup utilisé par Adobe PhotoShop. « Il connaît un grand nombre de

---

[13] Durand, Nicola. La compression des images numériques : Le JPEG (ou Joint Photographique Expert Group). http://www.chez.com/nico77/ 21 pages. (p.1)
[14] Cleveland, Gary. Selceting electronic document formats. July 1999, 13pages (p.7)
http://www.ifla.org/VI/5/op/udtop11/udtop11.htm

variantes pour tous les types d'images (au trait, en niveaux de gris ou en couleurs) et différents algorithmes de compression pour les stockages »[15]. L'image sur TIFF est très fidèle à l'original et conserve la mise en page, le graphique et les propriétés de toutes sortes de documents. Le fichier TIFF n'est pas éditable (c'est à dire, qu'il ne donne pas la possibilité de faire des changements sur le texte déjà existant). À cause de sa grande taille, il est difficile de visualiser le format TIFF par un navigateur Web, mais il est bien adapté pour la capture d'image et comme format d'échange et d'archivage. Il s'accorde bien avec la plupart des logiciels d'image.

1.3. Comment stocker ?

La gestion des images numérisées, bien évidemment, nécessite de stocker les documents dans un endroit à partir duquel on peut gérer, consulter et récupérer ces images, il s'agit d'une base de données. «Une base de données est un ensemble de données modélisant les objets d'une partie du monde réel et servant de support à une application informatique».[16] Les documents de la base de données seront facilement interrogeable par le contenu (par les objets qui satisfont à un certain critère). Pour obtenir ce but, une certaine structure de ces documents est donc nécessaire. Ce structure on appelle dans le mode d'informatique «métadonnées» .

1.3.1. Comment accéder à l'information ?

La numérisation donne un document très fidèle à l'original non structuré, non accessible mais pour rendre le document numérisé plus utile, il doit donc « aller plus loin que le mode image elle doit nous permettre d'accéder à ce que l'on appelle le mode texte».[17] Il est indispensable de pouvoir accéder à l'information d'une façon moins rudimentaire que « je voudrais l'image de la page 152 ». La prise en compte de la structure du document est une approche un peu plus performante. Mais un accès par le

---

[15] Jacquesson, Alain et Alexis Rivier. Bibliothèques et documents numériques : concepts, composantes, techniques et enjeux. Paris : Electre - Editions du Cercle de la Librairie, 1999. 377p. (p.41)

[16] Gardage, Georges. Bases de données. Paris: Editions Eyrolles, 2003. 788p. (P.3)

[17] Soufi, Souad. Contribution à la reconnaissance des structures des documents écrits : approche probabiliste. Thèse préparée en cotutelle aux laboratoires : Reconnaissance de Formes et Vision, INSA, Lyon et la Vision et systèmes Numériques, Université Laval, Quebec, Canada. Sous la direction de Hubert Emptoz et Marc Parizeau., Souteneue le 21 septembre 2002., 189 pages (page. 15)

contenu est plus satisfaisant pour le chercheur. On peut distinguer deux sortes de contenus :

- Un contenu graphique (bandeaux, lettrines, ponctuations en forme de fleurs, cachets, illustrations, etc.),
- Un contenu textuel sur lequel on peut appliquer des procédures de traitement de texte (par exemple une recherche de mots-clés ou d'index, etc.). On dit alors qu'on passe du mode image au mode texte. Des algorithmes de reconnaissance optique des caractères OCR (Optical Character Recognition) sont alors utilisés.

1.3.2. Accès au contenu ?

Pour les manuscrits arabes ainsi que pour les manuscrits en langue latine, il est presque impossible d'accéder à ces documents en mode texte à cause de différents facteurs :

L'écriture arabe peut prendre plusieurs styles. On peut parler de cent styles différents, parmi eux, dix styles sont les plus utilisés. La diversité est aussi dans la taille de chaque genre d'écriture. Dans les manuscrits, l'écriture arabe est calligraphiée, c'est à dire que du fait des liaisons entre caractères, la forme des lettres varie suivant que ce caractère se trouve isolé ou lié à d'autre caractères (en début ou en fin de mot, ou au milieu du mot). De plus, on peut observer différentes formes (ajusté, courbé, décoré ou bien bouclé).

Par ailleurs, certains manuscrits ont subi, de par leur ancienneté et leurs conditions de conservation, des dégradations dues aux moisissures, aux insectes et aux caractéristiques chimiques du support et de l'encre.

Enfin, les textes de ces manuscrits sont souvent accompagnés d'annotations sur les marges ou parfois dans le corps même du texte, qui ont été ajoutées par les différents lecteurs ou par les auteurs eux-mêmes et qui ont, parfois, autant de valeur que le texte principal.

La différence peut être aussi due aux spécificités personnelles de l'écriture, au niveau d'études, à l'humeur, à la santé et aux autres conditions du calligraphe ou du copiste. D'autres facteurs, tels que l'instrument d'écriture, la surface consacrée pour l'écrit, peuvent changer la représentation du caractère. Tous ces facteurs rendent la reconnaissance des caractères par la machine très difficile.

Pour toutes ces raisons, nous estimons que la seule solution serait une numérisation en mode image malgré les contraintes relatives à ce mode. « Le mode image est indispensable lorsque des dessins, des formules scientifiques, des manuscrits, des travaux préalables à l'écriture interviennent dans l'appropriation du document »[18]. La numérisation en mode image est plus lourde à gérer et requiert, en fait, plus d'espace pour le stockage, comme le signale Julie Bouchard « pour un livre de 300 pages qui occupe 20 millions d'octets en mode image, alors que ce même livres n'occupe que 600.000 octets on mode texte »[19]. Cependant, si la numérisation en mode image présente ses inconvénients, elle a, d'un autre côté, ses avantages indéniables, surtout lorsqu'il s'agit de documents dont la valeur n'est pas uniquement textuelle, comme c'est justement le cas des manuscrits arabes. Julie Bouchard[20] précise que le mode image peut être aussi une solution que l'on préfère aux autres modes parce qu'il permet de rendre l'état originel du document. 90% des collections numérisées à la BNF ont été faites en mode image et seulement 10% sont en mode texte, car la numérisation des documents anciens en mode texte offre des performances peu satisfaisantes, ce qui favorise le mode image. Le résultat du mode image est une représentation fidèle de la page imprimée au niveau de sa mise en page mais aussi de sa forme et même des défauts du papier ou du parchemin. Ceci est d'autant plus important quand il s'agit de manuscrits, pour lesquels chaque détail de la forme, de la présentation, de l'état physique peut être riche en enseignements pour les chercheurs. Cela dit, la calligraphie arabe n'étant pas facile à déchiffrer, il serait très intéressant d'accompagner les textes numérisés en mode image, par une transcription dactylographiée de ces textes. Encore, faut-il que ce soit possible, car ceci pose le problème de la conformité au texte originel. Curieusement, comme il permet une reproduction fidèle et à large diffusion de la copie originale, le mode image peut être un outil qui facilite les travaux de révision.

Déjà la numérisation en mode texte des caractères arabes imprimés pose des problèmes de reconnaissance. Avec les textes manuscrits, les choses se compliquent beaucoup

---

[18] Bouchard, Julie. *Des puces, des livres et des hommes futuribles*, Oct. 1996. P26

[19] Ibid, p.26

[20] ibid., p.27

plus, au point qu'il devient impossible d'en extraire un texte déchiffrable. Les procédés d'OCR ne marchent bien que sur des documents très bien écrits au niveau de la forme, ce qui n'est pas le cas des imprimés anciens (livres du 16ème siècle, manuscrits européens ou arabes).

Pour pallier ce problème, nous avons donc demandé au laboratoire LIRIS-RFV de l'INSA de Lyon et dans un cadre de coopération, de développer un outil qui aide à la reconnaissance automatique de certaines descriptions à partir des images de manuscrits (le titre de chapitre ; le texte ; l'illustration, le cachet, etc.). Ce travail sera détaillé dans la page 232.

1.4. Les métadonnées et le catalogage électronique

1.4.1. Une description formelle opposée à une description du contenu. Pourquoi une description formelle ?

Dans une base de données, le document doit être recherché par la biais d'une description formelle. La description formelle fournit des informations assez abrévatives (informations bibliographiques tels que : le titre, l'auteur, le lieu et l'année de publication etc.) Ce type de description est essentiel pour montrer l'existence ou pas d'un document donné.

Mais les documents électroniques exigent une description autre que la description formelle. Une description plus fine du contenu est donc nécessaire.

1.4.1.1. Le catalogue traditionnel des manuscrits

Le catalogue traditionnel est un exemple de description formelle des documents. Dans le cas des manuscrits, on peut dire que le catalogage a souvent été fait de façons différentes qui varient d'un catalogueur à un autre, c'est-à-dire que le catalogage est plus au moins fait en fonction de l'intérêt et de la spécialité des catalogueurs. Leurs méthodes consistaient en la compilation de listes de notices, sous une forme réduite, d'une collection particulière ou encore de descriptions exhaustives avec des citations extensives.

En fait, cette liste donne un aperçu rapide d'une certaine collection et ne contient que les informations bibliographiques minimales nécessaires, à savoir: l'auteur, le titre, la date de la copie et les dimensions physiques. Par ailleurs, le catalogue est plus exhaustif et contient des descriptions plus détaillées sur l'apparence extérieure des manuscrits et sur la manière suivant laquelle ils ont été faits. Ces éléments de description sont aussi importants que ceux identifiant le contenu. Ils permettent d'avoir des informations sur la reliure, la décoration ou l'encre, qui peuvent être utilisés pour identifier la date de création de ce document (l'écriture du texte) en cas de l'absence d'une date précise. La description du document, que ce soit physique ou intellectuel, a toujours été faite selon les interprétations des catalogueurs qui se basent sur les caractéristiques du document à traiter, sans avoir aucune référence normative fixe.

Autrement dit, la description des documents est effectuée suivant les différentes normes faites par le catalogueur, suite à ses recherches, afin de répondre à ses besoins propres de description.

Le catalogue traditionnel des manuscrits a été sous forme d'un livre imprimé ayant ses propres caractéristiques et sa propre présentation. Le catalogage des documents en général et des manuscrits en particulier souffre du manque de règles normatives.

1.4.1.2. Le catalogue informatisé

L'informatisation des catalogues est une technologie qui a amélioré l'utilisation du contenu des anciens catalogues manuels et traditionnels. Ils ont rendu possible les recherches sur tous les types de documents, que ce soit texte, manuscrits numérisés, images et même le son et les séquences vidéos.

Le catalogue électronique n'est plus le catalogue traditionnel propre à une bibliothèque précise dans un endroit précis qui demande le déplacement des utilisateurs pour pouvoir y accéder. Il a changé de prestations mais aussi de fonctions. Ces fonctions ne sont plus limitées à un lieu physique (salle de catalogue) par exemple, au contraire et grâce à l'Internet, le catalogue électronique ou encore le catalogue interconnecté peut être interrogé par des internautes du monde entier.

Au cours de ces dernières années, presque tous les services des bibliothèques ont été informatisés. Citons principalement le cas des catalogues d'ouvrages et de périodiques qui forment un outil performant et très utilisé pour un accès rapide aux différents documents. En effet, ces catalogues sont devenus de plus en plus multimédias avec l'insertion des nouveaux types de documents tels que l'image, le son, le texte numérique et les manuscrits.

Les nouvelles technologies ont beaucoup aidé les bibliothécaires et les archivistes dans le sujet de traitement et d'accès aux documents rares (manuscrits, cartes, etc.). Et bien évidemment, il y a eu la création d'un format électronique normalisé qui a servi de base pour faciliter le traitement et l'accès à ces documents. Grâce à la normalisation, l'internaute peut consulter le catalogue d'une bibliothèque en parcourant une structure familière, ce qui n'était pas le cas quand chaque site présentait son catalogue de façon différente.

1.4.1.3. L'évolution du catalogage consécutif à l'arrivée de l'Internet

La catalogue électronique est la première initiative menée par les institutions qui conservent les documents (ouvrages, manuscrits, documents multimédias etc.) Le but de ces institutions a été en premier lieu de faciliter l'accès distant à leurs fonds, et en deuxième lieu, de faciliter la diffusion de l'information.

La Bibliothèque du Congrès à Washington a été la première à mettre en place un catalogue électronique. D'autres instituts (privés ou publics) ont suivi son chemin. Selon Fabienne Queyroux, «un traitement électronique du catalogue permettrait de ne plus se poser la question du volume et de constituer automatiquement des index aussi fournis que nécessaire, tout en permettant une diffusion de l'information selon toute probabilité plus efficace, car d'emblée plus universelle grâce à l'Internet »[21].

Mais, le catalogue électronique demande la mise en oeuvre de formats différents de ceux utilisés dans le catalogue sur papier. Plusieurs organisations ont participé à la mise en place des formats qui correspondent à certains types de documents à partir du format

[21] Queyroux, Fabienne. L'informatisation des catalogues de manuscrits : rapport à la suite d'un voyage d'étude dans cinq bibliothèques nord-américaines, septembre- octobre 1999. « 36pages » P.3
http://www.sup.adc.education.fr/bib/Acti/Coop/Fulb/Queyroux.htm

MARC (*Machine Readable Catalogue or Cataloguing*) dont la première version a été créée en 1962. Le format MARC définit une structure logique des données saisies dans un système informatique de gestion de bibliothèque. Le format a subi des modifications selon les besoins des bibliothèques et des pays. On a assisté ainsi à la définition d'USMARC, CANADAMARC, UKMARC, SWEMARC etc. USMARC signifie le format MARC des Etats Unis (United States Marc), CANADAMARC est celle de Canada etc., Cependant, aux Etats-Unis, grâce à un système bibliographique national du au format MARC AMC conçu dans les années 1970, la diffusion de l'information sur les collections des bibliothèques s'est faite à une grande échelle. La création de notices sur MARC AMC a fourni la possibilité aux chercheurs de trouver les informations, avec une facilité pareille à celle du catalogue sur papier. Malgré l'avancement de MARC AMC, les notices MARC n'étaient que des notices abrégées et ne permettaient pas la mise en œuvre de stratégie fine de recherche d'information. Les chercheurs étaient assez insatisfaits de cette situation.

L'existence de formats MARC nationaux différents ne facilitait pas les échanges internationaux. Pour résoudre ce problème, le format UNIMARC qui unifie tous les autres formats (*Unified Format*) a été défini. Il correspond à la partie commune de tous les formats concernés. Comme format d'échange universel, il contient un grand nombre de champs et sous-champs qui reflète la diversité des usagers et des formats MARC.

Avec l'émergence de l'Internet et l'évolution du document électronique en texte intégral, d'autres genres de formats sont apparus, comme le format SGML•, suivi par le format HTML et XML•, etc.)

1.4.2. Les métadonnées

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 »[22].

