

école nationale supérieure
des sciences de l'information
et des bibliothèques



Université
Claude Bernard
LYON 1

DESS en Ingénierie Documentaire

Rapport de stage

Etude d'une chaîne de publication web de texte intégral au format XML

Emilie ROUSSEAU

Effectué sous la direction de Jacqueline LAVANDIER
Service informatique de l'enssib – Villeurbanne

Septembre 2001

Remerciements

A Jacqueline Lavandier pour m'avoir accueillie dans son service et m'avoir fait confiance pour la conduite de cette étude,

A son équipe pour sa sympathie et sa disponibilité,

A Emilie Romand-Monnier qui m'a épaulée durant ce stage et qui m'a patiemment expliqué les rouages du XML ...

Etude d'une chaîne de publication web de texte intégral au format XML

Le projet de publication du BBF électronique sur le Web au format XML nécessite la mise en place d'une logistique informatique et humaine. Le comparer à d'autres projets similaires basés sur XML, aidera le service informatique à évaluer le travail réalisé et à se positionner dans l'édition électronique.

Descripteurs chaîne de publication, revue électronique, XML, DTD.

Study of a web publication process in full text with extensible Markup Language

Electronic BBF XML web publication project requires to set up data processing and human logistics. A comparison of this project with others XML-based ones, will help the computing department in estimating the work done, and in positioning itself in relation to the electronic publishing field.

Keywords publication process, electronic journal, XML, DTD

Table des matières

Introduction	7
Le projet d'une revue électronique : le <i>Bulletin des Bibliothèques de France</i>	9
1. De l'ancienne version du <i>Bulletin des Bibliothèques de France</i> électronique.....	9
1.1. Caractéristiques communes aux formes électronique et papier	9
1.1.1 La politique éditoriale	9
1.1.1.1 <i>Equipe de rédaction</i>	10
1.1.1.2 <i>Comité de rédaction</i>	10
1.1.1.3 <i>Conseil scientifique</i>	10
1.1.2 Les auteurs et les contenus	11
1.1.2.1 <i>Les auteurs</i>	11
1.1.2.2 <i>Les contenus</i>	12
1.1.3 Le protocole de rédaction	12
1.2. Caractéristiques du <i>BBF</i> électronique	13
1.2.1 L'équipe chargée de la chaîne de production	13
1.2.1.1 <i>L'équipe du BBF à Paris</i>	13
1.2.1.2 <i>L'équipe à l'enssib</i>	14
1.2.2 Les rubriques et les formats proposés	14
2.à la nouvelle version en XML.	15
2.1. Le standard XML	15
2.1.1 Caractéristiques du XML	15
2.1.2 Avantages du XML	16
2.2. Le projet et ses évolutions	17
2.2.1 Une revue électronique enrichie	17
2.2.1.1 <i>Contenus : enrichissement et nouvelles offres</i>	17
2.2.1.2 <i>Ergonomie des contenus</i>	18
2.2.1.3 <i>Financement</i>	18
2.2.2 Réalisations depuis 2000	19
2.2.3 Etat du projet en juin 2001	20
3. Définition du stage	21
3.1. Mission du stagiaire	21
3.2. Cadre du stage	21
3.2.1 Activités et membres du Département informatique et de l'Ingénierie documentaire	22
3.2.2 Topologie du réseau informatique de l'enssib	23
3.3. Déroulement du stage et méthodes de travail	24
3.3.1 Rencontres avec Araxe et comptes-rendus	24
3.3.2 Recherches sur internet	25
3.3.3 Etude de la procédure de balisage	25
La chaîne de production du <i>BBF</i> électronique	26
1. La récupération des textes	26
1.1. Récupération des fichiers PDF et HTML	26
1.1.1 Traitement	26

1.1.2	Solutions envisagées	28
1.2.	Amélioration des procédures	30
1.2.1	Format RTF	30
1.2.2	Utilisation de macros	30
2.	Les différents formatages	32
2.1.	Formatage en XML à l'aide des DTDs-BBF	32
2.1.1	Présentation des DTDs	32
2.1.2	Création de l'article en XML et indexation	33
2.2.	Formatage en PDF	34
2.3.	Formatage en HTML	35
2.3.1	Les pages statiques	35
2.3.2	Les pages dynamiques	35
	2.3.2.1 Affichage en HTML sur le web	35
	2.3.2.2 XSL et XSL-FO	36
	2.3.2.3 CSS	37
3.	Droits d'auteur et traitement des images	38
3.1.	Droits d'auteur sur les images	38
3.2.	Traitement des graphiques et des tableaux	38
4.	Système de gestion des données pour le BFF	40
4.1.	Gestion des fichiers sur le serveur des Editions	40
4.2.	La base de données BBF	41
4.3.	Arborescence du site	42
5.	Les nouveautés du BBF électronique	43
5.1.	Accès et abonnement	44
5.1.1	Conditions d'accès et abonnement	44
5.1.2	Traitement informatique	44
5.2.	Modalités de recherches et affichage des documents	45
5.2.1	Généralités	45
5.2.2	Les différents types de recherche	46
5.3.	Les nouvelles rubriques et les nouveaux services	47
5.3.1	Les profils	47
5.3.2	la gazette électronique	47
5.4.	Nouveau graphisme	48
	Autres projets de textes électroniques au format SGML/XML	52
1.	XML et l'échange d'informations bibliographiques	53
2.	XML et l'archivage de documents	55
3.	XML et la structuration de thèses, de travaux d'étudiants et de chercheurs	56
4.	XML et la structuration de revues et journaux électroniques	58
5.	XML et la structuration de textes littéraires	62
6.	XML et les plate-formes documentaires	63
7.	Commentaires	64
	Bilan sur le BBF électronique et son traitement informatique	68

1. Observations au terme de ces quatre mois	68
1.1. Un choix éditorial "prédéterminé" dès le départ	68
1.2. Un choix technique judicieux	68
1.2.1 Réalisations du SI	68
1.2.2 Autres outils disponibles	69
1.3. Une expertise sur un système d'information liant la technologie XML	72
2. Pour un balisage plus humain	72
2.1. Aide au balisage	72
2.2. Révision des DTDs	73
Conclusion	74
Abréviations	75
Références bibliographiques	77
BIBLIOGRAPHIE	77
WEBLIOGRAPHIE	77
Sites web de référence	79
Portails XML	79
DTDs normalisées	79
ANNEXES	80

Introduction

Né en 1956 de la fusion entre le *Bulletin d'information de la Direction des bibliothèques de France* et le *Bulletin de documentation bibliographique*, le *Bulletin des Bibliothèques de France*, *BBF*, son appellation courante, est une revue francophone d'information et de réflexion qui "s'intéresse à tous les sujets concernant le fondement, l'organisation et la gestion des bibliothèques"¹.

Il joue "un rôle important dans la formation continue des professionnels, (...) traite de toutes les questions concernant le développement des bibliothèques et de la documentation : place et rôle des bibliothèques dans les politiques culturelles, éducatives et de recherche, évolution des métiers et des formations, contenu, constitution et disposition des collections, mutations suscitées par les nouvelles technologies et les réseaux, modalités du traitement documentaire, etc."

Il s'agit avant tout d'une revue papier, publiée depuis 1980 par l'enssib². Elle connaît une large diffusion au sein des professionnels des bibliothèques et de la documentation dans le monde entier.

Quant à sa version électronique, créée en 1996 et hébergée par le site de l'enssib, elle est le fruit d'une étroite collaboration entre les successifs rédacteurs en chef de la revue, et les services des Editions et Informatique de l'enssib. Le *BBF* a en effet bénéficié de la renommée du site web de l'école qui a ainsi saisi l'opportunité de s'offrir une expérience en terme d'édition électronique d'une revue scientifique.

L'école participe par ailleurs à d'autres projets de documentation électronique puisqu'elle s'intéresse à la mise en ligne de travaux d'étudiants et de thèses soutenues dans son établissement ou ailleurs, en sciences de l'information.

¹ CALENGE B., Politique éditoriale, *BBF* [On line]. sept.2000 [visité le 06/06/01]. Available from internet : <url: www.enssib.fr/bbf/page_accueil_BBF.htm>

² Ecole nationale supérieure des sciences de l'information et des bibliothèques, Villeurbanne (69) : www.enssib.fr

Les responsables du *BBF* ont rapidement compris les avantages qu'offraient les nouvelles technologies en édition électronique. Ils ont donc donné à leurs utilisateurs une première version électronique du *BBF*, copie conforme de la version papier.

Depuis 2000, un groupe de travail a mené un long travail de réflexion afin de réviser cette version. Sans vouloir se détacher de l'imprimé, il souhaitait que le *BBF* électronique bénéficie des avantages du langage XML³. C'est donc un produit nouveau qui sera en ligne sur le site de l'enssib dès le mois de septembre 2001, parallèlement à l'ancienne version. Et c'est la raison pour laquelle Jacqueline Lavandier, responsable du département informatique, m'a demandé de m'intéresser durant mon stage à la chaîne de publication du *BBF* électronique.

L'introduction d'un nouveau format a effectivement modifié la chaîne de production de la revue électronique et il m'a fallu la décrire : traitement des fichiers, balisage, mise en ligne...

L'intérêt du stage était également de faire de la veille sur les outils qu'offre sur le marché, ainsi que sur d'autres projets de documents formatés en XML, voire SGML⁴. Ceci afin de mieux positionner le service informatique dans la réalisation du projet de revue au format XML.

³ eXtensible Markup Language, langage à balises étendu.

⁴ Standardized General Markup Language, norme ISO 8879 depuis 1986. Très utilisé dans le monde documentaire, il permet de différencier par des jeux de balises la structure logique de la sémantique d'un document pour le formater.

Le projet d'une revue électronique : le *Bulletin des Bibliothèques de France*

Le projet de *BBF* électronique s'est développé en deux étapes majeures, à l'initiative des rédacteurs en chef successifs de la revue.

Même si certaines informations valent aussi bien pour la revue électronique que pour la revue papier, il ne sera question ici que du *Bulletin des Bibliothèques de France* électronique⁵.

1. De l'ancienne version du *Bulletin des Bibliothèques de France* électronique....

La réflexion sur le *BBF* électronique existe depuis plusieurs années. En 1996, la rédactrice en chef orienta la revue dans une logique d'édition électronique. Une coopération avec le service des Editions de l'enssib a donc permis au *BBF* en ligne de se développer et c'est actuellement, cette version qui est sur le site de l'enssib.

Les articles en ligne sont les mêmes que ceux présents dans la revue papier.

1.1. Caractéristiques communes aux formes électronique et papier

1.1.1 La politique éditoriale

La politique éditoriale du *Bulletin des bibliothèques de France* est établie par l'équipe de rédaction, avec la participation d'un comité de rédaction

⁵ La version papier est traitée dans ROMAND-MONNIER E., *Migration d'une revue professionnelle vers un modèle structuré en ligne* [On line]. Villeurbanne : ENSSIB-LYON1, 2000, p. 12-17. Available from internet :
<url : www.enssib.fr/bibliotheque/documents/dessid/rsromandmonnier.pdf>

et d'un conseil scientifique. Leurs membres sont *"des professionnels du livre et de la documentation, de l'édition, des chercheurs [qui] sont les premiers " grands lecteurs " de la revue et font part de leurs propositions et remarques"*.

La publication de la revue sur papier et sur le site web est sous la direction de F. Dupuigrenet-Desroussilles.

1.1.1.1 Equipe de rédaction

L'équipe de rédaction est située à Paris et est composée de plusieurs membres permanents :

- Bertrand Calenge, rédacteur en chef
- Annie le Saux, rédactrice en chef adjointe
- Frédérique Ienemanivong
- Celestino Avelar
- Isabelle Masse
- Annie-Claude Bindyk, cyber-bibliothécaire

1.1.1.2 Comité de rédaction

Le comité de rédaction se réunit trois fois par an et précise les sommaires des différentes livraisons de la revue.

Membres :

Dominique Arot
Anne-Marie Bernard
Nic Diamant
Anne Kupiec
Pierre Le Loarer
Frédérique Saby

1.1.1.3 Conseil scientifique

Le conseil scientifique est réuni une fois par an, sous la présidence de Jean-Luc Gautier-Gentès.

