

**L'intérêt des DTDs pour le livre ancien :**  
**Signalement d'un fonds d'ouvrages de**  
**botanique au Service Commun de la**  
**Documentation de l'Université Claude Bernard**  
**Lyon 1**

**Béatrice Delestre**

Sous la direction d'Isabelle Westeel

Conservateur des bibliothèques

Responsable de l'informatique et des nouvelles technologies à la  
Bibliothèque municipale de Lille

Stage effectué sous la direction de Dominique Wolf

Conservateur des bibliothèques

Responsable de la Bibliothèque des sciences du SCD Lyon 1

## **Résumé :**

On assiste depuis plusieurs années aux développements de bibliothèques numériques ayant recours à des standards XML, et notamment les Définitions de Type de Document (DTDs). L'objectif est d'étudier en quoi réside l'intérêt de ces DTDs pour le livre ancien. De telles applications permettent d'en améliorer le signalement et l'accès, mais aussi, de manière plus générale, d'aller dans le sens d'une normalisation croissante et de faciliter les échanges.

Descripteurs :

Bibliothèques -- Fonds spéciaux -- Sciences

Bibliothèques -- Fonds spéciaux -- Livres rares

Bibliothèques universitaires -- Développement des collections -- France

Catalogage -- Livres anciens

XML (langage de balisage)

XSL (langage de balisage)

DTD (Définition de type de document)

Toute reproduction sans accord express de l'auteur à des fins autres que strictement personnelles est prohibée.
---

**Abstract :**

Digital libraries are still in development, especially by using XML standards and Document Type Definition (DTDs). The purpose is to study what is the DTD's advantage for Rare book. Such specifications should increase rare book's description and access, but also normalisation and exchanges on the Web.

**Keywords :**

Libraries -- Special collections -- Scientific literature

Libraries -- Special collections -- Rare books

Academic libraries -- Collection development -- France

Cataloging Early works to 1900

XML (Document markup language)

XSL (Computer markup language)

DTD (Document Type Definition)

## Remerciements

Je tiens à remercier Jean-Jacques Flahaut, conservateur responsable du SI au SCD Lyon 1 et David Jouve, technicien, pour leur assistance logistique pendant le stage ; Louis-Bertrand Duc, ingénieur informaticien et Emilie Romand-Monnier, informaticienne à l'Ensib pour leurs conseils informatiques ; Françoise Francillon, chargée du fonds ancien de la Bibliothèque des sciences du SCD Lyon 1, Aline Chareyron, administrateur du SIGB, et Dominique Wolf, conservateur responsable de la Bibliothèque des sciences, pour leur aide scientifique, technique, et leur soutien. Je tiens également à remercier Monique Hulvey, chargée de l'administration des bases documentaires de la Bibliothèque Municipale de Lyon, pour l'expérience acquise en sa compagnie, qui m'a permis de m'engager dans cette étude. Enfin, mes remerciements s'adressent tout particulièrement à Isabelle Westeel, conservateur responsable de l'informatique à la Bibliothèque Municipale de Lille, qui m'a guidée par ses précieux conseils tout au long de cette étude.

# Sommaire

## INTRODUCTION ..... 7

## PARTIE 1 : LIVRE ANCIEN ET XML, DÉFINITIONS ET THÉORIE ..... 9

1. Structure du livre ancien et formats actuels de description bibliographique	9
1.1. Structure du livre ancien et implications sur la description bibliographique	9
1.1.1. L'objet-livre : aspects physiques du document	10
1.1.2. Mise en texte, mise en page, et identité du livre : aspects physiques du contenu	11
1.1.3. Spécificités de la description bibliographique du livre ancien	13
1.2. Les formats actuels de description bibliographiques : MARC	16
1.2.1. La structure des formats MARC	16
1.2.2. Conséquences de la structure de MARC	17
1.2.3. MARC et la description bibliographique du livre ancien	18
2. Présentation de XML et des DTDs	19
2.1. Qu'est-ce que l' <i>eXtensible Markup Language</i> (XML) ?	20
2.1.1. Un exemple de description bibliographique	20
2.1.2. Explicitation de la dénomination <i>eXtensible Markup Language</i>	22
2.1.3. Elaborer un document en XML	23
2.2. Qu'est-ce qu'une Définition de Type de Document (DTD) ?	25
2.2.1. Explicitation du concept	25
2.2.2. Avantages et limites des DTDs	26
2.2.3. Les schémas XML	27
3. DTDs et ressources numériques patrimoniales	29
3.1. TEI, Text Encoding Initiative	30
3.2. EAD, Encoded Archival Description	31
3.3. BiblioML	32
3.4. Dublin Core	34

## PARTIE 2 : APPLICATION D'UNE DTD AU FONDS DE BOTANIQUE DU SCD DE L'UNIVERSITÉ CLAUDE BERNARD LYON1 ..... 36

1. Justification du choix du corpus et de la DTD	36
1.1. Choix de la thématique du corpus	36
1.2. Réception du projet : quelques pistes	37
1.3. Choix des ouvrages du corpus : méthodologie	38
1.4. Choix de la DTD	39
2. Elaboration des documents structurés	40
2.1. Construction de la DTD	40
2.2. Ressources bibliographiques pour la description des ouvrages du corpus	45
3. Retour d'expérience : éditeur XML et feuilles de style	45
3.1. Un éditeur XML : XMetaL	45
3.2. Les feuilles de style	48

<b>PARTIE 3 : DE LA DESCRIPTION BIBLIOGRAPHIQUE À L'ÉCHANGE DE MÉTADONNÉES.....</b>	<b>50</b>
1. Mise en perspective	50
1.1. Bases de données .....	50
1.2. Modes de recherche, navigation et indexation .....	51
1.2.1. Navigation .....	51
1.2.2. Indexation .....	52
2. Intérêts des formats bibliographiques XML	53
2.1. BiblioML.....	53
2.2. Autres formats bibliographiques XML.....	55
2.2.1. Formats dérivés de MARC .....	55
2.2.2. Formats intermédiaires : vers la prise en compte de l'utilisateur.....	56
2.2.3. Formats de conception nouvelle .....	57
2.3. Description bibliographique : apports et limites de XML et des DTDs ..	58
2.3.1. Constats généraux .....	58
2.3.2. Retour sur le livre ancien.....	59
3. Description fine et description minimale : normalisation et échanges de données	61
3.1. Liberté de description et normalisation .....	61
3.2. Echanges de données .....	63
<b>CONCLUSION .....</b>	<b>65</b>
<b>BIBLIOGRAPHIE.....</b>	<b>67</b>
<b>TABLE DES ANNEXES (TOME SÉPARÉ) .....</b>	<b>78</b>

## ***Introduction***

Le patrimoine suscite toujours un vif intérêt, et les bibliothèques sont particulièrement concernées du fait de leur vocation de conservation. En témoignent toutes les publications récentes sur le sujet, comme le numéro 49 du *Bulletin des Bibliothèques de France* 2004 et son dossier « Patrimoines »<sup>1</sup>, ou encore les projets de bibliothèques numériques qui ont vu le jour ces derniers temps, par exemple celle de l'Ecole centrale de Lyon<sup>2</sup>. Comme le montre l'intitulé des rencontres « Numérisation et Patrimoine » organisées lors de la Semaine du Document Numérique de la Rochelle en juin 2004, les débats professionnels autour du patrimoine associent fréquemment des réflexions sur la numérisation. En effet, le numérique a pu apparaître comme la solution au dilemme communication/conservation. Aujourd'hui, sans en remettre en cause les apports, les professionnels semblent plus réalistes sur la numérisation, notamment sur son caractère partiel et les coûts de mise en œuvre.

Les bibliothèques numériques patrimoniales affichent pour objectifs principaux de mettre à disposition de tous via le Web des documents rares, fragiles, très demandés ou susceptibles de l'être. Elles ont en outre l'ambition d'améliorer la qualité de la consultation de ces ouvrages, au-delà de l'accès à distance et non limité dans le temps, en offrant des images plus lisibles que le document original et des modes de recherche et de navigation qui se veulent performants et attrayants. Certaines bibliothèques numériques ont fait le choix d'XML pour décrire les documents et ont utilisé ainsi certaines DTDs (Définition de Type de Document). Ces choix de description associés à des outils de gestion et de navigation sont susceptibles d'apporter de la valeur ajoutée.

Ces modèles XML ne sont plus aujourd'hui des standards émergents<sup>3</sup> mais font encore l'objet de nombreux développements. Aussi peut-on se demander, dans le cadre des bibliothèques numériques patrimoniales, en quoi réside l'intérêt de ces dernières pour le livre ancien, et notamment pour son signalement. Car il convient de préciser que « [...] La numérisation n'est pas le catalogage ». Ces récents propos de Henri-Jean Martin<sup>4</sup> illustrent l'objectif de l'étude menée ici : à

---

<sup>1</sup> Dossier « Patrimoines », *BBF* 2004, t. 49, n° 5, p. 5-78.

<sup>2</sup> Bibliothèque numérique patrimoniale de l'Ecole Centrale de Lyon : <<http://bibli.ec-lyon.fr/pages/catalogue/patrimoine.html>>.

<sup>3</sup> F. Queyroux, « EAD, la description archivistique encodée », *in* : *La numérisation des textes et des images*, I. Westeel et M. Aubry (dir.), Lille, 2003, p. 86.

<sup>4</sup> H.-J. Martin, « Une vision totale du livre », *BBF* 2004, t. 49, n° 5, p. 23.

l'heure où les projets et les réalisations de bibliothèques numériques concernant les fonds patrimoniaux se font plus nombreux, qu'en est-il du traitement du livre lui-même, et de la description qui en est proposée ? Mais alors, comment résoudre le paradoxe qui existe entre le livre ancien, document dont la structure et donc la description sont spécifiques, et un document patrimonial numérisé, dont le signalement se dirige quant à lui vers davantage de normalisation ? En effet, les DTDs, en tant qu'applications XML, sont partie intégrante du contexte de coopération instauré par le Web, et donc d'échanges de données, d'interopérabilité, et de normalisation.

Ainsi, une étude préalable sur la structure du livre ancien et les formats de description existants – MARC et XML - permettront d'analyser l'application d'une DTD, en l'occurrence BiblioML, à un fonds d'ouvrages de botanique du SCD de l'Université Claude Bernard Lyon 1, pour enfin aborder la question du choix à opérer entre une description optimale du livre ancien, qui vise la finesse, et une description d'ordre minimale offerte par les métadonnées permettant normalisation et échanges sur le Web.

# **Partie 1 : Livre ancien et XML, définitions et théorie**

Avant d'entrer dans le vif du sujet concernant l'intérêt des DTDs pour le livre ancien, il est nécessaire de préciser préalablement quelques concepts. Ainsi, cette première partie présentera la structure du livre ancien et ses conséquences sur la description bibliographique, puis les notions relatives au langage XML et aux applications que sont les DTDs, et enfin des exemples concrets de DTDs utilisées pour la construction de bibliothèques numériques, en particulier celles contenant des ressources patrimoniales.

## **1. Structure du livre ancien et formats actuels de description bibliographique**

Les spécificités de la structure du livre ancien en conditionnent la description bibliographique telle qu'elle est définie dans les normes. Comment des formats bibliographiques actuels, tel que MARC<sup>5</sup>, prennent en compte ces données ?

### **1.1. Structure du livre ancien et implications sur la description bibliographique**

Avant d'aborder toute question de structure, il convient de rappeler ce qu'est une monographie, puisque tel est l'objet de l'étude. Une monographie est un « ouvrage qui forme un tout, en un seul ou plusieurs volumes, soit qu'il paraisse en une seule fois, soit que sa publication s'étende sur une durée limitée selon un plan établi à l'avance »<sup>6</sup>. L'opposé est la publication en série.

Par ailleurs, nous considérons ici comme livre ancien tout livre imprimé avant 1801<sup>7</sup>, date à laquelle le processus de fabrication s'industrialise et perd son caractère manuel. Le processus de fabrication manuel explique l'importance de la matérialité du livre ancien. Cette matérialité est double. Elle concerne d'une part les aspects

---

<sup>5</sup> *MAchine Readable Cataloging*.

<sup>6</sup> AFNOR Z 44-050, *Norme de description des monographies imprimées*, 1989.

<sup>7</sup> Cf. la norme AFNOR Z 44-074 de *Catalogage des monographies anciennes*, 1986, p. 5, « [qui] s'applique à l'ensemble des monographies imprimées artisanalement des origines à 1800 ».

physiques du document, le support, mais aussi les aspects physiques du contenu, la mise en page et la mise en texte, pour reprendre l'expression de Henri-Jean Martin<sup>8</sup>.

### 1.1.1. L'objet-livre : aspects physiques du document

L'étude des aspects physiques du document est ici ciblée sur les notions de format et de reliure, qui donnent une idée précise de la complexité du livre ancien.

## Le format

Le format est le rapport entre la feuille vierge originale et la feuille pliée un certain nombre de fois, le résultat est dénommé cahier. Selon le mode de pliage, le nombre de feuillets par cahier varie. Par exemple, dans les cas ordinaires, un livre de format in-folio est composé de cahiers de deux feuillets, un quarto de cahiers de quatre feuillets, et un octavo de cahiers de huit feuillets. Ce n'est pas tant la dimension en centimètres qui est à noter, mais le format ainsi défini, car il permet de comprendre la structure du livre. Le format n'est pas anodin, il a souvent un rapport avec le contenu. Par exemple les formats "de poche", in-8 à in-18, étaient souvent utilisés pour les livres de dévotion. En revanche, les in-folio étaient le format d'usage pour les sommes, telles celles de Thomas d'Aquin<sup>9</sup>.

## La reliure

La notion de cahier n'induit pas seulement celle du format, mais aussi la question de la reliure. La pagination – souvent multiple et parfois erronée - n'aurait été d'aucun recours pour le relieur. D'autres marques de repère sont davantage significatives. Il s'agit du registre, des réclames, et des signatures. Le registre est un relevé des premières lettres de chaque feuillet, il est imprimé en fin d'ouvrage. La réclame consiste à reproduire les premières lettres d'un cahier ou d'un feuillet à la fin du cahier ou du feuillet précédent. Quant aux signatures, elles sont l'attribution d'une lettre à chaque cahier, indiquée dans le coin inférieur droit des feuillets, suivie d'un chiffre pour numéroter les feuillets au sein du cahier. Ces dernières reflètent la complexité de la structure du livre ancien. Tous ces repères, présents simultanément ou non, sont utiles pour plier les feuillets et assembler les cahiers sans faire d'erreurs nuisibles pour la lecture et la compréhension du texte.

---

<sup>8</sup> Cf. le sous-titre de l'ouvrage de H.-J. Martin, *Naissance du livre moderne : mise en page et mise en texte du livre français (XVI<sup>e</sup>-XVII<sup>e</sup> siècles)*, Paris, 2000.

<sup>9</sup> Exemple d'édition in-folio de la Somme théologique de Thomas d'Aquin : [Summa theologiae. 2<sup>nde</sup> partie. 2. 1522]. Venise : Lucantonio de Giunta.

### 1.1.2. Mise en texte, mise en page, et identité du livre : aspects physiques du contenu

Ces aspects physiques du contenu reflètent en particulier la présentation logique du texte et sa présentation générale.

#### Mise en texte et mise en page

La rupture n'est pas brutale entre la présentation du texte du manuscrit et celle du livre imprimé ; le livre imprimé conserve la structure du livre manuscrit, qui est celle du codex. Le manuscrit a en effet d'abord été un modèle pour les premiers imprimeurs. La densité de la présentation du texte demeure, qu'il s'agisse d'un texte à longues lignes ou sur deux colonnes. Mais il existe cependant des marques qui en signalent les divisions, comme la lettre d'attente, et surtout le pied de mouche. Ce dernier disparaît pour un simple retour à la ligne avec retrait, l'alinéa, à partir du XVI<sup>e</sup> siècle. La structuration en paragraphes, les *capitula*, intervient au XVII<sup>e</sup> siècle : ils sont regroupés en chapitres, eux-mêmes rassemblés en livres. La glose, encadrante ou interlinéaire, se raréfie. Les manchettes et les notes en fin de volume cèdent la place aux notes de bas de page à partir du XVII<sup>e</sup> siècle<sup>10</sup>.

Le repérage au sein du texte ne s'est pas fait aisément dès les premiers temps. En effet, comme nous l'avons noté ci-dessus, les signatures et les réclames ne servaient pas tant à guider le lecteur que l'artisan relieur. Ce n'est que plus tard que la foliotation puis la pagination apparaissent pour devenir systématiques à partir du XVI<sup>e</sup> siècle. On peut d'ailleurs remarquer dans ce cas un glissement de fonction de ces éléments de repérage : ils relèvent des aspects physiques du document mais aussi des aspects physiques du contenu. Tout ce qui n'est pas indispensable à la compréhension du texte a tendance à disparaître, et la logique de repérage se modifie pour ressembler à ce qu'elle est encore aujourd'hui.

Les pièces liminaires se formalisent : privilège, dédicace, avis au lecteur, tables et index. A l'origine, une table des matières sommaire servait de guide pour le travail des relieurs<sup>11</sup>. Au XVI<sup>e</sup> siècle, les *capita* consistent à résumer les matières présentes dans l'ouvrage<sup>12</sup>. Les tables et les index se constituent progressivement, grâce aux « méta-descripteurs »<sup>13</sup>, comme la pagination ou la numérotation des paragraphes et des chapitres. Ils favorisent ainsi une lecture non cursive. Pourtant, les index se raréfient au XVIII<sup>e</sup> siècle<sup>14</sup>.

<sup>10</sup> H.-J. Martin, *La naissance du livre moderne*, Paris, 2000, p. 310.

<sup>11</sup> H.-J. Martin, *id.*, p. 31.

<sup>12</sup> R. Chartier, H.-J. Martin (dir.), *Histoire de l'édition française*, Paris, 1989, t. 1 p. 77.

<sup>13</sup> H.-J. Martin, *Les métamorphoses du livre*, Paris, 2004, p. 253.

<sup>14</sup> H.-J. Martin, *Histoire et pouvoirs de l'écrit*, Paris, 1996, p. 289.

Nous pouvons remarquer ici que les aspects physiques du contenu recourent en partie les aspects logiques. En effet, ces aspects physiques que sont le titre, le chapitre, le paragraphe, mais aussi des illustrations par exemple, permettent de dégager la structure logique du document.

## « L'état civil du livre »<sup>15</sup>

Le livre a construit son identité au fil des siècles - même si l'on peut considérer qu'elle est acquise au milieu du XVI<sup>e</sup> siècle<sup>16</sup> -, selon trois éléments principaux : l'*incipit* qui se transforme en page de titre, la marque typographique (au titre : marque du vendeur)<sup>17</sup>, et le colophon (mention du fabricant, en particulier s'il diffère du vendeur) que remplacera ensuite l'achevé d'imprimer.

La page de titre apparaît au XVI<sup>e</sup> siècle, et demeure un élément essentiel du livre d'aujourd'hui, où figurent les indications principales sur la publication et son contenu. Auparavant, c'est l'*incipit*, la première ligne de texte, qui donnait le titre, complété parfois du nom de l'auteur. Pour protéger ce premier feuillet de texte, lors du stockage et des transports, les artisans du livre laissaient un premier feuillet blanc. Petit à petit, une brève mention du titre y a été imprimée pour une identification rapide de l'ouvrage, devenue au fur et à mesure une véritable page publicitaire, portant notamment la marque de l'imprimeur-libraire, en plus de la mention du titre, de l'auteur, et de l'adresse bibliographique<sup>18</sup>. Au XVIII<sup>e</sup> siècle l'habitude de laisser de nouveau un feuillet blanc pour protéger le titre se généralise, feuillet sur lequel est ensuite imprimé un titre abrégé : il s'agit du faux-titre ou de l'avant-titre<sup>19</sup>.

Le colophon, jusqu'au XVIII<sup>e</sup> siècle, venait en complément de la page de titre, notamment pour mentionner le nom de l'imprimeur s'il différait du libraire<sup>20</sup>. Son usage a été remplacé par celui de l'achevé d'imprimer<sup>21</sup>.

<sup>15</sup> L. Febvre, H.-J. Martin, *L'apparition du livre*, Paris, 1999, p. 122.

<sup>16</sup> H.-J. Martin, *Histoire et pouvoirs de l'écrit*, Paris, 1996, p. 289.

<sup>17</sup> Cf. Annexe 4, notice 1 p. XXXIV, *Illustrations of the Nueva Quinologia of Pavon ... by John Eliot Howard ...* London : Lovell Reeve and Co, 1862 : marque du libraire Lovell Reeve au titre.

<sup>18</sup> Cf. Annexe 4, notice 19 p. LXXXVIII, *De historia stirpium ... Leonharto Fvschio ... autore*. Basileae : in officina Insigniniana, anno Christi M.D.XLII. : cet ouvrage donne un bon exemple de page de titre classique.

<sup>19</sup> Cf. Annexe 4, notice 6 p. XLIX, *Filices horti botanici Lipsiensis bearbeitet von Dr Georg Mettenius ...* Leipzig : verlag von Leopold Voss, 1856, pour un exemple de faux-titre.

<sup>20</sup> Cf. Annexe 4, notice 5 p. XLVI, *Icones plantarum cryptogamicarum ... collegit et descripsit Carol. Frideric. Philip. de Martius ...* Monachii : impensis auctoris, 1828-1834. Cette édition du XIX<sup>e</sup> siècle reprend l'usage du colophon tel qu'au XVI<sup>e</sup> s. : (in officina Caroli Wolf, finitum Idibus Decembris [i. e. 13 décembre] MDCCCXXXIV).