• SGML (Standard General Markup Language)

• XML (extensible Markup Language)

[22] Role, François. Représentation et exploitation de métadonnées complexes : le cas des documents anciens. – Document numérique. Vol.3. N°.1-2/1999. Pages 135-150 (P.136)

Les métadonnées désignent toute forme de données relatives à d'autres données (data about data) et servent à en faciliter l'utilisation « catalogues bibliographiques, inventaires, index, sommaires, etc. »[23] Elles sont également définies comme « *the content of a surrogate record that characterize an object* »[24]. Pour Alan Hopkinson, les métadonnées sont un élément de référence à une donnée utilisée pour aider à l'identification, à la description et à la localisation d'un réseau de ressources électroniques : « *It is used increasingly to refer to any data used to aid the identification, description and location of networked electronic resource* »[25]. Stuart Lee les a appelées aussi « *data about data* » ; pour lui c'est un matériel descriptif qui enregistre une gamme d'information attaché à l'objet numérique même, «*descriptive material that records a range of information attached to the digital object itself* »[26]. Ils sont une « pièce de données qui décrit une autre pièce de données »[27] du point de vue de la nature du document, de son créateur, du format utilisé pour stocker le document, de l'endroit de stockage etc.

Les métadonnées sont devenues récemment importante surtout pour ceux qui sont concernés par la création et la gestion de document électronique. Plusieurs organisations considèrent aussi que les métadonnées sont indispensables pour le stockage et la dissémination d'information sur Web.

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.

- 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

---

[23] ibid.

[24] Weibel, Suart, Jean Godby et Eric Miller. Workshop/ DCI: OCL/NCCSA metadata workshop report, march 30, 1999 11pages (P. 4) http://purl.org/dc.

[25] Hopkinson, Alan. UNIMARC and metadata: Dublin Core. 64th IFLA general conference, august 16- august 21, 1998. 5pages http://IFLA.org/IV/ifla/138-161e.htm (25/08/1999)

[26] Lee, Stuart D. Digital imaging: a practical handbook. New York: Neal-Schumen, Inc. 2000. 194pages (P. 104)

[27] Morrison, Alan, Michael Popham and Karen Wikander. Creating and documenting Electronic Texts: a guide to good practice. Chapter 6: Documentation and Metadata. 15pages (P.1) http://ota.ahds.ac.uk/documents/creating/chap4.html 02/03/2000

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.

Lors de la création des métadonnées on doit se poser la question suivante: «*what exactly one wishes to record about the individual item and how that will be used by the project and its end-users?*».[28]

1.4.3. Les métadonnées : un supercatalogue.

Le concept de métadonnées recouvre, bien sûr les éléments traditionnels d'un catalogue: description formelle du document et les informations de liens entre la notice et le document physique. Mais le changement intervenu avec l'usage des nouvelles technologies et des métadonnées fait apparaître trois aspects nouveaux : les potentialités offertes par les métadonnées permettent d'abord une description plus fine (et donc un accès plus précis) à l'information et ensuite un élargissement de la description à des éléments jusqu'à présent exclus des catalogues habituels. Enfin, l'usage de l'Internet a donné naissance à des documents purement électroniques pour lesquels les règles classiques de catalogage ne pouvaient pas s'appliquer avec efficacité.

[28] ibid (P. 109)

1.4.3.1. Un catalogage plus fin relatif aux parties de documents

L'usage des métadonnées permet un catalogage à un niveau plus fin que dans les catalogues traditionnels, notamment en ce qui concerne des parties du document. On peut décrire sa structure logique (chapitre, sous chapitre, tables de content, index etc.) et des éléments du contenu autre que le texte (illustration, cachet, figure, graphique etc.

1.4.3.2. Un catalogue généralisé incluant des éléments de contexte

Les métadonnées dans ce cas reprennent celles que l'on trouve habituellement dans le catalogue traditionnel. Autrement dit, il donne des informations générales sur le document comme l'auteur, le titre, le lieu et date de publication, les conditions de production, le coût, l'usage, etc., tout cela pour faciliter la recherche d'un document quelconque. Le format MARC est un exemple des métadonnées créées pour les bibliothèques, métadonnées qui consistent en un ensemble de règles appliquées par la communauté des professionnels.

1.4.4. Une description des documents accessible par le Web

Les documents numérisés ne sont pas tous homogènes dans leur catégorie (roman, poésie, texte scientifique etc). En conséquence la variation se trouve au niveau de la structure logique et physique de ces documents (chapitre, paragraphe, introduction etc.). Bien évidemment la variation de structuration n'est pas spécifique aux documents numériques mais elle relève de la nature même de ce document. Le texte poétique n'a pas la même structure qu'un roman ou qu'un texte scientifique. Ce dernier est souvent accompagné par des illustrations ou des graphiques. Faciliter l'accès à ces documents sur le web exige deux fonctionnements importants par les professionnels de l'information : le premier est la création des normes spécifiques à ces documents et le deuxième l'utilisation du langage de balise pour le rendre consultable par l'utilisateur.

Plusieurs initiatives déjà apparues visent à la création de métadonnées propres aux documents électroniques sur le Web afin de trouver un moyen d'accès plus simple et plus rapide. Citons comme exemple Dublin Core, TEI, le EAD etc.

1.4.4.1. Dublin Core

Vers l'année 1995[29] , un groupe d'experts, sponsorisé par *Online Library Center* (OCLC) et *National Center For Supercomputing Applications* (NCSA), s'est réuni à Dublin, Ohio. Leur but était de proposer un ensemble d'éléments d'identification permettant une description simple pour les documents électroniques écrits sur différents sujets. Cet ensemble de métadonnées vise à développer des mécanismes de description de différentes sources de document dans l'environnement électronique, pour permettre aux auteurs et aux éditeurs de texte électronique de décrire leur propre document sur l'internet. Ces éléments facilitent à la fois la recherche et la récupération d'information. En fait quinze[*] éléments d'identification principaux sont définis pour former le Dublin Core. Les quinze éléments se repartissent en trois catégories principales : les métadonnées relatives au contenu du document ; celles qui concernent la propriété intellectuelle et celles qui sont relatives à l'instance du document considéré. Le jeu d'éléments Dublin Core sont tous des éléments optionnels et répétables. Ils décrivent non seulement l'auteur du texte mais aussi d'autres informations de catalogage plus précises et plus complètes, comme le peintre, le photographe, le programmeur considérés comme créateurs de texte.

[29] Richy, Héléne. Métadonnées pour les documents sur l'Internet. *Document Electronique Dynamique (CIDE'2000)*, M. Gaio, E. Trupin (Eds.), Europia Productions, 141-151, July 2000.

[*] (Title; Author, Creator; Subject, Keywords; Description; Publisher; Other Contributor; Date; Resource Type; Format; Resource Identifier; Source; Language; Relation; Coverage; Rights Management)

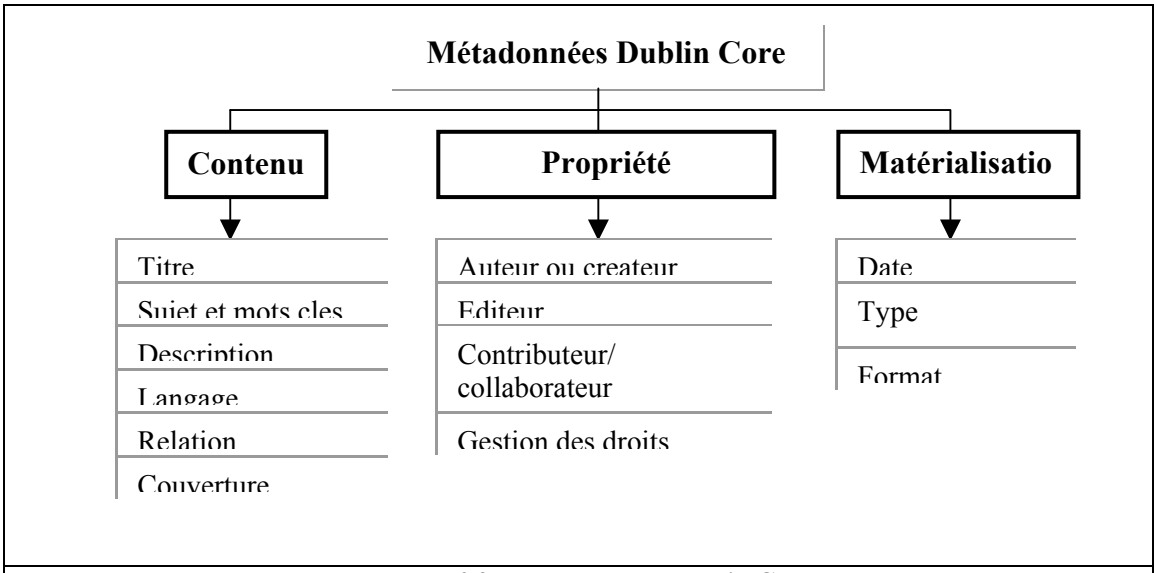

Tableau n°.2 : Métadonnées Dublin Core

1.4.4.1.1. Les inconvénients de Dublin Core

L'inconvénient de Dublin Core est le manque de clarté dans le mécanisme de la qualification. Le mécanisme de la qualification est l'élément qui enrichit la description dans Dublin Core.

Ce manque de clarté a conduit à des interprétations locales différentes, ce qui a rendu difficile l'interopérabilité de Dublin Core.

Actuellement un nombre très important d'utilisateurs utilise les métadonnées Dublin Core avec l'Hyper Text Markup Language (HTML) qui est la *lingua franca*[30] dans le majorité du World Wild Web.

1.4.4.2. Le métadonnées EAD (Encoding Archival Description)

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. Aux Etats Unis, il n'existe pas de fonds autonome des manuscrits mais il est classé avec la collection d'archives. Donc nous voudrions présenter les normes d'archivage EAD qui ont été conçues au cours de l'année 1993, en parallèle avec un projet de la bibliothèque de l'Université de Californie, Berkeley. Le but du projet de Berkeley est d'étudier la

[30] http://www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/

possibilité de développer des normes d'encodage standards et non-propriétaires, destinées aux documents lus par la machine, tels que : les *inventaires*, les *registres*, les *index* et à d'autres documents créés par les archives, les bibliothèques, les musées, pour aider à l'usage de leurs collections.

Les directeurs du projet ont voulu inclure d'autres informations que celles fournies par les notices traditionnelles du MARC. Dès le début, le développement du EAD DTD était un travail coopératif, avec des spécialistes du Berkeley en consultation avec des experts venus d'autres institutions. Daniel Pitti, le fondateur du projet de Berkeley, a établi des conditions pour la création des normes d'encodages standards qui incluent les critères suivants:

1) la capacité de présenter les éléments de recherche traditionnelle d'archive, de manière exhaustive,
2) la capacité de conserver la relation hiérarchique existante dans les différents niveaux des descriptions,
3) la capacité de représenter les informations descriptives obtenues dans les différents niveaux hiérarchiques,
4) la capacité de déplacer l'information dans une structure hiérarchique,
5) l'aide à l'indexation et à la récupération des éléments de descriptions d'archive très spécifiques.

La norme SGML a été choisie parmi les autres formats existant à cause de certaines caractéristiques qu'elle possède.

En mars 1995, la première version du Projet du Berkeley est apparue sous le nom (BFAP) DTD[*], aussi connu par le FINDAID DTD. Dans cette version, l'équipe de travail a tenté de définir une catégorie de documents qui, en général, consiste en une page de titre facultative, une description d'une unité de matériel d'archive, etc.. Ces DTD ont été testées par un groupe d'experts.

[*] DTD = Document Type Definition. C'est une liste des types d'éléments utilisés dans la description des documents.

L'équipe de Berkeley définit deux segments pour le document : le premier segment fournit l'information sur les moyens de recherches principaux (le titre, le compilateur et la date de compilation) et le deuxième segment fournit l'information sur le corps des documents d'archive (une collection, un ensemble des documents ou un feuilleton).

En suivant l'exemple de Texte Encoding Initiative (TEI), le groupe a défini le segment de "l'en-tête" qui est divisé en deux types d'informations:

- L'information hiérarchiquement organisée qui décrit une partie des notices bibliographiques ou un article avec ses différentes parties ou divisions.
- L'information associée qui ne décrit pas des notices ou des articles directement mais qui facilite leur usage par les chercheurs (par exemple, une bibliographie).

Préparation de la version 1.0 du EAD DTD

La première version de EAD a été modifiée par l'introduction de format XML. Les documents déjà encodés selon SGML seront transformés vers le format XML par l'introduction de quelques modifications exigées par XML. Pour avoir un document valide XML, il faut les modifications suivantes :

- Les noms des éléments et des attributs : dans XML, les éléments et les attributs doivent être écrits en minuscule et non en majuscule comme dans SGML. Le groupe de travail de EAD était conscient de ce problème. Donc tous les éléments et les attributs ont été mis en minuscule.
- Les éléments vides « empty » : Dans XML il existe des éléments vides qui ne contienne pas d'élément filles. Ces éléments sont marqués très explicitement par le signe « /> ». Parmi les éléments EAD, il y a sept éléments vides qui doivent être modifiés pour être en accord avec XML.
- Les attributs : dans SGML, seuls les attributs doivent être mis entre guillemets, alors que dans XML tous les mots doivent l'être.

Les créateurs d'EAD essaient d'établir une relation entre les champs de description existant dans le format MARC et leurs correspondants dans EAD.

1.4.4.2.1. Le structure des métadonnées EAD

145 éléments sont réunis pour construire les métadonnées EAD qui a pour but principal de décrire deux types d'informations concernant les documents d'archive :

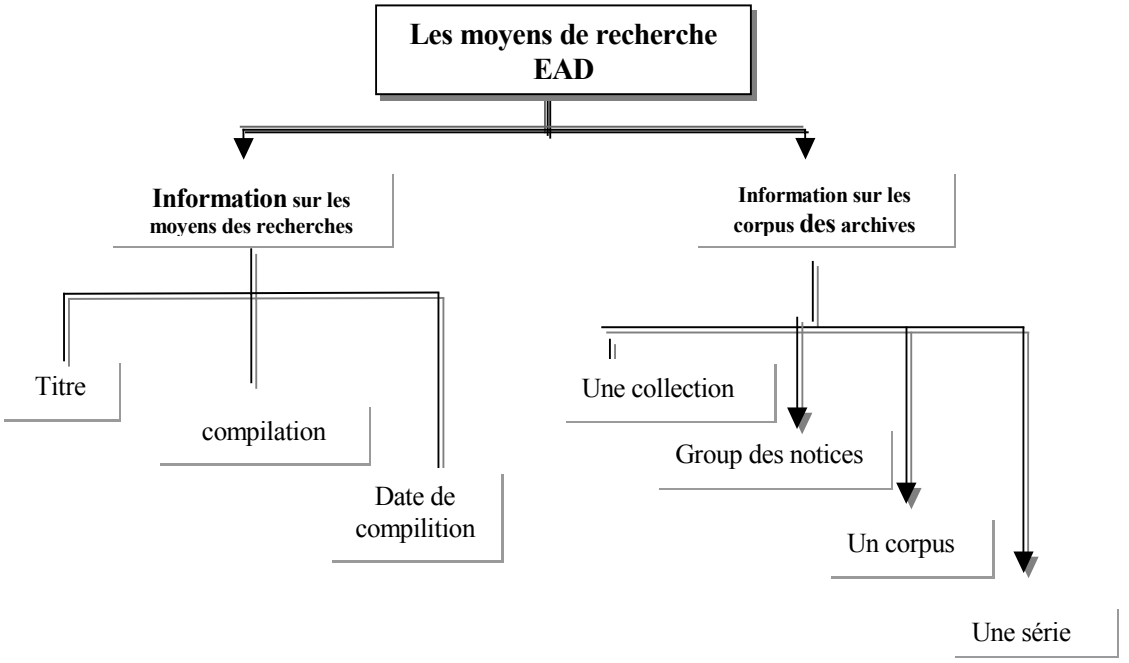

Tableau n°.3 : Les moyens de recherche EAD

Pour faciliter l'utilisation de leurs métadonnées, les ingénieurs de EAD ont défini trois grands segments de la DTD : le ead header <eadheader>, le Front Matter <frontmatter> et l'Archival Description <archdesc>.