Membres :

Jean-Luc Gautier-Gentès, président	
Dominique Arot	Almuth Grésillon
Patrick Bazin	Hélène Grognet
Anne-Marie Bertrand	Gérard Grunberg
Pierre Carbone	Jacques Kériguy
Roger Chartier	Benoît Lecoq
Ghislaine Chartron	Jean-Michel Paris
Guy Cobollet	Jacques Perriault
Robert Damien	Daniel Renoult
François Dupuigrenet Desroussilles	Odile Riondet
Jean-Pascal Foucher	Guy Saez
Catherine Gaillard	Laurence Tarin
Aline Girard-Billon	Valérie Tesnière

1.1.2 Les auteurs et les contenus

1.1.2.1 Les auteurs

"Les auteurs des articles proposés dans le Bulletin des bibliothèques de France sont des professionnels des bibliothèques et de la documentation, mais aussi des responsables politiques ou administratifs (élus et responsables des collectivités territoriales, présidents d'université, par exemple), des chercheurs (historiens, sociologues, chercheurs en sciences politiques ou en sciences de l'information), des professionnels de l'édition, de la culture ou de tout secteur dans lequel s'intègre l'activité des bibliothèques".

Le *BBF* bénéficie en plus de la présence de correspondants à l'étranger qui participent à la rédaction de la revue.

1.1.2.2 Les contenus

Les contenus du *BBF* qui composent l'ancienne version ont connu des modifications en ce qui concerne leur accès, leurs rubriques mais également le type de formats avec lesquels ils ont été traités.

Nous ne retiendrons que la structure actuelle du *BBF* telle qu'elle se présente depuis le numéro 2000-1 conformément à la version imprimée. Chaque numéro du *BBF* comporte donc un dossier consacré à un thème et il comprend également quatre autres rubriques :

- *Varia* présente des articles divers sur les bibliothèques et sur les disciplines qui concourent à leur activité,
- *Outils* propose des articles méthodologiques utiles aux bibliothécaires dans l'exercice de leur profession,
- *Chroniques* rend compte de la plupart des colloques et rencontres tenus en France ou à l'étranger. Cette rubrique accueille aussi des informations ponctuelles sur des sujets variés,
- *Ressources* rend compte d'ouvrages nouvellement parus, mais aussi de titres électroniques et de revues européennes récentes, tous documents contribuant à la formation des professionnels et à leur capacité à se situer dans les évolutions techniques ou sociales.

1.1.3 Le protocole de rédaction

Le protocole de rédaction donne aux auteurs les instructions pour la présentation de leurs textes. Il figure en 3^e de couverture de chaque numéro :

1) Les manuscrits sont adressés au comité de rédaction du *BBF* sur disquette par les auteurs aux formats DOC⁶ ou .RTF⁷, accompagnés d'une version papier. La frappe au kilomètre sans enrichissement est impérative. En effet, le traitement des fichiers informatiques est effectué

⁶ Format propriétaire de Microsoft Word.

⁷ Rich Text Format. Format d'échange proposé par Microsoft, pour des environnements différents.

à Paris, où une équipe est chargée de la mise en page et de la mise en forme des articles garantissant ainsi une homogénéité de présentation.

L'article doit également présenter l'institution à laquelle est affilié l'auteur ainsi que ses coordonnées, physiques et électroniques.

Les articles doivent être accompagnés du résumé en français rédigé par l'auteur (environ 100 mots), la traduction de ce résumé en anglais, allemand et espagnol est laissée à la charge du *BBF*. Les résumés présentent le contenu et les rapides conclusions de l'article.

2) Les notes infrapaginales, signalées dans le texte en appel de notes doivent être regroupées en fin d'article et numérotées de façon continue.

3) Les références bibliographiques figurent en fin d'article : les appels dans le texte sont mis entre parenthèses.

1.2. Caractéristiques du *BBF* électronique

1.2.1 L'équipe chargée de la chaîne de production

*1.2.1.1 L'équipe du *BBF* à Paris*

La production du *BBF* électronique commence à Paris avec la mise en page des fichiers source. Issus de logiciels de traitement de texte et ils sont ensuite formatés avec le logiciel de PAO, XPress pour MAC de Quark, leur mise en page dépend d'une feuille de style qui comprend une quinzaine de styles : grand-titre, sous-titre, auteur, titulature, note, bibliographie, tableau, corps de l'article. Les fichiers ainsi créés sont envoyés à l'imprimeur et à l'enssib sur CD-ROM.

Cette équipe est donc chargée de la réception des textes, de leur mise en forme et de leur homogénéisation. Le *BBF* a en effet un graphisme et une typographie à ne pas négliger, il ne suffit pas de produire que du texte mais également de restituer l'"empreinte *BBF*" avec la version électronique.

1.2.1.2 L'équipe à l'enssib

Le service des Editions⁸ est chargé de la publication web du *BBF*. R. Grenier, magasinier en chef, est responsable de la création des fichiers destinés à la mise en ligne (la question est traitée dans le rapport d'E. Romand-Monnier)⁹.

Cette version du *BBF* électronique est alimentée à chaque nouveau numéro et la mise en ligne d'un nouveau numéro se fait 2 mois après la sortie de la version imprimée du *BBF*.

1.2.2 Les rubriques et les formats proposés

Le *BBF* tel qu'on le trouve en ligne entre les numéros 95-1 et 2001-6 n'est pas homogène en termes d'organisation de la revue et des formats utilisés. En voici un tableau récapitulatif et succinct :

Numéros	Rubriques	Formats
95-1 à 95-4	Articles seulement	Word
95-5 à 96-6	Edito Résumés (français, anglais et allemand), Bibliothèque du bibliothécaire (Bibdubi)	HTML ¹⁰
	Articles Tour d'horizon (sauf n° 95-6 à 96-1 en html)	PDF ¹¹
97-1 à 99-6	Edito Résumés (français, anglais et allemand), Bibliothèque du bibliothécaire (Bibdubi) Nous avons reçu (NAR)	HTML
	Articles Tour d'horizon Nous avons reçu (NAR) (sauf n° 97-6 à 98-1 en html)	PDF
2000-1 à 2001-6	Résumés, plus un en espagnol Edito Ressources : Bibdubi + NAR	HTML
	Articles : Varia + Dossier + Outils Chroniques : Tour d'horizon	PDF

⁸ Ce service, dirigé par L. Miremont, est rattaché au département des Ressources Documentaires, sous la responsabilité d'A Meyer.

⁹ *Ibid.*

¹⁰ Hyper Text Markup Language. Langage de balises pour la présentation de pages web.

A partir de septembre 2001, la version telle qu'elle est décrite pour les numéros 2000-1 à 2001-6 coexistera avec la nouvelle version, proposée par Bertrand Calenge jusqu'à fin 2001.

En mars 2000, B. Calenge propose en effet au conseil scientifique du *BBF*, d'étudier la question de nouveau produit électronique qui permette la coexistence de la version papier et d'une version électronique enrichie. Pour Bertrand Calenge, *"un nouveau produit électronique issu du BBF et respectant son projet éditorial doit apporter une nouvelle plus-value, tenant notamment aux possibilités offertes par la technologie et aux nouvelles pratiques de lecture autorisées par l'hypertexte."*¹²

Un groupe de travail sur le projet *BBF* électronique est alors constitué.

2.à la nouvelle version en XML.

2.1. Le standard XML

Le XML est le fruit d'un groupe de travail du W3C¹³. Depuis février 1998, les spécifications XML 1.0 sont devenues des recommandations du W3C, ce qui signifie qu'il est recommandé de s'y référer pour utiliser le XML dans des applications informatiques. Tous les documents liés à XML (documentation, spécifications...) sont consultables et téléchargeables sur le site www.w3c.org/XML/.

2.1.1 Caractéristiques du XML

Sous-ensemble de SGML, il possède la plupart de ses caractéristiques tout en étant plus facile à manipuler :

- Extensible : il peut décrire n'importe quel type de données,

¹¹ Portable Document Format.

¹² *Ibid.*

¹³ World Wide Web Consortium, fondé en 1994, crée des standards pour le Web et a pour principale mission d'optimiser le potentiel du Web tout en développant des technologies qui favorisent l'échange d'information, le commerce, le libre-arbitre et la compréhension collective.

- Format de description de données auto-descriptif (métalangage) : modélisation du document grâce à des balises qui décrivent la structure et le contenu du document et non sa présentation (comme le fait le HTML). Les DTDs¹⁴ et les schémas XML sont des grammaires qui définissent la structuration d'un document mais elles ne sont pas obligatoires :
 - o Le document XML seul est dit « bien formé »
 - o Le document XML lié à une DTD est dit « valide »
- Les balises correspondent à des éléments qui peuvent comprendre des attributs. Les éléments ont des valeurs, ils sont définis par des types de données et caractérisés en terme de fréquence.

La sortie, papier (imprimante) ou électronique (CD-ROM, écran, fichiers...) du document est prise en charge par d'autres langages : CSS, XSL, XSLT, technologies également développées au sein du W3C et développées plus loin. Le XML se présente donc comme un nouveau format universel pour le stockage des données.

2.1.2 Avantages du XML

L'enthousiasme autour de XML est justifié en raison de ses nombreuses qualités.

XML représente avant tout un format d'échanges standardisé. Reconnu par le W3C et libre de droits, il est plus stable et plus lisible que les formats propriétaires et c'est ce qui en fait aussi un format pérenne. Cette liberté facilite en plus sa diffusion et son utilisation puisqu'il suffit d'un parseur (analyseur, ou parser en anglais) pour valider et visionner un document XML.

Sa structure arborescente, grâce aux DTDs entre autres, permet de modéliser les contenus et la majorité des situations de gestion informatiques. Des documents ainsi structurés permettent de créer des ensembles de documents homogènes et rendent possibles les recherches très précises sur le contenu d'un document, en fonction de sa structure et des informations qui s'y rattachent.

¹⁴ Document Type Definition. "Moule" qui correspond à la structure syntaxique du document.

Sa lisibilité et sa portabilité le rendent accessible à tous puisque différents jeux de caractères sont pris en compte et que les principes du langage XML sont abordables par tous.

XML est également adaptable à tous les protocoles transférant du texte puisqu'un document XML ne contient que des caractères ASCII.

Son extensibilité enfin le rend opérationnel dans n'importe quel domaine d'applications.

2.2. Le projet et ses évolutions

2.2.1 Une revue électronique enrichie

2.2.1.1 Contenus : enrichissement et nouvelles offres

Le nombre de consultations de la version électronique déjà existante n'a cessé d'augmenter. Le paysage éditorial a évolué dans un contexte électronique où les nouvelles technologies, telles que celles liées à XML, permettent d'améliorer les services éditoriaux et d'ajouter une plus-value à l'édition électronique. Ce sont les raisons pour lesquelles B. Calenge a revu la version électronique du *BBF* pour un nouveau produit proposant un contenu riche et une offre innovante et utile.

De nombreuses questions ont été abordées depuis 2000 sur les solutions à mettre en place mais les principes de base sont, durant 1 an, à peu près, restés les mêmes :

- garder le format pdf, issu d'XPress et destiné à l'impression
- formater en xml, conformément à une DTD, pour l'accès aux documents sur le web
- offre simultanée d'une revue enrichie et d'une base thématique.

Le *BBF* électronique s'enrichit dès le n° 2001-5 grâce à 3 rubriques d'actualités signalant et analysant l'actualité des bibliothèques sur internet :

- o "*Panorama international du Web*" correspond au thème de chaque dossier

- o "*Actualités des textes officiels*" signale les textes réglementaires et rapports officiels récemment parus et disponibles en ligne
- o "*Débats et publications électroniques*" liste les débats en cours sur des listes de discussion et rend compte des contenus de plusieurs lettres électroniques.

Ces rubriques ont nécessité l'embauche, par l'équipe du *BBF*, d'une "cyberbibliothécaire" en mai 2001, Annie-Claude Bindyk.

La base thématique, quant à elle, sera composée de tous les numéros du *BBF* et organisée autour de 3 types d'indexation :

- o indexation traditionnelle de type auteurs, mots du titre, matière Rameau
 - o indexation via des concepts, thèmes, combinables entre eux
 - o indexation de tous les mots pour une recherche en texte intégral
- proposer des services personnalisés en adéquation avec les besoins des utilisateurs (profils, annonce d'événements...).