<sup>21</sup> Cf. Annexe 4, notice 13 p. LXX, *La statique des végétaux ou l'analyse de l'air ... par M. Hales ...* A Paris : chez Jacques Vincent ..., M.DCC.XXXV. Dans ce cas l'achevé d'imprimer reprend les informations de la p. de titre : "De l'imprimerie de Jacques Vincent".

Ces éléments constituent ce que Henri-Jean Martin a appelé la « morphologie » du livre<sup>22</sup>. Ils ont pu être recueillis au cours de nombreuses études qui ont forgé l'histoire du livre, qui renseignent sur les lieux, les processus de fabrication, et les artisans du monde du livre<sup>23</sup>.

### 1.1.3. Spécificités de la description bibliographique du livre ancien

Cette matérialité du livre ancien, que l'on peut ici qualifier de structure, à savoir « une forme observable et analysable que présentent les éléments d'un objet »<sup>24</sup>, permet de saisir en quoi la description bibliographique du livre ancien est spécifique. De manière générale nous pouvons reprendre la définition suivante<sup>25</sup> : « la description bibliographique est une sélection des données extraites du document primaire telles que le titre, la mention de responsabilité, les données relatives à l'éditeur, la description physique. Elle est parfois complétée par des commentaires rédigés par le catalogueur. Ces informations sont créées en suivant les principes des règles de catalogage ».

Dans le cadre du livre ancien la description bibliographique se fonde sur la distinction entre la notion d'édition et celle d'exemplaire, qui reflète l'unicité de chaque livre en tant qu'objet. Ces types de données sont d'égale importance, alors que, dans le cas du livre contemporain, les données d'exemplaire se limitent à la création d'un accès pour signaler la cote de l'ouvrage, afin de localiser le document. De plus, il est évident de remarquer que les aspects physiques du document ne sont plus guère pris en compte dans la description bibliographique du livre contemporain, seule une brève collation est indiquée. Cette remarque permet de noter d'une part que cette distinction édition/exemplaire ne recoupe pas exactement celle faite ci-dessus entre aspects physiques du document ou du contenu, et d'autre part, qu'il y a parfois risque de confusion entre les données d'exemplaire et celles d'édition dans la description du livre ancien et dans les accès à cette description. Par exemple, lorsqu'il est impossible de comparer plusieurs exemplaires éparpillés géographiquement, si l'un porte un papillon, comment savoir s'il s'agit d'une donnée d'exemplaire ou d'une donnée d'édition ? De même, il se peut qu'un cahier comporte un nombre impair de feuillets, et cela pour toute l'édition, sans que ce soit un cas d'exemplaire au feuillet manquant.

---

<sup>22</sup> H.-J. Martin, *Les métamorphoses du livre*, Paris, 2004.

<sup>23</sup> Par exemple, les études de Martin Lowry sur Alde Manuce s'inscrivent dans cette lignée. Cf. M. Lowry, *Le monde d'Alde Manuce : imprimeurs, hommes d'affaires et intellectuels dans la Venise de la Renaissance*. Paris, Ed. du Cercle de la librairie, 1989.

<sup>24</sup> Définition du *Grand Robert*.

<sup>25</sup> C. Lupovici, « L'information bibliographique des documents électroniques », *BBF* 1998, t. 43, n° 4, p. 42.

Nous proposons donc à présent une liste synthétique des éléments spécifiques à la description bibliographique du livre ancien, ceux qui permettent d'identifier le texte et de décrire l'édition, classés selon les deux types de données, édition ou exemplaire :

### **Données d'édition :**

- Auteur
- Titre
- Edition
- Adresse bibliographique
- Collation : description physique de l'ouvrage (importance matérielle, format)
- Empreinte
- Notes sur le contenu, mais aussi sur la structure, par exemple si le dernier feuillet est blanc.

### **Données d'exemplaire :**

- Reliure
- Décoration
- Appartenance à un recueil factice
- Marque(s) de provenance
- Annotations manuscrites
- Notes sur l'exemplaire : détails de l'exemplaire pouvant renseigner sur l'édition (par comparaison entre plusieurs exemplaires) ou non, incomplétude de l'exemplaire.

Les informations nécessaires pour la description bibliographique du livre ancien, comme pour le livre moderne, se trouvent en grande partie sur la page de titre, à partir du moment où elle existe, mais pas seulement. Le colophon et l'ensemble du document jouent également leur rôle, bien davantage cette fois que dans un livre contemporain. D'autre part, il faut préciser malgré l'évidence du propos que toutes ces informations n'ont pas été imprimées sur le livre ancien de manière normalisée comme cela l'est aujourd'hui. Si les pratiques ont eu tendance à s'unifier et à s'harmoniser au fil des siècles, les divergences dans les notations sont extrêmement fréquentes. Aussi est-il d'une importance primordiale pour un bibliographe travaillant sur l'identification des éditions de transcrire le plus fidèlement possible ces informations, par exemple d'indiquer la date en chiffres romains si tel est le cas dans l'ouvrage, de respecter la graphie des u et des v, des i et j, mais aussi d'indiquer entre parenthèses toute

information trouvée ailleurs dans l'ouvrage que sur la page de titre, et entre crochets toute information externe au livre, ajoutée par le catalogueur. La norme de catalogage des monographies anciennes donne des recommandations en ce sens, en particulier pour les zones de transcription que sont les zones du titre, de l'édition, de l'adresse et de la collection<sup>26</sup>.

Cependant on ne peut que constater, en particulier dans les catalogues collectifs et les outils de localisation, que les pratiques sont loin d'être unifiées, tant au niveau international que national. Les différentes manières de renseigner la description physique dans la base Hand Press Book<sup>27</sup>, par exemple, est à ce sujet significative. Cela peut paraître minime - car la recherche n'est pas altérée dans le sens où l'on retrouve les titres voulus -, mais l'imprécision peut conduire à des difficultés d'identification. Dans plusieurs notices décrivant une même édition, le format in-douze peut être relevé soit (12mo) soit 12°, quand ce n'est pas uniquement la dimension en cm qui est donnée<sup>28</sup>. Des remarques similaires pourraient être faites pour la pagination, les signatures, ou encore la présence d'illustrations. Au-delà de ces pratiques de saisie, sont également en cause des questions d'ordre technique, comme la nécessité d'éviter la création de doublons lors des échanges de notices. Par exemple, insérer des crochets dans un titre, parce qu'il apparaît partiellement sur l'exemplaire catalogué en local, peut être sans conséquence dans le SIGB de la bibliothèque, mais entraîne la création de doublons dans le CCFr<sup>29</sup>.

Il s'agissait ici de donner un aperçu de la complexité et des problèmes d'insuffisance rencontrés dans les catalogues du fait de la spécificité du livre ancien. Ces quelques remarques amènent à préciser la structure des formats bibliographiques utilisés par les bibliothèques, pour construire leurs catalogues délivrant l'information secondaire, que cette dernière concerne le livre ancien ou les documents modernes. Nous verrons donc d'une part la structure de ces formats MARC puis nous affinerons les limites rencontrées dans le cas du livre ancien.

---

<sup>26</sup> Cf. « Langue et écriture de la description », *in* : AFNOR, *Catalogage des monographies anciennes : rédaction de la description bibliographique*, Norme Z 44-074, 1986, p. 8.

<sup>27</sup> Base de données du CERL, *Consortium of European Research Libraries* : <<http://www.cerl.org>> (nécessité de consulter dans une bibliothèque membre, cf. annexe 1-3 p. XI pour la liste des bibliothèques membres françaises).

<sup>28</sup> Recherche effectuée en saisissant le titre "curiositez de la nature" (cf. annexe 4, notice 17 : Vallemont, P. Le Lorrain, abbé de, *Curiositez de la nature et de l'art sur la végétation, ou l'agriculture et le jardinage ... A Paris* : par la sociétés [sic], 1711).

<sup>29</sup> Catalogue Collectif de France : <<http://www.ccf.fr/>>.

## 1.2. Les formats actuels de description bibliographiques : MARC

### 1.2.1. La structure des formats MARC

MARC est un format de codage de données bibliographiques qui a vu le jour dans les années 60. L'objectif était de créer un format normalisé permettant la gestion et l'échange de données bibliographiques entre centres nationaux. La structure de ce format se calque sur celle de l'ISBD<sup>30</sup>, à savoir les différentes zones de la description bibliographique. MARC a été décliné selon une cinquantaine de versions nationales<sup>31</sup>.

Ces formats, au-delà des données descriptives issues du pavé ISBD, intègrent les accès à ces données, et les données supplémentaires propres au format pour en permettre l'échange. Il s'agit donc d'une structure particulière de l'information secondaire incluant la description bibliographique pure et la valeur ajoutée que sont les points d'accès.

La structure du format d'échange est définie selon la norme ISO 2709 : label de notices / répertoire / zone de données de longueur variable / code de fin de notice. Quant à la structure générale d'une notice MARC, elle est la suivante :

- Label de notice et zones codées, composés de champs à longueur fixe. Ces données ne sont pas destinées à être affichées.
- Champs – ou zones, au nombre de 999 - répartis en dix blocs, annoncés par une étiquette comportant trois caractères numériques, suivie de deux indicateurs qui donnent des informations sur le contenu de la zone. Les champs sont divisés en sous-champs, eux-mêmes précédés de séparateurs – ou codes de sous-zones - sous forme codée numérique ou alphabétique. Etiquettes et séparateurs sont des marqueurs et peuvent être considérés comme l'équivalent des balises XML. Ces champs à longueur variable contiennent les données descriptives ainsi que les accès, données textuelles destinées à être affichées.

Enfin, notons que les formats MARC ne permettent qu'une arborescence à deux niveaux seulement : champs et sous-champs.

---

<sup>30</sup> *International Standard Bibliographical Description.*

<sup>31</sup> IFLA, *UNIMARC : an introduction*, 1996 : <<http://ifla.inist.fr/VI/3/p1996-1/unimarc.htm>>.

### 1.2.2. Conséquences de la structure de MARC

Nous anticipons quelque peu sur le chapitre suivant, qui reviendra sur la définition des DTDs. En effet, de la structure de MARC que nous venons d'évoquer résultent les réflexions qui suivent.

Nous avons évoqué l'insuffisance des catalogues en MARC, il est temps à présent d'en préciser les raisons. DTDs et formats MARC définissent des formats de description de données. Cependant, MARC n'a permis qu'une normalisation partielle des métadonnées, dans la mesure où les versions se sont multipliées. Par exemple, en Amérique du Nord c'est le format MARC 21 qui est utilisé, alors qu'en France les bibliothèques ont recours à INTERMARC et UNIMARC. De plus, conçus pour les bibliothécaires, ils sont demeurés un outil utilisé uniquement par cette profession.

D'autre part, alors que les DTDs permettent de produire des documents primaires structurés incluant des métadonnées, les formats MARC n'ont longtemps donné que l'information secondaire, du moins à l'origine. Deux champs récemment développés ont comme fonctionnalité d'associer à la description bibliographique des données primaires. Le champ 856<sup>32</sup> a été instauré pour créer un lien pointant vers une ressource numérique. Par exemple, la description du document primaire structurée en MARC, et des représentations intégrales ou non du document primaire numérique intégrées dans un logiciel de Gestion Electronique du Document (GED) peuvent ainsi être reliées. Cependant cela suppose un stockage des données éclaté, car un système de GED classique juxtapose aux notices du catalogue informatisé une base de données contenant des fiches documentaires, auxquelles sont associées des images numérisées. Cela peut être un obstacle à la recherche, notamment en cas d'absence de moteur permettant de fédérer une recherche. Le second développement concerne le champ 359, qui rend possible la présence d'une table des matières dont l'arborescence peut atteindre huit niveaux. MARC et DTDs proposent donc deux niveaux d'architecture<sup>33</sup> : séparation, mais jonction possible par un lien, entre l'information secondaire et le document primaire – par exemple une notice à laquelle serait relié un sommaire numérisé -, et documents primaires contenant leurs propres métadonnées – comme différents ensembles d'une collection iconographique encodée en EAD, incluant les images numérisées.

---

<sup>32</sup> Dans le Sudoc, c'est le champ 325 qui est utilisé.

<sup>33</sup> A. Jacquesson, *Bibliothèques et documents numériques*, Paris, 1999, p. 232-233, et C. Lupovici, « L'information bibliographique des documents électroniques », *BBF* 1998, t. 43, n° 4, p. 47.

### 1.2.3. MARC et la description bibliographique du livre ancien

Concernant plus précisément la description bibliographique du livre ancien, quelques limites des formats MARC à cette description et donc à la communication des collections patrimoniales de monographies imprimées peuvent être notées. En effet, étant donné la structure complexe d'un livre ancien, les informations à renseigner sont nombreuses, et cela d'autant plus qu'il n'existe dans ce domaine aucun document similaire à un autre, même au sein d'une édition. La principale difficulté provient de l'ambivalence qui peut parfois exister entre données d'édition et données d'exemplaire. C'est le cas du champ 608 du format UNIMARC. Ce champ permet de renseigner et d'indexer les particularités d'un document, que ce soient des données d'édition de genre comme le fait que l'ouvrage soit un manuel scolaire ou un traité, des données d'édition de description physique comme la présence de musique gravée, ou encore des données d'exemplaires, par exemple une reliure allemande du XVI<sup>e</sup> siècle en peau de truie. Cependant, notons que les champs 316 "Note sur l'exemplaire", et 317 "Note sur la provenance", permettent de renseigner en texte libre les données d'exemplaire. Une indexation plein texte est possible selon le paramétrage du système<sup>34</sup>.

D'autre part, pour affiner les quelques remarques avancées ci-dessus<sup>35</sup>, l'étude des notices<sup>36</sup> d'un catalogue collectif comme l'ICCU<sup>37</sup>, qui propose une seule notice par édition suivie des localisations et des données d'exemplaires, révèle la difficulté d'intégrer les données spécifiques du livre ancien dans un format MARC. Pour simplifier l'uniformisation, le choix a été fait d'intégrer la majeure partie de ces informations dans les "Note generali" - équivalent du champ 300 en UNIMARC : les caractères typographiques, l'emplacement et la description des marques typographiques, le type d'illustration comme les portraits, les notes de contenu, les références bibliographiques, mais aussi les signatures, qui pourraient être indiquées dans la collation. Chaque type d'information est saisi dans une note séparée. En revanche, l'empreinte fait l'objet d'une entrée spécifique "Numeri". C'est le même principe qui prévaut dans la base HBP, à quelques différences près. Dans ce catalogue, cette fois spécifique au livre ancien et qui juxtapose les notices correspondant à une même édition, l'empreinte est également incluse dans les notes,

---

<sup>34</sup> Dans le catalogue de la Bibliothèque Municipale de Lyon, la recherche par mots dans les champs 316 et 317 est possible grâce à l'index NOEX (note d'exemplaire). D'autre part, certaines bibliothèques utilisent le champ 905 du bloc des données locales pour renseigner ces particularités d'exemplaires.

<sup>35</sup> Cf. § 1.1.3. "Les spécificités de la description bibliographique du livre ancien" p. 13-15.

<sup>36</sup> Deux types de recherche simple, effectués dans l'ICCU et HPB : recherche par auteur "Fuchs Leonhard" (cf. annexe 4, notice 19) et recherche par titre "instruction pour les jardins fruitiers et potagers" (cf. annexe 4, notice 16).

<sup>37</sup> OPAC de l'Istituto Centrale per il Catalogo Unico, catalogue collectif italien : <<http://opac.sbn.it/>>.

alors que les références bibliographiques et les notes sur l'édition possèdent leur propre entrée.

Enfin, la description à plusieurs niveaux proposée par les formats MARC n'est pas satisfaisante. Elle est fastidieuse à renseigner par le bibliothécaire, qui doit saisir ou copier tout ou partie de l'information dans toutes les notices concernées, et complexe à lire pour le lecteur effectuant ses recherches dans le catalogue. Sans compter que le paramétrage des logiciels de gestion est rarement optimal sur ces questions marginales de description, par rapport à l'ensemble des collections d'une bibliothèque. Deux exemples peuvent être cités. Celui des titres collectifs d'abord, qui nécessitent une analyse précise afin de pouvoir renseigner les champs 423 "Publié avec" ou 464 "Niveau de dépouillement". Une notice mère avec le titre principal doit faire référence aux notices filles. Ensuite, le cas des recueils factices, parfois très présents dans certains fonds anciens. La complexité de ce cas de description n'est pas due à l'identification d'éditions indépendantes comme dans le cas précédent, mais plutôt à la fréquente hétérogénéité de ces recueils. En effet, lorsque ce sont des libraires qui les ont constitués, ils ont souvent davantage d'unité de sens que lorsqu'ils ont été relié par leur possesseur. Dans cette situation, chaque édition fait l'objet d'une notice propre. Il est alors nécessaire de renseigner le champ 481 "Est relié avec" dans la notice mère – c'est-à-dire celle de l'édition située en début du volume -, à multiplier par le nombre de titres qui constituent le recueil, et un champ 482 "Relié avec" dans chaque notice fille. Le résultat est d'autant plus complexe que le nombre d'éditions dans le recueil est important. Les administrateurs des catalogues recommandent généralement de limiter le nombre de 481 à trois ou cinq, pour ne pas altérer la lisibilité et conserver des notices de taille raisonnable. Notons par ailleurs que la dénomination des champs dans les OPAC, qui reprennent souvent celle proposée par UNIMARC, n'est pas très limpide pour un usager quelconque. Dans ces deux cas de description à plusieurs niveaux, l'inconvénient majeur réside dans l'impossibilité d'obtenir une vue globale de la hiérarchisation des notices.

## **2. Présentation de XML et des DTDs**

Les limites de MARC avaient déjà été relevées à différentes reprises par la profession. Nous venons de préciser celles qu'implique la rigidité de sa structure sur le livre ancien. A présent, une synthèse sur des questions d'ordre général sur XML et les spécifications qui sont associées à ce langage est proposée, axée dans la mesure du possible sur le livre ancien.

XML et ses applications ayant une structure formée par des balises ainsi qu'une syntaxe particulière, nous utiliserons dans ce mémoire, et en particulier au sein de ce chapitre, des conventions typographiques. Les balises sont notamment indiquées en caractères gras et les attributs en italique<sup>38</sup>.

## 2.1. Qu'est-ce que l'eXtensible Markup Language (XML) ?

Avant de définir ce qu'est XML, nous commencerons par donner un exemple comparatif afin de permettre une visualisation de la structure de ce langage.

### 2.1.1. Un exemple de description bibliographique

L'exemple présente une description bibliographique minimale du *De historia stirpium* de Leonhard Fuchs, publié à Bâle en 1542<sup>39</sup>, sous trois formes différentes : ISBD(A)<sup>40</sup>, UNIMARC, et XML. Les marqueurs sont indiqués en caractères gras.

#### ISBD(A)

De historia stirpivm commentarii insignes, maximis impensis et vigiliis elaborati, adiectis earvndem vivis plvsqvam quingentis imaginibus ... Leonarto Fvschio medico ... autore [Texte imprimé]. – Basileae : in officina Isingriniana, M. D. XLII. – [28], 896, [4] p. (sig. [alpha]6, [beta]8, A-Z6, a-3f6) : ill. gr. s. b. ; 2° (39 cm). - Demi-reliure basane XVIIIe siècle, plats recouverts de papier marbré.

Points d'accès :

Fuchs, Leonhard, (1501-1566)

Isengrin, Michael, actif en 1542. Impr.

Les marqueurs sont dans ce cas les signes typographiques séparateurs des différentes zones de la description bibliographique : point, espace, tiret, espace.

---

<sup>38</sup> La signification de ces termes est donnée dans ce premier chapitre.

<sup>39</sup> SCD Lyon 1 Université Claude Bernard, Bibliothèque de sciences, n° d'inventaire 1646 (cf. annexe 4, notice 19).

<sup>40</sup> *International Standard Bibliographic Description for Older Monographic Publications (Antiquarian)* : Description bibliographique internationale normalisée des monographies anciennes, Paris, 1982, 2de éd. 1991.

## UNIMARC

- 200 1 0** \$aDe historia stirpivm commentarii insignes, maximis impensis et vigiliis elaborati, adiectis earvndem vivis plvsqvam quingentis imaginibus ... Leonarto Fvschio medico ... autore
- 210 . .** \$eBasileae :\$gin officina Isingriniana,\$dM. D. XLII
- 215 . .** \$a[28], 896, [4] p. (sig. [alpha]6, [beta]8, A-Z6, a-3f6) :\$cill. gr. s. b. ;\$d2° (39 cm)
- 316 . .** \$aDemi-reliure basane XVIIIe siècle, plats recouverts de papier marbré.
- 700 . 1** \$aFuchs,\$bLeonhard,\$f(1501-1566)
- 702 . 1** \$alsengrin,\$bMichael,\$cactif en 1542\$4160

En UNIMARC, les marqueurs sont plus nombreux et correspondent à des codes informatiques, avec d'une part les étiquettes des champs et d'autre part les séparateurs de sous champs, introduits par un \$.