- Le premier est l'*eadhead* et vient du modèle de la TEI, il décrit l'information qui concerne le moyen de recherche eux-même
- Le deuxième est *frontmatter* qui décrit les éléments nécessaires à l'affichage et à la publication des moyens de recherche.
- Le troisième est l'*archdesc* qui contient des descriptions sur les notices d'archive et les recueiles des manuscrits.

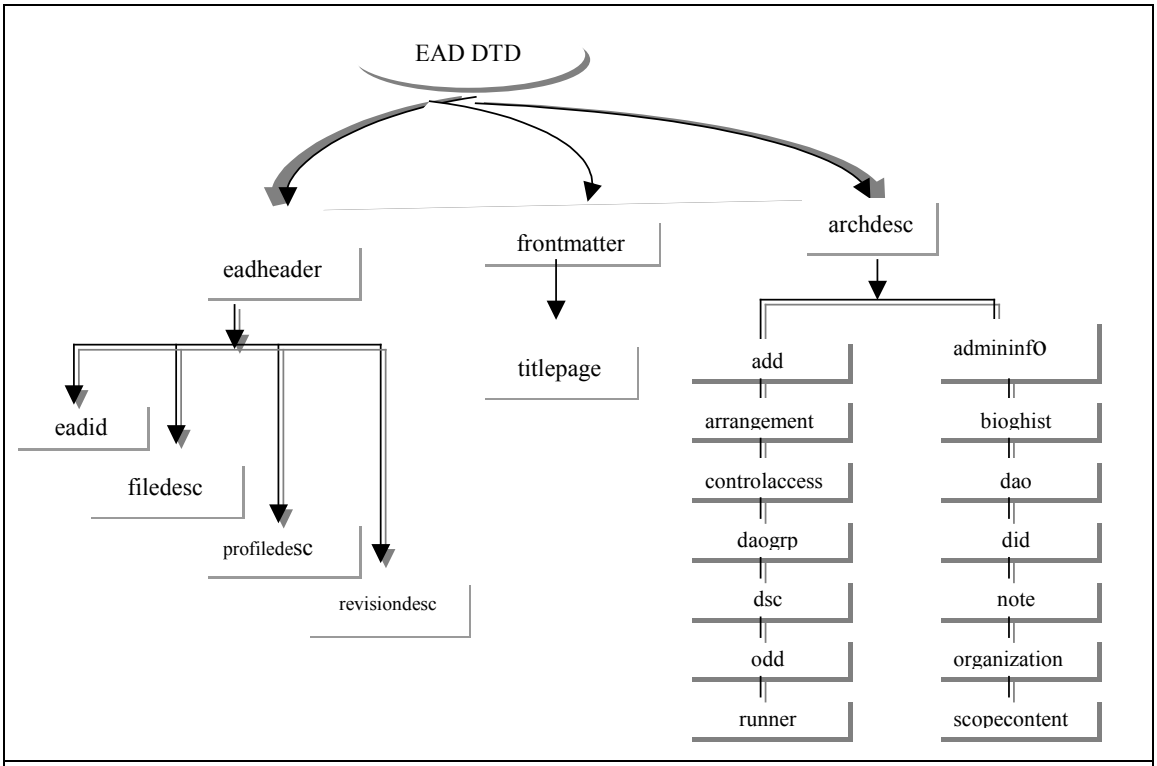

Tableau n°.4 : EAD DTD

Les éléments mentionnés dans la figure ci-dessus sont la base à partir de laquelle les autres éléments fils sont construits. Chacun des éléments mentionnés dans la figure ci-dessus a été désigné pour une tâche bien spécifique. Pour une liste détaillée de tous les éléments EAD voir le site[•] dont l'adresse est mentionnée en bas de page.

1.4.4.3. TEI (Text Encoding Initiative)

L'idée de créer la TEI est née en novembre 1987 à Vassar College, Poughkeepsie, New York, lors d'une conférence organisée par *US National Endowment for the Humanities NEH,* une agence fédérale indépendante. La création de texte électronique considérée du point de vue de la recherche d'informations était le thème principal de cette conférence sponsorisée par : *The Association for Computers and the Humanities (ACH), the Association for computational Linguistics (ACL) et the Association for Literary and Linguistics Computing (ALLC).*

La première version (document TEI P1) apparut en juin 1990 à la suite du travail mutuel fait par les Américains, notamment le *NEH* et les Européens (*La communauté*

[•] http://www.loc.gov/ead/tglib1998/tlindex.html

européenne) et la fondation Andrew W. Mellon. La première version contenant les directives a été suivie en 1993 de une deuxième version. Une nouvelle version de ces directives TEI P4[31] a vu le jour en avril 2002 : elle apporte des modifications sur la TEI pour qu'elle puisse s'adapter aux formats SGML et XML.

La TEI fait référence à des codes de caractères standards reconnus par ISO 624 et ISO 10646 et Unicode ISO 646. Cette dernière définit une norme de caractères fondée sur un codage à sept bits qui a été le plus répandu dans le monde au début de la TEI et permettait un échange d'information au niveau mondial. Les jeux de caractères présents dans d'autres langues que l'anglais pouvaient être utilisés grâce à un codage indirect par le biais de tables associées « *also make reference to character encoding standards such as ISO 646, ISO 10646 and Unicode. ISO 646 defines a standard seven-bit character set in terms of which recommendations on character-level interchange are formulated; this is the most portable character set for broad interchange, but requires indirect encoding of many characters*" [32].

Comme le EAD, la TEI utilise la grammaire formelle de SGML pour définir l'encodage. Au contraire du SGML/DTD, les groupes, travaillant sur la TEI, visent à créer une DTD, flexible pour répondre aux besoins de la majorité des chercheurs. Exemple : au lieu d'avoir trois étiquettes (tags) séparées (*FOOTNOTE*>, <*ENDNOTE*>, <*SHOULDERNOTE*>, la TEI donne la possibilité d'avoir un seul élément <*Note*> avec un attribut *TYPE* pour identifier tous les types de note, que ce soit, *FOOTNOTE*, *ENDNOTE* ou *SHOULDERNOTE* etc.

L'originalité de TEI

- Étant un encodage cohérent, la TEI a été acceptée par les spécialistes dans le domaine des sciences humaines, comme la norme de balisage non-propriétaire.
- Il fournit trois éléments principaux : le jeu d'étiquettes complet ; la méthodologie et l'ensemble des DTD pour décrire en détail la forme des documents du point de vu intellectuel, structurel et typographique.

[31] http://www.tei-c.org/P4X/AB.html

[32] http://www.tei-c.org/P4X/AB.html

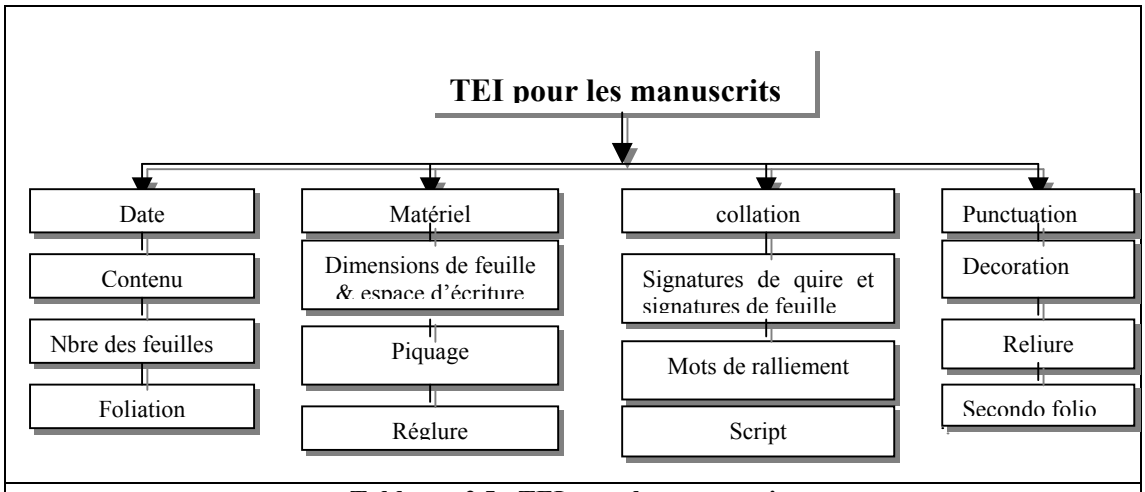

Tableau n°.5 : TEI pour les manuscrits

- La TEI a créé un système de base pour lire le texte en permettant le traitement informatique. Ce système doit faciliter l'échange et le partage de textes numérisés.
- Il produit un modèle général qui devrait pouvoir s'appliquer à des textes de tout langage et de tout genre, répondant aux besoins actuels des chercheurs.
- La TEI est une norme universelle qui peut être mise en œuvre avec des notices, de la plus simple à la plus complexe, et dans tous les domaines.
- Elle fait un lien entre les métadonnées et les parties de texte.

1.4.4.3.1. La TEI et les manuscrits

La norme d'encodage TEI pour le catalogage des manuscrits médiévaux est actuellement fondée sur les seize[33] éléments de descriptions, proposées par Neil Ker de la bibliothèque britannique « British Library », qu'il s'agisse de la date, du contenu, du nombre de feuilles, de la foliation, du matériel, de la dimension de la feuille, de l'espace d'écriture, du pricking (percer le papier afin de le préparer pour l'écriture), de la réglure, du feuilletage (collation), de la signature du cahier et signature de la feuille, des mots vides (catchwords), des types d'écriture, de la ponctuation, de la décoration, de la reliure, du *secundo folio*.

Malgré ses limites, les seize recommandations sont largement acceptées, pour préparer les notices des manuscrits dans un catalogue compréhensif. Il faut noter ici que le projet MASTER, que était la base de notre travail pour l'élaboration de métadonnées pour les

[33] http://www.stg.brown.edu/conferences/tei10/tei10.papers/gartner.html

manuscrits arabes, est aussi fondé sur le TEI. Mais, d'autres éléments supplémentaires ont été ajoutés par le projet MASTER.

La TEI a défini des encodages spécifiques aux descriptions de manuscrits mais un peu limité. «La TEI *P3* avec son format actuel fournit une facilité très limitée dans l'encodage détaillé de manuscrits. Par exemple, l'élément <header>, <hand> et <handshift> ne contient pas suffisamment d'éléments spécifiques aux manuscrits, surtout pour décrire les informations concernant les scribes et le style d'écriture» [34]. En conséquence et pour apporter une solution à ce problème, les experts des manuscrits doivent modifier et ajouter des DTD pour des applications spécifiques comme dans le cas de projet MASTER et la « Bibliothèque Bodléienne « Bodleian Library » à Oxford.

En 1996 la Bibliothèque Bodléienne commence à travailler sur la possibilité de créer des TEI supplémentaires pour ajouter des métadonnées plus détaillées, afin de répondre au besoin de leur catalogage des manuscrits. C'est un projet de quatre ans, dont le but est de donner accès aux descriptions des manuscrits médiévaux occidentaux qui font partie de la collection de la Bibliothèque Bodléienne mais qui ne sont pas encore catalogués.

Il est indispensable de mentionner d'abord le structure de base de TEI avant de décrire la TEI de la Bibliothèque de Bodleian.

1.4.4.3.2. Les jeux des étiquettes de TEI Lite header

L'entête <*Header*> est le commencement de la description d'un manuscrit. Elle contient des informations essentielles sur le document lui-même, sur la source originale et l'encodage utilisé. Elle fournit un outil pour décrire, d'une manière complète, tous les aspects électroniques du texte. Cette entête est une source d'information essentielle pour l'utilisateur du texte, pour le logiciel qui va gérer les métadonnées et pour les catalogues de bibliothèque, de musée et de centre d'archives. On considère, en général,

[34] Gartner, Richard et Lou Burnard. A TEI extension for the description of medieval manuscripts. Abstract for TEI 10. August 20, 1997. http://www.stg.brown.edu/webs/tei10/reviews/papers/gartner.html 4 pages (P.1) 21/09/1999.

que l'entête apporte les mêmes informations que celles fournies par la page du titre d'un ouvrage imprimé. Voir l'exemple suivant[35] :

```
< !DOCTYPE tei.2 PUBLIC "-//TEI//DTD TEI Lite 1.6//EN"><tei.2>
<teiheader>
[Détail sur l'entête se trouve ici]
</teiheader>
<text>
<front>…..</front>
<body>
```

Le <teiHeader> peut être un objet complexe ou simple ; il doit contenir quatre composants majeurs des éléments descriptifs dont seul, le <fileDesc> est obligatoire, alors que le reste <encodingDesc>, <profileDesc> et <revisionDesc> est optionnel. Chacun de ces éléments contient d'autres sous-éléments et quelquefois, des attributs qui servent à décrire l'information bibliographique, la relation entre le texte et sa source, la langue du document, les modifications qui ont été apportées au document, etc.

Cette entête est tellement souple qu'elle donne aux informaticiens la possibilité d'extraire et d'ajouter d'autres normes de catalogage comme les notices de bibliothèque faites selon le format MARC. Par exemple: l'étiquette <auteur> qui est à l'intérieur de <titleStmt> est analogue à l'étiquette 100 du format Marc ou à l'élément <CREATOR> dans la norme de Dublin Core. La normalisation des champs est un avantage qui permet de réunir les différents formats des métadonnées existants.

1.4.4.3.3. Les éléments de TEI ajoutés par la Bibliothèque Bodleian

Le projet de la Bibliothèque Bodleian propose huit éléments supplémentaires et les sous-éléments et les attributs associés, destinés à l'entête et à la référence bibliographique. Les huit éléments principaux sont : <décoration>, <area>, <leaves>, <foliation>, <collation>, <scriptDesc>, <rubrication> et <secundo folio>. Le dernier élément a été proposé par Neil Ker de la British Library parmi l'ensemble des seize éléments déjà mentionnés dans la section précédente.

---

[35] Morrison, Alan, Michael Popham and Karen Wikander. Creating and documenting Electronic Texts: a guide to good practice. Chapter 6: Documentation and metadata. 15pages (P.4)
http://ota.ahds.ac.uk/documents/creating/chap5.html 02/03/2000

Ces éléments supplémentaires proposés par le projet ne sont pas toujours applicables à tous les catalogues de manuscrits. « Il y a un manque monumental dans le domaine d'inscription et dans le domaine des manuscrits non–latin »[36].