2.2.1.2 *Ergonomie des contenus*

Le travail sur le *BBF* électronique consistait non seulement à offrir une plus importante richesse de contenus mais aussi à offrir de nouvelles possibilités d'accès, une nouvelle ergonomie telle que :

- la possibilité pour les utilisateurs de se construire des profils
- la génération automatique de courriers électroniques à parution d'un nouveau numéro en fonction du profil de l'utilisateur
- proposer des modes d'édition variés
- présenter les notes et la bibliographie de façon ergonomique
- développer les possibilités d'une lecture hypertextuelle

2.2.1.3 *Financement*

Dès décembre 2000, le *BBF* dispose d'un financement accordé par le ministère pour mener à bien le projet à hauteur de 400 KF, puis 220 KF

par an recouvrant la réalisation de nouvelles rubriques et la gestion de l'intégration de chaque numéro en XML.

Il était envisagé de rétro-convertir le fonds depuis 1995 jusqu'à 2001, hors le budget accordé ne permet d'agir que sur les années 2000 et 2001.

2.2.2 Réalisations depuis 2000

E. Romand-Monnier, stagiaire DESSID l'an passé et actuellement contractuelle au service informatique de l'enssib, a proposé lors de son stage, un scénario de traitement informatique pour la migration de la revue sous forme papier vers l'électronique¹⁵. Ce traitement s'appuyait sur une DTD-BBF et les attentes concernaient le logiciel avenue.quark, capable de générer automatiquement un fichier XML lié à une DTD à partir du document Quark XPress.

Le problème était double :

- les styles utilisés pour la mise en page des documents n'étaient pas les mêmes que ceux développés dans la DTD et il aurait été trop complexe d'adapter la feuille de style XPress aux balises de la DTD-BBF.
- la même démarche a été faite à partir de Word et avec l'outil XMLSpy, il garantissait la conversion d'un document Word stylé en XML. Après une phase de tests, il s'est avéré que XMLSpy ne savait pas traduire ces styles.

Suite à cela, E. Romand-Monnier n'envisage plus qu'un balisage manuel pour structurer les documents en XML : cette tâche est totalement confiée à un vacataire embauché par le service informatique pour une durée de 5 mois et ayant pour mission, le balisage des numéros des années 2000 et 2001 du *BBF* en XML.

Elle a donc perfectionné le procédé de balisage pour le rendre plus facile au baliseur mais surtout elle lui a expliqué chacune des DTDs-BBF (structure de la DTD, définition des éléments et des attributs, validation du document grâce au parseur ...).

¹⁵ *Ibid.*

2.2.3 Etat du projet en juin 2001

Devant le manque de stabilité des outils, les nombreuses solutions envisagées n'ont pu aboutir. Le service informatique de l'enssib a donc demandé des devis à plusieurs prestataires concernant le développement de la base du nouveau *BBF* électronique à partir des différentes DTDs-BBF. E. Romand-Monnier a en effet éclaté la DTD-BBF en plusieurs, en fonction des rubriques pour lesquelles elles sont utilisées, sur le conseil d'un consultant.

Les devis ne devaient pas excéder 105 KF.

Les demandes de devis concernaient également le graphisme et la typographie du site et le budget était plafonné à 30 000 francs.

Araxe¹⁶, a proposé à l'enssib d'effectuer les développements ainsi que la partie graphique du site du *BBF* électronique pour une somme correspondant au budget. Le service informatique a alors validé la participation d'Araxe aux développements mais le graphisme devait être soumis au jugement de l'équipe à Paris.

Dès lors, des réunions eurent lieu dans les locaux d'Araxe, durant tout le mois de juin. Elles ont permis de rédiger un dossier de conception sur les développements informatiques de la base du *BBF* et de ses fonctions. L'équipe de Paris ayant choisi de faire confiance à Araxe pour le graphisme, il fallait donc mener de front l'intégration du graphisme et les développements en cours.

Parallèlement, le baliseur traitait 30/40 articles en une semaine à temps plein.

¹⁶ Société spécialisée dans les services pour le web et le commerce électronique. Adresse : 53 rue de l'Etang – 69760 LIMONEST. URL : www.araxe.fr

3. Définition du stage

3.1. Mission du stagiaire

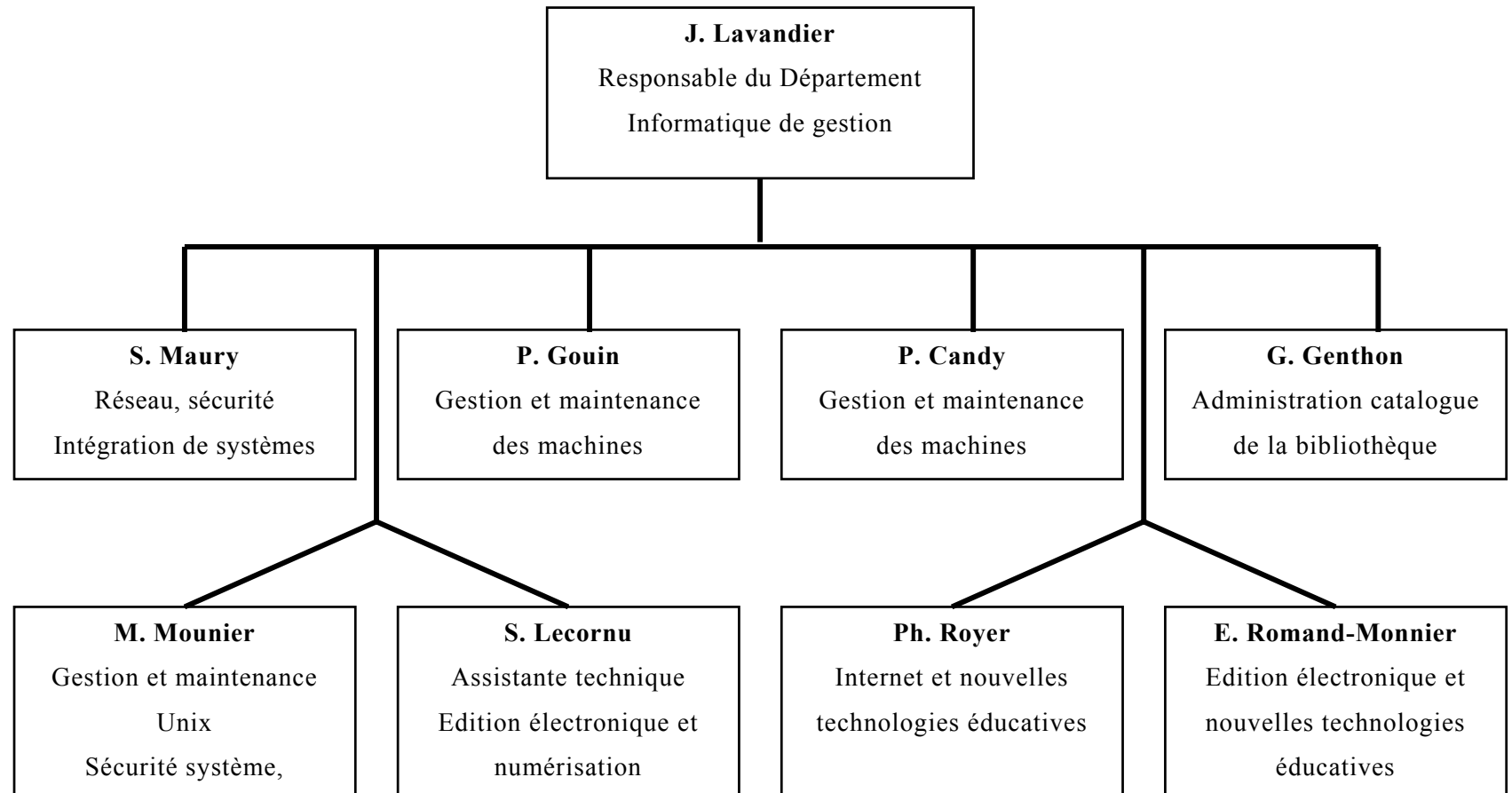
L'intitulé du stage tel qu'il m'a été proposé en avril/mai 2001 a évolué en raison du manque de stabilité des outils utilisables.

En juin, J. Lavandier m'a donc proposé de reprendre la chaîne éditoriale du *BBF* et chacune de ses étapes. Ce stage avait un double intérêt :

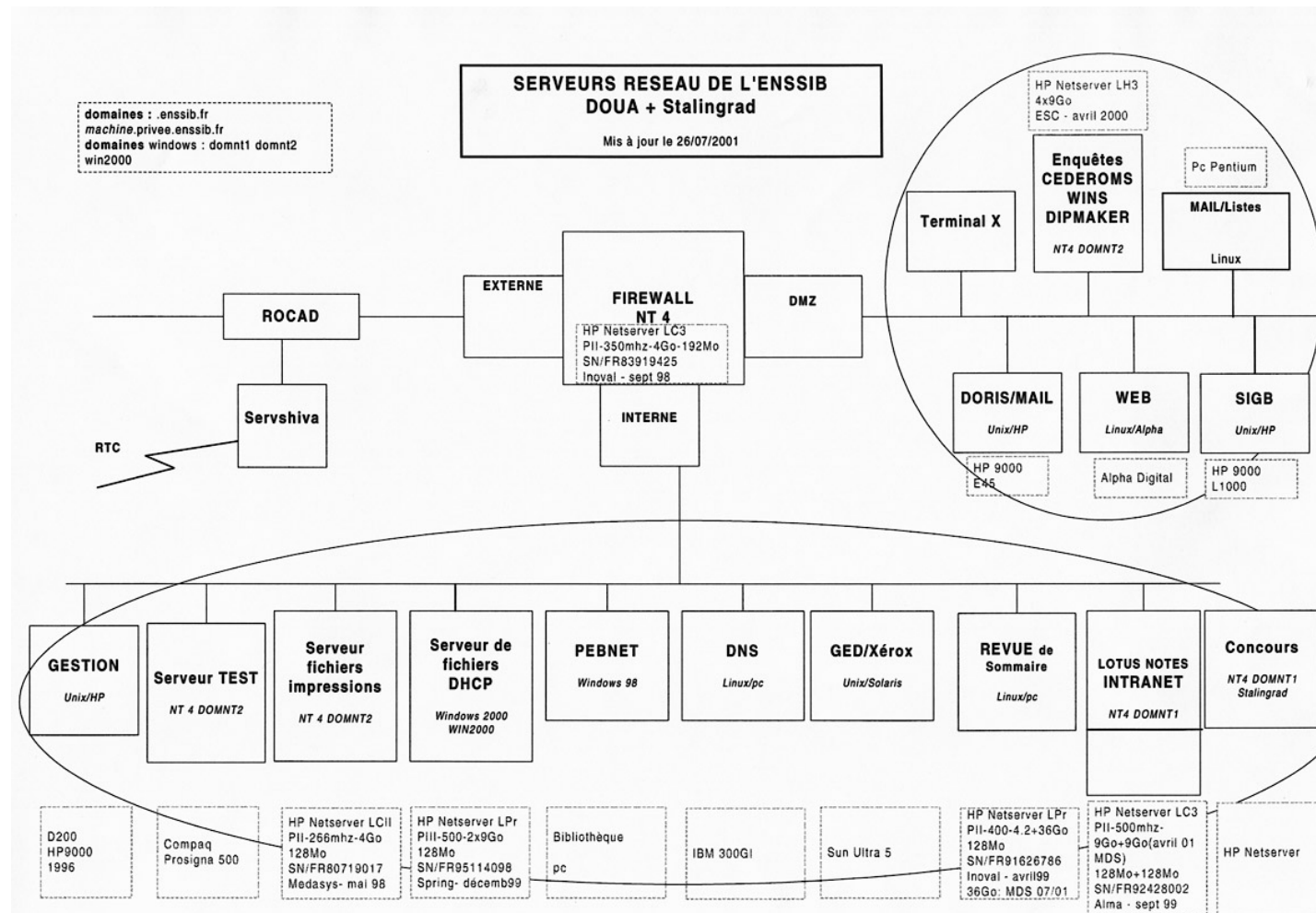
- situer les processus d'édition par rapport à ce qu'il est possible de faire avec d'autres outils s'appuyant sur la technologie XML,
- positionner le service informatique par rapport à d'autres projets d'édition électronique en XML et leurs réalisations

Elle m'a donc demandé de faire un état de l'art d'outils et de projets basés sur la technologie XML. Elle souhaitait également que j'offre des pistes de travail sur la structuration future des mémoires d'étudiants en XML afin d'évaluer quelles seraient les implications de son service sur cette question.