## XML

### <BibliographicRecord>

<Title> De historia stirpivm commentarii insignes, maximis impensis et vigiliis elaborati, adiectis earvndem vivis plvsqvam quingentis imaginibus ... Leonarto Fvschio medico ... autore </Title>

#### <Publisher>

<Place> Basileae </Place>

<Address> In officina Isingriniana </Address>

<Date> M. D. XLII </Date>

#### </Publisher>

#### <PhysicalDescription>

<PagesAndSignatures> [28], 896, [4] p. (sig. [alpha]6, [beta]8, A-Z6, a-3f6) </PagesAndSignatures>

<Illustrations Value="yes"/>

<Dimensions> 2° (39 cm) </Dimensions>

#### </PhysicalDescription>

<LocalNote> Demi-reliure basane XVIIIe siècle, plats recouverts de papier marbré. </LocalNote>

#### <ControlAccess>

<Author>

```

      <Name> Fuchs </Name>
      <Forename> Leonhard </Forename>
      <Dates> 1501-1566 </Dates>
    </Author>
    <SecondaryAuthor>
      <Name> Isengrin </Name>
      <Forename> Michael </Forename>
      <Dates> actif en 1542 </Dates>
    </SecondaryAuthor>
  </ControlAccess>
</BibliographicRecord>

```

Dans le cas de XML, ce sont les différentes balises qui jouent le rôle des marqueurs. Elles sont indiquées par les caractères inférieur et supérieur.

### 2.1.2. Explication de la dénomination *eXtensible Markup Language*

A présent qu'un exemple simplifié<sup>41</sup> de document XML a pu être visualisé, explicitons ce que l'expression XML recouvre.

XML peut être en effet défini de plusieurs manières. Commençons par un bref rappel historique afin de comprendre les objectifs de ce langage. Mis au point par un groupe de travail du World Wide Web Consortium (WC3) institué en 1996, les principes de XML remontent en réalité aux années 70, lorsque des chercheurs réfléchissaient aux questions d'échanges de données et d'interopérabilité, c'est-à-dire la communication entre systèmes hétérogènes. Ils ont abouti à la notion de document structuré. L'objectif du groupe de travail du WC3 était de créer un langage de balisage, plus simple que SGML<sup>42</sup>, mais en conservant les fonctionnalités de ce dernier. Le résultat a fait de XML un sous-ensemble de la norme SGML. Les avantages de XML sont certains, notons par exemple l'indépendance vis-à-vis des outils du marché informatique, et l'utilisation de l'encodage Unicode pour les caractères, qui permet la manipulation des

---

<sup>41</sup> Cet exemple est une conception personnelle. Le nom des balises est inspiré de celles de BiblioML. L'exemple vise à illustrer le vocabulaire défini dans le paragraphe 2.1.3. "Elaborer un document en XML" p. 23-25.

<sup>42</sup> *Standard Generalized Markup Language*, langage de balisage défini par la norme Iso 8879.

documents XML par un simple éditeur de texte. Les deux éditions de la version 1.0 ont fait l'objet d'une *Recommandation XML*<sup>43</sup>.

Clarifions à présent sa dénomination, *eXtensible Markup Language*. En français, cela signifie Langage de balisage extensible. Mais XML n'est pas un simple langage. En effet, il peut être qualifié de **métalangage**, car il donne la structure qui permet de créer d'autres langages. De plus, le terme de **balisage** est employé, car XML, comme SGML, permet de structurer de l'information, et cela à l'aide de marqueurs, les balises. Enfin, XML est **extensible**, ce qui signifie d'une part qu'aucune balise n'est prédéfinie, et d'autre part que tout un chacun peut créer de nouvelles balises, dont le choix du nom est libre, d'où la souplesse de ce métalangage.

### 2.1.3. Elaborer un document en XML

Il n'est pas possible et de faible intérêt dans le corps de ce mémoire de donner une description détaillée de la syntaxe de XML<sup>44</sup>. Nous présentons ici seulement les quelques notions clés, ainsi que les règles de base utiles pour construire un document en XML.

#### La syntaxe : vocabulaire

- Les **balises** sont les outils qui permettent de matérialiser la structuration des données du document en XML, dont les composantes sont ainsi facilement interprétables, tant par l'homme que les ordinateurs. Elles sont toujours entourées des caractères inférieur et supérieur < >. XML utilise toujours deux balises en association, une ouvrante et une fermante : < **balise ouvrante** >, < / **balise fermante** >. Ce couple de balises et leur contenu textuel s'appelle un **élément**. Si le contenu entre les deux balises n'est pas un autre couple de balises, il s'agit de données caractères dites analysables, ou PCDATA, Parsed Characters Data.
- Il existe cependant des cas où l'élément n'appelle ni saisie de texte, ni sous-éléments. Dans ce cas, une seule balise, à la fois ouvrante et fermante, suffit. Cela s'appelle un **élément vide**, il est représenté par une

---

<sup>43</sup> *Recommandation XML* 1.0. <<http://www.w3.org/TR/2004/REC-xml-20040204/>>. La version 1.1 était candidate à la recommandation fin 2002-début 2003. Des modifications minimales sont proposées, comme le changement de version Unicode (3.2).

<sup>44</sup> De nombreux ouvrages ou cours en ligne peuvent être consultés pour cela, en particulier l'ouvrage de Cyril Vincent, *XML et XSL*, Nantes, 2001, et la version française en ligne du manuel d'Eric Lease Morgan, *Getting started with XML : XML pour les bibliothécaires*, 21 avril 2003 : <<http://morinn.free.fr/xml/getting-started.htm>>.

balise unique contenant le signe / à la fin. Dans l'exemple donné ci-dessus figure un élément vide : `<Illustrations Value="yes"/>`.

- Un document XML comporte toujours un **élément racine** et un seul. Il est significatif et représente le contenu du document. Dans l'exemple ci-dessus, l'élément racine est le couple de balises `<BibliographicRecord>` `</BibliographicRecord>`. Les premiers éléments-fils s'imbriquent dans ce dernier, et eux-mêmes peuvent contenir d'autres éléments-fils.
- Une balise est toujours constituée d'un **identifiant**, c'est-à-dire du nom de l'élément, par exemple "Title" ou encore "Date".
- Une balise peut accessoirement être qualifiée par un **attribut** : `<identifiant Attribut="Valeur de l'attribut">`. L'attribut est inclus dans la balise ouvrante, et est toujours constitué de son nom, suivi du signe égal et de la valeur de cet attribut entre guillemets. Dans notre exemple, il s'agit de l'élément `<illustration Value="yes"/>`, qui indique la présence d'illustrations dans l'ouvrage décrit.

## Document "bien formé"

Un document XML doit être "**bien formé**", c'est-à-dire qu'il doit respecter un certain nombre de règles définies dans la *Recommandation XML* :

- Le document doit posséder un élément racine et un seul.
- Les éléments doivent être correctement imbriqués (pas de chevauchement) et fermés (toute balise ouvrante doit posséder une balise de fermeture, à moins qu'elle soit un élément vide).
- Les valeurs des attributs sont entre guillemets ou apostrophes.
- Certains caractères appartenant à la syntaxe XML sont interdits dans la saisie des données. Il s'agit des caractères inférieur et supérieur, des apostrophes, des guillemets, du &. Il faut remplacer ces cinq caractères par des **entités prédéfinies**. Par exemple la caractère & doit être saisi : `&amp;`.
- XML est sensible à la casse.

Le résultat est un document présentant une arborescence, c'est ce que l'on appelle l'**arbre d'éléments**. Cependant, si l'imbrication des balises représente les relations hiérarchiques entre les différentes composantes du document, l'indentation n'a aucune signification pour la lecture par le processeur XML. Il s'agit en effet d'une simple commodité de présentation et de lecture pour l'œil

humain. Précisons enfin qu'un document XML n'est qu'une chaîne de caractères, définis par la norme ISO/IEC-10646.1, *Universal Character Set*.

Pour clore cette présentation sur XML, il est important de préciser que ce langage ne peut être utilisé seul. Avant même de recourir à des outils pour l'exploitation des données créées, il existe différents besoins pour lesquels d'autres spécifications sont nécessaires, comme la transformation des structures, ou encore le contrôle de validité. Pour expliciter le premier point, il faut préciser que XML a l'avantage de séparer la structuration des données de leur représentation, favorisant de la sorte non seulement leur pérennité mais aussi leur échange. Par conséquent, une spécification permettant la présentation de ces données est nécessaire. Citons par exemple le langage XSL<sup>45</sup>, langage de feuilles de style qui permet de transformer les données XML pour obtenir des formats de sortie différents (PDF, page web ...). Le contrôle de validité est un autre cas de spécification, nécessaire pour que le document ne soit pas seulement "bien formé", mais aussi "valide". C'est ce que nous proposons d'approfondir dans le paragraphe suivant, en expliquant le rôle des DTDs.

## **2.2. Qu'est-ce qu'une Définition de Type de Document (DTD) ?**

L'explicitation de la notion de DTD, ainsi que les avantages et les limites de ces dernières conduiront à aborder ensuite les nouveaux modèles XML que sont les schémas.

### **2.2.1. Explicitation du concept**

Pour situer les DTDs<sup>46</sup> par rapport à XML, reprenons l'analogie linguistique proposée par Cyril Vincent<sup>47</sup>, qui distingue trois types d'analyses effectuées lors de la réception d'un message. Ces analyses sont les suivantes :

- l'analyse syntaxique, c'est le rôle de XML,
- l'analyse grammaticale, c'est celui des Définitions de Type de Document ou DTDs,
- l'analyse sémantique, qui fait depuis quelques années l'objet de nombreuses réflexions dans le cadre du Web, comme par exemple le

---

<sup>45</sup> *eXtensible Stylesheet Language*. Cf. Partie 2, ch. 3, § 3.2. "Les feuilles de style" p. 48-49.

<sup>46</sup> Les expressions "modèle XML" ou "application XML" remplacent souvent le terme de DTD, plus spécifique.

<sup>47</sup> C. Vincent, *op. cit.*, p. 35-36.

*Ressource Description Framework* (RDF), format de description de métadonnées<sup>48</sup>.

L'idée principale à retenir concernant les DTDs, c'est qu'elles permettent de valider un document XML bien formé, c'est-à-dire syntaxiquement correct. En effet, pour que le document soit aussi "valide", c'est-à-dire grammaticalement correct, il est nécessaire de recourir à une structuration formalisée, donnée par une grammaire elle-même incarnée par une DTD. En d'autres termes, un document est dit valide s'il respecte la grammaire choisie. Cette grammaire définit la liste des éléments qui peuvent être utilisés, ainsi que les relations entre ces éléments. Une DTD fait donc référence à deux notions de structure. D'une part la structure spécifique, celle du document traité, et d'autre part la structure générique, donnée par la DTD. On dit alors que le document est une instance de la DTD. Toutes les instances de la DTD appartiennent à une même classe de documents. Le lien entre ces deux notions de structures générique et spécifique se crée lors du contrôle de validité, effectué par un analyseur syntaxique ou parseur. Ce programme lit le document et le confronte à la DTD référencée dans la déclaration du prologue du document XML, dans le cas d'une DTD externe<sup>49</sup>. La condition pour que la validation par le parseur soit possible est donc que le document fasse référence à la DTD externe choisie par une commande du type `<!DOCTYPE nom (PUBLIC "Id Publique"/SYSTEM "URL")>`.

### 2.2.2. Avantages et limites des DTDs

Dans le monde du Web et plus particulièrement dans celui des bibliothèques, plusieurs DTDs ont été construites, dont nous donnons quelques exemple ci-dessous<sup>50</sup>. Elles ont été conçues avec l'objectif de faciliter non seulement l'échange de données, mais aussi l'accès à ces données. Elles ont été élaborées avec le souci de respecter les normes en vigueur, d'être réutilisables, et libres de droit. Cependant, le risque, tout comme il n'existe pas de grammaire universelle<sup>51</sup>, réside dans la multiplication de créations de ce type de langage, bien qu'on ne puisse nier que chaque communauté d'utilisateurs a des besoins spécifiques. En effet, il est important de préciser que chacun peut créer sa propre DTD, c'est une conséquence des fonctionnalités de XML.

---

<sup>48</sup> RDF : <http://www.w3.org/RDF/>.

<sup>49</sup> Une DTD peut également être interne au document : dans ce cas elle est incluse à l'intérieur du document XML lui-même : `<!DOCTYPE racine [<!ELEMENT racine ...]>`.

<sup>50</sup> Cf. ch. 3 "DTDs et ressources numériques patrimoniales" p. 29-35.

<sup>51</sup> C. Vincent, *op. cit.*, p. 36.

D'un point de vue informatique, signalons que les DTDs ont été conçues pour être non-propriétaires, indépendantes des plates-formes, matériels, et logiciels informatiques. D'autre part, tous les outils XML sont compatibles avec les DTDs. Elles sont relativement simples d'utilisation, et peu coûteuses en termes de conception.

Toutefois, les DTDs apparaissent comme des outils rigides. Héritières directes de SGML, elles ne requièrent une syntaxe XML que par adaptation. Ces considérations ont ainsi conduit à l'élaboration depuis 2001 des schémas XML, définis par la spécification du W3C *XML Schema*<sup>52</sup>, comme alternative. Ils sont plus complets, mais aussi plus complexes.

### 2.2.3. Les schémas XML

Les schémas comme les DTDs sont des modèles qui permettent de définir des structurations, c'est-à-dire les éléments utilisés ainsi que les relations entre ces éléments. Afin de définir les schémas sans entrer dans le détail de leur complexité<sup>53</sup>, puisque notre étude se concentre sur les DTDs, nous proposons un tableau synthétisant les principales différences entre les deux types de modèles, DTD et schéma<sup>54</sup>.

	<b>DTD</b>	<b>Schéma</b>
<b>Langage de balisage originel</b>	SGML, adaptation à XML avec simplification syntaxique.	XML, s'exprime donc directement en syntaxe XML bien formée. Il en résulte une simplification de l'analyse syntaxique.
<b>Typage des données</b>	Données non typées, il s'agit de simples chaînes de caractères.	Typage des données : permet de contrôler la forme lexicale des données textuelles contenues dans les éléments et les attributs.
<b>Espaces de nom</b>	La déclaration <DOCTYPE ... !> qui permet de référencer une DTD par un identifiant logique	La déclaration <DOCTYPE ... !> est remplacée par une spécification d'espace de noms

<sup>52</sup> *XML Schema* : <<http://www.w3.org/XML/Schema>>, mai 2001.

<sup>53</sup> En réalité la *Recommandation XML Schema* est comparable à la *Recommandation XML*, et les modèles de documents appelés schémas XML correspondent à ce que sont les DTDs.

<sup>54</sup> Synthèse réalisée à partir de A. Mkadmi, « Partager les modèles XML : quel intérêt ? », *BBF* 2003, t. 48, n° 5, p. 68-73, et J.-J. Thomasson, *Schémas XML*, Paris, 2003, ch. 9, p. 291 et suivantes.

	<b>DTD</b>	<b>Schéma</b>
	(concept XML correspondant à celui d'espace de noms dans <i>XML Schema</i> ) ne permet pas d'obtenir des fichiers XML autonomes et inclus les uns dans les autres. Il est impossible de construire une DTD à partir d'éléments provenant de différents espaces de noms.	sur l'élément racine du document XML : généralisation de l'usage des espaces de noms à tous les niveaux d'un document XML.  Les espaces de noms auxquels appartiennent les éléments et attributs d'un schéma sont parfaitement identifiés.  Le caractère modulaire des schémas résulte de cette meilleure intégration des espaces de noms .
<b>Validation</b>	Utilisée à la fois comme formalisme grammatical et comme outil de modélisation de structures. Structure orientée document.	Nouveau formalisme de validation. Au-delà du formalisme grammatical des DTDs : technique de modélisation. Orienté objet.
<b>Extensibilité</b>	N'offre qu'un ensemble de données relativement pauvre et non extensible.	Possibilité de créer ses propres types de données de deux manières :  - en créant un type complexe par agrégation d'autres types (extension),  - en limitant un type existant (restriction)  (notion d'héritage).
<b>Eléments racine</b>	Obligation d'utiliser la déclaration <code>&lt;!DOCTYPE ...&gt;</code> pour déclarer l'élément-racine.	Possibilité d'utiliser comme élément racine n'importe quel élément global d'un schéma, sans déclaration : on parle d'éléments racine libres.

	DTD	Schéma
<b>Modèles de contenu locaux</b>		Possibilité de définir des éléments dont le modèle de contenu diffère selon leur position dans le document.

Plusieurs schémas XML ont été élaborés à des fins d'utilisation bibliothéconomiques. Citons par exemple MARC 21 XML Schema et MODS, de la Bibliothèque du Congrès, ou encore XOBIS, de la Lane Medical Library de l'Université de Stanford.

D'autre part, depuis l'adoption par le W3C en mai 2001 de la *Recommandation XML Schema*, certaines DTDs ont été converties en schémas XML. C'est le cas de la DTD MARC XML DTD, convertie en MARC 21 XML Schema, ou encore du Dublin Core<sup>55</sup>.

Ce paragraphe entendait donner une vision centrée sur les aspects informatiques, permettant d'aborder la relation entre XML et les DTDs, l'intérêt et les limites de ces dernières, ainsi que les transformations et dépassements très rapides que subissent ces modèles.

### 3. DTDs et ressources numériques patrimoniales

A présent que les différents formats et structures ont été évoqués, il est possible d'examiner quelques exemples les plus courants des DTDs utilisées dans le monde des bibliothèques, avec leurs objectifs et leurs particularités. Afin de montrer ce que peuvent permettre les différentes applications XML, cette présentation est illustrée par quelques exemples de projets de bibliothèques numériques de collections patrimoniales qui y ont recours. Certains projets sont recensés dans le catalogue des fonds culturels numérisés du Ministère de la Culture et de la Communication<sup>56</sup>. Un catalogue similaire qui recenserait les réalisations des bibliothèques universitaires et coordonné par le Ministère de

<sup>55</sup> Pour les références précises et pour de plus amples détails sur les exemples cités dans ce paragraphe, cf. Partie 1, ch. 3 "DTDs et ressources patrimoniales" p. 29-35 et Partie 3, ch. 2, § 2.2. "Autres formats bibliographiques XML" p. 55-58.

<sup>56</sup> Numérisation du patrimoine, catalogue : <[http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f\\_02.htm](http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f_02.htm)>.

l'Education Nationale est à souhaiter. La consultation de plusieurs bibliothèques numériques a conduit à centrer les exemples sur les DTDs qui semblent les plus couramment utilisées, et qui sont aussi reconnues comme des standards, à savoir la TEI, l'EAD, BiblioML, et le Dublin Core.

### 3.1. TEI, Text Encoding Initiative<sup>57</sup>

L'objectif qui a donné naissance en 1994 à la TEI était de créer un standard de codage et d'échange de données textuelles numérisées, quelle que soit la langue et le genre de ces données. La TEI à l'origine a été écrite en SGML, car ce langage permet une structuration des données, puis a été convertie en XML. Une de ses fonctionnalités principales est de baliser l'intégralité d'un texte, afin de mettre à jour les différentes caractéristiques de ce texte, de sa structure. Elle est principalement utilisée en sciences humaines, ainsi que pour la description des corpus littéraires, ou encore le traitement des thèses. Cette DTD a fait l'objet d'une simplification, appelée TEILite.

Un document encodé en TEI se compose de deux sections, inclus dans l'élément racine **<TEI.2>** : l'en-tête, annoncé par la balise **<teiHeader>**, et le corps de texte, introduit par la balise **<text>**. Dans cette seconde section, les trois éléments-fils principaux sont **<front>** (description de la couverture), **<body>** (corps de texte), et **<back>** (quatrième de couverture). L'en-tête contient les métadonnées, c'est-à-dire des données à propos d'autres données : l'information bibliographique mais aussi les données qui décrivent le document électronique et en permettent le traitement. Ces informations sont introduites par la balise **<fileDesc>**, qui donne des indications sur le fichier TEI<sup>58</sup>.

Les balises de la TEI se décomposent en trois ensembles :

- Balises obligatoires structurant un socle commun, pour tout type de texte.
- Balises optionnelles, selon le type de texte à encoder (poésie, théâtre, prose, dictionnaire).
- Balises additionnelles et facultatives permettant l'enrichissement de l'encodage (tables, appareil critique ...).

Le schéma de l'ISBD a été pris en compte dans l'élaboration de la DTD. Il existe d'autre part une table de correspondance entre les champs du format MARC et les balises de la TEI. L'Electronic Text Center<sup>59</sup> de l'Université de Virginie propose des textes encodés à l'origine selon la TEI, puis selon la TEILite.

<sup>57</sup> TEI : <http://www.tei-c.org>.

<sup>58</sup> A. Jacquesson, *Bibliothèques et documents numériques*, Paris, 1999, p. 177 et suivantes.

<sup>59</sup> ETC : <http://etext.lib.virginia.edu/>.

L'en-tête TEI de ces documents est automatiquement converti en MARC 21 pour permettre l'enrichissement parallèle du catalogue<sup>60</sup>.

### 3.2. EAD, Encoded Archival Description<sup>61</sup>

L'EAD est un modèle de description adapté aux archives, initialement conçu avec SGML puis adapté à XML, et publié en 1998. L'EAD permet une description à plusieurs niveaux, comme le propose la norme de description archivistique ISAD(G)<sup>62</sup>. Elle permet ainsi d'encoder les instruments de recherche des archives, comme ceux de l'American Memory<sup>63</sup>, projet collectif élaboré sur les années 1990-1994. Cette collection historique regroupe un vaste ensemble de documents appartenant à différentes institutions et de tout type, comme des enregistrements sonores et des images fixes ou animées. L'EAD permet également d'encoder des catalogues de collections. Elle est utilisée par exemple dans le cadre du projet du *Catalogue général des manuscrits des bibliothèques publiques de France*. Elle peut être mise à profit également pour des collections iconographiques structurées en fonds.