1.5. Comment automatiser cette description à partir du document lui-même ? Les langages à balises.

En soit, l'ensemble des éléments définis par les différents acteurs mentionnés dans les parties précédents reste inutile, sans l'intervention d'un format électronique. Un langage à balises est donc nécessaire à la fois pour définir ces éléments, pour établir un lien hiérarchique entre eux et pour rendre les documents saisis sous ces formats accessibles par le web.

1.5.1. Les langages à balises (markup)

1.5.1.1. Les balises, leur rôle et leur définition.

Il est nécessaire de préciser les éléments mis en jeu :

- J'appelle document source le document initial numérisé ; c'est une séquence de photographies numériques, chacune étant l'image d'une page.
- J'appelle document de description du document concerné l'ensemble des étiquettes renseignées permettant l'accès au document source. Ce document de description peut être considéré comme un texte (par exemple, c'est un bordereau de saisie), mais il doit être utilisé en ordinateur. La machine doit donc être capable d'identifier les étiquettes et leur contenu automatiquement.

La solution à ce problème passe par l'utilisation d'un système de balises. Ce système est ancien et il est utilisé depuis un certain temps dans l'imprimerie pour donner des directives de composition à l'imprimeur. En informatique, les balises vont servir à délimiter les composants du document de description. Il existe deux grandes classes de balises:

[36] Gartner, Richard et Lou Burnard. A TEI extension for the description of medieval manuscripts. Abstract for TEI 10. August 20, 1997. http://www.stg.brown.edu/webs/tei10/reviews/papers/gartner.html 4 pages (P.4) 21/09/1999.

1. Les balises qui sont relatives à la présentation du document par exemple à l'écran ou sur une imprimante. Ces balises sont dites « propriétaires » parce qu'elles sont imposées par le système de visualisation ou d'impression utilisé. Elles permettent de décrire toutes les opérations de mise en forme du document (police de caractères, taille, italique, gras, centrage des paragraphes, etc.)
2. Les balises descriptives à la disposition de l'utilisateur et qui lui permettent de décrire la structure ou les éléments de contenu du document source. Ces balises sont libres (et non-propriétaires) en ce sens que l'utilisateur peut les définir à son gré en fonction de ses besoins.

1.5.1.2. Un exemple de balises ciblées pour servir à l'impression

Ce sont le *PostScript* et *Portable Document Forma*t (PDF), deux langages créés consécutivement en 1985 et 1993 par Adobe. Les deux formats appartiennent à la catégorie des balises de présentation de document plutôt qu'à la catégorie des balises descriptives. Le *PostScript* a pour objectif d'établir un langage entre l'ordinateur et l'imprimante. Autrement dit, ce langage donne à l'imprimante les instructions concernant l'apparence de la page (le texte, les graphiques, les couleurs, les images) afin que le document conserve sa présentation au moment de la transmission d'un ordinateur à l'autre. Ce langage est reconnu par les imprimantes de certaines marques comme *Epson*, *IBM*, *Hewlett-Packard* etc. Le langage *PDF* a été créé comme un langage compatible avec le *PostScript* mais avec une différence majeure : il permet aux utilisateurs de visualiser le document d'une manière intégrale qui ressemble presque à une image scanner de source. Le format *PDF* garde le document intact grâce à sa capacité de faire la combinaison entre le texte, le graphique et l'image. Un document *PDF* peut être lu par l'intermédiaire d'*Acrobat Reader*.

1.5.2. Document Type Definition (DTD)

La DTD, comme son nom l'indique, sert à définir un type de document donné. Le mot document est compris dans un sens très général ; il peut comprendre différents

contextes associés au document considéré. La DTD a pour but de déclarer la structure d'un type de documents en précisant ses différentes composantes c'est dans la DTD que sont définis le nom des éléments descriptifs et leurs types de contenus ainsi que la relation entre ces éléments. C'est aussi là que sont définis les attributs ainsi que leur type. Chaque élément logique peut comporter un ou plusieurs attributs qui ont pour objectif de fournir des informations sur les éléments en question. La définition des différentes relations structurales entre l'élément et l'attribut est donc nécessaire.

Il existe plusieurs types de DTD qui chacun définit un certain type de document, à savoir la TEI DTD (*Text Encoding Initiative*) pour les textes littéraires. Daniel Pitti a développé l'EAD-DTD (Encoding Archival Description) propre aux inventaires d'archives, alors que le MARC- DTD a été créé pour participer à la conversion des notices MARC en SGML ou encore en HTML, etc.

1.5.2.1. L'utilisation du Document Type Definition DTD

La DTD aide à schématiser un document encodé par les langages à balises comme SGML, HTML et XML. Il traduit les règles de SGML utilisées dans le document. Selon Alan Morrison, Michael Popham et Karen Wikander « *Often, when people talk about 'using SGML', they are actually talking about using a particular DTD* ».[37] La DTD établit les normes que le balisage doit utiliser et il définit le mode d'interaction entre les différents balisages. Les syntaxes utilisées dans la DTD sont simples à lire et à écrire.

La création d'une DTD pour un document quelconque exige une bonne expérience de la part du créateur et une familiarité avec le besoin qu'en a l'utilisateur final. Le but principal du créateur de DTD est toujours d'établir les mécanismes possibles qui permettront aux utilisateurs d'adapter la DTD à leurs besoins et de lui donner un maximum d'extension.

Donc, la DTD insiste sur le fait que le document doit contenir certains éléments et pour indiquer que ces éléments sont placés dans un certain ordre et finalement pour définir

[37] Morrison, Alan, Michael Popham and Karen Wikander. Creating and documenting Electronic Texts: a guide to good practice. Chapter 5: SGML/XML and TEI. 21pages (P.6) http://ota.ahds.ac.uk/documents/creating/chap5.html 02/03/2000

les types d'informations que ces éléments doivent contenir. A partir de ce mécanisme, on a la possibilité, grâce à l'analyse syntaxique avec la DTD source, de créer un DTD qui définit une norme et d'assurer que chaque document doit nécessairement être conforme à cette norme.

La déclaration de type de document et la définition de type de document sont deux choses différentes. Le rôle principal de la déclaration de type de document est de donner une référence à la DTD à laquelle un document devrait être comparé pour être valide. Il se trouve dans le prologue du document SGML/XML.

L'exemple qui suit est une déclaration faite par la « déclaration de type de document » qui accompagne le texte encodé en HTML pour indiquer que le document utilise le HoTMetaL PRO 4.0 DTD[38] :

```
<DOCTYPE HTML PUBLIC "-//SoftQuad//DTDHOTMetaL PRO 4.0::
19970714 ::extensions to HTML 4.0//EN" "hmpro4
dtd">
```

1.5.3. SGML (Standard General Markup Language).

1.5.3.1. Une présentation générale

Le SGML est une norme internationale d'un métalangage d'origine américaine apparue vers l'année 1996 mais l'origine de ce travail remonte au début des années 1980. Un comité international d'experts, dirigés par Charles Goldfarb, a été l'auteur de ce travail. Le SGML n'est pas un nouvel arrivant mais il est le résultat d'un travail commencé dans les années soixante qui se dénommait GML (*General Markup Language*). Le SGML identifié comme étant la norme ISO 8879 a été normalisé au niveau international vers l'année 1986 ; pour la France, depuis 1991, c'est la norme NF EN 28879, 1990. Le SGML a été la première tentative systématique de créer de réels documents électroniques. Au depart la norme a été utilisée seulement dans le milieu scientifique nord-américain mais, actuellement elle a pris une dimension plus large et

[38] Morrison, Alan, Michael Popham and Karen Wikander. Creating and documenting Electronic Texts: a guide to good practice. Chpter 4: Markup: the key to reusability. 8pages (P.6) http://ota.ahds.ac.uk/documents/creating/chap4.html 02/03/2000

est devenue utilisable par tous les secteurs de l'édition, surtout avec le format XML (eXtensible Markup Language) qui est un sous-ensemble de SGML.

Selon Pierre-Yves Duchemin, « l'objectif principal de SGML est de définir des structures logiques (définition de type de document = DTD) ou des structures hiérarchisées de données selon une architecture arborescente, a priori infinie, qui permet, entre autres, la représentation de données en format MARC»[39] . Au contraire de HTML, la balise du SGML a une structure strictement hiérarchique. Ainsi, un ouvrage doit avoir un chapitre et chaque chapitre doit avoir un paragraphe etc. Pour cette raison, le SGML peut donc s'appliquer à n'importe quel type de texte. Il est possible aussi de décrire avec un format SGML les notices MARC ou les notices d'une base de données. Egalement, le SGML vise à établir une technologie de base pour l'Internet.

Etant un langage de balise descriptive, le SGML peut servir comme encodage de base pour le développement d'un projet général de système de description électronique du manuscrit. En conséquence, l'encodage SGML est plus avantageux pour les descriptions de manuscrits électroniques que celle du système utilisé dans une base de données.

L'exploitation automatique de document par la norme SGML « permet, entre autre d'éditer le document avec une structure identique ou de repérer les informations nécessaires à l'établissement de la description bibliographique, on peut parler de « catalogue automatique»[40].

1.5.3.2. La structure du document de description SGML

Trois parties majeures sont réunies pour former un seul document de description SGML, à savoir :

- *LA DECLARATION SGML* (*SGML declaration*): Son rôle principal est d'informer le logiciel que le document en train d'être traité est un document SGML. Il déclare également toutes les contraintes de système ou de logiciel,

[39] Duchemin, Pierre-Yves. L'art d'informatiser une bibliothèque : Guide pratique. Paris : Electre-éditions du cercle de la librairie, 1996 ,424pages (P.153)
[40] Guinchat, claire et Yolande Skouri. Guide pratique des techniques documentaires-Vanves : EDICEF, 1996. vol.1, p.97. (4 volumes)

comme la longueur du nom de la balise, etc. Il informe le logiciel qui gère un document SGML de toutes les informations nécessaires comme les jeux des caractères utilisés dans le document (Unicode ou bien ISO 1046) etc.

- *LE PROLOG* : définit la structure du document. Il doit se conformer aux spécifications présentes dans la norme formelle de SGML ainsi qu'aux syntaxes trouvées dans le « SGML déclaration ». Il faut qu'il contienne au moins une Déclaration de type de Document qui, à son tour contient aussi une Définition de type de Document (DTD). La DTD est une série de déclarations qui définit un langage de balise bien particulier qui sera utilisé dans le « document instance ». Il précise comment les différentes parties d'un langage peuvent être liées ensemble (par exemple, il précise toutes les balises qui doivent être obligatoires ou optionnels, le contexte dans lequel elles seront utilisées, etc.)
- *LE DOCUMENT INSTANCE* : c'est la partie du document dans laquelle les matières premières (le contenu même du document) et les balises sont associées. Le SGML déclaration et le contenu de Prolog (spécialement la déclaration dans le DTD) ont beaucoup d'influence sur le contenu du document instance. En effet, c'est le document instance qui contient les éléments clés pour la création d'un document électronique.

Pour arriver à encoder un document par le SGML, il faut commencer par d'autres procédures, car en soi le SGML n'est pas directement utilisable. Un document doit passer par l'analyse logique qui est possible grâce à la définition de type document (DTD - Document Type Definition).

Pour créer un document, un éditeur SGML (*parser*) est nécessaire ; il facilite le tâche. L'éditeur est un logiciel qui peut lire la « SGML déclaration » et le « document Prolog ». Il comprend aussi les déclarations trouvées dans la DTD et assure que la balise SGML utilisée dans le document est conforme à ce document. L'éditeur SGML est un outil nécessaire à l'utilisateur pour qu'il puisse créer un document à partir de zéro.

1.5.3.3. L'utilisation de SGML pour la description des documents

La norme SGML, comme nous l'avons mentionné ci-dessus, est un langage de balisage, c'est-à-dire qu'il distingue, à l'aide de certains repères créés et lus automatiquement par l'ordinateur, le document et la manière dont il a été conçu. La norme SGML fonctionne selon trois éléments principaux : l'élément de description, le balisage et les attributs.

1. **L'élément de description:** l'unité principale du SGML qui peut être une partie du texte (un paragraphe complet, une tête, une note etc.) .
2. **Le balisage**: chaque élément doit être mis entre deux caractères ; le premier est le caractère d'ouverture « **<** » et le deuxième est le caractère de fermeture « **>** », le tout s'appelle une balise. Selon Claire Guinchat, et Yolande Skouri, « Le balisage est un codage spécifique qui permet de préciser l'ordre des éléments et si une donnée est obligatoire, facultative, répétable ou non, grâce à l'utilisation de « connecteurs » et « indicateurs d'occurrence» ; il existe 13 classes de déclaration de balisage »[41]. La lettre <**P**> mise entre parenthèses est un exemple d'un code (balise) qui désigne le commencement d'un **Paragraphe** alors que la même lettre avec deux parenthèses et une *barre oblique* </**P**> représente la fin du même paragraphe.

Un document SGML peut aussi consister seulement en séries d'éléments successifs, exemple :

```
<P>C'est un paragraphe</P>
<Tête>C'est un élément « tête »</Tête>
<Note>C'est une note</note>
```

3. **L'attribut** : Pour donner plus de valeur aux éléments principaux, le SGML donne la possibilité d'ajouter à ces éléments des attributs. L'objectif majeur des attributs est d'établir des liens entre les données et les éléments, ce qui permet la création de liens hypertexte. Les attributs ajoutés au paragraphe permettent de spécifier les différentes positions du paragraphe par rapport au texte. Par exemple, le paragraphe peut être:

[41] Guinchat, claire et Yolande Skouri. Guide pratique des techniques documentaires-Vanves : EDICEF, 1996. vol.1, P.154. (4 volumes)

- Centré : <P justification=centre>C'est un paragraphe centré</P>
- Justification à droite <P justification=Droit>C'est un paragraphe avec une justification droite</P>
- Les attributs et leurs valeurs peuvent être mis dans le même élément, comme dans l'exemple suivant, où la lettre n signifie que le paragraphe se situe en premier dans le texte.
- <P>justification=Droit n=1> C'est le premier paragraphe avec une justification droite</P>

Le SGML donne la possibilité d'ajouter aux attributs et à leurs valeurs des informations (seules ou combinées) pour faciliter un encodage plus complexe.

Une autre possibilité est aussi fournie par le format SGML : il s'agit de permettre à un élément donné d'englober d'autres éléments : par exemple l'élément <Chapitre> peut contenir l'élément <Tête>, et les séquences ou les éléments des paragraphes <P>. Exemple :

```
<Chapitre>
<Tête> C'est la tête de chapitre 1 </Tête>
<P> C'est le premier paragraphe</P>
<P> C'est le deuxième paragraphe</P>
<Chapitre>
```

Le peut aussi être utilisé au lieu de la balise <chapitre> pour montrer une division textuelle, l'attribut *type* serait utilisé ici pour indiquer que cette partie est un chapitre.

```
<Div type=Chapitre>
<Tête> C'est la tête de chapitre 1 </Tête>
<P> C'est le premier paragraphe</P>
<P> C'est le deuxième paragraphe</P>
</Div>
```

Dans l'exemple mentionné ci-dessus, la DTD est utilisé pour obliger chaque division à avoir l'attribut « *type* » et au moins un élément <head> suivi par au moins un élément paragraphe <p>.