3.2. Cadre du stage



3.2.2 Topologie du réseau informatique de l'enssib



3.3. Déroulement du stage et méthodes de travail

3.3.1 Rencontres avec Araxe et comptes-rendus

Le premier jour de mon stage coïncidait avec la première réunion chez Araxe. Ces réunions ont duré tout le mois de juin : E. Romand-Monnier a été nommée chef de projet et j'étais chargée des comptes-rendus et de leur diffusion à l'ensemble des membres du projet.

En l'absence de synthèse récapitulative, la difficulté était de rendre compréhensible au développeur d'Araxe les différentes caractéristiques de structure et de graphisme du *BBF*. Nous leur avons donc fournis :

- une maquette, créée par P. Royer et élaborée pour la validation du nouveau produit par les membres du *BBF*, donnant une idée sur l'interface graphique du *BBF* (onglets "Rechercher", "Feuilleter", "Profils", couleurs ...)
- un schéma sur papier de l'arborescence du site avec les caractéristiques des pages telles que statiques/dynamiques, informations récurrentes à certaines pages, modalités de recherche (texte intégral, titre-auteur-résumé...)
- différents écrans, un pour chacune des rubriques, reprenant les éléments graphiques et typographiques particuliers à inclure ainsi que leur positionnement sur les pages ("imassettes" et notes de bas de pages dans la colonne de gauche, titres en bleu ou en rouge, boutons destinés à l'impression html ou pdf ...)

Araxe a enfin rédigé un dossier de spécification au cours du mois de juin et nous avons dû régler les imprévus de dernière minute comme revoir certaines DTDs, discuter des modalités de recherche et d'indexation mais également de connexion et d'identification.

3.3.2 Recherches sur internet

Parallèlement je réalisais l'état de l'art demandé par J. Lavandier. J'en propose un tableau récapitulatif accompagné de schémas dans les parties 2 et 3.

J'ai d'abord recensé des sites ressources sur le XML m'informant ainsi de ses dernières évolutions et des produits qui lui sont dédiés. E. Romand-Monnier m'a aussi fourni les noms et adresses d'initiatives connues. Le référencement s'est principalement fait à partir de :

- moteurs de recherche avec des mots-clés tels que "XML", "édition électronique", "documentation électronique" ...
- pages web de projets qui présentent les outils et applications qu'ils ont utilisés
- sites des entreprises enfin qui donnent en ligne les caractéristiques techniques de leurs produits mais sans le prix.

J'ai également pris contact avec des éditeurs tels que EDP Sciences et les Editions Masson pour connaître plus en détails la chaîne de production de leurs revues en XML : seul EDPSciences m'a répondu. C'est la raison pour laquelle je l'ai retenu dans la description que je fais de certains projets en partie 3.

3.3.3 Etude de la procédure de balisage

Le baliseur a été, jusqu'au mois de juillet, a été le même. Il a travaillé sur tous les numéros de l'année 2000 (sauf le 3 et les numéros 1 et 2 de 2001). Il était donc la seule personne ressource pouvant décrire précisément les différentes étapes de balisage du *BBF*, de XPress vers XML, ainsi que les difficultés qui en découlent.

Grâce à son expérience de balisage et aux compétences techniques d'E. Romand-Monnier, il a été possible d'envisager d'autres traitements à mettre en place pour faciliter la chaîne de production.

La chaîne de production du *BBF* électronique

Je signifierai ici les méthodes de travail qui ont été exploitées pour le balisage. Des changements ont été amorcés au cours de mon stage en fonction des différents tests que nous avons faits sur le traitement de certains fichiers.

Les observations faites pendant les deux premiers mois ont permis de mettre en place une méthodologie de balisage et de l'intégrer dans les tâches du service des Editions.

1. La récupération des textes

Il y a eu 2 procédés :

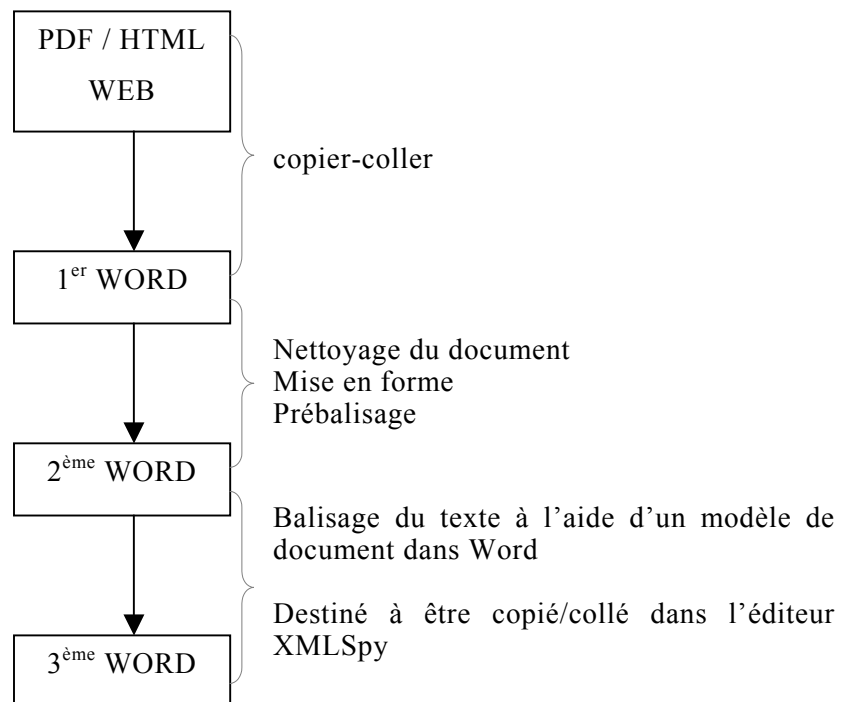
- Récupération des fichiers PDF en ligne depuis 2000 sur l'ancien site du *BBF*
- Récupération de fichiers RTF envoyés par l'équipe de Paris et balisage grâce à des macros sous Word

1.1. Récupération des fichiers PDF et HTML

1.1.1 Traitement

Rappelons que le format initial des articles fournis par les auteurs est de type traitement de texte. Les documents ont été mis en page avec XPress pour Mac alors que le baliseur travaille sur PC.

Il s'agit de rétro-convertir des numéros au format Word alors qu'ils étaient destinés au départ, à un affichage sur le web, en PDF ou HTML. Ce procédé de rétro-conversion est lourd et comprend le fonds 2000/2001-2.



Le nettoyage du document consiste à supprimer les espaces qui se glissent entre les caractères, les retours chariot dus à l'importation de colonnes depuis le PDF et supprimer les éléments graphiques caractéristiques de la version papier et inutiles pour l'électronique.

La mise en forme consiste à séparer les titres du corps du texte et restituer les paragraphes conformément à la version papier.

Il faut également récupérer des informations éparpillées dans le document et les insérer de façon à respecter l'ordre de la structure de la DTD. Effectivement lors de la validation du document XML, XMLSpy¹⁷ balaie la DTD du début à la fin du document, si l'ordre n'est pas respecté, il ne valide pas le document. Il vaut donc mieux ordonner les différents éléments dès le début , par exemple :

¹⁷ Editeur et parseur XML, il permet de créer des DTDs, des schémas et des documents XML et de les valider par rapport à des structures préalablement déclarées. Téléchargeable pour une durée limitée de 30 jours : www.tucows.com

- au début du document insérer la date sous la forme "Décembre 1999", en 3^{ème} ligne, et la pagination sous la forme : "p. 10-17" en dessous.
- La bibliographie de l'auteur à insérer sous les coordonnées de l'auteur

Le prébalisage consiste à repérer les notes de bas de pages et les citations puisqu'elles subissent un traitement particulier au moment du balisage XML :

- les notes de bas de page sont regroupées en fin de document
- grâce à la fonction Rechercher/Remplacer, on recherche tous les caractères italiques que l'on remplace par ses mêmes caractères d'une couleur différente.

Ces procédés permettent de repérer la structure syntaxique de l'article et d'anticiper sur le balisage.

Le balisage du texte se fait à partir de modèles dans Word, il y a un modèle par rubrique. Chacun de ces modèles est composé de balises, copie des différents éléments "pères" d'une DTD et des raccourcis clavier permettent d'insérer des éléments "fils". Le baliseur a utilisé un tel système de façon à travailler plus rapidement.

Ce fichier Word final est destiné à être copié/collé dans l'éditeur XML, XMLSpy pour le balisage final.

1.1.2 Solutions envisagées

L'une des solutions aurait été de traiter les fichiers directement à partir du fichier XPress mais il y a plusieurs inconvénients liés à l'incompatibilité des systèmes MAC et PC.

Premièrement, Laurent doit changer de poste de travail car les fichiers XPress sont destinés au MAC et il travaille sur PC.

Deuxièmement, il est certes possible de convertir automatiquement un format XPress pour Mac vers un format Word pour PC. Mais les paramètres de mise en page et les caractéristiques typographiques du

BBF sont tellement élaborés qu'il est impossible de traiter de tels blocs automatiquement et qu'il convient de charger systématiquement les nouvelles polices.

L'autre solution aurait été d'utiliser le logiciel *avenue.quark*, module spécifique se greffant à *Quark XPress*. Il permet la conversion automatique d'un document natif *XPress* en document XML à partir d'une DTD. Dans leur article¹⁸, G. Beaudry et G. Boismenu en recommandent l'utilisation. Certaines revues destinées au papier et qui migrent vers l'électronique suivent selon eux le modèle de "*mise en page vers XML*" et utilisent ce logiciel.

Nous l'avons testé et il présente quelques inconvénients liés à la mise en page faite sous *XPress* et à la structure des DTDs-BBF.

XPress permet une mise en page par blocs de texte. Le bloc est ensuite inséré automatiquement dans un seul élément de la DTD alors que ce bloc comprend le titre, l'auteur, informations sur l'auteur, le texte, la bibliographie autant d'informations qui rentrent dans des balises particulières de la DTD et qui ne peuvent être reconnues que par celui qui balise.

Les blocs de texte sont donc un découpage trop grossier par rapport aux possibilités d'affinement des DTDs-BBF permettent en effet de baliser des citations, des appels de note, ou autres éléments très précis dans le texte.

Les logiciels de PAO ou de traitement de texte ne stylent que dans "deux dimensions" : les styles de caractères et les styles de paragraphes. Le XML permet de multiplier les attributs sur un seul élément, c'est en cela qu'il fournit des données sur les données. Les DTDs-BBF ont largement exploité cette possibilité, le traitement automatique ne peut se faire et le balisage manuel est pour le moment indispensable.

Il était donc préférable de récupérer une version pdf "finie" et de nettoyer ensuite le fichier Word.

¹⁸ BEAUDRY G., BOISMENU G. *Expertise technique et organisationnelle* [On-line] Villeurbanne, [consulté le 02/07/01].
Available from internet : <url: <http://revues.enssib.fr/Index/indextecnic.htm>>

Pour les numéros à venir, il a par contre été demandé à l'équipe de Paris de nous fournir des fichiers RTF.

1.2. Amélioration des procédures

1.2.1 Format RTF

Depuis le n° 2001-3, l'équipe de Paris envoie les fichiers en RTF sur CD-ROM accompagné des impressions 1^{er} jet du *BBF* version papier.

L'avantage est que la mise en forme du texte est correcte (titres séparés du texte par exemple). Les styles ne sont pas préservés mais le nouveau baliseur qui a commencé à baliser fin août s'appuie sur la version papier pour baliser. Les étapes de conversion du texte en Word et de nettoyage ont été ainsi supprimées.

Par ailleurs, la procédure de balisage a été automatisée à l'aide de macros.

1.2.2 Utilisation de macros

Les macros sont des programmes informatiques développés grâce au module Visual Basic intégré à Word. Ces macros sont incluses dans un modèle de document appelé *bbf_balisage*. Dans Word, on peut créer une barre d'outils composées de commandes (Affichage – Barre d'Outils – Personnaliser) qui permettent d'appeler les macros attachées au modèle.

Il existe des macros pour chaque type d'articles : *Varia*, *Dossier*, *Chroniques*, *Outils*, *Bibliothèque du bibliothécaire*, *Nous avons reçu et Gazette électronique*.