La structure d'une instance EAD se compose de trois sections incluses dans l'élément racine **<ead>** : l'en-tête (**<eadheader>**) qui reprend celui proposé par la TEI, les préliminaires (**<frontmatter>**), et la description archivistique à plusieurs niveaux du fonds en question (**<archdesc>**).

Les balises incluses dans ces sections sont optionnelles et n'ont pas d'ordre prédéfini, d'où une grande souplesse d'utilisation.

Selon les concepteurs, en particulier l'Université de Californie à Berkeley et la Bibliothèque du Congrès, cinq fonctionnalités devaient être assurées :

1) Présenter de manière extensive l'information descriptive d'un instrument de recherche archivistique. Un des avantages de l'EAD est de permettre une description du fonds jusqu'à douze niveaux (**<c01>** à **<c12>**), jusqu'à donner accès à l'unité de base, qui peut être le document primaire, par exemple une image numérisée, et cela à l'aide de la balise **<dao>**, digital archival object. Auparavant, le format MARC AMC<sup>64</sup> proposait des notices descriptives de l'ensemble du fonds ; le lien peut être fait à présent avec l'inventaire détaillé.

<sup>60</sup> C. Lupovici, « L'information bibliographique des documents électroniques », *BBF* 1998, t. 43, n° 4, p. 46.

<sup>61</sup> EAD : <http://www.loc.gov/ead/>.

<sup>62</sup> *General International Standard Archival Description*. Il s'agit de la première norme de description archivistique, publiée en 1994 par le Conseil international des archives.

<sup>63</sup> American Memory : <http://memory.loc.gov/ammem/>.

<sup>64</sup> *MARC Archival and Manuscript Control*.

- 2) Préserver les relations hiérarchiques entre les différents niveaux de description.
- 3) Représenter l'information descriptive héritée d'un niveau hiérarchique à un autre. L'EAD évite toute redondance d'information : ce qui est décrit à un niveau donné n'est pas répété aux niveaux inférieurs.
- 4) Possibilité de naviguer au sein de la structure de l'information hiérarchique.
- 5) Possibilité de créer des index facilitant la recherche.

Au-delà de ces fonctionnalités, l'EAD a permis une normalisation des descriptions archivistiques<sup>65</sup>. Réaliser des encodages de sources collectives est devenu envisageable, la base de données Archives Canada France<sup>66</sup> en témoigne. Elle donne accès à la description ainsi qu'à la reproduction de documents concernant la Nouvelle France et conservés par le Centre des archives d'Outre-Mer à Aix-en-Provence, le Centre historique des Archives nationales à Paris et la Bibliothèque et Archives Canada à Ottawa. D'autres établissements sont appelés à coopérer.

Enfin, il est possible d'établir des correspondances avec les formats MARC, notamment au moyen de l'attribut *encoding analog*, qui indique l'équivalent des champs MARC afin de faciliter la conversion des données encodées en EAD. La conversion génère une notice MARC appauvrie. Il est possible également de créer un lien à partir du champ 856 de MARC vers l'inventaire encodé.

### 3.3. BiblioML<sup>67</sup>

Cette DTD<sup>68</sup>, dont la première version date de mai 2000, est le fruit des réflexions menées au Ministère de la Culture et de la Communication sur les nouveaux formats pour élaborer un moyen de gérer et diffuser les références bibliographiques. Ses balises permettent d'intégrer les données descriptives et les accès que propose une notice en UNIMARC. Cette DTD permet ainsi de récupérer de manière optimale les données saisies et enregistrées en UNIMARC. Les éléments directement inclus dans l'élément racine **<BiblioRecord>** reprennent les différents blocs d'UNIMARC. Toutefois la DTD apporte des extensions majeures. En effet, la longueur des données étant illimitée, intégrer des tables des matières, des index, des illustrations, ou même du texte intégral

---

<sup>65</sup> L'EAD est un standard traduit par l'AFNOR : Traduction française de la DTD EAD 2002 : <[http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/EAD%202002\\_Complet\\_20040930.pdf](http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/EAD%202002_Complet_20040930.pdf)>.

<sup>66</sup> Archives Canada France : <<http://www.archivescanadafrance.org/francais/accueil.html>>.

<sup>67</sup> BiblioML : <<http://www.adnx.org/BiblioML>>.

<sup>68</sup> Cette DTD fait l'objet d'une étude plus détaillée dans la suite du mémoire. Cf. Partie 3, ch. 2, § 2.1. "BiblioML" p. 53-55, et annexe 2 p. XV-XXVI.

devient envisageable. La bibliothèque numérique d'histoire régionale du Nord-Pas-de-Calais, Nordnum, a utilisé cette DTD pour intégrer les tables et les index des ouvrages numérisés, et permettre une recherche plein texte dans ces outils de repérage. De plus, pour la réalisation de ce projet, une récupération des notices du SIGB a pu être effectuée. Pour cela, deux étapes de conversion ont été nécessaires : conversion des notices USMARC exportées du SIGB en UNIMARC par l'outil USEMARCON, puis conversion de l'UNIMARC en BiblioML par un outil conçu dans le cadre du projet.

Signalons à propos de ces exemples de DTDs, que les sites et les diverses présentations des bibliothèques numériques n'évoquent souvent que le principe de la structuration en XML, sans donner davantage de précision. D'autre part, de nombreuses réalisations ont élaboré une DTD spécifique, comme dans le cadre du projet d'édition sur le Web des manuscrits de *Madame Bovary* conservés à la Bibliothèque Municipale de Rouen. L'application XML, portant le nom de GUSTAVE-ML permet deux spécificités : le couplage du texte et de l'image, et la description précise des caractéristiques typographiques du manuscrit original<sup>69</sup>. En effet, certaines DTDs, qu'elles soient "maison" ou standards comme la TEI ou l'EAD, permettent de refléter la structure logique du livre ancien, mais aussi d'en intégrer les aspects physiques, ceux du document comme ceux du contenu. Cela provient du fait même de l'utilisation de certaines balises qui définissent de manière bien spécifique la structure de ces modèles XML.

Par ailleurs, des réflexions similaires à celles conduites sur les formats bibliographiques XML sont menées sur les autorités. Ce sont des DTDs en lien avec les DTDS de type bibliographiques. Citons par exemple l'EAC, *Encoded Archival Context*<sup>70</sup>, conçue pour compléter la DTD EAD. Elle est utilisée dans le cadre du projet LEAF, *Linking and Exploring Authority Files*, projet européen visant à faciliter l'accès partagé aux autorités, qu'elles relèvent du monde des archives ou de celui des bibliothèques<sup>71</sup>. EAC, tout comme la norme ISAAR(CPF)<sup>72</sup> permet des correspondances entre notices d'autorités bibliographiques et archivistiques. Autre exemple, Authorities-ML<sup>73</sup>, élaborée par les concepteurs de BiblioML.

<sup>69</sup> Cf. « Le projet Bovary », in *Rencontres Numérisation et Patrimoine*, T. 1, SDN La Rochelle, 2004, p. 9.

<sup>70</sup> EAC : <<http://jefferson.village.virginia.edu/eac/>>.

<sup>71</sup> LEAF : <<http://www.crxnet.com/leaf/>>.

<sup>72</sup> *International Standard Archival Authority Record for Corporate bodies, Persons and Families*.

<sup>73</sup> Authorities-ML : <[http://www.biblioml.org/dtds/authoritiesml\\_030/html/](http://www.biblioml.org/dtds/authoritiesml_030/html/)>.

### 3.4. Dublin Core<sup>74</sup>

Le Dublin Core a été instauré en 1995 par un groupe de réflexion de l'OCLC, travaillant sur les questions du catalogage à la source des documents électroniques. Il est aujourd'hui devenu un standard pour la description de ce type de documents, et a fait l'objet en 2002 d'un schéma XML<sup>75</sup>.

Le Dublin Core, format de métadonnées, a un objectif quelque peu différent des premières DTDs évoquées. Sa conception n'est pas fondée sur l'ISBD, et se compose de 15 éléments, tous facultatifs et répétables. Ces derniers permettent de donner une description minimale consensuelle des documents électroniques pour en faciliter la recherche dans des environnements applicatifs hétérogènes. Les 15 éléments du Dublin Core (titre, créateur, sujet, description, éditeur, contributeur, date, type de ressource, format, identifiant de la ressource, source, langue, relation, localisation, droits) peuvent être classés selon trois types de métadonnées : les métadonnées relatives au contenu, les métadonnées relatives à la propriété intellectuelle, et les métadonnées relatives à l'instance documentaire<sup>76</sup>. Des attributs, utilisables pour chacun de ces éléments, ont été créés par la suite afin d'enrichir ce type de description. Lorsqu'ils sont utilisés, on parle de Dublin Core qualifié. L'un des grands intérêts de ce format de métadonnées, du fait de la simplicité de l'ensemble des éléments, est de permettre aux créateurs de documents électroniques de créer eux-mêmes leurs métadonnées, afin de rendre possible une identification automatique de leurs documents sur Internet.

Le Dublin Core est utilisé notamment dans le format de description des métadonnées RDF, mais aussi dans le cadre de l'OAI, Initiative pour les Archives Ouvertes. Il s'agit d'un protocole d'échange de métadonnées, dénommé OAI-PMH, *Protocol for Metadata Harvesting*<sup>77</sup>, dont la version 2.0 date de 2002. Trois niveaux d'information sont définis : la ressource, l'item (l'ensemble des métadonnées qui identifie de manière unique la ressource) et l'enregistrement (description dans un format spécifique issue d'un item). Deux types de fournisseurs fonctionnent dans le cadre de ce protocole : les *data providers*, fournisseurs de métadonnées, et les *service providers* qui collectent les archives

<sup>74</sup> Dublin Core : <<http://www.dublincore.org/>>.

<sup>75</sup> Simple DC XML Schema version 2002-03-12 : <<http://www.dublincore.org/schemas/xmls/simpledc20021212.xsd>>.

<sup>76</sup> C. Lupovici, « L'information bibliographique des documents électroniques », *BBF* 1998, t. 43, n° 4, p. 45.

<sup>77</sup> Initiative Archives Ouvertes – Protocole pour la Récupération de Métadonnées : <<http://www.openarchives.org/OAI/openarchivesprotocol.html>>.

des *data providers*<sup>78</sup>. Pour cela six types de commandes selon des requêtes HTTP sont formulées, en utilisant le format de métadonnées Dublin Core. L'utilisation d'autres formats de métadonnées comme l'EAD est possible, mais en plus du Dublin Core, obligatoire.

Cette première partie a tenté de clarifier des notions théoriques voire techniques sur les divers formats utilisés dans la profession en les mettant en regard des spécificités du livre ancien. Cela a permis de revenir sur les limites connues des formats actuels tout en précisant les points qui concernent le livre ancien, et de donner quelques pistes sur l'utilité des nouveaux formats pour la mise sur le Web des ressources patrimoniales. Il s'agit à présent de proposer un exemple d'application en XML, centré sur l'intérêt des DTDs par rapport à cet aspect précis qu'est la description bibliographique du livre ancien.

---

<sup>78</sup> S. Aubry, « Archives ouvertes », *in* : *Publier sur Internet, cours INRIA 2004*, Paris, 2004, p. 255.

## ***Partie 2 : Application d'une DTD au fonds de botanique du SCD de l'Université Claude Bernard Lyon1***

Cette partie propose d'approfondir les points théoriques précédents en montrant un exemple d'application de DTD à un corpus bien défini. Lors de l'étude, différentes étapes d'ordre méthodologique se sont succédées : le choix de la thématique et des ouvrages du corpus, le choix de la DTD, l'élaboration des documents XML. Nous proposons en outre quelques remarques sur le logiciel utilisé et les principes de la feuille de style.

### **1. Justification du choix du corpus et de la DTD**

#### **1.1. Choix de la thématique du corpus**

Le fonds ancien de la Bibliothèque des sciences du SCD de l'Université Claude Bernard Lyon 1 comprend 8480 titres, en 12480 volumes d'imprimés, monographies et périodiques. Sa constitution progressive a fait de ce fonds un ensemble disparate. La plupart des ouvrages datent du XIX<sup>e</sup> siècle, le plus ancien de 1516. Le fonds dit "ancien" au SCD comprend cependant tous les ouvrages du XX<sup>e</sup> siècle jusque 1965, date du transfert des livres de sciences de l'ancienne Bibliothèque Centrale, aujourd'hui SICD Lyon 2 – Lyon 3, à la Bibliothèque de la Faculté des sciences. Une réflexion serait d'ailleurs à mener sur la délimitation précise du fonds ancien, prenant en compte, au-delà du critère de la date de parution de la Bibliographie nationale, à savoir 1811, les risques de détérioration rapide du papier pour les ouvrages du début du XX<sup>e</sup> siècle. Cette expertise devrait de plus être réalisée par un spécialiste de l'histoire des sciences.

Le choix du corpus à examiner dans le cadre de cette étude n'allait pas de soi. En effet, l'intérêt du SCD se portait sur plusieurs collections. D'une part, le fonds Gallavardin<sup>79</sup>, fonds d'homéopathie qui inclut une partie importante de la

---

<sup>79</sup> Don par le Docteur E. Gallavardin de la bibliothèque de son père Jean-Pierre Gallavardin (1825-1898), thématiques dominantes : histoire de l'homéopathie, de la médecine, de la pharmacie.

bibliothèque de Sébastien Des Guidi, médecin homéopathe lyonnais<sup>80</sup>. Ce fonds est en cours de numérisation, cependant il ne pouvait convenir pour l'étude car premièrement il relève de la Bibliothèque santé, ce qui compliquait l'organisation du travail, mais surtout, au moment du choix du corpus, il était question d'en numériser les périodiques et non les monographies. Le SCD était également intéressé par l'étude du fonds Claude Bernard, mais ce dernier est trop limité, en nombre d'ouvrages et en variété, pour permettre une étude significative.

Le choix s'est finalement porté sur un ensemble d'ouvrages de botanique. Cette étude s'inscrit ainsi dans le prolongement de la politique de vulgarisation du SCD, puisque plusieurs expositions et publications - antérieures à 1990 - ont été réalisées sur la botanique. Ce thème est encore à l'honneur au SCD lors de la Fête de la science d'octobre 2004<sup>81</sup>. Un projet de bibliothèque numérique serait alors en quelque sorte un aboutissement. Notons par ailleurs que ce corpus se prête particulièrement bien à cette étude du fait de sa relative hétérogénéité, du nombre conséquent d'ouvrages, de sa transversalité sur les bibliothèques sciences et santé, impliquant ainsi l'ensemble du SCD. Signalons également que ces ouvrages sont très souvent illustrés de belles gravures, qu'elles soient en couleur ou non. Enfin, ce corpus ayant été choisi selon une thématique, et de nombreuses réalisations ayant lieu sur ce sujet, des collaborations avec d'autres institutions pourraient être entreprises. Une dernière précision, qui a toute son importance dans le cadre d'un projet de numérisation : ces ouvrages sont libres de droit.

## 1.2. Réception du projet : quelques pistes

Tout projet de mise sur le Web de documents patrimoniaux doit débiter par une étude sur la réception que pourrait avoir la bibliothèque numérique envisagée. En effet, de cela dépendra d'une part la sélection des documents, et d'autre part des fonctionnalités recherchées de cette bibliothèque. De ces dernières résulte le choix d'utiliser telle ou telle application XML.

Dans le cas de notre étude, le public susceptible d'être intéressé concerne tout d'abord les chercheurs en botanique et en histoire des sciences. Si, au-delà des péritextes d'édition les notes manuscrites sont retenues dans la bibliothèque numérique, les historiens de la lecture pourraient être également intéressés. Comme

---

<sup>80</sup> Sébastien Des Guidi (1769-1863). Cf. la notice de la Base *Provenances des livres anciens* de la Bibliothèque municipale de Lyon : <<http://sgedh.si.bm-lyon.fr/dipweb2/apos/possesseurs.htm>>.

<sup>81</sup> Thématique générale : *Des expéditions et des plantes* déclinée à la Bibliothèque des sciences sous le titre "Voyages et botanique", avec exposition de livres anciens et conférences.

l'a montré la conférence de M. Pinault-Sorensen lors de la Fête de la science, intitulée « Voyages et botanique », les historiens de l'art pourraient être satisfaits par un corpus de botanique richement illustré de planches hors-texte. Seraient évidemment concernés les étudiants en sciences, public premier du SCD, qui devraient pouvoir s'intéresser à l'histoire de leur discipline. Mais le SCD peut aussi envisager de toucher un public plus vaste, comme les étudiants en épistémologie. Ce projet serait une opportunité, vis-à-vis des chercheurs en sciences, pour donner davantage de crédit au livre ancien de sciences. En effet, pour ce type de chercheurs, l'historicité d'un livre de sciences le rend avant tout obsolète.

Quant au grand public, il requiert une politique particulière. En effet, la vulgarisation, malgré la beauté de certains de ces ouvrages, ne peut se faire spontanément. Ne serait-ce que du point de vue du vocabulaire scientifique, difficulté amplifiée par le fait qu'un certain nombre d'ouvrages sont en latin ou en allemand.

Contenter tous ces publics est difficilement envisageable, du moins les satisfaire à part égale. Les chercheurs et les étudiants dans ces disciplines - ce qui est déjà différent du fait d'un souci de pédagogie plus important -, sont privilégiés, car c'est la vocation première d'une bibliothèque universitaire.

### **1.3. Choix des ouvrages du corpus : méthodologie**

Le corpus retenu, du fait de la thématique, peut être considéré comme une collection. Cependant, dans le temps imparti, il n'était pas envisageable de la traiter dans son ensemble. Une sélection d'ouvrages a donc été entreprise.

Cette dernière s'est effectuée en plusieurs étapes. En effet, comme tout fonds ancien rangé en magasin, les ouvrages sont classés par format et par ordre d'entrée, portant ainsi un numéro d'inventaire comme cote. Ils ne sont donc aucunement organisés selon les sujets traités. Les étapes de la sélection furent les suivantes :

- visites avec la responsable du fonds ancien pour une première connaissance du fonds et des locaux,
- recherches à partir du fichier papier matières<sup>82</sup>,
- recherches dans le module recherche professionnelle du SIGB du SCD<sup>83</sup> pour avoir accès aux index professionnels (sujets, auteurs, cotes) et éditer des listes<sup>84</sup>,

---

<sup>82</sup> Le catalogue sur fiches papier (auteurs et matières) est complet, contrairement au catalogue informatisé : 8039 notices ont été rétroconverties par OCLC sur les 8480 titres en 12480 volumes. Ce catalogue papier a été réalisé dans les années 60 et est donc antérieur à toute normalisation. Cf. L. Richard, *Évaluation et mise en valeur d'un fonds ancien scientifique ...* Villeurbanne, Enssib, 1994.

<sup>83</sup> Horizon, version 7.2, Société Dynix.

- comparaison avec la documentation culturelle produite sur le sujet<sup>85</sup>,
- liste des ouvrages en réserve,
- enfin, les notices des ouvrages de la liste obtenue ont été rapidement examinées afin de pouvoir envisager les différents cas particuliers rencontrés lors des descriptions bibliographiques : titres parallèles, titres collectifs, traduction d'une œuvre originale, ouvrage en plusieurs volumes ..., mais aussi des données d'exemplaires particulières autre que la reliure.

Les ouvrages sélectionnés<sup>86</sup>, à savoir une centaine, recouvrent donc notamment les différents siècles de la production imprimée - du XVI<sup>e</sup> siècle au début du XX<sup>e</sup> siècle, avec une majorité d'ouvrages des XVIII<sup>e</sup> et XIX<sup>e</sup> siècles. Cette sélection, si le projet de bibliothèque numérique se concrétise, devrait être soumise à la validation de chercheurs spécialistes de botanique et d'histoire des sciences. A partir de cette liste, une vingtaine d'ouvrages, principalement de la réserve, ont fait l'objet d'une sélection affinée<sup>87</sup>, et ont été traités avec la DTD choisie.

#### 1.4. Choix de la DTD

L'étude porte sur l'intérêt des DTDs pour la description bibliographique d'ouvrages qui constituent ici une collection de structure simple et non hiérarchisée. Les intérêts principaux d'EAD<sup>88</sup> n'étaient pas opérationnels pour cette collection thématique qui ne nécessite pas de description imbriquée sur de nombreux niveaux. Quant à la TEI, son avantage le plus significatif consiste à baliser le corps de texte des ouvrages. Une brève étude des principaux éléments qu'elle propose a montré qu'ils étaient trop sommaires pour envisager de décrire le corpus de manière approfondie. En TEI, la description bibliographique proprement dite est répartie principalement entre les éléments **<teiHeader>**, qui permet une brève description du document encodé, et **<titlePage>** qui transcrit toutes les informations que l'on peut trouver sur une page de titre, source principale d'information.

Le choix s'est rapidement porté sur BiblioML. Cette DTD, adaptation d'UNIMARC au langage XML, est idéale pour mieux affiner l'étude des limites du format MARC, et permettre une comparaison entre ce format et les fonctionnalités proposées par le langage XML. D'autre part, BiblioML est intéressante en ce qui concerne la récupération des données des notices bibliographiques, plus complète que les

---

<sup>84</sup> Requête SQL.

<sup>85</sup> Notamment le catalogue de l'exposition réalisée à la Bibliothèque des sciences *Livres illustrés de botanique*, 1978.

<sup>86</sup> La liste des ouvrages sélectionnés se trouve en annexe 1-1 p. VIII.

<sup>87</sup> La liste des ouvrages traités figure en annexe 1-2 p. IX.