L'encodage SGML est différent de la structure de la base de données dans le type de caractères utilisés pour créer la balise : par exemple, à l'intérieur de la base de données, les informations sont mises dans des délimiteurs 'head' 'p' 'note', alors que dans l'encodage SGML, les informations sont mises dans les éléments nommés <head>,

<p>, et <note> avec des délimiteurs différents. On conclut que le SGML est un langage parent de HTML et XML.

1.5.3.4. Les inconvénients de SGML

- Le SGML est réputé difficile et très coûteux car il impose un travail intellectuel prohibitif et le logiciel nécessaire et une certaine facilité de manipulation manquent.
- « Le SGML, comme le HTML, permettait aux éléments de ne pas fermer leurs balises, ce qui nécessitait un effort de codage important pour arriver à déterminer précisément où était supposé se terminer un élément»[42].
- A cause de ce problème, il y eut souvent des résultats inattendus lors de la conversion de ce format en XML.

1.5.4. Le langage de Hyper Text Markup Language (HTML)

Le HTML, qui s'inspire beaucoup dans sa construction de SGML, est un langage non-propriétaire, utilisé pour publier l'hypertexte sur le World Wide Web. Il a été créé en 1992 par un groupe de chercheurs qui travaille dans le Centre européen de Recherche nucléaire (CERN) à Genève. Il a pour but principal d'établir la structure logique d'un document en indiquant ses différents niveaux d'organisation ainsi que les informations relatives à la disposition des paragraphes, et les images à l'intérieur du document source. HTML joue un rôle fondamental dans la diffusion et la présentation d'information sur le web. Un des avantages principaux de HTML se présente dans ses jeux de balise fixes et limités, jeux destinés à créer, de façon souple, des documents hyper textuels. Le HTML est facile à comprendre et facile pour ceux qui veulent l'implanter dans leur logiciel. En conséquence, ces avantages aident à l'expansion rapide du langage HTML sur le web et à l'acceptation de ce langage à l'échelle internationale.

Quatre versions de HTML sont sorties jusqu'à présent, à savoir : 1.0, 2.0, 3.2 et 4.0. Chaque version a sa propre DTD mais c'est la version 4.0 qui est la plus recommandée

---

[42] St. Laurent, Simon. Introduction au XML.- Localisation par Arnaud Petitjean. Paris: Osman Eyrolles Multimedia, 2000. P.364 (P. 317)

par le W3C•. La version 4.0 qui a été créée le 18 décembre 1997 a pour but de normaliser le format HTML. HTML version 4.0 est une application de SGML conforme à la norme Internationale ISO 8879. Cette version est un langage normalisé pour la publication sur le World Wide Web. Une nouvelle version plus développée que la version 4.0 a été créée le 24 août 1999 par le W3C sous le nom XHTML 1.0[43].

Le HTML est un langage de balise conventionnel qui fournit aux utilisateurs un ensemble de balises prédéfini, à partir duquel les utilisateurs peuvent sélectionner celles dont ils ont besoin. Les nouvelles balises qui ne font pas partie des spécifications de HTML sont toujours rejetées par le logiciel ou par le navigateur HTML. Ce dernier point est un des points faibles de HTML.

Le format HTML a été considéré par les chercheurs comme le langage le plus répandu pour la publication sur le Web, parce qu'il permet aux utilisateurs de créer en-ligne des documents en texte intégral avec des éléments multimédias (comme les images, les sons et les vidéo clips). Les documents créés par le format HTML sont mis dans un environnement spécial pour faciliter la publication immédiate et la recherche facile d'information. Le langage HTML, à l'origine, a été ciblé pour servir à la diffusion de documents scientifiques et techniques.

1.5.4.1. Les avantages de HTML

1- Les documents encodés par l'HTML peuvent être recherchés par n'importe quel navigateur.

2- Le HTML peut traverser facilement les différentes passerelles.

3- Très simple à manipuler, facile pour la conversion des documents et facile pour la création de document par l'intermédiaire de logiciel comme *Notepad* par exemple.

4- Il a résolu le problème de complexité du document SGML, en créant des jeux d'étiquettes simples pour simplifier la structure et la sémantique du document

---

• World Wide Web Consortium.

[43] http://www.w3.org/TR/1999/PR-xhtml1-19990824/

1.5.4.2. Les inconvénients de HTML

A cause des balises non identifiées dans le format HTML, les taux de bruit dans les résultats de requête sur web sont très élevés.

1.5.5. Extensible Markup Language (XML)

La complexité de SGML et les limites de HTML ont amené le World Wide Web Consortium (W3C) à développer un nouveau standard connu comme le XML. Le format XML, qui a été initialisé officiellement en avril 1998, est un format plus avancé que les deux formats précédents (SGML, HTML) mais il est basé sur le format SGML avec quelques changements positifs.

Contrairement au HTML qui présente des jeux de balises très limités et orientés sur la présentation (titre, paragraphe, image, lien hypertexte, etc.), XML est un métalangage qui permet aux utilisateurs et/ou créateurs des documents électroniques de créer, à volonté, de nouvelles balises pour isoler toutes les informations élémentaires (titre d'ouvrages, prix d'articles, numéro de sécurité sociale, référence de pièce) ou des agrégats d'informations élémentaires qu'une page Web peut contenir. Autrement dit, dans le XML, le choix de DTD est plus flexible.

XML est un langage universel d'échange de données particulièrement performant : il est simple et peut être véhiculé grâce à des protocoles standards de transport Web comme http•. L'utilisateur de XML peut définir la DTD de son choix, de façon plus facile et plus rapide qu'en SGML. Il est même possible d'intégrer dans le même document plusieurs DTD simultanément. L'importance de XML vient de la possibilité d'englober plusieurs fonctions déjà utilisées dans les deux formats précédents (SGML et HTML). Il permet de diffuser et de recevoir des hyper-documents sur le Web, comme la notion de liens hypertextes et hypermédias de l'HTML.

Selon Alain Jacquesson et Alexis Rivier, « l'ambition de XML est de rendre plus simple et plus accessible l'utilisation de SGML sur le Web »[44]. En effet le XML réunit

---

• (HyperText Transfer Protocol)

[44] Jacquesson, Alain et Alexis Rivier. Bibliothèques et documents numériques : concepts, composantes, techniques et enjeux. Paris : Electre - Editions du Cercle de la Librairie, 1999. 377p. (p.61)

l'avantage visuel de HTML avec l'avantage contextuel de SGML/TEI. Autrement dit, le XML vise à avoir le privilège générique de SGML qui est offert par l'arbitraire de SGML DTD et la simplicité optionnelle de HTML. «Le XML n'est pas encore ISO standard qui le fait distinguer de la SGML mais il a gardé les bonnes caractéristiques de SGML (pouvoir, richesse et flexibilité) mais sans prendre la complexité de ce dernier»[45]. N'importe quelle application peut interpréter un document XML en utilisant un interpréteur et retirer l'information qu'il contient. XML est un langage orienté sur le contenu et non sur la présentation du texte.

1.5.5.1. Pourquoi XML?

L'idée générale derrière la création de XML est de donner satisfaction aux créateurs de document électronique sur Web, ce que les deux formats précédents n'arrivent pas à faire. Malgré sa richesse en sémantique, le SGML est un langage de balise normalisé, qui est lourd à mettre en œuvre et inadéquat sur Web. Par contre, l'HTML est un langage parfaitement adapté au Web mais dont les applications sont limitées par une bibliothèque de balises figée et réduite. Par sa structure, HTML mélange le contenu et la présentation. Le concept à l'intérieur de HTML est à peine présent ou en tout cas, il est difficile à modifier. XML est un langage plus riche que celui de HTML mais il ne le remplace pas (le format XML permet de décrire les données alors que HTML permet de décrire leur présentation).

Les caractéristiques générales du format XML, que chaque créateur de document XML doit prendre en considération, sont les suivants:

- La simplicité : au contraire de SGML, le XML est un format simple à utiliser sur l'Internet. La création de document XML est très simple à mener et simple à écrire par le logiciel qui gère le document XML.
- La souplesse : XML est capable d'être utilisé dans de grandes variétés d'applications.
- Pérennité : on peut penser que XML est un format qui a plus de perspectives d'avenir que SGML et HTML.

---

[45]MASTER : a gentle introduction. http://www.cta.dmu.ac.uk/projets/master/gentintr.html (dernière révision 14 jan, 2001) p.4 (05/03/2001)

- La compatibilité : avec le format SGML.
- La lisibilité, le format XML est lisible et clair.
- La structure de XML est formelle et concise.
- XML est extensible car l'ajout de nouvelles balises est à tout moment possible pour prendre en compte un nouvel élément d'information dans un document.
- La netteté de balise XML est de niveau minimal.
- Le format XML décrit le fond et pas la forme.
- L'utilisation de unicode permit de stocker n'importe quelle langue.

Les points précédents donnent au format XML son originalité. Les créateurs de XML essayent d'établir un concept de données bien formées, « *well-formed data* », qui exige de mettre les données balisées entre deux étiquettes d'ouverture (<) et de fermeture (>), stockées dans un endroit spécifique pour faciliter la structure hiérarchique des éléments des données trouvées dans le document.

L'objectif majeur de la structuration des éléments dans un document sur format XML est de faciliter l'échange d'information entre un format de base de données et un document sur un serveur Web ou vice versa. Pour aboutir à ce but, il faut que l'application source et l'application récepteur aient la même balise et la même structure.

1.5.5.2. Les différences entre le XML et le HTML

- Au contraire de XML, HTML n'a pas la notion d'identification des champs.
- HTML n'a pas de structure sémantique comme dans le XML
- La structure en HTML n'est pas aussi claire que dans XML.

1.5.5.3. Les formats composant le XML

Le format XML se compose de plusieurs formats qui à chacun assignent une tâche particulière :

1) Les formats techniques et standards de support

- La DTD et le schéma pour définir la grammaire de document.
- Le CSS et XSL ou XSLT deux fichiers de feuilles des styles; la deuxième est plus complexe que la première qui permet de construire un tableau par exemple.

- Xpath et Xlink pour faire passer le document.

2) Les formats d'encodage horizontal

Les formats d'encodage horizontal basés sur XML sont les suivants :

- XHTML
- XSL-FO
- SVS
- MATHML

3) Le format d'encodage vertical

Le format d'encodage vertical basé sur XML est le suivant :

- Schéma/DTD à valeur sémantique métier

1.5.5.4. Les éditeurs XML existant dans le marché

Plusieurs éditeurs XML, permettant à l'utilisateur d'écrire un document valide, existent sur le marché tels que :

- SOFTQUAD XMETAL:
- FM+SGML
- Plugin Word

- Emax
- XML Notepad
- XML spy : nous avons utilisé XML spy pour définir le DTD pour les manuscrits arabes et pour le saisie de l'information qui concerne ces manuscrits.

1.5.5.5. Convertisseurs XML

De nombreux documents sont déjà créés sur format Word au PDF qui ont besoin d'être convertis vers le XML. Pour convertir un document Word vers un document XML, il faut créer une *macro* avec la DTD déjà créé afin de baliser le document pour le préparer à être converti. Des logiciels sont déjà réalisés pour cette raison comme UPCAST qui est un convertisseur vers XML.

1.5.5.6. XML et la description de manuscrits arabes.

Au chapitre, *3.2 page 181,* sera détaillée l'utilisation de XML dans le cas des manuscrits arabes.

1.6. Choix des matériels et logiciels

Le choix du matériel dépend du type de document et de sa condition physique (bonne, moyenne, mauvaise, détériorée, etc.). Si la collection de documents à la main est très fragile, la numérisation par l'utilisation d'une caméra numérique avec un « scanner à plat» sera plus adéquat. La caméra numérique est un moyen indépendant pour capturer une qualité d'image numérique très performante. Ce genre de caméra est adapté à la numérisation des manuscrits ou des livres anciens, en raison de sa flexibilité. Du fait de leur fragilité, il est toujours difficile de numériser ces documents par l'utilisation des «scanner à plat»[46].

[46] Robinson, P. The digitisation of primary textual sources. Oxford: Office for Humanities Communication Publications, 1993. P.39.

Le choix du logiciel : il est très difficile de recommander un logiciel ou un matériel standard car il y a toujours des nouveautés dans ce domaine. Les logiciels existant sur le marché ont des capacités variables.

Néanmoins, on trouve que la plupart des projets de numérisation utilise le même logiciel pour certaines applications, notamment en ce qui concerne la numérisation du texte et la manipulation des images. Un exemple d'un logiciel associé à la numérisation de texte est Reconnaissance Optique de Caractères (*Optical Character Recognition* (OCR)) qui est souvent livré avec le numériseur. Son but principal est à la fois de reconnaître le caractère du document déjà numérisé et de passer du mode image au mode texte. Ceci donnera de très bons résultats avec une page imprimée récemment dans une langue donnée mais quand il s'agit d'un document manuscrit dans n'importe quelle langue, les résultats obtenus sont souvent très mauvais.

En ce qui concerne la base de données et sa diffusion sur le web, nous avons utilisé le logiciel SDX (**S**ystème **D**ocumentaire en **X**ML). SDX est un outil de recherche basé sur XML, composé de plusieurs logiciels qui permettent de *publier,* d'*indexer* et de *rechercher* des documents XML.

1.7. Les projets de numérisation

Notre but, dans cette partie, est de montrer les projets de numérisation des manuscrits déjà réalisés. Les projets de numérisation qui seront mentionnés dans ce chapitre, ont été précédés d'autres initiatives d'informatisation, en particulier la mise en oeuvre de catalogues électroniques de manuscrits. Les projets de numérisation des manuscrits ont tous tenté de créer un modèle de structuration standard international.

Ces projets peuvent être classés en deux catégories :

- La première est concernée par l'accès aux manuscrits écrits en langue latine. Citons comme exemple les projets EAMMS*, MASTER*, etc.

* Electronic Access to Medieval Manuscripts

* Manuscript Access through Standards for Electronic Records

- La deuxième traite de la numérisation des imprimés du XVIème siècle comme le projet européen DEBORA[*].

L'objectif principal de ces types de projets est la création des métadonnées comme moyens d'accès spécifiques à ces manuscrits.

1.7.1. Le projet DEBORA

***D**igital acc**E**ss to **BO**oks of the **R**en**A**issance* Accès numérique à des livres de la renaissance « Il s'agit d'un projet européen numéro (LB 5608/A) dont l'objectif est de concevoir un ensemble d'outils permettant l'accès à distance et collectif à des livres numérisés du 16ème siècle, sans passer obligatoirement par les bibliothèques dépositaires des originaux »[47].

« Le choix du 16ème siècle présente l'intérêt de se situer à un moment où le livre imprimé acquiert ses caractéristiques modernes (apparition et structuration de la page de titre, organisation interne, normalisation de la typographie, etc.) tout en conservant un mode de production artisanal qui durera les deux siècles suivants»[48].