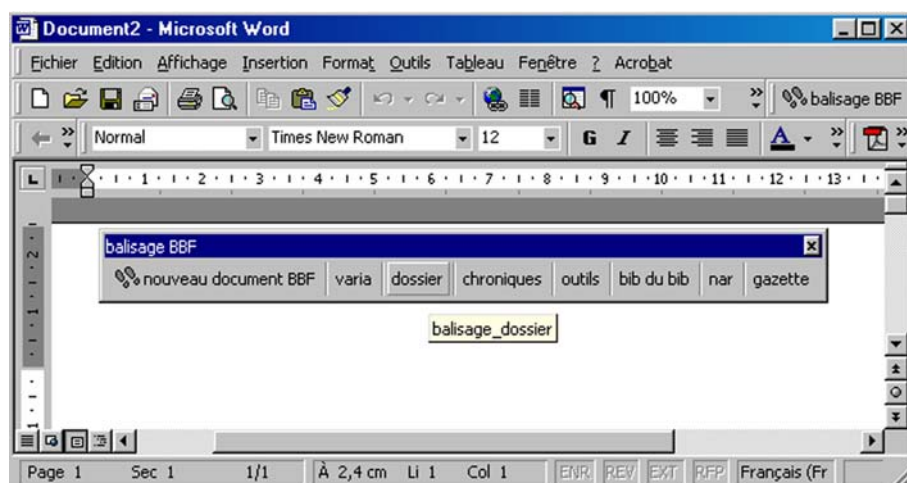


Figure 1 – Appel de la barre d'outils balisage_dossier.

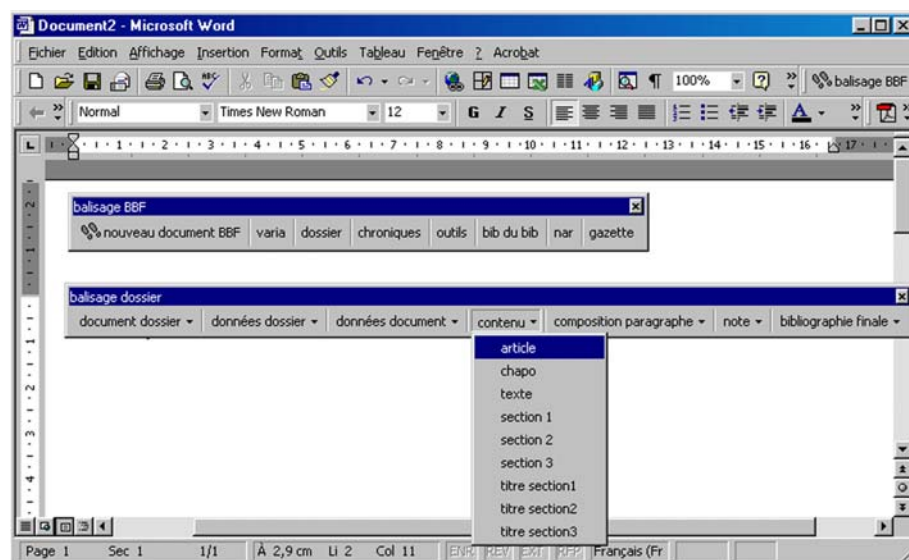


Figure 2 – Appel de la macro Article

Le principe de ces macros est de partir des éléments pères pour arriver aux éléments fils, par exemple des blocs de paragraphe aux citations contenues dans ces paragraphes. Il suffit pour baliser de surligner la partie du texte constituant un élément et de cliquer sur le nom de la macro adéquat. Des balises encadrent alors la sélection.

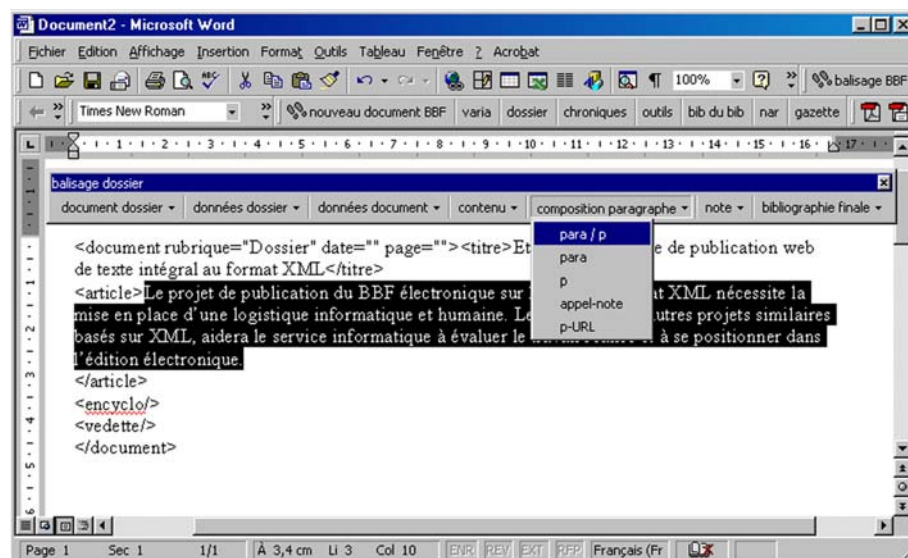


Figure 3 – Exemple de balisage

Le modèle et les macros ont donc été installés sur le poste du baliseur. Dès le premier numéro, ses remarques ont permis à E. Romand-Monnier d'apporter les modifications afin que les macros collent à la manière de travailler du baliseur.

Au final, le texte est balisé et enregistré en format Word (.doc).

2. Les différents formatages

2.1. Formatage en XML à l'aide des DTDs-BBF

2.1.1 Présentation des DTDs

Les documents XML, réalisés à partir d'une DTD et conformes à celle-ci, sont dits valides. La DTD est déclarée dans le document XML ou dans un fichier externe.

En voici une définition d'après Alain Michard : *"Les documents valides obéissent à une structure type prédéfinie. La définition de document type (Document Type Definition, DTD) est le mécanisme par lequel de telles*

structures sont spécifiées. Un document valide sera toujours une instance de la structure définie dans une DTD, et un élément de ce document sera lui-même une instance de la structure particulière définie dans la déclaration du type d'élément correspondant figurant dans cette DTD"¹⁹.

Les DTDs-BBF ont été réalisées par E. Romand-Monnier à l'aide du logiciel XMLSpy. Il existait au départ une seule et unique DTD-BBF mais sur les conseils d'un consultant, elle les a éclatées en autant de rubriques que comprend le *BBF* électronique.

Ces DTDs-BBF sont complexes et très structurées, elles reprennent au plus près la structure papier de la revue.

Les différentes DTDs sont :

DTDedito	DTDdossier
DTDchroniques	DTDoutils
DTDvaria	DTDnar
DTDbdb	DTDe-gazette

Afin de mieux se rendre compte des informations détaillées présentes dans les DTD, il est préférable de se référer à l'annexe 1 qui présente l'arborescence de la DTDgazette.

2.1.2 Création de l'article en XML et indexation

Il s'agit tout d'abord, pour les baliseurs de reprendre le fichier Word final et de le copier/coller dans un nouveau fichier XML, dans XMLSpy, lié à l'une des DTD.

XMLSpy permet alors de valider le document XML en vérifiant que la structure de l'instance suit l'arborescence de la DTD : au moment de la validation, il surligne les éléments incorrectement balisés. Le baliseur corrige alors les dernières imperfections du balisage. Il doit également insérer les liens, sous forme des chemins des fichiers, pour les images (adresse relative) et les fichiers PDF (adresse absolue).

¹⁹ MICHARD A. *XML, Langages et applications*. Paris, 1998, Ed ; Eyrolles, pp. 57

Le traitement des images est décrit plus loin, la gestion des fichiers également.

Une fois l'indexation thématique réalisée, le document XML est totalement balisé. En effet, B. Calenge a introduit ce nouveau type d'indexation qui donne lieu à une nouvelle fonctionnalité de recherche. Cette indexation s'applique à chaque document excepté ceux appartenant aux rubriques *Edito* et *Nous avons reçu*. Elle s'effectue à partir de 8 concepts-clés et de leur sous-concepts. La liste a été établie par l'équipe du BBF :

- Objet	- Contenus
- Etablissement	- Lieu
- Support	- Histoire
- Publics	- Forme

La liste détaillée des concepts est donnée en annexe 2.

2.2. Formatage en PDF

Cette manipulation existait pour la version en ligne précédente. Depuis la version 5 d'Adobe Acrobat, elle s'exécute automatiquement à partir d'Xpress et offre de nouvelles fonctionnalités. Parmi elles, la navigation hypertextuelle à partir de liens cliquables dans le texte ou à partir de signets créés en fonction de la feuille de style du document originel, et ainsi la hiérarchisation des styles permet la navigation de titres en titres.

L'intérêt de conserver un tel format est qu'il est avant tout destiné à l'impression et qu'il reprend, conformément à la version papier, les différentes caractéristiques typographiques, graphiques et de mise en page.

L'un des souhaits de B. Calenge est de garder l'apparence actuelle du document PDF qui reprend exactement les mêmes éléments graphiques, typographiques et de présentation que la version papier du *BBF*.

La version PDF des documents du *BBF* électronique – destiné à l’affichage et à l’impression - est généré de cette façon à partir des fichiers XPress.

2.3. Formatage en HTML

Le nouveau site du *BBF* électronique offrira deux types de pages web : des pages statiques et d’autres dynamiques.

2.3.1 Les pages statiques

Les fichiers statiques correspondent pour la plupart aux pages d’information, rédigées par B. Calenge pour expliquer les nouvelles modalités de recherche et les nouvelles fonctionnalités du site. Elles ont été développées en interne par l’enssib à l’aide de l’éditeur HTML Dreamweaver.

D’autres pages statiques telles que les pages de recherche et la page d’accueil ont été développées par Araxe.

Ces pages ne reprennent donc aucune donnée XML.

2.3.2 Les pages dynamiques

Elles sont développées par la société Araxe et sont issues de la collaboration entre leur graphiste et leur développeur.

2.3.2.1 Affichage en HTML sur le web

Les fichiers XML balisés s’affichent sur le web en HTML. En effet, l’équipe du BBF a pris le parti d’un affichage HTML plutôt que XML, car seules les dernières versions des navigateurs (version 5 pour Internet Explorer et version 6 de Netscape) lisent le XML. Or dans le cas du *BBF*, il est apparu que la plupart des utilisateurs n’en possèdent qu’une ancienne version. Il existe tout de même des plug-in qui, ajoutés au navigateur, permettent la lecture des documents XML.

2.3.2.2 XSL et XSL-FO

Définition

Les feuilles de styles XSL permettent, d'une part de définir le langage de transformation XSLT²⁰, d'autre part, de spécifier une sémantique de formatage. C'est grâce à XSL que peuvent être réalisées des présentations de document XML, quelles soient liées au papier, au Web, ou à tout autre support électronique. Ce langage a été proposé comme recommandation du W3C, le 28 août 2001 et la réponse sera donnée le 25 septembre.

Objectif

L'objectif de XSL est de définir un langage de présentation de document, indépendant des systèmes et des logiciels.

Principes

Une feuille de styles XSL est définie comme un document XML contenant des spécifications de transformation et de formatage d'objets. Elle permet de transformer un document XML d'entrée en un autre document XML, dont les éléments de structure sont tous liés à des éléments typographiques représentant des intentions de formatage : des pages, des fenêtres, des paragraphes, des listes, etc.

Le document XML résultant de la transformation doit ensuite être pris en charge par un outil de formatage, qui créera une version papier, PDF, RTF, HTML ou tout autre format. La spécification différencie donc deux processus : la transformation d'arbre (*tree transformation*) et le formatage (*formatting*).

Le document présenté pouvant être structurellement très différent du document XML d'origine, tout le pouvoir de transformation d'XSLT doit pouvoir être utilisé dans la première partie du processus, pour ajouter, par exemple, des tables de matières ou encore filtrer et réordonner des informations. XSLT est donc le langage spécifique de transformation en vue de la mise à page d'un document, sur papier, sur le Web...

²⁰ eXtensible Style Sheet Language Transformation.

Dans le cas du *BBF*, le fichier XML est transformé "à la volée" et formaté dans un fichier de type XSL-FO²¹, puis de XSL-FO vers du HTML pour être lisible sur le web.