<sup>88</sup> Cf. Partie 1, ch. 3, § 3.2. "EAD" p. 31-32.

conversions qui peuvent être obtenues à partir de documents encodés en TEI ou EAD. En effet, depuis quarante ans des notices sont structurées à l'aide des formats MARC, et il n'est pas question de revenir sur le travail ainsi réalisé. Cependant, au-delà du traitement des ouvrages du corpus, et consciente que BiblioML n'est plus actuellement développée par ses concepteurs, une comparaison sera menée avec d'autres formats bibliographiques XML<sup>89</sup>.

Enfin, précisons que l'étude porte davantage sur la description bibliographique proprement dite, que sur les notices d'autorité, cela pour des questions de temps imparti. Si les accès ont été mentionnés dans la description, l'étude de DTDs permettant de créer des notices d'autorité n'a pas été menée.

Nous avons vu que la sélection des ouvrages devait être soumise à validation si le projet de numérisation est poursuivi. Il en est de même pour le choix de la DTD. Les objectifs de l'étude ont fait que le choix s'est porté sur BiblioML, mais l'analyse des publics visés et les fonctionnalités souhaitées qui en découlent devront conditionner le choix du modèle XML, au-delà des pistes avancées dans cette étude.

## 2. Elaboration des documents structurés

L'élaboration des documents peut être résumée en deux étapes : l'appropriation de la structure de la DTD, et le "catalogage livre en main" proprement dit, avec la consultation des ressources nécessaires.

### 2.1. Construction de la DTD

Une fois la DTD choisie, il convient de se l'approprier. Pour cela, outre la connaissance du langage XML nécessaire pour déchiffrer le texte de la DTD<sup>90</sup>, il a fallu se faire une idée de la forme que peut prendre la structure d'une description complète de livre ancien encodée avec BiblioML<sup>91</sup>. Cette DTD se présente en effet dans l'ordre alphabétique des éléments, à la différence des formats MARC qui reprennent la structure de l'ISBD pour lister leurs blocs, champs et sous-champs qui les composent. En revanche, MARC DTD<sup>92</sup> par exemple, standard de la Bibliothèque du Congrès, se présente selon l'ordre des champs MARC. Il est vrai qu'une fois la DTD référencée

---

<sup>89</sup> Cf. Partie 3, ch. 2, § 2.2. "Autres formats bibliographiques XML" p. 55-58.

<sup>90</sup> Fichier texte de la DTD : <[http://www.biblioml.org/dtds/biblioml\\_030/biblioml\\_030.dtd](http://www.biblioml.org/dtds/biblioml_030/biblioml_030.dtd)>.

<sup>91</sup> Il existe bien des notices proposées dans les documents explicatifs des pages sur BiblioML du site du Ministère de la Culture, mais ces notices sont partielles par rapport à l'objectif visé ici.

<sup>92</sup> Cf. Partie 3, ch. 2, § 2.2.1. "Formats dérivés de MARC" p. 55-56.

dans le prologue, l'éditeur XML donne la structure générale de la DTD. Mais, en attendant l'installation de l'éditeur, des tests préalables dans le Bloc-Notes<sup>93</sup> ont été réalisés, permettant de réaliser un important travail de construction et donc de familiarisation.

Les descriptions bibliographiques encodées réalisées avec l'éditeur XMetaL figurent en annexe<sup>94</sup>.

Pour la construction de ces documents, nous avons au préalable comparé les solutions proposées par les deux formats que sont UNIMARC et BiblioML. Nous en donnons ici une synthèse, en indiquant les données spécifiques de la description bibliographique du livre ancien, selon les deux types d'information secondaire que sont la description bibliographique proprement dite et les points d'accès. Pour des raisons de commodité, la présentation de la description bibliographique respecte l'ordre des différentes zones de l'ISBD(A). Voici tout d'abord un rappel général sur ces différentes structures.

### **Structure de l'ISBD(A)**

- Zone 1 : Titre et mention de responsabilité
- Zone 2 : Edition et histoire bibliographique de la publication
- (Zone 3 : Zone particulière à certains types de documents, numérotation : non utilisée)
- Zone 4 : Adresse
- Zone 5 : Collation
- Zone 6 : Collection
- Zone 7 : Notes
- Zone 8 : Empreinte
- Points d'accès : auteur personne physique, auteur collectivité, titres, et sujets

### **Structure d'UNIMARC**

- 0XX : Bloc des numéros d'identification
- 1XX : Bloc des informations codées
- 2XX : Bloc des informations descriptives
- 3XX : Bloc des notes

---

<sup>93</sup> Et non dans Word, afin d'obtenir un fichier ASCII pur.

<sup>94</sup> Cf. Annexe 4 p. XXXIII-XCIII.

- 4XX : Bloc des liens avec d'autres notices
- 5XX : Bloc des titres associés
- 6XX : Bloc de l'indexation matière
- 7XX : Bloc de la responsabilité intellectuelle
- 8XX : Bloc des données internationales
- 9XX : Bloc des données locales

## Structure de BiblioML

Comme toute DTD, BiblioML est constituée d'un élément racine, appelé ici **<BiblioRecord>**. Cet élément-racine comporte 10 éléments-fils.

1. **<Meta>** : Métadonnées administratives
2. **<Description>** : Métadonnées descriptives
3. **<CodedValues>** : Données codées
4. **<Notes>** : Notes sur l'ouvrage décrit
5. **<Relationships>** : Liens entre notices et titres associés
6. **<Subjects>** : Accès sujets
7. **<LocalData>** : Données locales
8. **<Copies>**
9. **<Tables>**
10. **<Text>**

Comme BiblioML a été conçue en fonction d'UNIMARC, la liste ci-après précise de quels blocs d'UNIMARC sont tirées les informations contenues dans les éléments de BiblioML.

1. **<Meta>** : Bloc 0XX, ainsi que quelques informations codées du bloc 1XX, et les données internationales du bloc 8XX
2. **<Description>** : Bloc 2XX et accès du bloc 7XX, quelques informations codées du bloc 1XX (langue du document, pays d'édition ou de production), et quelques informations du bloc 0XX (numéros d'identification)
3. **<CodedValues>** : Bloc 1XX
4. **<Notes>** : Bloc 3XX
5. **<Relationships>** : Bloc 4XX et 5XX
6. **<Subjects>** : Bloc 6XX
7. **<LocalData>** : Bloc 9XX

- 8. **<Copies>**
- 9. **<Tables>** : Extension par rapport à UNIMARC
- 10. **<Text>** : Extension par rapport à UNIMARC

De cette mise en regard, nous pouvons tirer quelques constats :

- Le format UNIMARC est présenté par blocs : zones codées, description bibliographique selon la structure de l'ISBD, puis accès. La DTD BiblioML se présente quant à elle selon l'ordre alphabétique des éléments. Cependant, dans le texte de la DTD, les éléments principaux de la structure de la DTD sont tous mentionnés dans la description de l'élément-racine **<BiblioRecord>**, et chaque description d'élément donne la liste des éléments-fils qui leur correspondent.
- Notons à l'évidence qu'il existe une différence de nature entre la structure des deux formats : UNIMARC est linéaire alors que BiblioML, comme tout document XML, propose une structure en indentation, ici sur cinq niveaux.
- Les contenus sont davantage transparents dans la DTD grâce aux noms d'éléments et d'attributs donnés sous forme textuelle signifiante, à la différence des étiquettes MARC codées selon trois caractères numériques annonçant les champs, et des \$ séparateurs de sous-champs, suivis d'un caractère alphanumérique.
- Ces deux formats font souvent référence aux mêmes codes. Certains trouvent leur origine dans le format MARC, par exemple "ba" signifie que le titre du document est en alphabet latin, d'autres sont normalisés, comme les codes de noms de pays selon la norme ISO 3166-1<sup>95</sup>. Ainsi, lors du travail avec BiblioML, en particulier pour renseigner les codes des valeurs des attributs, la connaissance d'UNIMARC est nécessaire, ainsi que la mise à disposition d'un manuel.
- Bien que BiblioML soit la traduction en langage XML du format UNIMARC, la correspondance ne se fait pas toujours entre les balises principales et les blocs. Certaines balises regroupent les informations données dans des blocs différents, et certaines informations de certains blocs sont réparties dans plusieurs balises. C'est notamment le cas des données codées<sup>96</sup>.
- Enfin, notons qu'il n'y a pas de correspondance exacte entre les champs, sous-champs, indicateurs du format UNIMARC, et les éléments et attributs de la DTD BiblioML. Par exemple, l'information d'un sous-champ peut se trouver soit dans un élément, soit dans un attribut<sup>97</sup>.

---

<sup>95</sup> "Codes pour la normalisation des noms de pays". Cf. IFLA, *Manuel UNIMARC*, München, 1999, p. 366.

<sup>96</sup> Cf. Annexe 2-2 "Tableaux de comparaison UNIMARC-BiblioML" p. XIX-XXVI.

<sup>97</sup> Cf. Annexe 2-2 pour des exemples.

## Tableau de synthèse des correspondances

ISBD(A) Zone	UNIMARC Bloc et champ	BiblioML Elément-père, élément-fils
Zone 1 : Titre et mention de responsabilité	Bloc des informations descriptives 2XX, champ 200	<Description> <Title and Responsibility>
Zone 2 : Edition	Bloc des informations descriptives 2XX, champ 205	<Description> <EditionGroup>
Zone 4 : Adresse	Bloc des informations descriptives 2XX, champ 210	<Description> <PublicationGroup>
Zone 5 : Collation ou description technique	Bloc des informations descriptives 2XX, champ 215	<Description> <PhysicalDescription>
Zone 6 : Collection	Bloc des informations descriptives 2XX, champ 225	
Zone 7 : Notes	Bloc des notes 3XX	<Notes>
Zone 8 : Empreinte	Bloc 0XX, champ 012	<Description> <IdentificationNumbers> <FingerprintID>
	Bloc des informations codées 1XX	<CodedValues>, mais aussi <Meta> et <Description>
Accès : auteurs, titres, sujets	- Bloc des titres associés 5XX - Bloc de l'indexation-matière 6XX - Bloc de la responsabilité intellectuelle 7XX	- <RelatedTitles> - <Subjects> - <Description> <IntellectualResponsibility>
	- Bloc 8XX des données internationales	- <Meta>

## 2.2. Ressources bibliographiques pour la description des ouvrages du corpus

Comme lors de tout catalogage de livres anciens, que ce soit sur fichier papier, en MARC, ou dans notre cas avec une DTD, des ressources bibliographiques sont nécessaires. La liste des ressources utilisées se trouve en annexe<sup>98</sup>.

L'étape de la construction de la DTD peut paraître artificielle, mais dans le cadre d'une construction de bibliothèque numérique, il ne faut pas oublier le paramétrage et l'adaptation nécessaires. L'optique est différente dans le cas du catalogage, car le métier reste alors inchangé.

## 3. Retour d'expérience : éditeur XML et feuilles de style

Cette étude a nécessité l'utilisation d'un éditeur XML dont nous présentons le fonctionnement, mais elle requiert aussi un prolongement de traitement par la construction d'une feuille de style, qui conditionnera la présentation des données.

### 3.1. Un éditeur XML : XMetaL

XMetaL Developer 4 propose trois outils XML : XMetaL Author, XMetaL Developer, et XMetaL for Active X. Nous avons travaillé avec XMetaL Author<sup>99</sup>, éditeur XML. Il est vrai que puisqu'un document XML est du texte, tout éditeur de texte permet de construire un document XML. Cependant, les éditeurs XML offrent de nombreuses facilités, notamment grâce à leur signalisation des balises, à l'intégration d'un parseur vérifiant le respect des contraintes de bonne formation et la validité des documents, et à l'application d'une feuille de style par référencement dans la déclaration du document.

XMetaL permet de créer des documents en SGML ou en XML, de travailler avec des DTDs (fichier.dtd) ou des schémas (fichier.xsd). Il permet d'utiliser les feuilles de style tant CSS que XSLT<sup>100</sup>.

L'écran de travail de l'éditeur XMetaL se présente sous la forme de quatre fenêtres interactives de travail, dont le contenu et l'apparence peuvent être adaptés<sup>101</sup> :

---

<sup>98</sup> Cf. Annexe 1-3 p. X-XIII.

<sup>99</sup> XMetaL Author 4.5, société Corel. Version d'évaluation de 30 jours. Editeur utilisé en anglais <<http://www.XMetaL.com/trials/XMetaL-author-4.5.9.016-eval.exe>>.

<sup>100</sup> Cf. paragraphe suivant p. 48-49.

<sup>101</sup> Cf. Annexe 3-2 "Présentation d'un nouveau document" p. XXIX, copie d'écran.

- La structure XML dans une colonne à gauche, fonctionnant sur le même principe qu'un explorateur de documents.
- La fenêtre de saisie.
- La liste des éléments dans une fenêtre à droite. Cette liste s'affiche en cliquant dans le menu sur Insert – Element, ou View – Element List.
- Les attributs correspondants à l'élément choisi dans une fenêtre à droite, qui s'affiche en cliquant dans le menu sur View – Attribute Inspector.

Pour créer un document XML qui soit une instance d'une DTD existante externe, il faut référencer cette DTD dans la déclaration du document. Dans notre cas, il s'agit d'enregistrer sur le poste de travail le fichier qui correspond au texte de la DTD BiblioML. Lors du téléchargement à partir du site de BiblioML, ce fichier a été nommé biblioml\_030.dtd.

Lorsque l'éditeur est ouvert, cliquer sur :

Tools

Options

File – choose by default

Et choisir comme fichier par défaut biblioml\_030.dtd.

Cette opération permet, lors de la création d'un nouveau document XML, le référencement à la DTD, et par conséquent l'affichage automatique des éléments et de leurs attributs dans la colonne de droite.

Pour toute création de document, il suffit ensuite de cliquer sur

File

New

Blank XML Document<sup>102</sup>.

Choisir ensuite le fichier.dtd qui correspond à la structure générique recherchée.

Le choix Blank Well-formed XML Document permet en revanche de créer un document XML sans référence à une DTD.

Le nouveau document s'ouvre alors avec la liste des éléments de la DTD choisie dans la fenêtre de droite. La déclaration est construite automatiquement. Il est nécessaire de connaître la DTD utilisée, notamment son élément-racine et les éléments-fils principaux qui en découlent. En effet, tout comme le texte d'une DTD, les éléments sont listés dans l'ordre alphabétique et l'indentation n'est pas immédiatement

---

<sup>102</sup> Cf. Annexe 3-1 "Ouverture d'un nouveau document" p. XXVIII, copie d'écran.

reflétée. De plus, renseigner les attributs de certains éléments nécessite de connaître leur signification et les codes adéquats. Pour insérer un élément, il suffit de double-cliquer dessus dans la liste, et les attributs correspondants s'affichent automatiquement dans la fenêtre des attributs. Si l'élément est saisi manuellement, la balise est invalide. Insérer l'élément-racine n'entraîne pas l'insertion des principaux éléments-fils. Par exemple, dans le cas de BiblioML, et sans paramétrage préalable de l'éditeur, seules les balises **<Meta>** et **<Description>**, avec quelques uns de leurs éléments-fils, sont insérées automatiquement lorsque l'on clique dans la liste des éléments sur **<BiblioRecord>**, élément-racine<sup>103</sup>.

La saisie des données textuelles contenues dans les éléments se fait ensuite manuellement dans la fenêtre de saisie, à l'intérieur des balises sélectionnées. La fenêtre de saisie propose trois affichages :

- « Plain Text View » : document XML "pur", qui seul permet de voir la déclaration,
- « Tags On View »<sup>104</sup>, où les balises sont signalées graphiquement,
- « Normal View »<sup>105</sup>, où le texte saisi s'affiche sans les balises.

Le choix entre ces trois modes d'affichage s'effectue grâce aux onglets situés en bas à gauche de la fenêtre de saisie. Seul le premier mode permet d'afficher les attributs et leurs valeurs. Dans les deux autres modes, un losange bleu apparaît pour signaler leur présence, sans délivrer aucune information sur le contenu.

Il est conseillé aux débutants de travailler avec les balises signalées graphiquement, et de relire ensuite leur document pour plus de clarté avec les données saisies sans les balises. Il est possible d'afficher plusieurs documents, en cliquant dans le menu sur View - Workbookmode.

Toute sauvegarde entraîne la mise en route du processus de validation du document par le parseur, opération équivalente si l'on clique sur l'icône de validation, icône de texte surmontée d'un V bleu. Un écran apparaît alors, soit indiquant que le document a été validé avec succès, soit listant les erreurs qui empêchent le document d'être validé<sup>106</sup>. Il est possible alors de placer directement le curseur sur l'endroit où se trouve l'erreur en cliquant sur « Go to ».

Une aide est disponible en appuyant sur la touche F1.

Dans le cadre précis de ce travail, à savoir l'élaboration de notices bibliographiques détaillées, le temps de construction est assez long, et si l'on choisit de

<sup>103</sup> Cf. Annexe 3-2 "Présentation d'un nouveau document" p. XXIX, copie d'écran.

<sup>104</sup> Cf. Annexe 3-4 "Mode d'affichage TagsOnView" p. XXXI, copie d'écran.

<sup>105</sup> Cf. Annexe 3-3 "Mode d'affichage NormalView" p. XXX, copie d'écran.

<sup>106</sup> Cf. Annexe 3-5 "Erreur de validation" p. XXXII, copie d'écran.

procéder par copier-coller, cela demande de longues vérifications, balise par balise, attribut par attribut, pour ne pas que cela soit source d'erreurs ou d'oublis. Car il faut préciser ici que ces derniers ne sont pas nécessairement remarqués par le parseur lors de la validation, notamment les erreurs de description bibliographique.

### 3.2. Les feuilles de style

Nous avons vu que XML ne permettait pas la mise en forme des données qu'il structure, en particulier leur disposition et la typographie. C'est le rôle de la feuille de style de transformer et de publier ces données, en indiquant les modalités de structuration physique du contenu encodé préalablement au moyen du modèle de document, DTD ou schéma XML.

Il existe deux types de feuilles de style applicables à XML et qui font l'objet d'une recommandation du W3C : CSS<sup>107</sup> et XSL<sup>108</sup>. CSS est d'utilisation relativement basique, en revanche XSL relève du langage de programmation.

XSL est un langage de feuilles de style, qui se compose de trois langages : XPath, langage d'adressage dans l'arbre XML, XSL-T, langage de transformation, et XSL-FO, un langage de formatage exploitant XSL-T. Pour expliquer le principe de XSL-T, il est nécessaire auparavant d'évoquer XPath en quelques mots, car il est à la base du langage de transformation d'arbres qu'est XSL-T. XPath est un langage d'adressage dans l'arbre XML, c'est-à-dire qu'il permet d'exprimer des chemins à l'intérieur de cet arbre. Selon XPath, un arbre XML se compose de sept types de nœuds, reliés par des relations<sup>109</sup> :

- le nœud racine
- le nœud élément
- le nœud texte
- le nœud attribut
- le nœud instruction de traitement
- le nœud déclaration du domaine nominal
- le nœud commentaire

Le langage XSL-T<sup>110</sup> est un langage déclaratif de manipulation d'arbres XML. Il permet d'obtenir différents formats de sorties : PDF, page web ... selon le type d'utilisation envisagé. Par le biais d'une feuille de style écrite selon ce langage, qui est elle-même un document XML nécessitant une construction, il devient possible de

---

<sup>107</sup> *Cascading StyleSheet.*

<sup>108</sup> *eXtensible Stylesheet Language. Recommandation XSL, version 2.0, octobre 2001 : <<http://www.w3.org/Style/XSL/>>.*

<sup>109</sup> A. Michard, *XML : langage et applications*, Paris, 2001, p. 165.

<sup>110</sup> *XSL Transformation, version 1.0, novembre 1999 : <<http://www.w3.org/TR/xslt>>.*

transformer un arbre source en arbre résultat. La feuille de style spécifie, pour chaque nœud rencontré dans l'arbre source, la structure à créer dans l'arbre résultat. L'élément-racine du fichier XSL-T est **<stylesheet>**, dont l'élément fils **<template>** définit les deux règles régissant la transformation d'arbre :

- La règle de sélection d'un fragment de l'arbre source selon des expressions XPath, définie par l'attribut *match*.
- La règle de création du fragment de structure à créer dans l'arbre résultat, donnée par le contenu de l'élément **<template>**. Cette règle génère la création des balises et attributs dans l'arbre résultat, et copie le contenu des éléments de l'arbre source dans l'arbre résultat.

Une même feuille de style peut s'appliquer à tous les documents conformes à une même DTD. Elle doit être référencée dans le prologue du document par une commande du type `<?xml-stylesheet type="text/xsl" href="fichier.xsl"?>`.

Il n'existe pas de feuille de style créée par les développeurs de BiblioML pour les utilisateurs de cette DTD, contrairement à l'EAD ou à la TEI<sup>111</sup>. XMetaL, lors du référencement à la DTD, a créé automatiquement deux feuilles de style CSS, mais dont le résultat ne diffère en rien de l'affichage proposé par le navigateur. D'autre part, une feuille de style doit être adaptée ou créée en fonction des besoins du public visé, et du format de sortie choisi.

Ce chapitre entendait montrer, au-delà de la nécessaire coopération entre bibliothécaires et informaticiens, les quelques obstacles de type informatique que nous pouvons rencontrer. En effet, des questions similaires doivent se poser pour tous ces types de projet, et des collaborations autour de discussions d'ordre technique sont les plus bénéfiques. Simplement, que ce soit pour le projet du stage ou dans un cadre plus large, ne figurent ici que les prémisses, et les développements sont à poursuivre.