Une centaine de livres a été choisie auprès des bibliothèques partenaires de projet, selon les critères suivants :

- Des critères culturels : il s'agit de l'intérêt de l'ouvrage quant à son contenu et à sa forme.
- Des critères d'usages : l'intérêt de l'ouvrage pour différents usages.
- Des critères de technique documentaire : l'intérêt de l'ouvrage quant à ses caractéristiques techniques.
- Des critères économiques : l'intérêt de l'ouvrage quant à son impact économique.
- Des critères de techniques pour étudier les contraintes de la numérisation.

---

[*] **D**igital acc**E**ss to **BO**oks of the **R**en**A**issance

[47] DEBORA : projet européen n°. LB 5608 A. Coordinateur R. Bouché, juin 2000. 179pages. (P.7)

[48] ibid.

1.7.1.1. Les partenaires du projet DEBORA

La composition du partenariat a consisté, d'une part, en bibliothèques possédant des collections importantes du 16ème siècle et, d'autre part, en laboratoires et équipes de recherche. Ces bibliothèques joueront un rôle important en ce qui concerne le choix des documents à numériser, l'harmonisation des moyens d'accès, la validation de la numérisation et l'évaluation des usages.

Les bibliothèques qui sont concernées par la description des documents sont les suivantes :

- Biblioteca Casannatense, Rome (Italie) qui a été choisie grâce à son fonds très riche de manuscrits de la Renaissance.
- Biblioteca Geral, Universidade do Coimbra, (Portugal), très célèbre par son fonds de manuscrits portant sur les grands navigateurs.
- Bibliothèque municipale de Lyon (France) : le fonds trouvé au sein de sa bibliothèque municipale est un témoin de l'époque du XVIème siècle où Lyon a été une ville où l'activité des imprimeurs a été très importante.

Les équipes de recherche et les laboratoires ayant fait partie de ce projet

- Département informatique de Lancaster University, (Grande-Bretagne) : spécialiste des outils de travail qui permettent la collaboration entre plusieurs équipes.
- Les équipes de recherche ERSICO, Université Lyon 3 (France) et LIRE, Institut des Sciences de l'Homme, Lyon (France) : concernées par l'analyse des besoins et des usages, par l'évaluation des systèmes et par l'analyse des coûts.
- Le laboratoire de Reconnaissance des Formes et Vision de l'INSA de Lyon (France) et l'Instituto Superior Tecnico, Lisbonne (Portugal) : leur coopération porte sur le développement des processus de traitement, de compression et d'indexation de l'image.
- SII-ENSSIB, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques, Villeurbanne (France) qui travaille sur les interfaces.

Les partenaires industriels associés au projet

- SGBI Enterprise, Lyon (France)

- XEROX, Grenoble (France) en association avec IIS, Bordeaux (France).

D'autres bibliothèques ont été intéressées à servir de site-test à DEBORA et à participer à son évaluation :

- Universidad de Zaragoza Biblioteca, Espana (Francisco Javier Garcia Marco)
- Bibliothèque Municipale et Universitaire de Genève, Suisse (Alain Jaquesson)

Certains de ces partenaires ont des expériences acquises grâce à leur participation à des programmes européens déjà existants dans ce domaine.

1.7.1.2. es éléments principaux du projet sont les suivants[49] :

- Une analyse des besoins et des attentes des usagers du livre du 16ème siècle, afin de leur offrir une station de travail correspondant, au mieux, à un usage distant et collaboratif via l'internet.
- La définition de la chaîne de production de documents numérisés et de ses spécificités techniques concernant les algorithmes et les procédés de traitement et de compression des images.
- La conception et la réalisation d'un ensemble de serveurs et de postes du travail (le système DEBORA). Cet ensemble doit permettre la gestion des collections numérisées (sous la forme d'un ensemble structuré d'images) et toutes les opérations nécessaires depuis l'intégration et l'indexation, jusqu'à la recherche d'information et un travail collectif.
- Une analyse économique des coûts de la numérisation.

1.7.1.3. Le mode de numérisation utilisé pour le projet

Le projet DEBORA a utilisé le mode image pour réaliser la numérisation des documents. Le choix du mode image a été adopté pour les raisons suivantes :

- Pour être plus fidèle aux documents originaux.
- Pour des raisons de production : les ouvrages de XVIème siècle ont comme caractéristiques une très grande variété de formes et d'information (l'utilisation de plusieurs langues dans le même manuscrit, l'état actuel des documents, etc.),

[49] ibid.

ce qui fait que la procédure de reconnaissance optique des caractères est très difficile à réaliser.

1.7.2. Le projet MASTER

Un projet européen en contrepartie des projets nord-américains existants. MASTER a été financé par le quatrième programme-cadre de recherche (1994-1996) de la Commission Européenne. Ce projet s'inscrit dans le contexte de mise en ligne de corpus numérisés et de mise en réseau des services des bibliothèques. L'idée de ce projet est née en novembre 1996 lors d'un séminaire d'inauguration qui a réuni environ trente experts européens et nord-américains à Studly Priory près d'Oxford mais la réalisation du projet a commencé le 1er janvier 1999. MASTER a pour but de « créer un système générique, suffisamment souple et robuste pour permettre son application dans différents domaines de la description du manuscrit et la technique choisie atteindre ce but plus ambitieux et fondé sur les normes international SGML et XML »[50]. Un groupe de travail dirigé par TEI, et dirigé par Consuelo Dutschke (Université de Columbia) et Ambrugio Piazzoni (Bibliothèque de Vaticane) a été chargé de développer une DTD (définition de type Document) spéciale pour la description de manuscrit qui serait compatible avec l'encodage concernant le texte électronique crée par la TEI (Texte Encoding Entiative).

1.7.2.1. Les objectifs de projet MASTER

Le projet vise à définir et à établir les fonctionnalités suivantes :

- De développer un standard accepté à l'échelle internationale, basé sur la TEI initiative qui a été mise en œuvre par SGML et qui concerne les notices de manuscrit.
- Développer des logiciels avec un format standard pour permettre aux bibliothèques à faibles ressources de construire elles-mêmes des notices simples (premier niveau) pour leurs manuscrits.

[50] Burnard, Lou et Peter Robinson. Vers un standard européen de description des manuscrits : le projet MASTER. Document numérique. Vol.3, N°.1-2/1999. Pages 151-169. (P.153).

- Publier sur l'Internet le prototype d'un catalogue collectif qui décrit en-ligne cinq mille notices de manuscrits. Ce catalogue est accompagné par des logiciels spécialisés pour la recherche qui ont des liens avec des images de manuscrits et de texte intégral. La mise en place d'un tel catalogue a pour but principal d'aider à la distribution des manuscrits au plus grand nombre d'utilisateurs possible.
- L'encodage standard crée par MASTER vise à faciliter l'échange des informations d'un ordinateur à l'autre qui utilise des formats différents.
- L'objectif des notices standardisées est d'établir le lien entre les descriptions d'un côté et les images de manuscrits de l'autre. Donc le catalogue électronique sera la première entrée pour la recherche de manuscrits.

1.7.2.2. Les partenaires de projet

MASTER est un consortium de bibliothèques et d'archives qui contient des centaines de milliers de manuscrits et dispose d'experts dans le domaine de l'encodage de texte (*Text Encoding*) et dans celui des bibliothèques numériques. Les partenaires sont répartis en deux catégories :

- Les partenaires financiers:
    - La Bibliothèque Royale des Pays-Bas (La Haye),
    - La Bibliothèque d'Hague,
    - La Bibliothèque Nationale Tchèque (Prague),
    - L'Institut de Recherche et l'Histoire de Textes (Paris).
- Les partenaires non-financiers sont :
    - La Bibliothèque Vaticane,
    - La Biblioteca Ambrosiana,
    - La British Library et la Bodleian Library (Oxford).

Au niveau technique le projet est étroitement lié aux projets des manuscrits nord-américains, EAMMS: Electronic Access to Medieval Manuscripts piloté par Université de Columbia, à Digital Scriptorium, ainsi qu'au Text Encoding Initiative TEI. EAMMS est un projet financé conjointement par les Etats-Unis d'Amérique et l'Union Européenne.

L'émergence de nombreux projets, à partir de l'année 1996, a favorisé l'apparition des descriptions alternatives de celles qui ont déjà été utilisées depuis les années 1980, dans la base de données. Le langage de balise SGML• a été utilisé pour l'encodage de texte électronique, alors que le TEI, depuis 1988, a créé un ensemble de liens directifs pour l'encodage d'un grand nombre de documents dans le domaine des sciences humaines.

Mais il reste un problème relatif à la compatibilité d'encodage créée par cette nouvelle technologie avec celle des systèmes de base de données traditionnelles. Pour résoudre ce problème, les projets EAMMS et MASTER «ont choisi de conserver deux systèmes de base de données, l'un (par exemple) utilisant le format MARC et l'autre le XML »[51].

EAMMS qui, depuis le début de son existence, vise à garder les deux structures (la base de données et la SGML/XML) a utilisé le format MARC comme base pour les descriptions de ces manuscrits. Cependant les encodages SGML/XML sont importants quand il s'agit de mettre les manuscrits sur les pages Web. La diffusion sur Internet exige des notices encodées avec une structure très simple pour accompagner les images manuscrites. De même, les descriptions les plus complexes des manuscrits peuvent aussi être hébergés dans l'encodage SGML/XML sans aucune crainte d'avoir des descriptions déformantes. Le développement de format SGML/XML comme format d'échange international permet le transfert de plusieurs notices de manuscrits, se trouvant dans les différentes bases de données, vers un format unique et plus malléable.

1.7.3. Le projet EAMMS

EAMMS (Electronic Access to Medieval Manuscripts) est un projet de trois ans financé par la fondation Andrew W. Mellon. Comme son nom l'indique, le projet a pour but de développer la recommandation pour l'encodage et le stockage électronique des descriptions de manuscrits médiévaux et de la Renaissance. Le projet vise à exploiter le développement de la nouvelle technologie pour fournir un moyen d'accès rapide et plus compréhensif aux informations trouvées à l'intérieur des manuscrits. Selon EAMMS, les avantages de ce moyen d'accès (catalogue électronique) est de permettre l'accès à

• SGML : 1986, HTML : 1992, XML : 1998

[51] Burnard, Lou et Peter Robinson. Vers un standard européen de description des manuscrits : le projet MASTER. Document numérique. Vol.3, N°.1-2/1999. Pages 151-169. p. 154.

distance, de faire le lien entre les images numérisées et le texte ; de faciliter la recherche qui est dès lors rapide et demande moins d'efforts qu'avec le volume imprimé : « *They allow remote access, can be linked to digitalized images and transcriptions of text, and can be searched more quickly and with less effort than printed volumes* »[52]

1.7.3.1. Les objectifs du projet EAMMS

Les objectifs principaux du projet EAMMS sont les suivants:

- Déterminer la norme d'encodage pour stocker les informations dans une base de données relationnelle avec des notices abrégées.
- Identifier et classer les informations trouvées dans un catalogue de notices complètes. EAMMS vise, également, à déterminer la façon d'encoder ces informations.
- Tester et raffiner la recommandation d'encodage par la création de base de données d'un catalogue électronique avec des notices encodées selon les recommandations et de fournir pour cette base de données un accès en-ligne.

1.7.3.2. La méthodologie suivie par le projet EAMMS

Le projet EAMMS utilise deux formats pour encoder ses manuscrits:

- La première est le format MARC (Machine-Readable Cataloging) : le format MARC a été désigné pour servir comme base de données bibliographiques. Mais ce format ne répond pas bien au besoin de l'encodage des manuscrits médiévaux.
- Le deuxième est le format SGML : pour facilité la diffusion de catalogues sur l'Internet et pour établir tous les types de lien hypertextuel.

Des logiciels spécialisés ont étés utilisés pour indexer et maintenir les données encodées par le SGML ; pour convertir le SGML de HTML dans le but de l'afficher sur le Web et

---

[52] Electronic Access to Medieval Manuscripts http://www.hmml.org/eamms/index.html (last revised August 02, 2000) 09/10/2000 4 pages.

pour publier les documents, encodés par SGML, comme un texte imprimé ou électronique sur format CD-ROM.

1.7.3.3. Les partenaires de EAMMS

- La bibliothèque de film du Vatican : sous sa direction, une équipe d'experts en format MARC a été organisée pour examiner les étiquettes et les protocoles déjà existants et pour recommander des modifications sur le format afin de bien accommoder le modèle du contenu *(content model)* des notices de premier niveau et des notices détaillées.
- La Bibliothèque du Congrès : le modèle mentionné ci-dessus est testé à la Bibliothèque du Vatican et révisé au sein de la Bibliothèque du Congrès, par le Comité Bibliographique Standard des Livres Rares et par la section des manuscrits de *l'Association of College and Research Libraries*. Selon l'article, la version finale des notices MARC relative au projet EAMMS a été publiée en juin 1999.
- La Bibliothèque de Hill Monastic Manuscript Library est l'endroit où l'encodage de SGML a été testé. La bibliothèque a créé une base de données des notices encodées pour sa collection de microfilms accessible par l'Internet.

1.8. Les métadonnées utilisés par les projets de numérisation

1.8.1. Le TEI/MASTER DTD

Dans cette partie nous voudrions présenter à la fois les métadonnées et la DTD créées par la TEI ainsi que celle ajouté par MASTER. Comme nous l'avons déjà mentionné ci-dessus, les objectifs de MASTER vise à créer, pour les manuscrits, des descriptions standardisés, plus flexible et plus professionnelles. Pour atteindre ce but, on a proposé deux types de notices :

- Un inventaire simple : il contient seulement de simples identificateurs des manuscrits avec des petites phrases décrivant les manuscrits ou les images.
- Une description plus complexe de manuscrit : notice avec une structure de balisage plus détaillée qui contient une transcription complète des manuscrits.

Quelle que soit la présentation générale de notice (simple ou plus complexe), il doit contenir l'élément « manuscrit description » encodé comme <m*sDescription*>. Chaque élément <m*sDescription*> doit avoir seulement une seule description. Ppar exemple, un volume qui est composé de plusieurs manuscrits est traité comme une seule entité dans l'élément <m*sDescription*>. Par contre, on utilise l'élément <*msPart*> à l'intérieur de l'élément <m*sDescription*> pour désigner qu'un tel manuscrit est composé de différentes parties et que chaque partie est conservée dans un endroit différent.

Dans MASTER, on trouve des descriptions spécifiques pour des tâches différentes :

- L'élément général de la description se trouve dans l'élément <msDescription> pour décrire le manuscrit.
- L'élément spécifique de la description pour décrire la phrase en détail comme le mot abrégé, le mot additionnel de texte original, la partie de texte endommagé, l'illustration trouvée dans le texte etc.