XSL Two Processes: Transformation & Formatting

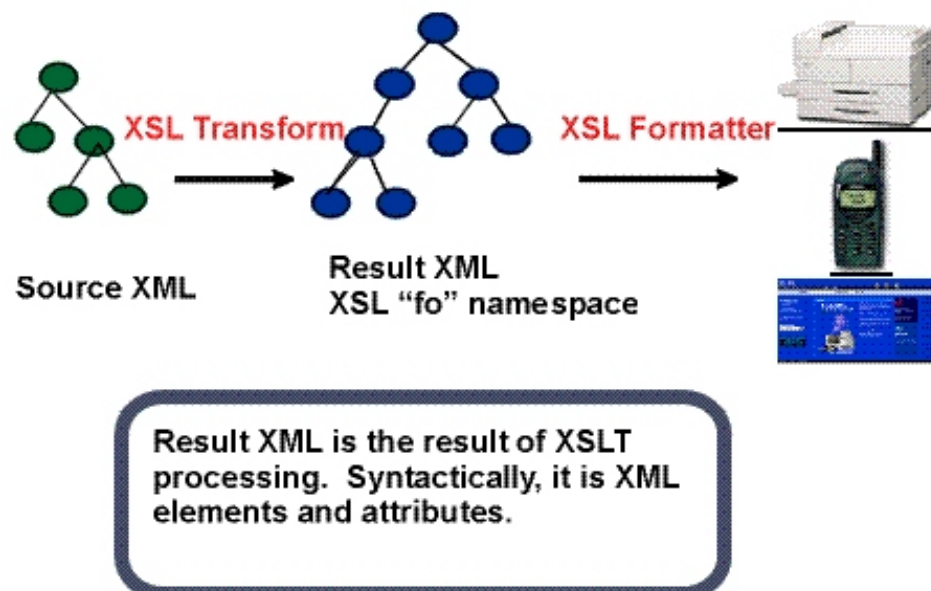


Figure 4 – Principes de fonctionnement des processeurs XSL²²

Le processus des feuilles de style XSL permet donc d'obtenir une sortie PDF. Si l'on décide de formater un document XML en PDF en utilisant XSL, il faut savoir que la mise en page sera épurée et très succincte. Or B. Calenge souhaitait surtout conserver la présentation graphique et les caractéristiques typographiques de la revue papier, ce qui permet le formatage depuis XPress vers le PDF. Nous n'avons donc pas utilisé cette potentialité de XSL-FO.

2.3.2.3 CSS

La mise en page finale du document sur le web est elle aussi générée "à la volée" et dépend des feuilles de style CSS²³, destinées au web et

²¹ eXtensible Style Sheet Language – Formatting Objects

²² Schéma téléchargé sur le site : www.mutu-xml.org/xml-base/shared/KEY-XSL.html

²³ Cascading Style Sheet, feuilles de style en cascade.

appliquées aux fichiers HTML. Ces feuilles de style sont des recommandations du W3C et sont désormais reconnues par tous les navigateurs.

Leur principe permet à des auteurs et à des utilisateurs d'attacher un modèle (par exemple, polices, interlignes, et caractères spéciaux) aux documents structurés (par exemple, des documents de HTML et des applications de XML). Cette feuille de style permet en fait de séparer le modèle de présentation des documents de leur contenu, et sa dernière recommandation, CSS2, simplifie la rédaction de pages Web et l'entretien de site.

Une page HTML épurée de cette feuille de style est proposée, elle est principalement destinée à l'impression.

3. Droits d'auteur et traitement des images

3.1. Droits d'auteur sur les images

Les images du *BBF* correspondent à des photos, des tableaux, des graphiques ou encore des encarts de texte. La diffusion des photos est soumise à l'autorisation de l'auteur, sauf si ce dernier les laisse libres de droits.

Le BBF négocie les droits pour une période de six mois puis il faut payer à nouveau pour prolonger l'exploitation des photos. I. Masse signalera alors à l'enssib, au terme de ces six mois, les images à retirer du site.

Cette réglementation sera appliquée dès le numéro 2001-5 ; pour les numéros antérieurs, il n'y a pas de photos disponibles.

Par contre, les tableaux et autres graphiques sont libres de droit et font parties intégrantes de l'article.

3.2. Traitement des graphiques et des tableaux

Les illustrations de la période 2000-2001 ont été traitées de deux façons :

- Copies d'écran à partir de l'affichage en ligne de l'article en PDF

- numérisation quand l'image est trop grande pour une copie d'écran

Pour l'un ou l'autre de ces traitements, il faut garder à l'esprit la taille de l'image pour l'écran (800*600 pixels) **et** pour l'impression (21*29,7 cm).

Les copies d'écran ont été ensuite collées dans le logiciel de traitement des images, Photoshop : nettoyage en rognant les bords de l'image.

Quant à la numérisation, elle s'est faite à partir d'un scanner Agfa. Les paramètres du logiciel étaient les suivants :

- type d'images : 256 couleurs ou Photo couleur (la qualité de l'image était meilleure en Photo couleur pour certaines)
- destination : Ecran
- résolution pour le web : 72 dpi
- augmenter les dimensions de sortie (400 %) : améliore la qualité de la numérisation

Le traitement se fait aussi sous Photoshop, il convient de choisir :

- Couleurs RVB
- Réglages au niveau des courbes (couche noire) : rendre le fond plus blanc et augmenter la couleur noire des pixels.

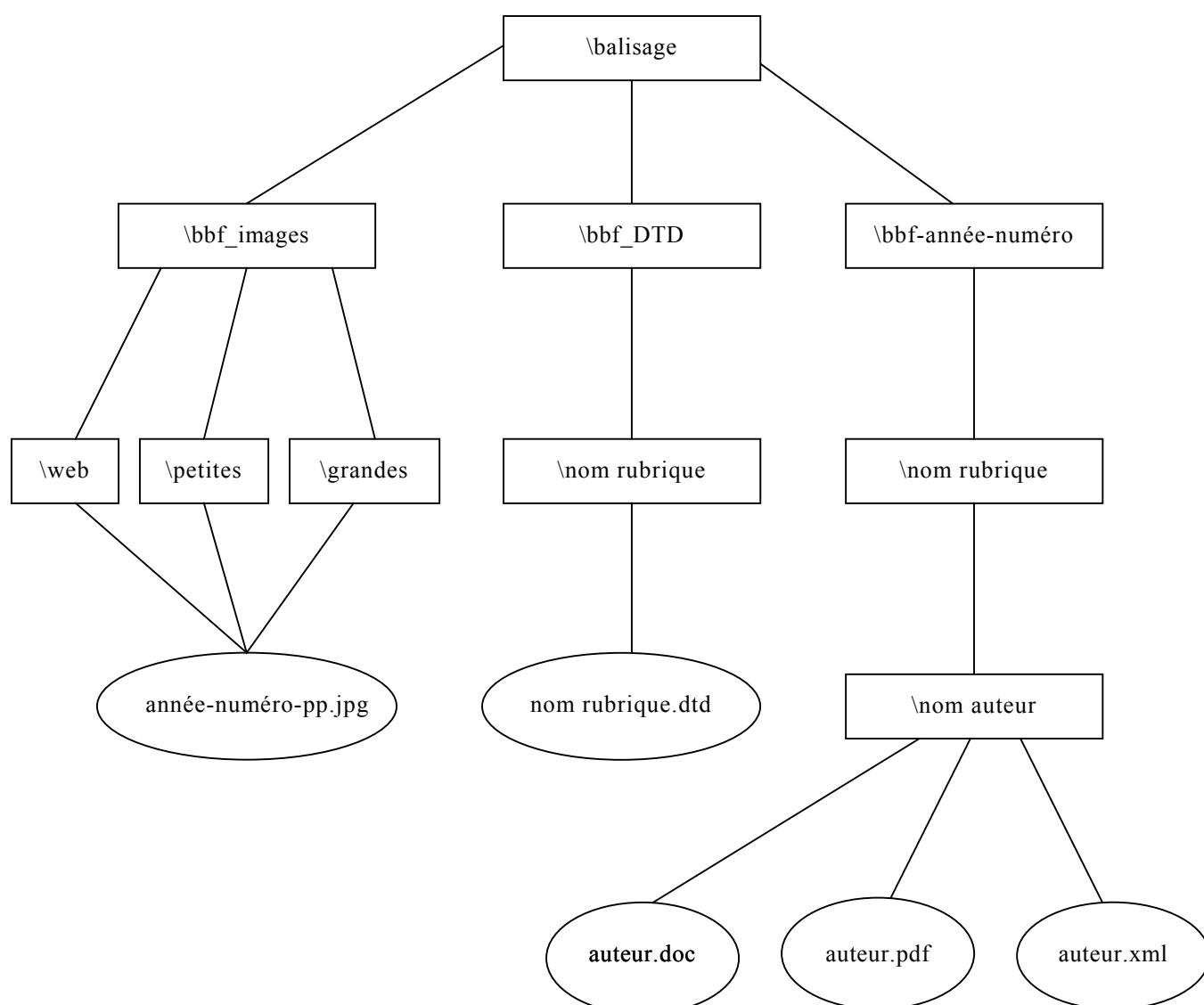
Le poids d'une image doit avoisiner 50 Ko, pour certaines 100 Ko. L'affichage sur le web doit en effet être rapide et adapté à l'équipement des utilisateurs. Après plusieurs tests, il s'est avéré que l'enregistrement en JPEG de qualité moyenne, niveau 5, était le plus intéressant car moins volumineux que le GIF.

L'équipe de Paris enverra pour les futurs numéros, les images en JPEG.

4. Système de gestion des données pour le BBF

4.1. Gestion des fichiers sur le serveur des Editions

Le stockage des fichiers pour leur création et leur traitement se fait sur le serveur des Editions dans un dossier "balisage" suivant cette structure :



Les grandes images correspondent aux photos, tableaux et images liées aux articles. Elles sont destinées à un affichage en pleine fenêtre sur le web.

Les petites images sont générées automatiquement, à l'aide de Photoshop, à partir des grandes. Elles sont destinées à l'affichage sur la même page que l'article, dans la colonne de droite.

Les images web sont celles que l'on retrouve sur toutes les pages comme les logos et autres pictogrammes.

4.2. La base de données BBF²⁴

Araxe a mis en place le site web du *BBF* électronique permettant la publication en ligne de la revue, sous forme HTML et PDF, ainsi que l'indexation des documents au sein d'une base offrant plusieurs types de recherche. Les différents développements informatiques ont été réalisés par étapes en fonction des problèmes rencontrés et des attentes de l'équipe du BBF.

L'enssib a opté pour un SGBDR²⁵, Microsoft SQLServer 2000. La structure de la base est détaillée en annexe 2. Elle recouvre un fonds de documents XML recouvrant les années 2000 et 2001, et sera alimentée par les numéros à venir.

Ces documents sont transmis vers le serveur Web par le biais d'un formulaire HTML (on inscrit le chemin d'accès aux fichiers sur le serveur des Editions).

Cette transmission déclenche les processus suivants :

- contrôle de la validité du document par rapport à la DTD,
- extraction des champs à indexer depuis le document XML et insertion dans les tables relationnelle de la base, des éléments du document XML,

²⁴ Les informations techniques suivantes dans le texte sont issues du dossier de conception rédigé par S. Baudet, développeur chez Araxe.

²⁵ Système de Gestion de Base de Données Relationnel

- transformation des fichiers en fichier de type XSL-FO puis HTML, pour l’affichage sur le web.

Pour ce faire, différents outils et composants de scripts ont été utilisés.

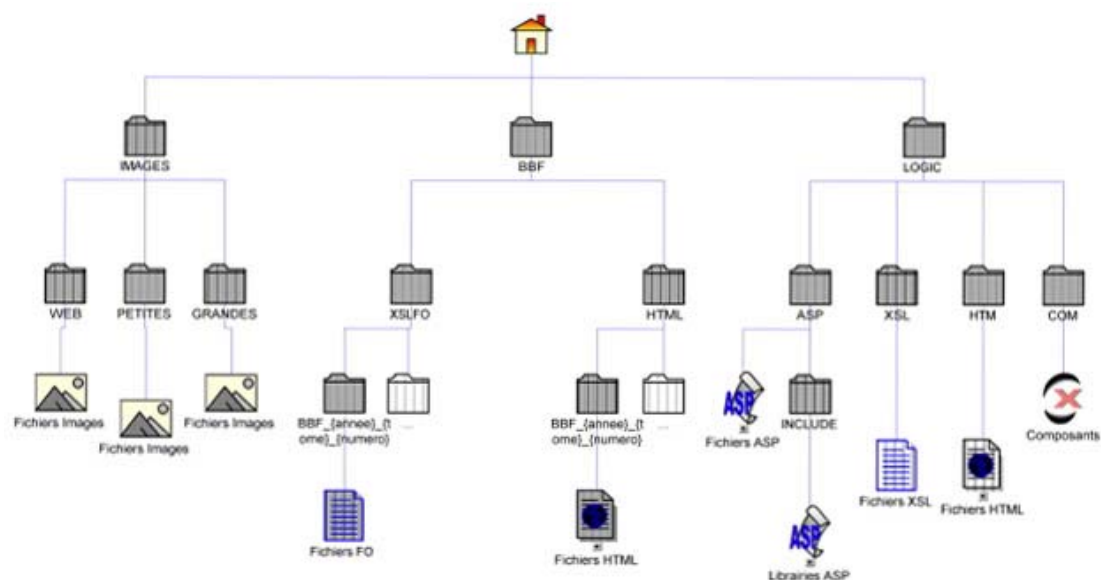
OPENXML est une fonction intégrée qui permet à SQLServer d’aller chercher les documents XML pour les indexer, en lui spécifiant à quelle table correspond quels champs XML.