---

<sup>111</sup> Le site web de la TEI propose des liens vers des feuilles de style XSL-T : <http://www.tei-c.org/Stylesheets/>.

## ***Partie 3 : De la description bibliographique à l'échange de métadonnées***

La présentation de l'étude menée, d'ordre essentiellement méthodologique, était nécessaire pour en clarifier la démarche et en illustrer les objectifs. Dans une perspective de bibliothèque numérique, nous proposons à présent d'en tirer les conclusions et d'élargir l'étude de l'intérêt des DTDs au-delà de l'aspect description bibliographique, en évoquant les échanges et les questions de normalisations qui en découlent, facettes non négligeables du contexte du Web.

### **1. Mise en perspective**

Nous voulons tout d'abord mettre en perspective les conclusions auxquelles peut aboutir cette étude. Deux questions principales sous-tendent les projets de numérisation : la constitution des bases de données et les moyens d'accéder aux informations qu'elles contiennent.

#### **1.1. Bases de données**

La numérisation ne consiste pas seulement à donner accès à la version électronique des documents, comme entendait le montrer cette étude. Il est effectivement possible de décomposer un projet de numérisation en quatre étapes :

- Numérisation
- Stockage et nommage des fichiers
- Description et structuration des documents
- Diffusion

L'étude menée correspond à la troisième phase, à savoir la description et la structuration des documents, avec l'élaboration de bases de données associées. Il existe ainsi tout un travail de fond à mener conjointement sur les métadonnées et sur le document primaire lui-même.

Les formats XML, qu'ils soient bibliographiques ou non, permettent d'envisager une reprise des données grâce à la conversion des notices MARC existant dans le catalogue du SIGB. Des outils permettent des migrations du catalogue vers la base de données de documents structurés en XML. Nous avons

évoqué le cas de BiblioML et un exemple d'outil comme USEMARCON dans le cadre du projet NordNum<sup>112</sup>. La Bibliothèque du Congrès propose également des outils de conversion qui correspondent aux standards<sup>113</sup> dont elle est responsable des développements.

Au-delà de cette reprise de données, l'intégration du document primaire peut être partielle, avec par exemple les illustrations, le traitement des index et des tables des matières, ou complète. Mais surtout, l'information secondaire peut être enrichie, par la conception de légendes pour les illustrations, et par l'élaboration d'introductions thématiques, de bibliographies, de listes de signets, de commentaires ... réalisés par des spécialistes des disciplines en question. Ce type d'enrichissement est primordial, en particulier dans une perspective de vulgarisation, car il présente le contexte des documents numérisés. Ces réalisations ont déjà été éprouvées par des bibliothèques numériques comme le CNUM, Conservatoire Numérique des Arts et Métiers<sup>114</sup>. Mais, la prise en compte du contexte peut aller plus loin, en offrant aux chercheurs la possibilité de créer leurs propres métadonnées : notes historiques, codicologiques, paléographiques ...<sup>115</sup>

## 1.2. Modes de recherche, navigation et indexation

### 1.2.1. Navigation

A chaque type de public correspondent des chemins de navigation particuliers, sur le site ou au sein d'un document numérisé : navigation superficielle, sélective, ou navigation plus approfondie, guidée par des modes de recherche diversifiés comme la recherche plein texte ou l'indexation par mots-clé. Il est nécessaire pour ces raisons de mener une réflexion sur ce que peut apporter XML dans la présentation de l'interface Web. En effet, les bibliothèques numériques patrimoniales se présentent souvent comme des outils de vulgarisation, avec la volonté de mettre à disposition du grand public des collections jusque-là réservées aux initiés. Mais qu'en est-il réellement ? Lorsque l'on visite ces bibliothèques en ligne, on se rend compte de la difficulté, pour celui qui n'est pas familiarisé avec les contenus, à naviguer entre les pages, les images et les textes. Il faut souvent savoir ce que l'on cherche, avoir une idée des mots à

---

<sup>112</sup> Cf. le site de BiblioML, page évoquant des outils de transcodage d'UNIMARC en XML : <<http://www.biblioml.org/fr/conversion.html>>

<sup>113</sup> Library of Congress, Standards : <<http://www.loc.gov/standards/>>.

<sup>114</sup> CNUM, bibliothèque numérique du CNAM : <<http://cnum.cnam.fr/>>.

<sup>115</sup> F. Role, « Métadonnées et documents anciens », *Document numérique : documents anciens*, vol. 3, n° 1-2, juin 1999, p. 135-150.

saisir dans les champs qui vont permettre l'interrogation du document, et cela malgré la présence de listes alphabétiques. Ne serait-ce que parce que les pratiques courantes de recherche sur le Web sont celles du langage naturel, favorisent la recherche plein texte, et diffèrent de la sorte de nos pratiques de bibliothécaires. Les efforts de présentation du contexte des documents, associés à des liens hypertextes, peuvent permettre une amélioration des chemins de navigation, en rendant le butinage plus aisé et significatif, au sens d'un parcours riche de sens.

En revanche, s'adresser aux chercheurs par le biais de bibliothèques numériques commence à être éprouvé, même si un enrichissement des fonctionnalités est à poursuivre, comme nous l'avons vu dans le paragraphe précédent. La bibliothèque numérique Colisciences par exemple, propose, au-delà de la diffusion d'un contenu culturel et patrimonial, une réflexion sur l'hypertextualité, dans un but clairement annoncé comme scientifique. Des liens sont conçus pour donner accès à des glossaires et des index, et les parcours de lecture sont étudiés.

### 1.2.2. Indexation

Au-delà des pratiques de navigation, les modes de recherche qui permettent d'accéder aux informations contenues dans ces bases de données sont conditionnés par la question des points d'accès et donc de l'indexation. Rappelons que « les points d'accès constituent la véritable valeur ajoutée [à la description bibliographique]. Ils sont contrôlés par des thesaurus ou des formes d'autorités pour les noms propres d'auteurs ou de collectivités, les noms géographiques ou les descripteurs »<sup>116</sup>. Ces points d'accès étaient déjà mentionnés sur les fiches papier, et le format MARC les avait intégrés. Ces données peuvent être reprises en XML. Par exemple, avec BiblioML, l'attribut *AuthorityRecord* permet de référencer la notice d'autorité.

Certains points d'accès sont primordiaux voire spécifiques au livre ancien, et ces informations ne devraient pas être abandonnées lors du passage au contexte du Web : les auteurs, indexés également pour les données d'exemplaire dans les cas fréquents des relieurs, possesseurs ..., l'indexation-matière, la ville d'édition ... Les accès titres sont primordiaux, du fait des difficultés de transcription de la page de titre<sup>117</sup>. Plusieurs accès titres sont envisageables : le titre uniforme, qui permet de regrouper toutes les versions et les variantes des titres d'une même oeuvre en créant un point d'accès unique, notamment pour les classiques, les

<sup>116</sup> C. Lupovici, *id.* p. 42.

<sup>117</sup> Cf. Partie 1, ch. 1, § 1.1.3. "Les spécificités de la description bibliographique du livre ancien" p. 14.

anonymes et les textes des Ecritures, la rubrique de classement, le titre de forme, et les variantes de titre qui permettent d'indexer les différentes possibilités de transcription de la graphie.

Le choix éventuel de renseigner certains points d'accès secondaires et la constitution de fichiers d'autorités associés, peut être lié à celui de privilégier telle ou telle spécificité du corpus. C'est le cas de la base européenne de livres anciens HBP, qui a développé un thésaurus regroupant les différentes formes des noms de lieux d'édition. Ce travail d'indexation doit permettre de faciliter l'opération d'identification, notamment en travaillant sur une présentation simplifiée et normalisée des données.

Cette mise en perspective montre que XML facilite la production des données, par leur conversion issue de MARC, par l'intégration de données "océrisées", mais aussi tout simplement parce que leur diffusion peut être simplifiée grâce au Web, après installation d'un outil d'exploitation et de diffusion des données<sup>118</sup>. Nous pouvons à présent nous attarder sur les conclusions générales de cette étude, relatives à la description bibliographique et aux questions d'échanges de données.

## 2. Intérêts des formats bibliographiques XML

Les conclusions sur l'application de la DTD BiblioML seront complétées par l'examen d'autres formats XML de type bibliographique, pour en venir aux conclusions sur l'intérêt des DTDs pour la description bibliographique du livre ancien.

Nous appelons ici formats bibliographiques XML des DTDs ou des schémas qui ont pour principal objectif de renseigner l'information secondaire qu'est la description bibliographique.

### 2.1. BiblioML

BiblioML apporte des extensions au format UNIMARC, dont voici les plus significatives :

- Il est possible d'encoder le texte du document en question, inclus dans l'élément **<Text>**. Cet élément reprend un sous-ensemble de la TEI.

---

<sup>118</sup> Cf. la plate-forme SDX- Système Documentaire XML - de logiciels libres développée par le Ministère de la Culture et de la Communication pour la publication sur le Web de documents XML : <http://adnx.org/sdx/>.

- Il est également envisageable d'intégrer des tables et des index, au moyen de la balise **<Tables>**, puis **<TableOfContent>** ou **<Index>**. Il s'agit d'une nette progression par rapport à la possibilité de référencer une table des matières au moyen du champ 359 puisqu'il n'y a plus de limitation de niveaux d'arborescence.
- En ce qui concerne les liens entre notices, indiqués par les éléments **<Relationships>** puis **<LinkedItem>**, plusieurs solutions sont possibles. Comme le propose MARC, un simple lien entre les notices peut être créé. L'élément **<LinkedItem>** est alors vide et l'attribut *RecordID* indique le numéro de la notice concernée. On peut également envisager d'inclure la seconde notice dans la première, en utilisant l'élément **<BiblioRecord>**. Dans tous les cas, l'attribut *Relation* permet d'indiquer le type de relation entre les deux ouvrages décrits. BiblioML apporte ainsi davantage de souplesse et de lisibilité que l'utilisation éventuelle des champs 4XX proposés par UNIMARC.
- L'information concernant les exemplaires locaux peut être donnée directement, alors qu'une notice d'exemplaire devait être liée à la notice bibliographique dans UNIMARC. Pour cela il suffit d'utiliser l'élément **<Copies>** et ses éléments-fils.

Cette expérience de traitement avec BiblioML montre que si toute information de type bibliographique aurait pu être saisie en MARC, BiblioML apporte davantage de clarté. Les extensions proposées par BiblioML ne sont donc pas si significatives pour la description bibliographique du livre ancien en tant que telle. Le véritable enrichissement, c'est de rendre possible l'intégration de tout ou partie du document primaire. D'autre part, une étude menée à la Bibliothèque municipale de Lille<sup>119</sup>, utilise BiblioML de manière particulière pour le référencement aux fichiers d'autorités. A l'intérieur de la saisie en texte libre dans la balise **<Notes>**, sont imbriqués des éléments comme **<persname>** et **<geogname>**, qui intègrent la forme normalisée, permettant ainsi un balisage sémantique. Ces avantages – intégration du document primaire et prise en compte normalisée du contexte - sont également fournis par des DTDs non strictement bibliographiques comme l'EAD<sup>120</sup>, tout comme les extensions listées ci-dessus, envisageables également avec la TEI.

---

<sup>119</sup> J. Westeel, « Patrimoine et numérisation : la mise en contexte des documents », *in* : *Le numérique: impact sur le cycle de vie du document*, Colloque EBSI/ENSSIB, Montréal, 13/15 octobre 2004 : <http://www.ebsi.umontreal.ca/ebsi-enssib/pdf/westeel.pdf>.

<sup>120</sup> A. Jacquesson, *Bibliothèques et documents numériques*, Paris, 1999, p. 299.

D'autre part, BiblioML permet certes une description bibliographique fine et complète, mais les possibilités de renseigner les métadonnées de gestion - techniques et administratives - sont bien moins riches que celles que proposent le Dublin Core ou l'en-tête de la TEI. En effet, l'encodage dans BiblioML de ces métadonnées se limite à l'élément **<Meta>**, et elles correspondent ni plus ni moins à celles qui figurent dans une notice MARC. Cela doit être perçu comme une limitation, car dans un contexte de numérisation et d'édition sur le Web<sup>121</sup>, le contexte du document ainsi que les conditions d'accès doivent être précisées<sup>122</sup>.

## 2.2. Autres formats bibliographiques XML

Afin d'élargir le champ de l'étude menée dans le cadre du stage, ce paragraphe propose d'étendre la réflexion à d'autres formats bibliographiques XML. Comme l'analyse Kevin S. Clarke<sup>123</sup>, il existe trois types de formats bibliographiques XML, qu'il s'agisse de DTD ou de schéma : les formats fondés sur une conversion de MARC en XML, les formats construits à partir des normes de description, et les formats qui sont élaborés en faisant abstraction des normes et des formats, leurs concepteurs souhaitant partir sur de nouvelles bases. Nous proposons ici une classification légèrement différente, résultant de l'examen des plus représentatifs de ces formats : les formats dérivés de MARC, les formats intermédiaires, toujours héritiers de MARC mais tentant de s'en détacher pour offrir des résultats plus lisibles pour l'utilisateur, et les formats de conception entièrement nouvelle.

### 2.2.1. Formats dérivés de MARC

Il existe plusieurs formats conçus à partir de MARC, dont fait partie BiblioML. MARC XML est un autre exemple de ce type de DTD. Elle a été conçue par la Bibliothèque du Congrès, initialement écrite en SGML puis convertie en XML<sup>124</sup>. L'objectif qui a présidé à son élaboration était tout simplement de pouvoir convertir de manière automatique standardisée et non-propriétaire du MARC en SGML sans perte de données, et inversement. La conception de l'information bibliographique demeure traditionnelle, puisqu'il s'agit d'une conversion littérale.

---

<sup>121</sup> F. Role, « Métadonnées et structuration des bibliothèques numériques », *in* : *Bibliothèques numériques, cours INRIA 2002*, Paris, 2002, p. 162.

<sup>122</sup> Cf. ch. 3, § 3.2. "Echanges de données" p. 62-64.

<sup>123</sup> K. S. Clarke, Lane Medical Library, Stanford University, *MARC and XML: A Choice or Replacement ?* Access2002 Conference, Oct. 22, Windsor, Ontario : <http://elane.stanford.edu/laneauth/ksc-access2002/index.html>.

<sup>124</sup> MARC XML DTD : <http://www.loc.gov/marc/marcxml.html>.

Cinq principes régissent MARC XML :

- Généralité plutôt que spécificité.
- Réversibilité de la conversion.
- Flexibilité : les éléments sont optionnels, exceptés les *Leader elements*, et répétables.
- Convivialité : introduction d'une hiérarchie qui n'existait pas avec MARC.
- Lien possible avec la TEI.

Les points 1 et 3 rejoignent les objectifs du Dublin Core.

La Bibliothèque du Congrès propose à présent un schéma, appelé MARC 21 XML Schema<sup>125</sup>. Ces modèles n'ont pas été instaurés pour remplacer MARC, mais pour permettre une récupération optimale des données lorsque l'on envisage de les adapter au contexte du Web. Par conséquent, leur avantage en est limité. En effet, il ne s'agit pas tant, comme nous l'avons montré déjà avec BiblioML, d'améliorer la description bibliographique et sa présentation, qu'il s'agisse du livre ancien ou de toute autre ressource, mais d'adapter l'information secondaire au Web. Les données ne sont donc pas nécessairement plus accessibles - au sens de plus significatives - pour les utilisateurs.

### 2.2.2. Formats intermédiaires : vers la prise en compte de l'utilisateur

Les formats bibliographiques XML conçus à partir de MARC reflètent ainsi toujours plus ou moins l'ISBD, malgré les opportunités laissées par les feuilles de style. Cela n'est pas surprenant, et provient de la structure même de MARC. Cependant, la souplesse offerte par les schémas XML permet de s'engager dans de nouvelles voies qui prennent davantage en compte les besoins des utilisateurs. En effet, MODS<sup>126</sup> par exemple, un autre standard développé par la Bibliothèque du Congrès, s'inscrit dans cette optique en tentant de simplifier l'accès à l'information secondaire.

MODS signifie *Metadata Objects Description Schema*. Il s'agit d'un schéma XML qui propose d'encoder des métadonnées sélectionnées à partir de notices en MARC 21. Il demeure de ce fait fidèle à la description normalisée mise en place dans le monde des bibliothèques, tout en se rapprochant de formats de métadonnées comme le Dublin Core, notamment par sa simplification et sa relative facilité d'utilisation. Par conséquent, son ensemble d'éléments est plus riche que celui du Dublin Core, tout en étant plus simple que le format MARC :

---

<sup>125</sup> MARC 21 XML Schema : <<http://www.loc.gov/standards/marcxml/>>.

<sup>126</sup> MODS : <<http://www.loc.gov/standards/mods/>>.

regroupement d'éléments, abandon des codes ... Par cette simplification, ce schéma comporte deux inconvénients : d'une part l'impossibilité de convertir une notice MODS en MARC 21, et d'autre part une perte probable d'informations lors de la conversion d'une notice MARC 21 en MODS.

### 2.2.3. Formats de conception nouvelle

L'objectif de ce dernier type de format consiste à tirer profit au maximum des opportunités offertes par XML. L'ensemble de l'information bibliographique est utilisé, mais l'approche traditionnelle des normes en vigueur qui transparaît dans les formats MARC n'est pas conservée. Cela consiste donc à mettre à plat la structure des données bibliographiques pour en élaborer une nouvelle. Les concepteurs insistent sur le fait qu'il est préférable de définir une structure qui soit réutilisable dans différents contextes, plutôt que d'envisager des structures modifiables, afin de simplifier l'accès aux données mais aussi leur présentation.

Quelques constats sur des limites précises de MARC sont à l'origine de XOBIS<sup>127</sup>, *XML Organic Bibliographic Information Schema*, de la Lane Medical Library de l'Université de Stanford. Ses développeurs ne remettent pas en cause l'efficacité reconnue de la structure de MARC pour les échanges réalisés entre les bibliothèques ces dernières décennies. Mais, ils partent du constat lui aussi avéré que la structure de MARC est difficilement adaptable aux nouveaux modes d'échanges, notamment en ce qui concerne les données codées<sup>128</sup>. D'autre part, si MARC permet de renseigner de nombreuses informations, certaines sont rarement utilisées, comme par exemple les mélanges, et en alourdissent ainsi la structure. Enfin, les concepteurs de XOBIS insistent sur quelques incohérences : la présence de types de données similaires dans différents champs, le nombre élevé de codes à utiliser pour le même type de données, et les différentes structures pour un même type de données, comme les dates par exemple. Les champs de notes sont également remis en cause. Ces incohérences sont pour la plupart liées au "type" de données. Nous rejoignons ainsi l'un des avantages principaux des schémas XML qui permettent justement de définir le type des données et de contourner ces difficultés qui entravent l'accessibilité aux informations.

---

<sup>127</sup> XOBIS : <<http://laneweb.stanford.edu:2380/wiki/medlane/schema>>. Cf. également les pages de la BNF : <<http://www.bnf.fr/pages/zNavigat/frame/infopro.htm>>.

<sup>128</sup> Malgré le rôle de ces dernières dans les programmes de dédoublement.

XOBIS se fonde sur quatre principes :

- Structure de données modulaire la plus simple possible.
- Insistance sur les accès et les autorités (*Concept Records*) et les relations entre les différents éléments. XOBIS propose des données fortement conceptualisées.
- Référencement des éléments de façon à ce qu'ils puissent être réutilisés dans des contextes différents.
- Création d'un système de référencement pour la validation des données.

L'élément principal de XOBIS est **RecordList**, qui inclut l'élément **Record**. A ce niveau trois éléments sont proposés :

- **ControlData**, dont les données ne correspondent pas aux données de contrôle de MARC. Seuls l'identifiant de la notice, la source de la notice, et la dernière mise à jour sont indiqués.
- **Principal Element**, qui propose 10 types d'information de base. Une notice XOBIS ne contient qu'un seul élément principal.
- **Relationships**, c'est-à-dire les relations entre l'élément principal d'une notice et celui d'une autre.

Ainsi, les réflexions sur l'utilisation de ces modèles XML s'orientent graduellement, non pas vers une description bibliographique détaillée, mais au contraire vers l'échange de métadonnées minimales et normalisées, ainsi que vers la prise en compte du contexte, que ce soit celui relatif au document primaire ou celui lié à la production des métadonnées.

### 2.3. Description bibliographique : apports et limites de XML et des DTDs

#### 2.3.1. Constats généraux

A partir de cet examen s'impose un premier constat. A l'évidence, les apports principaux de ces formats, qu'ils soient ou non calqués sur MARC, sont en fait purement et simplement ceux de XML : lisibilité, accès au document primaire, recherche structurée sur tous les éléments du document, adaptabilité au Web, indépendance vis-à-vis des matériels et logiciels informatiques, pérennité des données, facilitation des échanges, souplesse dans la présentation des données grâce aux feuilles de style ...<sup>129</sup>.

---

<sup>129</sup> S. Jolibois, « La gestion informatisée de corpus bibliographiques », *BBF* 2000, t. 45, n° 1, p. 101.

Cependant, l'utilisation des formats XML dérivés de MARC risque d'entraîner une logique de raisonnement calquée sur MARC et sa rigidité de structure. Cela limite les bénéfices que l'on peut tirer de la richesse de XML et de sa souplesse. C'est d'ailleurs ce qui justifie tant l'orientation prise par les concepteurs de formats comme XOBIS, que la logique similaire qui a présidé à l'adaptation des DTDs en schémas XML.