1.8.1.1. Les éléments composants de <msDescription>

Le <msDescription> contient des éléments de description pour un seul manuscrit ainsi que pour des parties de manuscrits. Le tableau ci-dessous montre bien les éléments des descriptions de manuscrit au niveau général[53] :

Les éléments trouvés dans le <MsDescription>	Le rôle des éléments.
<msIdentifier>	Il contient les éléments à partir desquels on peut identifier le manuscrit, comme le lieu, la cote etc.
<msHeading>	Il donne les informations principales sur les manuscrits comme celles qui apparaissent dans le simple catalogue (par exemple l'auteur, le titre, la date et le lieu d'achèvement du travail, et la langue du document)
<msSummary>	L'élément <msSummary> a été supprimé et amalgamé avec l'élément <msHeading> : une proposition suggérée par les TEI workgroup en septembre 1999.
<msContents>	Il contient des descriptions liées au contenu intellectuel du manuscrit soit par une série de paragraphes soit structurées par des sous-éléments bien définis.
<physiDesc>	Il contient les descriptions des aspects physiques des manuscrits structurés en sous-éléments bien définis.
<history>	Il contient des informations concernant l'histoire du manuscrit structurées en sous-éléments bien défini.
<additional>	Il contient des informations additionnelles reliées au manuscrit mais ne fait

[53] MASTER: A gentle introduction. http://www.cta.dmu.ac.uk/projects/master/gentintr.html. 11pages dernière révision le 14 Janvier 2001 (5/03/2001) p.8.

	pas partie du manuscrit lui-même, comme la bibliographie, les informations mises par les conservateurs des bibliothèques, et la disponibilité de manuscrits sur microfilm etc.
<msPart>	Utilisé pour les manuscrits composés de plusieurs parties. Il peut contenir tous les éléments mentionnés ci-dessus sauf le <summary>,car le <summary>est destiné uniquement à un manuscrit complet et non à un manuscrit composé de plusieurs manuscrits.

Tableau n°. 6 : Les metadonés de projet MASTER : Les éléments compsants de <MsDescription

L'élément <msDescription> contient aussi d'autres éléments qui servent à des objectifs spécifiques. L'attribut *status,* par exemple, définit la situation de la composition du manuscrit ou des parties du manuscrit (une unité complète, une unité composée de fragments ou de groupes isolés de fragments) qui contient les valeurs suivantes :

- *uni* (unitaire) : le manuscrit est une entité complète existant en une seule unité.
- *compo* (composite) : le manuscrit est une entité complète qui comprend plusieurs unité d'origines différentes.
- *frage* (fragmentary) : une feuille, une partie de feuille ou un manuscrit auquel manque la majorité des feuilles.
- *def* (defective) : un manuscrit auquel manque un petit nombre de feuilles.
- *unknown* : non connu, non mentionné.
- *type* : spécifie le type de manuscrit décrit, par exemple : « diplôma », codex, etc.
- Values : le type qui sera défini

1.8.1.2. Les éléments de description au niveau du paragraphe

La TEI crée des éléments de description standard pour décrire en détail des valeurs trouvées dans un paragraphe <p>. Parmi la liste trouvée dans le tableau ci-dessous, il y a des éléments TEI spécifiques à la transcription des manuscrits qui n'existent pas dans toute la TEI DTD. Le projet MASTER à son tour a aussi ajouté d'autres éléments qui correspondent à la spécificité d'un paragraphe du manuscrit. Le tableau ci-dessous contient tous les éléments standards détaillés, faits par la TEI et le projet MASTER[54] :

Les TEI éléments décrivant le paragraphe	
<abbr>	Il contient des abréviations de toutes sortes.
<add> (addition)	Il contient des lettres, des mots ou des phrases insérés dans le texte par l'auteur, par le copiste, par l'annotateur ou par le correcteur.
<addSpan>	Il marque le début des séries du texte ajouté par l'auteur, le copiste, l'annotateur

[54] http://hcu.ox.ac.uk/TEI/Master/Reference/ms.html

(added span of text)	ou le correcteur.
<bibl> (bibliographic citation)	Il contient une structure approximative de la citation bibliographique dans la quelle les sub-composants sont explicitement balisés ou non balisés
<corr> (coorection)	Il contient la forme correcte d'un passage dans le texte qui est apparemment erroné.
<damage>	Il contient une partie endommagée du texte.
<date>	Il contient une date dans n'importe quel format.
<dateRange>	Il contient deux dates ou une phrase qui délimite une période
<del> (deletion)	Il contient des lettres, des mots ou un passage du texte marqué par l'auteur, le copiste, l'annotateur ou le correcteur, comme : supprimé, inutile ou *spurious*.
<delSpan> (deleted span of text)	Il marque le début d'une longue série du texte supprimé qui a été marquée par l'auteur, le copiste, l'annotateur ou le correcteur comme : supprimé, inutile ou *spurious*.
<expan> (expansion)	Il contient l'annexe d'une abréviation.
<figure>	Il localise le graphique, l'illustration ou la figure dans le texte.
<foreign>	Il identifie les mots ou les phrases écrits dans une autre langue que celle du texte.
<formula>	Il contient des formules mathématiques ou d'autres.
<fw> (frame work)	Il contient un "en-tête"ou "en-pied" de page, « catchwords » ou autres, trouvés dans la page.
<gap> (omitted material)	Il indique les points qui manquent dans la transcription, soit pour des raisons éditoriales décrites dans l'entête « header » de TEI, ou parce que le matériel est illisible ou inaudible
<gloss>	Il identifie la phrase ou le mot utilisé pour fournir une annotation ou une définition de quelques mots ou phrase.
<handshift>	Il marque le début d'une série de texte écrit avec une autre main ou un changement dans le style d'écriture, d'encre etc.
<hi> (highlight)	Il marque un mot ou une phrase qui est distingué par son écriture du reste du texte.
<label>	Contient l'étiquette associée avec un article dans la liste. Dans les glossaires, il marque le terme qui doit être défini.
<list>	Contient n'importe quelle séquence d'article organisée dans une liste.
<listBibl> (citation list)	Contient une liste des citations bibliographiques de n'importe quel genre.
<name> (name propre noun)	Contient un nom propre ou (nom phrase)
<note>	Contient une note ou une annotation.
<num> (nombre)	Contient un nombre écrit dans n'importe quelle forme
<orig>(original form)	Contient la forme originale d'une lecture, une forme régularisée est mentionnée dans l'attribut *valeur*.
<ptr>	Définit un pointer pour une autre localisation dans le document en cours, qui prend la forme d'un ou plusieurs éléments identifiants.
<q> (quoted speech or thought)	Contient des cotations ou cotations apparentes.
<ref>	Définit une référence par un ou plusieurs éléments identifiables pour une autre localisation dans le document qui peuvent être modifiés par le texte ou le commentaire additionnel.
<reg> (regularisation)	Contient une lecture «reading » normalisée ou régularisée en quelque sorte.
<restore>	Indique la restauration du texte par l'annulation des marques éditoriales ou marques faites par l'auteur.
<sic>	Contient le texte déjà reproduit, qu'il apparaisse correct ou incorrect.
<space>	Indique la localisation d'un espace signifiant dans le texte.
<supplied>	Contient le texte fourni par le copiste ou l'éditeur à la place des écritures illisibles et endommagées.
<table>	Contient le texte présenté dans une forme de tableau (colonnes et lignes).

<term>	Contient un et/ou plusieurs mots ou représentations symboliques considérés comme des termes techniques.
<text>	Contient un seul texte de n'importe quel genre (unitaire, ou composite) par exemple, un poème ou drame, une collection d'essais, une nouvelle, un dictionnaire ou un simple corpus.
<title>	Contient un titre d'article, d'ouvrage, de journal ou de série qui contient aussi des titres ou sub-titres alternatifs.
<unclear>	Contient un mot, une phrase ou un passage difficile à transcrire car il est illisible et inaudible.
Tableau n°.7 : Les TEI éléments décrivant le paragraphe	

TEI était modulé pour servir le projet MASTER, mais MASTER DTD a définit d'autres éléments spécifiques à la description d'un paragraphe manuscrit. On peut dire que MASTER est beaucoup plus détaillé que TEI. Les éléments mentionnés dans le tableau ci-dessous sont les contributions de MASTER DTD[55]:

<catchwords>	Contient le « catchwords » trouvés dans les manuscrits.
<dimensions>	Contient n'importe quelle spécificité de dimensions, utilisée pour spécifier la mesure du document.
<heraldry>	Contient une formule héraldique ou une phrase rencontrée dans les composants de blason, d'armoiries, etc. trouvées dans le manuscrit.
<locus>	Définit un endroit à l'intérieur d'un manuscrit où on trouve une séquence ou une interruption dans le folio.
<material>	Il contient une phrase qui décrit le type de matériel, utilisé dans la fabrication des parties du manuscrit, et la reliure.
<msIdentifier>	Il contient les informations dont on a besoin pour identifier un manuscrit ou des parties de manuscrit uniquement à l'intérieur de l'institution.
<origDate>	Il contient n'importe quelle forme de datation utilisée pour identifier la date d'origine d'un manuscrit ou des parties du manuscrit.
<origPlace>	Il contient le nom de l'endroit utilisé pour identifier le lieu d'origine d'un manuscrit ou des parties du manuscrit fait dans n'importe quelle forme.
<secfol>	Le mot/mots utilisés par un catalogueur (médiéval ou moderne) pour indiquer le début d'un point fixe dans le codex. Par exemple : le début du deuxième folio ou la fin de l'avant dernier folio (pénultième) etc.
<signatures>	Il contient la discussion sur les signatures du folio ou les feuillets trouvés dans le manuscrit.
Tableau n°.8 : Les DTD ajoutés par le projet MASTER	

Le contrôle d'autorité.

MASTER utilise par défaut la Liste d'Autorité de la Bibliothèque du Congrès Américaine pour le nom, le lieu et le titre uniforme quant il y a un manque dans sa propre liste d'autorité.

[55] Brunard Lou Editeur. Reference manual for the MASTER document type definition discussion draft. Revised 6 January 2001. (12/03/2001) http://www.hcu.ox.ac.uk/TEI/Master/Reference/

1.8.2. Les métadonnées créées par le projet DEBORA

Les métadonnées créées par le projet et destinées aux documents du XVIème siècle occupent trois niveaux principaux ; le quatrième a été réservé à l'ajout des utilisateurs :

a) Les métadonnées du 1er niveau, contiennent les éléments suivants

Anglais	Francaise	Expliquation
Auteur	Author	La personne ou l'organisation responsable principale de la création du contenu intellectuel du document
Titre	Title	Nom donné au document par l'auteur ou le libraire.
Lieu de publication	Place of publication	Nom de la ville où a eu lieu la publication.
Date de publication	Publication date	Date à laquelle le document a été mis à disposition dans sa forme actuelle.
Imprimeur-libraire	Publisher	L'entité qui a mis à disposition du public le document dans sa forme actuelle.
Langue	Language	Langue du texte du document
Collation	Collation	Description physique du livre qui détaille la pagination, la numérisation des feuillets, les signatures, le format et les planches, les illustrations, Elle permet ainsi de s'assurer qu'il est bien complet.
Cote et localisation	Call Number	Identificateur dans la classification locale de la bibliothèque où se trouve l'ouvrage.
Empreintes	Fingerprint	Suite normalisée de caractères du livre permettant d'identifier une édition par référence à une base-source.
Notes	Notes	Toutes les remarques, que ce soit sur l'édition, sur les caractéristiques de l'exemplaire ou sur le contenu.
Auteurs secondaires	Secondary author	Tous les auteurs en dehors de l'auteur principal, préfacier, collaborateur, traducteur, etc.
Sujet	Subject	Le thème du document, exprimé habituellement par des mots-clés ou des phrases qui décrivent l'objet ou le contenu du document.

Tableau n°.9 : Métadonnées du projet DEBORA (Niveau 1)

b) Les métadonnées du 2ème niveau

Ils sont des éléments de structure interne d'un livre ancien constituant les éléments suivants :

Métadonnées du 2ème niveau		
Française	**Anglaise**	**Explication**
Les pages de titre	Title page	Souvent une seule page, parfois deux ou plus, les pages de titre comprennent généralement trois types d'information : le titre et le sous-titre, la mention de responsabilité (l'auteur), et l'adresse (nom du libraire, lieu d'édition et marque typographique du libraire ou de l'imprimeur ainsi que la date d'édition).
Le frontispice	Frontispiece	Gravure décorative placée en tête d'un livre, face au titre. Son usage est fréquent aux XVIe, XVIIe et XVIII siècles. Il a pris dans le livre imprimé des formes variées : portrait, allégorie, résumé en une image de l'œuvre entière ou sa représentation sous forme baroque, symbolique, allégorique, etc. Il ne faut pas le confondre avec le titre gravé qui, lui, contient des indications bibliographiques.
Les	Preliminary pages	Textes variés qui autrefois précédaient le texte principal du livre comme :

pièces liminaires	pages	Approbation	Approval	Accord donné pour l'impression d'un texte par les autorités civiles ou religieuses. Figurant en tête ou à la fin d'un volume, l'approbation mentionne quelquefois le nom de l'auteur qui peut ne pas figurer sur la page de titre.
		Avertissement	Foreword	Préface de peu d'étendue, écrite en général par l'auteur ou quelquefois par le libraire pour attirer l'attention du lecteur sur un point particulier.
		Avis au lecteur	Announcement	Texte contenant une explication ou un avertissement en début d'un livre, fait par l'auteur, le libraire ou le traducteur, et souvent présenté en italique pour attirer l'attention.
		Avis au relieur	Notice	Feuillet supplémentaire (isolé ou appartenant à un cahier) généralement placé à la fin d'un volume, destiné à faciliter le travail du relieur.
		Dédicace	Dedication	Texte en début d'ouvrage exprimant l'hommage que fait un auteur de son oeuvre à quelqu'un : comporte le nom du dédicataire, le texte, puis le nom de l'auteur de la dédicace, parfois le lieu et la date de la dédicace.
		Epitaphe	Epitaph	Se dit de certains éloges en prose ou en vers, à l'honneur d'un défunt pour en conserver la mémoire
		Floraison	Flowering	Petit texte sur l'ouvrage ou sur l'histoire littéraire
		Pièce de vers	Homage	Hommage à une personne vivante
		Portrait	Portrait	Représentation de l'auteur ou du dédicataire ou du héros.
		Préface	Preface	Texte de présentation en tête d'un livre écrit parfois par l'auteur pour définir son dessein. Il peut aller jusqu'à prendre la forme d'un manifeste. Mais, le plus souvent, la préface est écrite par une personnalité ayant autorité pour porter un jugement. Elle synthétise le contenu du livre.
		Privilège	Privilege	Texte faisant état de la permission et de la propriété des droits d'imprimer. Cette permission est attribuée par une autorité pour une période donnée à un personnage donné (typographe, libraire, auteur) parfois avec mention de rénovation du privilège et/ou son partage. Il peut aussi se trouver en fin de volume.
		Prologue	Prologue	Court avertissement en tête d'un livre.
Les pages de texte	The text	Ensemble des pages qui forment le contenu textuel du livre sous forme de chapitres ou de volumes.		
Le hors texte	Outside text	Toute page, ne faisant pas partie intrinsèque du corpus d'un livre, intercalée entre les pages du livre à des fins d'illustration ou de documentation et généralement non foliotée. Elle peut être une planche, un plan, une gravure, une carte repliée, une illustration collée		
Les index	Index	Liste alphabétique des noms, annotations ou sujets, etc. cités dans un ouvrage, avec un numéro de page. Elle peut prendre place au début ou à la fin d'un livre.		
Les tables	Tables	Récapitulation des matières traitées dans un ouvrage sous forme de répertoire qu'on met à la fin ou parfois au commencement d'un livre, pour permettre au lecteur de se repérer plus facilement. Dans les documents anciens, les tables peuvent porter sur n'importe quel élément et sont par conséquent très diversifiées :		
		Table des auteurs	Table of authors	
		Table des cartes	Table of maps	
		Table des chapitres	Table of chapters	
		Table des choses	Table of things	
		Table des citations	Table of citation	
		Table des épigrammes	Table of epigram	
		Table des livres	Table of books	
		Table des matières	Tables of content	
		Table des noms	Table of names	
		Table des oeuvres	Tables of works	
		Table des sommaires	Tables of summary	
		Textes variés qui se rencontrent dans les pages se trouvant vers la fin de l'ouvrage :		