L’affichage dynamique sur internet se fait à partir de pages ASP²⁶, elles servent d’interface entre le navigateur et la base.

La recherche dans la base SGBDR se fait à partir de requêtes SQL²⁷.

Un système de Backoffice, au gestionnaire du site, via une interface web depuis le web, permet la mise à jour des publications dans la base, et l’accès aux informations d’un abonné pour pouvoir retrouver son mot de passe par exemple.

4.3. Arborescence du site



²⁶ Active Server Pages. Les pages ASP se déclenchent côté serveur et servent à inclure les résultats donnés par les composants de la base de la donnée à l’intérieur d’une page HTML.

²⁷ Structured Query Language. Mode d’interrogation des bases de données relationnelles.

Description succincte des répertoires depuis la racine:

- Images : contient l'intégralité des images du site
- BBF : contient les fichiers des articles, séparés par type de fichiers
- LOGIC : contient les fichiers se rapportant à la partie intelligente du site, séparés par type de ressources

Répertoires	Sous-répertoires
Images	Web : contient les images des pages web
	Petites : contient les imagettes
	Grandes : contient les images correspondant aux imagettes
BBF	XSLFO : contient les fichiers au format XSL-FO
	HTML : contient les fichiers HTML visibles sur le site
LOGIC	ASP : contient les différents composants, Indexation, Recherche, Authentification dans une librairie ASP, et les pages ASP
	XSL : contient les feuilles de style XSL
	HTM : contient les pages de navigation du site, les en-têtes et pieds de page communs aux pages et les feuilles de stle CSS
	COM ²⁸ : contient les comoposants COM utilisés par les pages ASP

5. Les nouveautés du **BBF** électronique

Le site du *BBF* électronique se distingue de la version papier et du site actuel par les nouvelles rubriques et les services personnalisés. C'est à présent un site composé de pages dynamiques, exploitable grâce aux technologies précédemment citées.

²⁸ Component Object Model., créée à partir de la plateforme Windows COM.

5.1. Accès et abonnement

Jusqu'à présent, l'accès à la base BBF était totalement gratuit pour les abonnés à la revue papier comme pour les non-abonnés. Désormais, l'accès dépend de l'abonnement souscrit par l'utilisateur.

5.1.1 Conditions d'accès et abonnement

Abonnement	Caractéristiques	Tarifs	Accès
Complet	revue imprimée + version électronique enrichie	Normal : 87 € Etudiant ²⁹ : 60,90 € Etranger : 95 €	tout le site
Electronique	revue électronique enrichie seule	Unique : 53 €	tout le site
Option réseau	accès simultanés à la version électronique	Avoir souscrit à l'1 des 2 abonnements ci-dessus, 2 fois	tout le site
Aucun	Recherche possible mais accès aux contenus antérieurs aux 2 numéros les plus récents. Accès aux résumés. Pas d'accès aux services personnalisés		

5.1.2 Traitement informatique

Sur le site web, les abonnés devront fournir un nom d'utilisateur (login) et un mot de passe. L'authentification se fait à travers la table Tclients de la base de données. Cette table est alimentée par une base Access dédiée aux abonnements et gérée par l'enssib.

Un script générera en fait automatiquement les login et les mots de passe qui seront envoyés par courrier à tous les abonnés.

La connexion des abonnés sera facilitée, une fois la première connexion établie, par la possibilité de se connecter automatiquement. Une case à cocher sera donc sur la page d'accueil avec la mention "*Mémoriser le mot de passe*".

²⁹ Pour bénéficier de ce tarif, il faut être affilié à une formation concernant les bibliothèques ou les sciences de l'information.

En cochant cette case, un cookie est déposé sur le poste contenant les informations d'authentification encryptées grâce à un composant gratuit ANEI Base64³⁰. Il sera également utilisé pour le contrôle des abonnements, surtout en ce qui concerne l'option réseau. Un contrôle sera fait sur le nombre de sessions ouvertes pour un abonnement. Ce cookie aura enfin une durée de vie infinie, toujours en fonction des informations validées par le composant.

5.2. Modalités de recherches et affichage des documents

5.2.1 Généralités

Les recherches s'effectuent sur la base de données BBF telle qu'elle a été décrite en 4.2. Mais celle-ci a subi début septembre des changements qui offriront aux utilisateurs la possibilité d'interroger la totalité des numéros du *BBF* depuis 1956 avec une interface unique. Il était initialement prévu que la nouvelle base couvrirait les années 2000, 2001 et futures, et pointerait grâce à des liens sur l'ancienne base. Les notices bibliographiques, anciennement stockées dans le catalogue de la bibliothèque, seront donc reformatés en XML pour être intégrées dans la nouvelle base BBF et les anciennes pages web seront "rhabillées" avec la feuille de style du nouveau *BBF*.

Ces changements n'affectent en rien le nouveau site, ils l'enrichissent au contraire en alimentant la base de nouveaux documents et en offrant un affichage graphique homogène.

Voici un tableau sommaire qui récapitule les types de recherche en fonction des numéros, l'affichage des documents et leur format d'exploitation :

³⁰ Ce composant est téléchargeable à l'adresse suivante : www.anei.com/aneibase64.asp

Numéros	Recherches	Affichage	Documents
depuis 2000	Toutes	HTML issus du XML PDF, issus de l'XPress	Tous
1995-1999	Toutes sauf : - plein texte - illustrations-citations	HTML "rhabillé" PDF, issus de l'XPress Word	Sommaires, Articles, Résumés,
1956-1994	auteur-titre seulement	Documents HTML, issus du XML	Notices

Voyons maintenant les caractéristiques communes aux différents types de recherches :

- indifférence entre majuscules et minuscules
- troncature à droite à l'aide du signe : *
- opérateur implicite entre les termes de la requête : ET

5.2.2 Les différents types de recherche

La recherche Auteur-Titre-Résumé est une recherche classique et la recherche sur les résumés n'est possible qu'à partir de 1995.

La recherche thématique permet d'obtenir les documents indexés à partir de 8 concepts-clés et leur sous-concepts déjà évoqués et détaillés en annexe 2.

Si la recherche s'avère infructueuse, il est possible de l'élargir.

La recherche plein texte est restreinte aux années 2000 et postérieures.

La recherche de citations et d'illustrations exploite les possibilités offertes par le balisage effectué dans les documents XML Elle permet de trouver :

- les citations et leurs auteurs présentes dans les textes du *BBF* à partir d'un ou de plusieurs mots de cette citation,
- les occurrences des titres des livres ou revues cités.

Cette recherche permet également de trouver un tableau, un graphique ou une illustration à partir d'un ou de plusieurs mots de sa légende.

5.3. Les nouvelles rubriques et les nouveaux services

5.3.1 Les profils

La fonctionnalité "Mes profils" constitue un nouveau service réservé aux abonnés. Pour créer un profil, l'abonné doit renseigner son adresse électronique et son mot de passe qu'il aura lui-même choisis (ce traitement est indépendant de celui de la connexion au site). Là aussi, il y aura possibilité d'authentification automatique à l'aide d'un cookie qui aura une durée de vie de 3 mois.

Les profils sont des ensembles d'enregistrements de requêtes attachés à un utilisateur. Ils permettent non seulement de sauvegarder des requêtes intéressantes mais surtout d'offrir un service de push qui :

- la connexion de l'abonné actualise automatiquement les résultats de requête par rapport à sa précédente consultation,
- permet l'envoi d'un courrier électronique à l'abonné, quand le nouveau numéro contient les références qui correspondent à la requête enregistrée

Ce service est exécuté depuis le Backoffice : le service informatique lance un processus qui balaie tous les profils et qui vérifie si des nouveautés sont apparues depuis la dernière date de consultation du profil.

Les profils sont gérés par l'abonné qui peut les modifier à sa guise.

5.3.2 La gazette électronique

La rubrique "Gazette électronique" contribue également à l'enrichissement du *BBF* électronique. Elle débutera avec le numéro 5 de 2000 et est composée de 3 sous-rubriques, renouvelées à chaque nouvelle

parution. Cette gazette *"étend l'information offerte par le BBF à l'actualité disponible sur internet"*³¹:

- le panorama du Web : *"sur le thème de chaque dossier est proposé un diaporama de lectures complémentaires disponibles sur le web"*³²
- Actualités des textes officiels : *"tous les deux mois, un signalement des textes officiels intéressant les bibliothèques est recensé (lois, décrets, rapports, etc.), avec accès en ligne à ces ressources"*³³
- Débats et publications électroniques : *" dans chaque numéro également, le BBF signale les principaux débats en cours sur de nombreuses listes de discussion en France et dans le monde, et commente les analyses majeures parues dans une sélection de lettres électroniques s'intéressant aux bibliothèques"*³⁴

L'accès à la gazette est désactivé au bout d'un an mais continue d'être accessible sur le site grâce à la rubrique "Archives".

Cette gazette est balisée comme les autres articles, par rapport à une DTD indépendante et donc interrogée au même titre les articles via les pages de recherche.

5.4. Nouveau graphisme

Le site a entièrement été refait par le graphiste d'Araxe. Je vous présenterai donc différentes pages.

De nouveaux logos identifiant les différents types d'articles et de comptes-rendus (Varia, Outils, Bibliothèque du bibliothécaire...), les types d'impression (HTML et PDF). La présentation et l'affichage des résultats participent également de la nouvelle signature du site.

³¹ CALENGE B. *Page d'information. Le "BBF électronique"*. Document interne.

³² *Ibid.*

³³ *Ibid.*

³⁴ *Ibid.*



Figure 5 - Page d'accueil du site



Figure 6 – Page de recherche thématique

Bulletin des Bibliothèques de France **BBF**

- Accueil
- S'abonner
- Acheter un numéro
- Ecrire au BBF

Recherche Feuilleter Vos profils

Modifier la requête
Enregistrer cette requête

Les résultats de vos requêtes

Nombre de réponses : 10

Page 1/ 1

Articles Chroniques Ressources Gazette électronique

"Les bibliothèques universitaires : évaluation des nouveaux bâtiments (1992-2000)", Daniel Renault, 2001 - N° 5 -
"Les usages sociaux de la bibliothèque universitaire", Dendani Mohamed, Reyssat Pascal, 2001 - N° 1 -
"Savoirs, techniques et pouvoirs", Gleyze Alain, 2001 - N° 1 -
"Bibliothèque universitaire, bibliothèque publique?", Dujardin Brigitte, Jullien Madeleine, 2000 - N° 5 -
"Classements, mises en espace, départementalisations : voyage dans quelques bibliothèques", Moragas Montserrat, Jensen Mogens, Heranz Esther, Duperrier Alain, Icardo Marie-Noëlle, 2001 - N° 1 -
"Per piacer di porle in lista", Masson Francine, 2001 - N° 1 -

Figure 7 – Ecran de résultats

Bulletin des Bibliothèques de France **BBF**

- Accueil
- S'abonner
- Acheter un numéro
- Ecrire au BBF

Recherche Feuilleter Vos profils

Accès au sommaire de ce numéro

Sommaire du dossier : Les nouveaux objets de lecture

- Lecteurs et livres électroniques

Le Loarer Pierre

Pierre Le-Loarer...
Directeur du centre de documentation, Chargé de mission TICE

Titre et poids des Rocket eBook et Softbook

Rocket eBook
Taille : 10 cm de hauteur sur 12 cm de largeur
Poids : 50 g
Cran : 12 cm de hauteur sur 8 cm de largeur
Diagonale d'écran : 6,5"

Softbook
Taille : 20 cm de hauteur sur 25 cm de largeur

BBF 2000 – Paris, t. 45, n° 6, pp. 24-36.