### 2.3.2. Retour sur le livre ancien

Pour en revenir au livre ancien et à sa description, certains problèmes rencontrés avec MARC, comme la question du niveau de dépouillement, n'est résolu de manière satisfaisante que par des DTDs très spécifiques, comme l'EAD et son encodage à plusieurs niveaux. Avec BiblioML, si la hiérarchie entre notices est clarifiée, l'organisation des données demeure complexe à saisir. Il est vrai toutefois que cette difficulté semble davantage liée à une question de norme que de format ; son aspect problématique fait déjà l'objet d'un développement spécifique dans la norme AFNOR<sup>130</sup>.

Ensuite, les formats comme UNIMARC et leur traduction en XML ne permettent pas en soi de transcrire fidèlement et précisément la physionomie du livre ancien, que reflète notamment l'ornementation typographique (lettrines, bandeaux ...). Pourtant, ces éléments jouent un rôle dans l'identification des éditions. Doit-on en conclure de la nécessité de recourir à l'image, au-delà de l'encodage du texte du document ? Si l'image peut venir en complément, l'importance de la description complète en est relativisée. Les modèles XML permettent justement cet enrichissement, comme nous l'avons montré à plusieurs reprises.

Ainsi, créer une DTD – ou un schéma XML - spécifique à la structure et à la description bibliographique du livre ancien ne semble pas être la solution. Il y a en effet de grandes chances à ce que tout corpus comporte des spécificités. Si l'accent est mis avec insistance sur ces particularités, on risque de se diriger vers la création de DTDs "maison". Il semblerait plus judicieux, et de moindre coût, de prendre le temps d'étudier finement les spécificités du corpus à numériser, ainsi que le public visé – bibliographes, chercheurs de telles disciplines, grand public -, et de choisir en fonction de cela l'un des modèles XML pré-existants et reconnus comme standards. Quelles richesses doivent être mises en valeur en fonction de ces objectifs préalablement définis ? Quel niveau de description choisir alors en

---

<sup>130</sup> Cf. "catalogage à niveaux", *in* : Norme AFNOR Z 44-074, p. 38.

fonction de ces critères ? L'accent doit-il être mis sur les illustrations et la constitution de légendes assorties ? Sur les caractéristiques telles que l'ornementation typographique support de l'identification des éditions ? Sur l'intégration et la structuration des tables et des index ? Sur les notes de lecture manuscrites ? Par exemple, le corpus de botanique étudié présente en fait les spécificités qui sont celles des livres anciens, avec une accentuation particulière sur les nombreuses illustrations en planches hors-texte, les tables et les index. A ce niveau de structuration, le choix de l'application peut-être multiple, puisque nous avons vu que des DTDs comme la TEI, l'EAD, ou BiblioML permettent ce type d'intégration.

D'autre part, si choisir l'une de ces DTDs reconnues comme standards est une première étape pour limiter les facteurs d'hétérogénéité, cela peut être poursuivi au sein d'un établissement – ou entre établissements menant un projet commun – par l'élaboration de guides d'utilisation de la DTD choisie, adaptés au contexte. Citons par exemple celui réalisé par Florence Clavaud, sur l'EAD au Centre Historique des Archives Nationales<sup>131</sup>. De telles recommandations, rédigées à partir de l'expérience des institutions, permettent d'unifier les usages, de gagner du temps ensuite sur les choix à opérer, en renseignant les balises obligatoires ou facultatives permettant la structuration du document, en définissant le niveau de finesse de la description, la gestion des accès, les feuilles de style adaptées, en donnant des indications sur les outils d'exploitation ... Les implications de telles initiatives sont multiples : création d'un véritable corpus homogène de documents structurés, meilleure performance de la consultation en ligne dans le cadre d'un système d'information, meilleure visibilité pour le public.

Car, en fin de compte, XML ne réduit pas le manque d'uniformisation de la description bibliographique énoncé précédemment, à moins d'opter pour une description minimale. Et cela, notamment parce que les divers modèles en vigueur, associés à des outils d'exploitation et de diffusion des données, proposent les fonctionnalités nouvelles, souples, riches, et multiples que nous avons évoquées. Comment concilier alors une description bibliographique détaillée à laquelle aspirent les spécialistes des ressources patrimoniales, et les notices minimales qui permettent une première identification mais surtout l'échange de métadonnées et l'interopérabilité ?

---

<sup>131</sup> F. Clavaud, *Guide pour la mise en œuvre de la DTD EAD au CHAN*, version 4 - octobre 2004 : <<http://www.archivesnationales.culture.gouv.fr/chan/chan/infopro/fr/>>.

### 3. Description fine et description minimale : normalisation et échanges de données

Nous voudrions donc revenir à présent sur le dilemme qui existe entre description fine et description minimale. Ce problème transparait en effet en filigrane dans les normes de description bibliographique<sup>132</sup>. Il refait surface avec les fonctionnalités différentes proposées selon les DTDs retenues, et les limites des modèles XML que nous venons d'énoncer. Nous aborderons cette question en deux temps. D'une part en évoquant la liberté de description laissée par les normes, et d'autre part, en nous concentrant sur l'un des objectifs principaux qui régissent depuis l'origine de leur conception les modèles XML : l'interopérabilité et l'échange de données.

#### 3.1. Liberté de description et normalisation

Le constat de Jean-François Gilmont, dans *Le livre et ses secrets*, signalant que « la normalisation des descriptions catalographiques des livres anciens n'est pas satisfaisante pour le bibliographe »<sup>133</sup>, est toujours d'actualité. Cette difficulté qui semble inhérente à notre travail de bibliothécaire se rencontrait déjà avec les fiches papier, puis avec les formats MARC, et nous y sommes encore confrontés avec les formats XML, qu'ils soient de type bibliographique ou non.

En effet, comme nous l'avons vu plus haut<sup>134</sup>, les normes laissent une certaine liberté aux catalogueurs. En outre, plus la description est détaillée, plus les problèmes liés à cette uniformisation des pratiques se posent. Ce sont alors souvent les zones de notes qui recueillent une grande quantité d'information, saisie en texte libre. Les schémas XML bibliographiques en vigueur ne reviennent pas sur ce mode de saisie. Par exemple, XOBIS propose le typage de données suivant : attribut type="UnstructuredText". C'est pourquoi les études s'orientent vers la question du balisage sémantique à l'intérieur des saisies en texte libre, comme nous l'avons évoqué précédemment.

D'autre part, description fine et description minimale ne sont pas en soi incompatibles, mais complémentaires. Jean-François Gilmont rappelle à ce sujet que les descriptions détaillées sont précieuses, mais qu'une notice de type *Short Title Catalogue* préalable serait également utile<sup>135</sup>. S. Jolibois, sans s'en tenir

<sup>132</sup> Cf. « description allégée », *in* : Norme AFNOR Z 44-074, p. 15.

<sup>133</sup> J.-F. Gilmont, *Le Livre et ses secrets*, Louvain ; Genève, 2003, p. 117.

<sup>134</sup> Cf. Partie 1, ch. 2, § 1.2.3. "MARC et la description bibliographique du livre ancien" p. 18-19.

<sup>135</sup> J.-F. Gilmont, *id.*

cette fois au domaine du livre ancien, évoque la notion de *core record*, c'est-à-dire l'élaboration d'une notice minimale, portant mention des champs essentiels à l'identification du document<sup>136</sup>. On rejoint ici les réflexions menées à l'IFLA sur les FRBR.

Les *Functional Requirements for Bibliographical Records* résultent d'une modélisation conçue pour aller au-delà de la description ISBD. Cette étude porte en particulier sur deux aspects principaux : les relations qui existent entre les éléments de la description bibliographique, et l'essence des informations qu'une notice doit transmettre à l'utilisateur. Les réflexions ont abouti à la conception de trois groupes d'entités. Nous nous intéresserons dans cette étude davantage au premier, composé de quatre concepts :

- L'Oeuvre : la création intellectuelle abstraite,
- L'Expression : le texte original,
- La Manifestation : matérialisation de ce texte original,
- Le Document ou "Item" : un exemplaire de cette matérialisation.

Ces notions sont plus aisées à saisir lorsque l'on a été familiarisé avec les spécificités du livre ancien, en particulier avec la différenciation édition et exemplaire, ou encore avec le besoin - plus fréquent que lors du catalogage du livre moderne - de renseigner le titre uniforme, qui permet de mieux appréhender le concept d'Oeuvre des FRBR.

Mais en réalité, l'application des FRBR au livre ancien est loin d'être évidente, du fait de la distinction fondamentale entre données d'édition et données d'exemplaire. Le réservoir de notices *RedLightGreen* du *Research Library Group*<sup>137</sup>, non spécifique au livre ancien, propose une présentation issue du modèle FRBR<sup>138</sup>. Les informations renseignées sont assez succinctes, et sont organisées à partir de l'œuvre. L'auteur, le nombre d'éditions, ainsi que le sujet principal sont immédiatement renseignés dans une première notice minimale, et permettent des renvois à d'autres notices grâce aux liens hypertextes. Il n'est aucunement question de particularités d'exemplaires ; ce réservoir n'est pas un outil de localisation.

---

<sup>136</sup> S. Jolibois, *id.*

<sup>137</sup> RedLightGreen du *Research Libraries Group* : <<http://www.redlightgreen.org/>>

<sup>138</sup> Recherche effectuée à partir du nom d'auteur "Leonhard Fuchs".

## 3.2. Echanges de données

L'examen de nouveaux formats fondés sur les principes en vigueur de la description bibliographique, ainsi que celui sur les réflexions aboutissant à une nouvelle présentation normalisée de cette description<sup>139</sup>, montrent que recourir à une description fine n'est pas toujours envisageable. En effet, comme le révèle l'étude conduite au cours du stage, le catalogage livre en main, quel que soit le format utilisé, mobilise beaucoup de temps, et induit de ce fait un coût non négligeable.

En outre, nous avons vu que plus les descriptions sont détaillées, plus les sources d'hétérogénéité sont importantes et sont autant d'obstacles à l'échange de données. Plusieurs problèmes se posent en effet, ne serait-ce que le risque de créer des doublons, ou la perte, lors des échanges ou d'une mise en commun, des données spécifiques à l'institution comme les particularités d'exemplaires.

Si l'on peut choisir entre description fine et description minimale, la question des métadonnées de gestion ne peut quant à elle être évincée. Nous avons évoqué plus haut les limites de BiblioML concernant cet aspect puisque l'encodage de métadonnées dites de gestion y était très peu développé, au contraire de la TEI, de l'EAD, ou du Dublin Core. Par conséquent, ce sont ces dernières initiatives qui seraient à privilégier dans le cadre de catalogues collectifs ou de projets de portails. En effet, les indications fournies par ces métadonnées doivent permettre non seulement de faciliter la pérennité des données de l'ensemble du document structuré, mais aussi de les échanger, évoluant ainsi vers des concepts non strictement bibliographiques. Toutefois, la gestion de ces données est rendue plus ardue du fait de la multiplicité des projets, et seuls des protocoles visant à l'échange et l'interopérabilité peuvent y remédier. C'est l'un des buts de METS, *Metadata Encoding and Transmission Standard*<sup>140</sup>, format XML dont l'objectif, par la normalisation des métadonnées, est la transmission d'objets numériques entre systèmes différents.

Au niveau européen, le programme Minerva s'est concentré sur l'élaboration d'un Dublin Core Culture. Davantage qu'un nouveau format de métadonnées, il s'agit d'une adaptation du Dublin Core aux objets numériques patrimoniaux. En effet, le constat de départ est le suivant : le Dublin Core, grâce à ses 15 éléments, est le « plus grand dénominateur commun pour connecter des ressources

---

<sup>139</sup> Même si ces réflexions ne sont pas strictement fondées sur le livre ancien.

<sup>140</sup> METS : <<http://www.loc.gov/standards/mets/>>.

hétérogènes »<sup>141</sup>. Le Dublin Core Culture en propose une extension, par une qualification précisée pour certains des éléments, notamment ceux relatifs à la couverture temporelle et géographique. L'objectif est de faciliter la recherche de ressources patrimoniales hétérogènes, notamment grâce au protocole OAI-PMH : « Le Dublin Core Culture répond à la nécessité de créer des points d'accès standards aux ressources patrimoniales, quelle qu'en soit la nature et l'origine, afin de faciliter la mise en œuvre de portails patrimoniaux »<sup>142</sup>.

Le projet MICHAEL, *Multilingual Inventory of Cultural Heritage in Europe*, initiative du programme européen Minerva<sup>143</sup>, projet de plate-forme d'accès à des contenus culturels et scientifiques, s'inscrit dans cette optique. Il prend en compte l'apport de métadonnées telles que celles du Dublin Core Culture et les protocoles d'échanges, comme l'OAI-PMH. Cette plate-forme sera multilingue, et un moteur de recherche permettra de fédérer des recherches sur des serveurs dispersés, donnant ainsi accès aux documents numérisés dans des lieux divers.

Deux types d'initiatives semblent ainsi se développer en parallèle. D'une part, de nombreux projets patrimoniaux visent la finesse de la description afin de renseigner un maximum d'informations. La reconnaissance de modèles XML en tant que standards - c'est le cas de la TEI et de l'EAD -, ou encore la prise en compte normalisée du contexte – comme dans XOBIS ou le balisage sémantique, peuvent permettre alors de relativiser l'hétérogénéité de la structuration des documents. Un second axe de développement privilégie nettement les protocoles d'échanges, au-delà des principes strictement bibliothéconomiques, comme par exemple l'initiative du Dublin Core Culture, dont les métadonnées demeurent minimales malgré leur qualification.

---

<sup>141</sup> M. Foulonneau, *Un Dublin Core Culture pour accéder à des ressources patrimoniales hétérogènes*, avril 2003 : <[http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/dc\\_culture.pdf](http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/dc_culture.pdf)>.

<sup>142</sup> *Ibid.*

<sup>143</sup> J.-P. Dalbéra, D. Malbert, « La numérisation du patrimoine culturel en Europe : le projet européen MICHAEL ... », *in* : *Rencontres Numérisation et Patrimoine*, T. 1, SDN La Rochelle, 2004, p. 19.

## **Conclusion**

Plusieurs paradoxes constatés au cours de cette étude peuvent être relevés. Le premier est inhérent à un sujet qui entend comparer la structuration des documents XML, présentant une certaine abstraction - pas de représentation physique, ces documents ne sont rien sans les outils d'exploitation et de diffusion des données - et la structure du livre ancien, alors que ce dernier se définit essentiellement par sa matérialité et son contexte de production et d'utilisation. Ensuite, nous avons évoqué le dilemme entre une description fine, réalisée livre en main, facteur d'hétérogénéité, et qui ne permet de traiter qu'une masse infime de documents, et une description minimale, souvent issue de la conversion automatique des fichiers, couvrant un maximum de documents, mais imprécise et source d'erreurs. Enfin, l'élaboration des métadonnées, qui oscillent entre la continuité dans les pratiques bibliothéconomiques traditionnelles ou l'intégration au contexte du Web, comme celles du Dublin Core.

Les modèles XML, même ceux dits bibliographiques, ne sont pas en soi, du moins actuellement, la solution idéale pour le signalement du livre ancien. Cependant, leurs possibilités permettent de prendre en compte certaines de ses spécificités, selon les choix opérés en fonction des objectifs du projet de valorisation. Mais surtout, le grand avantage des DTDs pour le livre ancien réside dans l'opportunité d'intégration du document primaire et de prise en compte de son contexte. Le Web intègre des données de types hétérogènes, c'est irréductible, mais les réflexions et collaborations autour des métadonnées, mais aussi sur la modélisation des données, ouvrent de nouvelles perspectives de normalisation de signalement des ressources.

L'objectif de cette étude n'est pas d'être pérenne, les développements liés à XML sont trop rapides pour cela, comme entendait le montrer l'évocation des schémas. Conscients, comme l'annonce Jean-Claude Le Moal, que « Les outils changent mais les métiers restent [et qu']il ne faut pas confondre technique et métiers »<sup>144</sup>, nous espérons avoir rassemblé et approfondi certaines remarques dispersées sur ces modèles XML, et montré quels pouvaient être leurs intérêts pour le livre ancien. L'appropriation de ces données par les professionnels est

d'autant plus nécessaire que les bibliothèques numériques, si elles offrent l'opportunité de conserver les modes de consultation et de lecture des textes imprimés grâce à la prise en compte des péritextes, peuvent apporter de nouvelles pratiques documentaires par l'utilisation de l'hypertextualité et la prise en considération de la contextualité.

S'engager dans ces voies de réflexion permettrait aux bibliothèques de conserver un rôle de qualité, quand se profilent à l'horizon des projets de bibliothèques virtuelles de grande ampleur. Google vient en effet d'annoncer un plan massif de numérisation d'ouvrages libres de droits conservés dans les fonds de grandes bibliothèques américaines et britanniques<sup>145</sup>. Un contrat a été signé avec les Universités du Michigan et de Stanford ainsi qu'avec les Archives de Harvard, la Bodleian Library d'Oxford et la New York Public Library. Ces institutions considèrent cette numérisation<sup>146</sup> comme une opportunité pour conquérir un public plus large, notamment parmi les chercheurs. Google, en revanche, espère plus classiquement offrir ainsi les richesses patrimoniales qui demeurent encore aujourd'hui inaccessibles au grand public.

---

<sup>144</sup> J.-Cl. Le Moal, *La recherche d'information sur les réseaux, cours INRIA 2002*, Paris, 2002, avant-propos p. 10.

<sup>145</sup> Information du 14 décembre 2004. Cf. BBC news : <<http://news.bbc.co.uk/2/hi/technology/4094271.stm>> et Reuters : <<http://www.reuters.com/newsArticle.jhtml?storyID=7082710>>.

<sup>146</sup> Numérisation de plus de huit millions d'ouvrages en intégralité, prévue sur six ans.

# **Bibliographie**

## **STRUCTURE DU LIVRE ANCIEN**

**Chartier, Roger, Martin, Henri-Jean (dir.).** *Histoire de l'édition française : Le livre conquérant, du moyen âge au milieu du XVIIe siècle.* 2<sup>e</sup> édition. Paris : Fayard ; Ed. du Cercle de la Librairie, 1989. Tome 1.

**Chartier, Roger, Martin, Henri-Jean (dir.).** *Histoire de l'édition française : Le livre triomphant, 1660-1830.* Paris : Ed. du Cercle de la Librairie, 1984. Tome 2.

**Chartier, Roger, Martin, Henri-Jean (dir.).** *Histoire de l'édition française : Le temps des éditeurs, du romantisme à la belle époque.* 2<sup>e</sup> édition. Paris : Fayard ; Ed. du Cercle de la Librairie, 1990. Tome 3.

**Febvre, Lucien, Martin, Henri-Jean.** *L'apparition du livre.* 3<sup>e</sup> édition. Paris : Albin Michel, 1999.

**Fouché, Pascal, Martin, Henri-Jean, Pechoin, Daniel, et al.** *Dictionnaire encyclopédique du livre.* Paris : Ed. du Cercle de la Librairie, 2002. Tome 1.

**Gilmont, Jean-François.** *Le livre et ses secrets.* Genève : Droz ; Louvain : Presses universitaires, 2003.

**Harris, Neil.** *Analytical bibliography : an alternative prospectus.* Site de l'Institut d'Histoire du Livre [en ligne]. Disponible sur : <<http://ihl.enssib.fr/siteihl.php?page=55>> (consulté le 19/09/04).

**Martin, Henri-Jean.** *Histoire et pouvoirs de l'écrit.* 2<sup>e</sup> édition. Paris : Albin Michel, 1996.

**Martin, Henri-Jean (dir.).** *Les métamorphoses du livre : entretiens avec Christian Jacob et Jean-Marc Châtelain.* Paris : Albin Michel, 2004.

**Martin, Henri-Jean (dir.).** *La naissance du livre moderne : XVe – XVIIe siècles : mise en page et mise en texte du livre français.* Paris : Ed. du Cercle de la Librairie, 2000.

**Henri-Jean Martin.** *Une vision totale du livre.* Bulletin des Bibliothèques de France, 2004, t. 49, n°5, p. 21-23.

**Mercier, Alain (dir.).** *Les trois révolutions du livre : catalogue de l'exposition du Musée des arts et métiers, 8 octobre 2002 - 5 janvier 2003.* Paris : Musée des arts et métiers, 2002.

### DESCRIPTION BIBLIOGRAPHIQUE DES LIVRES ANCIENS

**Association française de normalisation.** *Catalogage des monographies anciennes : Rédaction de la description bibliographique : Octobre 1986 (Norme Z 44-074).* Paris : AFNOR, 1986.

**Dureau, Jeanne-Marie, Merland, Marie-Anne.** *Catalogage des livres anciens recueil d'exemples destinés à illustrer l'I.S.B.D. (A) [international standard bibliographic description, antiquarian].* Villeurbanne : ENSB, 1982.

**Fédération internationale des associations de bibliothécaires.** *Manuel Unimarc : format bibliographique.* München : K. G. Saur, 1999.

**Galantaris, Christian.** *Manuel de bibliophilie.* Paris : Editions des Cendres, 1997. Tome 2 : dictionnaire.

**IFLA.** *International Standard Bibliographic Description for Older Monographic Publications (Antiquarian) : Description bibliographique internationale normalisée des monographies anciennes.* 2<sup>e</sup> édition. Paris : Bibliothèque nationale, 1991.

**Minault-Richomme, Emmanuelle.** *Cataloguer les livres anciens, rares et précieux dans le contexte du déploiement du SUDoc et de la révision de la norme Z 44-074.* Diplôme de conservateur de bibliothèque, mémoire d'étude : Enssib. Villeurbanne, 2002.