		Textes variés qui se rencontrent dans les pages se trouvant vers la fin de l'ouvrage :		
		Catalogue	Catalogue	Liste de livres écrits par le même auteur ou édités par les mêmes imprimeurs. Figurant en tête ou à la fin d'un volume, l'approbation, du monde l'imprimeur et la date de l'impression.
		Colophon	Colophon	Indication, souvent, du nom de l'auteur et la date de l'impression qui peut ne pas figurer sur la page de titre.
		Errata	Errata	Erreurs à corriger
		Explicit	Explicit	Rappel du titre de l'ouvrage et de l'auteur
		Hors-texte	Outside text	Elément non compris dans les signatures ou la pagination ; peut être dépliant.
		Marques typographiques	Typographic marks	Marque du libraire ou de l'imprimeur
		Privilège	Privilege	Texte faisant état de la permission et de la propriété des droits d'imprimer. Cette permission est attribuée par une autorité pour une période donnée à un personnage donné (typographe, libraire, auteur) parfois avec mention de rénovation du privilège et/ou et son partage. Il peut aussi se trouver en début du volume.
		Registre	Register	Table des signatures (indication d'assemblage des feuillets)
		Autres textes variés d'auteurs variés	Other various texts for various authors	(excipit, quatrain, élégie, poésies diverses, etc.)
Tableau n°.10 : Métadonnées du projet DEBORA (Niveau 2)				

c) Les métadonnées du 3ème niveau

Il s'agit de la description des caractéristiques particulières des pages. Des procédures de reconnaissance de formes permettent d'identifier certains éléments, surtout ceux à caractères graphiques (bandeaux, lettrines, foliation, autres types d'illustration etc.). Cette « indexation » peut être faite automatiquement ou non. Les métadonnées de niveau 3 sont des éléments de base pour aider les chercheurs (surtout les « seizièmistes ») à étudier ces livres du XVIème siècle et à identifier, avec une certaine précision, la période de leur impression. Une page d'un livre du XVIème siècle peut contenir un ou plusieurs éléments de la liste suivante :

Les métadonnées du 3ème niveau		
Française	Anglais	Explication
Bandeau	Bandeau	C'est un élément décoratif non figuratif en forme de bande, qui orne le haut des pages. Il sert à séparer les chapitres et à marquer le début des paragraphes.
Colonnes	Columns	Une ou plusieurs colonnes
Cul-de Lampe	Vignette	Motif décoratif ou figure gravée ou typographique centrée en fin de chapitre.
Enluminure	Illuminator	C'est une lettre ornée et de petite dimension illustrant les feuillets d'un livre ou d'un manuscrit.
Estampe	Print	C'est toute image réalisée au moyen d'un élément d'impression (en creux, en relief, etc.)
Filet	Filet	Trait imprimé, simple ou décoratif
Fleurons	Fleuron	Eléments décoratifs utilisés pour signaler la fin des chapitres. Ils peuvent également être sur la page de titre à la place de la marque.
Foliation	Foliation	Numérotation des feuillets d'un livre. Seul le recto du feuillet est numéroté.
Frontispice	Frontispiece	Illustration ou une gravure qu'on trouve généralement en regard de la page de titre.
Garnitures	Decoration	Bandes de diverses largeurs servant à l'espacement en impression typographique.

Illustration	Illustration	Images, gravures, cartes, médaillons, etc.
Incipit	Incipit	Terme qui désigne les premiers mots d'un manuscrit ou d'un livre.
Lettre d'attente	Temporary lettre	Lettre minuscule mise en attente dans un manuscrit pour que le rubricateur dessine une initiale plus importante.
Lettrine	Dropped initial	Une lettre en grande capitale souvent ornée en début de paragraphe ou de chapitre.
Lettre de pied	Title abridger	Abrégé du titre, souvent de type Tome II
Manchette	Marginal notes	Elles peuvent être sous forme de références ou de notes explicatives brèves, placées en marge du texte
Marque typographique	Typographiques marks	Une composition emblématique ou héraldique adoptée par un imprimeur ou un libraire comme marque commerciale
Miniature	Miniature	C'est un terme qui signifiait à l'origine, toute lettre ornée et colorée en rouge au minium
Notes	Notes	Texte écrit en bas de page ou dans la marge
Pagination, numérotation	Pagination Page numbering	Deux façons de numéroter les feuillets coexistent. La première, celle des signatures en deux parties en bas à droite, est formée d'une ou de plusieurs lettres indiquant la place de la feuille dans le cahier. Le deuxième mode de pagination, l'utilisation des chiffres au recto de la page, en haut à droite (le verso ne porte longtemps pas de chiffre), apparaît progressivement au cours du XVIe siècle.
Pied de mouche	Flies leg	Signe de ponctuation équivalent au signe de paragraphe (¶)
Réclame	Reclame	Inscription, dans le coin inférieur droit de la page, du mot ou du début du mot de la page suivante.
Signature	Signature	Lettre ou signe que l'on met dans les marges inférieures au-dessous de la dernière ligne des premières pages de chaque feuillet ou de chaque cahier d'un livre.
Texte	Text	Le corpus textuel du livre
Titre courant	Running headlines	Titre de l'ensemble du livre ou du chapitre porté au haut de chaque page et se répétant d'une page à l'autre, dans la marge de tête ou la marge de pied.
Titre de départ	Head title	Titre se trouvant dans la marge de tête de la première page du texte d'un livre.
Typographie	Typography	Le type des caractères s'exprime en haut de 20 lignes, du bas de la queue au haut du hast, suivi de G= gothique, R=romain et I=italique, capitales ou bas de casse.
Vignette	Vignette	Illustration gravée, qu'on trouve généralement sur la page de titre.
Tableau n°.11 : Métadonnées du projet DEBORA (Niveau 3)		

d) Les métadonnées de 4ème niveau : les annotations

A ce niveau, c'est l'usager qui a le droit d'ajouter les descriptions dont il a besoin par le moyen des annotations. Un usager, spécialiste d'un certain aspect du livre aussi bien dans sa forme que dans son contenu, dispose d'une fonctionnalité lui permettant d'attacher des annotations à chaque page concernée. DEBORA donne aux chercheurs la possibilité de travail collectif par leur contribution à l'annotation nécessaire après la consultation de documents en-ligne. « L'usager pourra non seulement créer et structurer son propre corpus local, mais aussi l'enrichir d'annotations qu'il pourra partager avec d'autres usagers »[56]. Autrement dit, la base de données DEBORA peut servir «comme un outil de recherche en histoire du livre »[57]. En particulier les « seizièmistes », au cours de leurs travaux, sont conduits à écrire des notes relatives aux ouvrages qu'ils

[56] DEBORA : projet européen n°. LB 5608 A. Coordinateur Prof. R. Bouché, juin 2000. 179pages. (P.4)

[57] DEBORA : projet européen n°. LB 5608 A. Coordinateur Prof. R. Bouché, juin 2000. 179pages. (P.4)

étudient et peuvent ainsi apporter des connaissances nouvelles utiles à une autre lecture. Les données, sauvegardées sur le serveur d'annotation, sont traitées, par la suite, afin de permettre des recherches sur leurs contenus. Selon les spécialistes du projet DEBORA, la recherche dans les annotations peut se faire par des requêtes « plein texte ».

1.8.3. Les métadonnées EAMMS :

Les ingénieurs des métadonnées EAMMS créent trois niveaux des descriptions pour les manuscrits.

- La première s'intéresse beaucoup à la description des collections de manuscrits.
- La deuxième décrit notamment le partie de texte alors que
- La troisième concerne la description du texte.

Le premier niveau « **A** » qui concerne plutôt la collection de manuscrit même et est repartie en neuf éléments tels que:

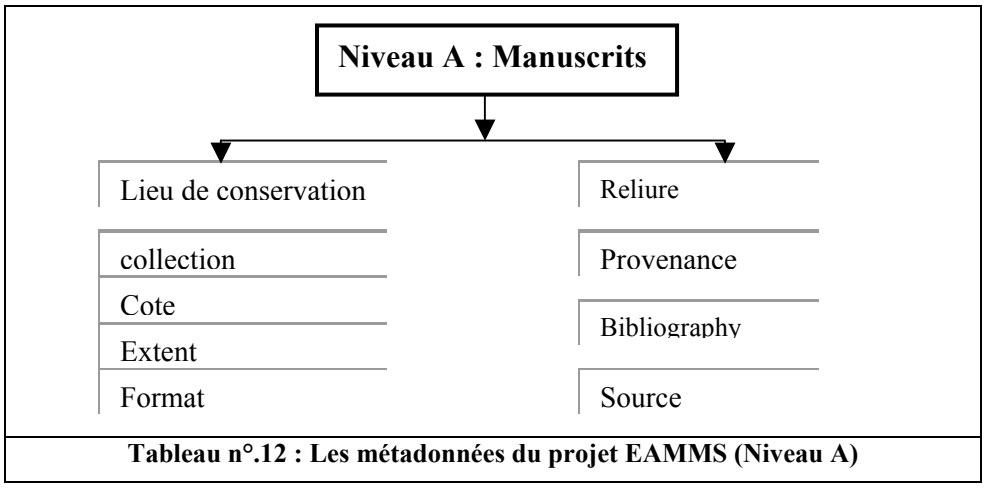

Tableau n°.12 : Les métadonnées du projet EAMMS (Niveau A)

Dans le deuxième niveau « **B** » se situent les métadonnées qui décrivent les parties du document. Il consiste en 14 éléments principaux à savoir :

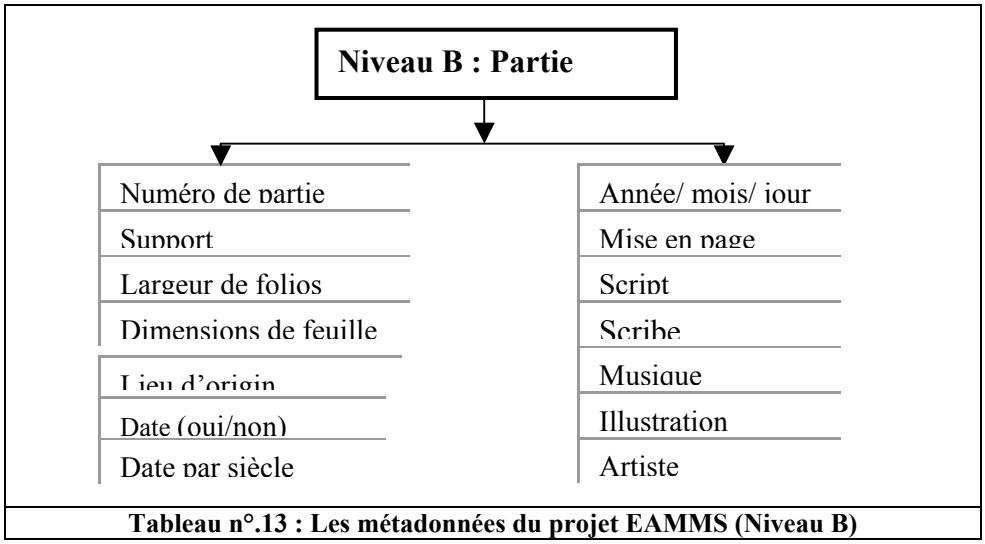

Tableau n°.13 : Les métadonnées du projet EAMMS (Niveau B)

Le troisième niveau « **C** », qui traite surtout des descriptions de texte, est consiste en huit éléments principaux :

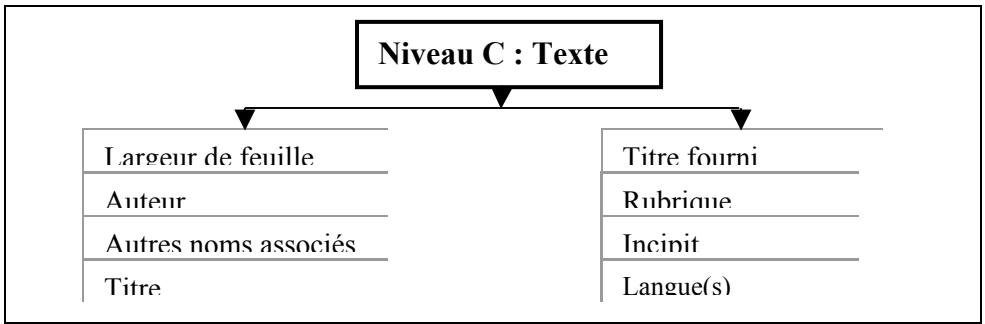

Tableau n°.14 : Les métadonnées du projet EAMMS (Niveau C)

1.9. Conclusion

Les documents électroniques ont besoin une structure spécifique pour les rendre plus accessibles et plus manipulables. Les métadonnées créées par les experts des bibliothèques et/ou des projets de numérisations montrent bien l'exigence de ces types de moyens d'accès que dans quelque sort ressemble beaucoup aux champs de catalogage d'un catalogue traditionnel mais avec plusieurs modifications. Les modifications arrivées à ces documents correspondraient biens aux caractéristiques spécifiques des documents électroniques. Les métadonnées DEBORA et MASTER sont un bon exemple de ce type du travail. L'étude de ces métadonnées et les formats de balisage qui coïncident avec, nous amenént à la création des métadonnées pour les manuscrits arabes mentionnés dans la cinquième partie de thèse page 180.

Les métadonnées exigent une structure hiérarchique spéciale pour structurer le document d'une part et pour définir les éléments, et les attributs et la relation entre eux d'autre part. Le but de cette structure est d'améliorer la recherche et la diffusion d'informations à l'intérieur de ces documents. Plusieurs DTD ont été créées par différents auteurs pour chacun sont propre but, citons comme exemple la DTD TEI, la DTD SGML et DTD XML et le dernier et plus récent est la DTD METS.

Le schéma de METS fournit un mécanisme flexible pour les métadonnées descriptifs, administratifs, et structurelles pour l'encodage d'un objet numérique de bibliothèque, et pour exprimer les liens de complexe entre les diverses formes des métadonnées. Il peut donc fournir une norme standard pour l'échange des objets numériques entre les différentes bibliothèques.

Le format qui servir comme porteurs de ces documents ont aussi nombreuse débutés par SGML suit par HTML et finit jusque maintenant par XML.