**Russo, François.** *Éléments de bibliographie de l'histoire des sciences et des techniques.* Paris : Hermann, 1969.

**Védrine, Mireille.** *200 références pour le livre ancien : du manuscrit à 1900.* 2<sup>e</sup> édition revue et augmentée par Caroline Durand et Pierre Guinard. Villeurbanne : ENSB, 1990.

#### **FONDS ANCIEN DU SCD DE L'UNIVERSITE CLAUDE BERNARD LYON1**

**Richard, Lisa.** *Evaluation et mise en valeur d'un fonds ancien scientifique : l'exemple de la bibliothèque de l'université Lyon I Claude Bernard.* Diplôme de conservateur de bibliothèque, mémoire d'étude : Enssib. Villeurbanne, 1994.

**Fondation des banques CIC pour le livre.** *Patrimoine des bibliothèques de France : un guide des régions.* Paris : Payot, 1995. Vol. 5, Auvergne, Bourgogne, Rhône-Alpes. Bibliothèque santé p. 126-127, Bibliothèque des sciences p. 128-129.

**Agence de la coopération régionale pour la documentation Rhône-Alpes.** *Patrimoine écrit en Rhône-Alpes : itinéraires.* [Valence] : ACORD, 1992, p. 148-149.

**Bibliothèque interuniversitaire de Lyon Sciences.** *Livres illustrés de botanique.* Villeurbanne, 1978 [catalogue d'exposition dactylographié].

**Bibliothèque interuniversitaire de Lyon Sciences.** *Voyages d'exploration scientifique (XVIIIe – XIXe siècles).* Villeurbanne, 1980 [catalogue d'exposition dactylographié].

#### **HISTOIRE DES SCIENCES**

**Allorge-Boiteau, Lucile, Ikor, Olivier.** *La fabuleuse odyssée des plantes : les botanistes voyageurs, les jardins des plantes, les herbiers.* Paris : JC Lattès, 2003.

**Bedel, Charles.** *La curiosité scientifique au XVIIIe siècle : cabinets et observatoires.* Paris : Hermann, 1986.

**Blay, Michel, Halleux, Robert, Barthélémy, Georges.** *La science classique, XVI<sup>e</sup>-XVIII<sup>e</sup> siècles : dictionnaire critique.* Paris : Flammarion, 1998.

**Lecourt, Dominique, Bourgeois, Thomas.** *Dictionnaire d'histoire et de philosophie des sciences.* Paris : Presses Universitaires de France, 2003.

### CATALOGUES EN MUTATION

**Chouanière, Dominique, et al.,** *La gestion informatisée de corpus bibliographiques.* Bulletin des Bibliothèques de France, 2000, t. 45, n° 1, p. 98-108 [en ligne]. Disponible sur : <[http://bbf.enssib.fr/bbf/html/2000\\_45\\_1/2000-1-p98-jolibois.xml.asp](http://bbf.enssib.fr/bbf/html/2000_45_1/2000-1-p98-jolibois.xml.asp)> (consulté le 19/09/04).

**Desrichard, Yves.** *Vers la convergence des formats bibliographiques ? Onyx, application XML du monde de l'édition.* Bulletin des Bibliothèques de France, 2004, t. 49, n° 3, p. 55-63.

**Duchemin, Pierre-Yves.** *There is a life beyond MARC* [en ligne]. Disponible sur : <[http://www.abf.asso.fr/IMG/ppt/2003-05-20\\_ABFPARIS.ppt](http://www.abf.asso.fr/IMG/ppt/2003-05-20_ABFPARIS.ppt)> (consulté le 14/09/04).

**Groupe XML « Enrichi » de l'Association des Bibliothécaires Français.** *Concertation sur l'information bibliographique enrichie* [Liens : dernière mise à jour 15 mai 2001] [en ligne]. Disponible sur : <<http://www.abf.asso.fr/enrichi/xmldoc.htm>> (consulté le 19/09/04).

**IFLA.** *Functional Requirements for Bibliographic Records : final report* [en ligne]. Disponible sur : <<http://www.ifla.org/VII/s13/frbr/frbr.htm>> (consulté le 03/11/04).

**Jacquesson, Alain, Rivier, Alexis.** *Bibliothèques et documents numériques : concepts, composantes, techniques et enjeux.* Paris : Ed. du Cercle de la Librairie, 1999.

**Lahary, Dominique.** Des bibliothèques traditionnelles aux « bibliothèques virtuelles ». *In : La recherche d'information sur les réseaux : cours INRIA, 30 septembre-4 octobre 2002, Le Bono (Morbihan).* Le Moal, Jean-Claude, Hidoine, Bernard, Calderan, Lisette (dir.). Paris : ADBS éd., 2002, p. 169-201.

**Lahary, Dominique.** *Informatique et bibliothèques*. Bulletin des Bibliothèques de France, 2002, t. 47, n° 1, p. 60-67 [en ligne]. Disponible sur : <[http://bbf.enssib.fr/bbf/html/2002\\_47\\_1/2002-1-p60-lahary.xml.asp](http://bbf.enssib.fr/bbf/html/2002_47_1/2002-1-p60-lahary.xml.asp)> (consulté le 19/09/04).

**Lupovici, Catherine.** *L'information bibliographique des documents électroniques*. Bulletin des Bibliothèques de France, 1998, t. 43, n° 4, p. 42-47 [en ligne]. Disponible sur : <[http://bbf.enssib.fr/bbf/html/1998\\_43\\_4/1998-4-p42-lupovici.xml.asp](http://bbf.enssib.fr/bbf/html/1998_43_4/1998-4-p42-lupovici.xml.asp)> (consulté le 19/09/04).

**Lupovici, Christian.** *La chaîne de traitement des documents numériques*. Bulletin des Bibliothèques de France, 2002, t. 47, n° 1, p. 86-91 [en ligne]. Disponible sur : <[http://bbf.enssib.fr/bbf/html/2002\\_47\\_1/2002-1-p86-lupovici.xml.asp](http://bbf.enssib.fr/bbf/html/2002_47_1/2002-1-p86-lupovici.xml.asp)> (consulté le 19/09/04).

**Samain, Thierry.** *Du catalogue enrichi au métacatalogue*. Bulletin des Bibliothèques de France, 2001, t. 46, n° 2, p. 82-86 [en ligne]. Disponible sur : <[http://bbf.enssib.fr/bbf/html/2001\\_46\\_2/2001-2-p82-samain.xml.asp](http://bbf.enssib.fr/bbf/html/2001_46_2/2001-2-p82-samain.xml.asp)> (consulté le 19/09/04).

## **XML : GÉNÉRALITÉS**

**W3C.** *Extensible Markup Language (XML)* [en ligne]. Disponible sur : <<http://www.w3.org/XML/>> (consulté le 16/09/04).

**W3C.** *Extensible Markup Language (XML) 1.0 (third edition), Recommendation 04 February 2004* [en ligne]. Disponible sur : <<http://www.w3.org/TR/2004/REC-xml-20040204/>> (consulté le 16/09/04).

**W3C.** *Langage de balisage extensible, Recommendation du W3C, 10 février 1998* [Traduction française de XML 1.0] [en ligne]. Disponible sur : <http://pages.videotron.com/fyergeau/w3c/xml10/REC-xml-19980210.fr.html> (consulté le 16/09/04).

**Chahumeau, François.** XML : un langage universel pour la représentation textuelles des données structurées. *In : Bibliothèques numériques : cours INRIA, 9-13 octobre 2000, La Bresse.* Le Moal, Jean-Claude, Hidoine, Bernard (dir.). Paris : ADBS éd., 2000, p. 119-141.

**Hunter, David.** *Initiation à XML : avec trois études de cas détaillées.* Paris : Eyrolles, 2001.

**Michard, Alain.** *XML : langage et applications.* Paris : Eyrolles, 2001.

**Morgan, Eric Lease.** *XML pour les bibliothécaires : un manuel et un atelier* [en ligne]. Disponible sur : <<http://morinn.free.fr/xml/getting-started.htm#id246749>> (consulté le 19/09/04).

**Role, François.** XML et la documentation structurée. *In : La recherche d'information sur les réseaux : cours INRIA, 30 septembre-4 octobre 2002, Le Bono (Morbihan).* Le Moal, Jean-Claude, Hidoine, Bernard, Calderan, Lisette (dir.). Paris : ADBS éd., 2002, p. 71-97.

**Vincent, Cyril.** *XML et XSL : les feuilles de style XML.* Nantes : ENI, 2003.

## FEUILLES DE STYLE

**W3C.** *Extensible Stylesheet Language family (XSL)* [en ligne]. Disponible sur : <<http://www.w3.org/Style/XSL/>> (consulté le 28/10/04).

**W3C.** *Transformation XSL (XSLT), Version 1.0, Recommandation W3C 16 Novembre 1999* [en ligne] <<http://xmlfr.org/w3c/TR/xslt/>> (consulté le 28/10/04).

## DTDS : GENERALITÉS

**Fréon, Marie-Elise.** Les métadonnées : accès aux ressources électroniques. *In : La recherche d'information sur les réseaux : cours INRIA, 30 septembre-4 octobre 2002, Le Bono (Morbihan).* Le Moal, Jean-Claude, Hidoine, Bernard, Calderan, Lisette (dir.). Paris : ADBS éd., 2002, p. 99-135.

**Mkadmi, Abderrazak, Bouhaï, Nasreddine, Langlois, Marc.** *A propos : partager les modèles XML quel intérêt ?* Bulletin des Bibliothèques de France, 2003, t. 48, n° 5, p. 68-73 [en ligne]. Disponible sur : <[http://bbf.enssib.fr/bbf/html/2003\\_48\\_5/2003-5-p68-mkadmi.xml.asp](http://bbf.enssib.fr/bbf/html/2003_48_5/2003-5-p68-mkadmi.xml.asp)> (consulté le 19/09/04).

**Role, François.** Métadonnées et structuration des bibliothèques numériques. *In : Bibliothèques numériques : cours INRIA, 9-13 octobre 2000, La Bresse.* Le Moal, Jean-Claude, Hidoine, Bernard (dir.). Paris : ADBS éd., 2000, p. 143-170.

### XML SCHEMA : GENERALITES

**W3C.** *XML Schéma* [en ligne]. Disponible sur : <<http://www.w3.org/XML/Schema>> (consulté le 28/10/04).

**W3C.** *XML Schema tome 0 : Introduction, Recommandation du W3C du 2 Mai 2001* [Traduction française] [en ligne]. Disponible sur : <<http://xmlfr.org/w3c/TR/xmlschema-0>> (consulté le 28/10/04).

**W3C.** *XML Schema tome 1 : Structures* [Traduction française] [en ligne]. Disponible sur : <<http://xmlfr.org/w3c/TR/xmlschema-1/>> (consulté le 28/10/04).

**Thomasson, Jean-Jacques.** *Schémas XML.* Paris : Eyrolles, 2002.

### DTDS ET LIVRE ANCIEN

**Role, François.** *Métadonnées et documents anciens.* Document numérique : Documents anciens, vol. 3, n° 1-2, juin 1999, p. 135-150.

**Romand-Monnier, Emilie.** *XML/SGML et la numérisation des textes anciens.* DESSID, rapport de recherche bibliographique : Enssib. Villeurbanne, 2000.

**Westeel, Isabelle.** Patrimoine et numérisation : la mise en contexte du document. *In : Le numérique : impact sur le cycle de vie du document pour une analyse interdisciplinaire.* Colloque EBSI/Enssib : Montréal, 13-15 octobre 2004. [en ligne]. Disponible sur : <<http://www.ebsi.umontreal.ca/ebsi-enssib/pdf/westeel.pdf>> (consulté le 30/11/04).

## MODELES XML

### **EAD, Encoded archival description :**

**Library of Congress.** *Encoded Archival Description (EAD), official EAD Version 2002 website* [en ligne]. Disponible sur : <<http://www.loc.gov/ead/ead.html>> (consulté le 19/09/04).

**Chartron, Ghislaine.** *La DTD EAD dans les archives et les bibliothèques.* Bulletin des Bibliothèques de France, 2003, t. 48, n° 2, p. 112-114 [en ligne]. Disponible sur : <[http://bbf.enssib.fr/bbf/html/2003\\_48\\_2/2003-2-p112-chartron.xml.asp](http://bbf.enssib.fr/bbf/html/2003_48_2/2003-2-p112-chartron.xml.asp)> (consulté le 19/09/04).

**Queyroux, Fabienne.** EAD, la description archivistique encodée. *In : La numérisation des textes et des images.* Westeel, Isabelle, Aubry, Martine (dir.). Lille : Ed. du Conseil scientifique de l'Université Charles-de-Gaulle Lille 3, 2003, p. 75-89.

### **TEI, Text encoding initiative :**

**TEI Website.** *Text Encoding Initiative* [en ligne]. Disponible sur : <<http://www.tei-c.org>> (consulté le 19/09/04).

### **BiblioML :**

**Ministère de la Culture et de la Communication.** *BiblioML* [en ligne]. Disponible sur : <<http://www.biblioml.org/>> (consulté le 19/09/04).

**Association pour la documentation numérique en XML.** *BiblioML* [en ligne]. Disponible sur : <<http://vernier.gamsau.archi.fr:8080/JSPWiki/Wiki.jsp?page=BiblioML>> (consulté le 19/09/04).

**Bottin, Michel.** BiblioML : une application XML pour les données bibliographiques et d'autorités, compatible avec le format (uni)marc. *In : La numérisation des textes et des images.* Westeel, Isabelle, Aubry, Martine (dir.). Lille : Ed. du Conseil scientifique de l'Université Charles-de-Gaulle Lille 3, 2003, p. 69-74.

**Autres formats bibliographiques XML :**

**Library of Congress.** *MARC XML DTD* [en ligne]. Disponible sur : <<http://www.loc.gov/marc/marcxml.html#marcdtd>> (consulté le 03/11/04).

**Library of Congress.** *MARCXML : MARC 21 XML Schema* [en ligne]. Disponible sur : <<http://www.loc.gov/standards/marcxml/>> (consulté le 03/11/04).

**Library of Congress.** *MODS, Metadata Objects Description Schema* [en ligne]. Disponible sur : <<http://www.loc.gov/standards/mods/>> (consulté le 03/11/04).

**Lane Medical Library, Stanford University Medical Center.** *XOBIS, The XML Organic Bibliographic Information Schema* [en ligne]. Disponible sur : <<http://laneweb.stanford.edu:2380/wiki/medlane/schema>> (consulté le 03/11/04).

**Clarke, Kevin. S.** *MARC and XML: A Choice or Replacement ?* [en ligne]. Disponible sur : <<http://elane.stanford.edu/laneauth/ksc-access2002/index.html>> (consulté le 03/11/04).

**Dublin Core :**

**OCLC.** *The Dublin Core metadata initiative* [en ligne]. Disponible sur : <<http://purl.oclc.org/dc>> (consulté le 19/09/04).

**Foulonneau, Muriel.** *Un Dublin Core Culture pour accéder à des ressources patrimoniales hétérogènes* [en ligne]. Disponible sur : <[http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/dc\\_culture.pdf](http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/dc_culture.pdf)> (consulté le 19/09/04).

**Protocole OAI**

**OAI.** *The Open Archives Initiative Protocol for Metadata Harvesting* [en ligne]. Disponible sur : <<http://www.openarchives.org/OAI/openarchivesprotocol.html>> (consulté le 19/09/04).

**Aubry, Sara.** Archives ouvertes : nouveaux modèles de publication et de diffusion et nouveaux outils. *In : Publier sur Internet : séminaire INRIA, 27 septembre – 1<sup>er</sup> octobre 2004, Aix-les-Bains.* Le Moal, Jean-Claude, Hidoine, Bernard, Calderan, Lisette (dir.). Paris : ADBS éd., 2004, p. 209-236.

### **METS, Metadata Encoding Transmission Standard :**

**Library of Congress.** *METS, Metadata Encoding Transmission Standard* [en ligne]. Disponible sur : <<http://www.loc.gov/standards/mets/>> (consulté le 03/11/04).

## **VALORISATION DE COLLECTIONS : NUMERISATION**

**Arot, Dominique.** La bibliothèque numérique patrimoniale. *In : Bibliothèques numériques : cours INRIA, 9-13 octobre 2000, La Bresse.* Le Moal, Jean-Claude, Hidoine, Bernard (dir.). Paris : ADBS éd., 2000, p. 11-51.

**Calenge, Bertrand.** *Contenus des bibliothèques numériques et bibliothéconomie.* Bulletin des Bibliothèques de France, 2002, t. 45, n° 4, p. 117-119 [en ligne]. Disponible sur : <[http://bbf.enssib.fr/bbf/html/2000\\_45\\_4/2000-4-p117-calenge.xml.asp](http://bbf.enssib.fr/bbf/html/2000_45_4/2000-4-p117-calenge.xml.asp)> (consulté le 05/10/04).

**Chabin, Marie-Anne.** *Patrimoine numérique : révolution et racines.* Document numérique : Les bibliothèques numériques, 1998, vol. 2, n° 3-4, p. 243-258.

**Cubaud, Pierre, Deblock, Geneviève.** Le Conservatoire numérique des Arts et Métiers : historique du projet et organisation du site. *In : La numérisation des textes et des images : techniques et réalisations.* Westeel, Isabelle, Aubry, Martine (dir.). Lille : Ed. du Conseil scientifique de l'Université Charles-de-Gaulle Lille 3, 2003, p. 13-22.

**Dalbéra, Jean-Pierre, Malbert, Daniel.** La numérisation du patrimoine culturel en Europe : le projet européen MICHAEL d'un inventaire multilingue du patrimoine culturel numérisé en Europe. *In : Rencontres Numérisation et Patrimoine.* Semaine du document numérique : La Rochelle, 22/24 juin 2004. Tome 1, p. 17-19.

**Fondin, Hubert.** *Le traitement numérique des documents*. Paris : Ed. Hermès, 1998.

**Legendre, Françoise, et al.** Le projet Bovary. *In : Rencontres Numérisation et Patrimoine*. Semaine du document numérique : La Rochelle, 22/24 juin 2004. Tome 1, p. 49-58.

**MINERVA (Ministerial Network for Valorising Digitisation Activities / Réseau Ministériel pour la Valorisation des Activités de Numérisation).** *Recommandations techniques pour les programmes de création de contenus culturels numériques*, version 1.0 révisée le 07/05/2004 [en ligne]. Disponible sur : <[http://www.culture.gouv.fr/culture/mrt/numerisation/fr/eeurope/documents/guide\\_technique.pdf](http://www.culture.gouv.fr/culture/mrt/numerisation/fr/eeurope/documents/guide_technique.pdf)> (consulté le 19/09/04).

**Ministère de l'Éducation nationale, Sous-direction des bibliothèques.** *Panorama des projets de numérisation mis en œuvre par les SCD, SICD, et grands établissements en 2000* [en ligne]. Disponible sur : <<http://www.sup.adc.education.fr/bib/acti/Num/Eng/num1.htm>> (consulté le 19/09/04).

**Ministère de la Culture et de la Communication.** *Catalogue des fonds culturels numérisés* [en ligne]. Disponible sur : <[http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f\\_02.htm](http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f_02.htm)> (consulté le 19/09/04).

**Sévigny, Martin.** La numérisation dans un système d'information. Harmoniser la variabilité ? *In : La numérisation des textes et des images : techniques et réalisations*. Westeel, Isabelle, Aubry, Martine (dir.). Lille : Ed. du Conseil scientifique de l'Université Charles-de-Gaulle Lille 3, 2003, p. 173-178.

**Sévigny, Martin.** Nordnum : rapport technique. *In : La numérisation des textes et des images : techniques et réalisations*. Westeel, Isabelle, Aubry, Martine (dir.). Lille : Ed. du Conseil scientifique de l'Université Charles-de-Gaulle Lille 3, 2003, p. 27-46.

**Westeel, Isabelle.** Nordnum : bibliothèque numérique d'histoire régionale (Nord-Pas-de-Calais). *In : La numérisation des textes et des images : techniques et réalisations*. Westeel, Isabelle, Aubry, Martine (dir.). Lille : Ed. du Conseil scientifique de l'Université Charles-de-Gaulle Lille 3, 2003, p. 23-26.

## ***Table des annexes (Tome séparé)***

<b>SOURCES DES DOCUMENTS PRESENTES .....</b>	<b>I</b>
<b>ANNEXE 1 : ETUDE DU CORPUS DE BOTANIQUE .....</b>	<b>II</b>
Annexe 1-1 : Liste des ouvrages sélectionnés.....	III
Annexe 1-2 : Références des ouvrages traités en XML .....	IX
Annexe 1-3 : Liste des ressources utilisées pour le catalogage .....	XI
<b>ANNEXE 2 : BIBLIOML .....</b>	<b>XV</b>
Annexe 2-1 : Extraits du fichier texte de la DTD .....	XV
Annexe 2-2 : Tableaux de comparaison UNIMARC-BiblioML .....	XIX
<b>ANNEXE 3 : EDITEUR XML XMETAL .....</b>	<b>XXVII</b>
Annexe 3-1 : Ouverture d'un nouveau document.....	XXVIII
Annexe 3-2 : Présentation d'un nouveau document.....	XXIX
Annexe 3-3 : Mode d'affichage « NormalView » .....	XXX
Annexe 3-4 : Mode d'affichage « TagsOnView » .....	XXXI
Annexe 3-5 : Erreur de validation .....	XXXII
<b>ANNEXE 4 : NOTICES XML CREEES .....</b>	<b>XXXIII</b>