Dossier : Les nouveaux objets de lecture

Pierre Le Loarer

Lecteurs et livres électroniques (1)

Après la Biennale du savoir à Lyon fin janvier 2000, le dernier Salon du livre, qui s'est tenu à Paris du 17 au 22 mars 2000, fut l'occasion pour beaucoup de découvrir un nouvel objet nomade, le livre électronique, également dénommé *e-book*. Tout un espace, le *Village e-book*, avait été conçu à cette occasion (2). Antérieurement à ces manifestations, le livre électronique avait déjà suscité l'attention de nombreux médias, de la presse quotidienne (3) aux magazines spécialisés (4).

Avant même d'être disponible, le livre électronique fascine : une nouvelle période s'ouvrirait après l'ère Gutenberg. Les bibliothécaires, pour lesquels la culture – voire le culte ? – du livre est historiquement au cœur de leur métier (5), s'interrogent, soucieux de comprendre les changements que l'avènement de ces nouveaux dispositifs de lecture peut provoquer (6) dans leur profession. Avec l'apparition du livre électronique, parler aujourd'hui (7) d'évolution ou de révolution du livre nécessite d'interroger ce nouveau média, quant à sa nature, son ergonomie et ses usages.

De l'imprimé à l'écran

L'expression « livre électronique » est utilisée indifféremment pour désigner aussi bien ces nouveaux appareils présentés au Salon du livre (Rocket eBook (8), Softbook (9), Everybook (10), Cytale (11)) que les « livres numérisés » de toutes sortes disponibles

HTML PDF

Figure 8 – Ecran d'affichage d'un article

The screenshot shows the 'Bulletin des Bibliothèques de France' (BBF) website. The header includes the site title, a logo, and navigation links: 'Accueil', 'S'abonner', 'Acheter un numéro', 'Ecrire au BBF', 'Recherche', 'Feuilletage', and 'Vos profils'. The main content area is titled 'Gestion des profils' and contains instructions for logging in or creating a new profile. It includes input fields for 'E-mail' and 'Mot de passe', a checkbox for 'Créer un nouveau profil', and an 'Envoyer' button.

Bulletin des Bibliothèques de France

BBF

- Accueil
- S'abonner
- Acheter un numéro
- Ecrire au BBF

Recherche Feuilletage Vos profils

Gestion des profils

Renseignez votre adresse e-mail et votre mot de passe pour vous connecter à votre profil.
Cochez la case pour créer un nouveau profil.

E-mail :

Mot de passe :

Créer un nouveau profil ☐

Envoyer

Figure 9 – Page d'accès aux profils

Après avoir détaillé la chaîne de publication du *BBF* électronique et les services qu'offrent ce site, je commencerai le recensement de différents projets en les comparant aux différentes réalisations du service informatique.

Autres projets de textes électroniques au format SGML/XML

Les différents projets présentés ci-dessous ont été sélectionnés en raison de leur caractère scientifique d'une part, et parce qu'ils offrent une explication détaillée de la chaîne éditoriale de leur revue électronique, le plus souvent sur leur site web. J'ai pu ainsi mieux positionner le service informatique et ses accomplissements.

Au regard de cette liste, on remarque que les projets en sciences humaines et sociales (SHS) qui s'apparentent au *BBF* sont moins nombreux que ceux issus de sciences dures. Et ce, pour 3 raisons :

- les disciplines de sciences dures recouvrent sont plus attentives aux avancées technologiques et aux formats d'échange de données alors qu'en ce qui concerne les SHS, la mise en commun et l'échange d'informations sont des phénomènes récents
- les habitudes de travail rédactionnel en sciences dures ne sont pas les mêmes qu'en sciences « molles » : le balisage d'un texte, en LaTeX par exemple, est un exercice connu et pratiqué, qui s'avère bienvenu au moment du passage à l'électronique d'une revue en XML. Puisque le texte est déjà balisé au minimum, c'est toute la chaîne de production du document qui s'en voit simplifiée.
- Les SHS gardent des habitudes de travail "traditionnelles" où les références au papier sont fortes et où l'échange de données électroniques ne s'est pas encore généralisée.

1. XML et l'échange d'informations bibliographiques

Nom du projet et description	DTD utilisée	Traitement des données / Outils	Accès	Services / Fonctions
<p>BiblioML / AuthoritiesML http://www.culture.fr/BiblioML</p> <p>Formats basés sur XML pour l'échange d'enregistrements bibliographiques et d'autorités entre applications : BiblioML est une représentation XML du format Unimarc. Avec XML/BiblioML, l'intégration de la description, de la table des matières, du texte intégral, des images, devient naturelle.</p>	<p>BiblioML, Authorities ML développées par AJLSM³⁵</p>	<p>Saisie de la notice Saisie effectuée grâce à un module de saisie en ligne (navigateur Web) de certains types de références bibliographiques.</p> <p>Conversion automatique de BiblioML à Unimarc Outil de conversion sommaire permettant de récupérer des enregistrements Unimarc (bibliographiques ou autorités) et de les convertir en format BiblioML. Ecrit en Java, utilisable sur plusieurs plates-formes, et disponible sur le site principal de BiblioML.</p> <p>Diffusion Plate-forme SDX : outil de diffusion de bases de documents XML, développée par le ministère et composée de logiciels libres.</p>	<p>Accès gratuit aux DTDs et à l'outil de conversion UNIMARC vers XML</p>	<p>Fonctions</p> <ul style="list-style-type: none"> - peut être utilisée pour diffuser des bases de données bibliographiques en format BiblioML - permet d'effectuer des recherches - permet d'effectuer des listes de termes - permet d'afficher des documents en plusieurs formats
<p>CISMEF http://www.chu-rouen.fr/cismef/ Projet du Centre Hospitalo-Universitaire de Rouen.</p> <p>Objectif Catalogage et indexation des sites francophones sur la santé et les sciences médicales, dépassant la médecine proprement dite et accessibles sur l'internet.</p> <p>Evolution Depuis juin 2001, le standard XML est utilisé, pour permettre une interopérabilité avec d'autres catalogues ou serveurs de ressources comme les enregistrements bibliographiques de la base de données MEDLINE, disponibles au format XML.</p>		<ul style="list-style-type: none"> - le recensement des sites et des documents est effectué grâce à une veille quotidienne sur les annuaires multidisciplinaires francophones permettant une consultation de leurs nouveautés, - la validation des documents se fait par l'utilisation du référentiel des critères de qualité de l'information de santé sur l'Internet, Net Scoring, qui comprend 49 critères regroupés en huit classes principales : crédibilité, contenu, hyper-liens, design, interactivité, aspects quantitatifs, déontologie, et accessibilité. - le catalogage et l'indexation utilisent 2 outils standards pour organiser l'information : le thesaurus MeSH (Medical Subject headings) utilisé notamment pour la base de données bibliographique Medline et le format de méta-données du Dublin Core³⁶. 	<p>Accès gratuit : donne seulement un exemple de notice en xml, sinon tout le reste est en HTML statique</p>	<p>Services</p> <ul style="list-style-type: none"> - un total de plus de 10.000 ressources indexées en juillet 2001 - un classement thématique, - un classement alphabétique, tous deux inspirés du thesaurus de Medline. - une recherche multi-critères sur 15 champs en mode simple ou avancée

³⁵ AJLSM est une société spécialisée dans l'édition électronique via les formats SGML et XML : <http://www.ajlsm.com>

³⁶ le Dublin Core est une norme qui standardise l'écriture et les caractéristiques des méta-données de documents électroniques destinés au web.

Nom du projet et description	DTD utilisée	Traitement des données / Outils	Accès	Services / Fonctions
<p>MASTER, Manuscript Access through Standards for Electronic Records http://www.cta.dmu.ac.uk/projects/master/</p> <p>Lancé par l'union européenne en janvier 1999. Regroupe des bibliothèques nationales et des centres de recherche scientifiques tels l'IRHT, Institut de recherche et d'Histoire des Textes</p> <p>Objectif principal Créer un catalogue en ligne de descriptions de manuscrits médiévaux dans les bibliothèques européennes, pour une meilleure diffusion et un échange normalisé de références médiévales entre les différents acteurs (bibliothèques, centres de recherche...)</p>	<p>DTD MASTER</p> <p>Basée sur la DTD TEI³⁷ et adaptée aux besoins du projet</p>	<p>Outil avant tout destiné aux catalogueurs et aux bibliothèques en général</p> <p>Création d'une DTD MASTER</p> <ul style="list-style-type: none"> - Travail en collaboration avec le groupe de travail de la TEI pour créer une DTD MASTER, 1 destinée à xml et 1 autre à sgml, - Rédaction d'un manuel de références pour utiliser la DTD, - Mise en ligne sur le site³⁸ <p>Développement d'une suite d'outils pour préparer les descriptions conformément aux standards sgml et xml :</p> <ul style="list-style-type: none"> - création des documents avec NoteTab SP³⁹, éditeur de page html qui gère le format sgml - passage du document à l'aide de parseurs xml ou sgml, développés en interne⁴⁰, pour valider le document par rapport à la DTD <p>Affichage des descriptions Se fait grâce à un viewer développé en interne</p>	<p>Accès réservé mais pour certaines descriptions de manuscrits, accès pour tous</p>	<p>Services</p> <ul style="list-style-type: none"> - Recherche sur environ 80 éléments composant la DTD,. On peut y spécifier les attributs ainsi que les éléments père pour une recherche plus large. - recherche combinée avec opérateurs AND et OR. <p>Inconvénient</p> <ul style="list-style-type: none"> - base contenant peu de documents : 831 en août 2001.

³⁷ DTD de la Text Encoding Initiative. Née en 1987 de la collaboration entre chercheurs représentatifs d'un grand nombre de disciplines elle a pour objectif de faciliter l'échange et la circulation des documents électroniques au sein de la communauté scientifique. En 1994, ses spécifications sont devenues des recommandations et s'appuient sur la norme SGML. Il existe une version allégée TEILite. DTDs et documentation t téléchargeable sur le site : <http://www.tei-c.org/>

³⁸ Pour le manuel : <http://www.hcu.ox.ac.uk/TEI/Master/Reference/>, et pour la DTD xml : <http://www.hcu.ox.ac.uk/TEI/Master/Reference/DTD/masterx.dtd>

³⁹ Ce logiciel existe en version payante (\$19,95) ou gratuite (démonstration) et téléchargeable sur le site www.tucows.com

⁴⁰ Ce sont des prototypes à tester sur le site <http://www.cta.dmu.ac.uk/projects/master/notetabhelp.html>

2. XML et l'archivage de documents

Nom du projet et description	DTD utilisée	Traitement des données / Outils	Accès	Services / Fonctions
<p>ETEXT, Electronic text Centre http://etext.lib.virginia.edu</p> <p>Projet initié par la bibliothèque de l'Université de Virginie depuis 1992. Regroupe 51 000 textes et 350 000 images</p> <p>Objectifs</p> <ul style="list-style-type: none"> - Créer des archives en ligne de textes et d'images en sciences humaines, basés sur les standards, sgml et xml, tiff, us-marc. - Construire et aider la communauté des utilisateurs à la création et à l'utilisation des ressources en ligne dont le E-Book avec MS Reader et Palm Devices <p>Evolution</p> <ul style="list-style-type: none"> - passage de la norme sgml à la norme xml. - depuis 1997, utilisation de la DTD EAD pour l'archivage de leurs collections, s'associant aux projets d'autres universités. - en 2001, le résultat est le stockage d'un important volume de manuscrits du 17^e au 20^e siècle. 	<p>DTD TEILite</p> <p>DTD EAD⁴¹</p>	<p>Numérisation des textes et images</p> <ul style="list-style-type: none"> - les textes sont océrisés⁴² : conversion d'un format image à un format texte grâce au logiciel <i>OmniPage Pro</i>⁴³ - les images sont scannées⁴⁴ grâce au logiciel <i>ScanDesk</i> : <p>Enregistrement des documents ainsi numérisés</p> <ul style="list-style-type: none"> - les textes : format de type traitement de texte, RTF recommandé - les images : un format TIFF (archivage) et JPEG (web) <p>Balises</p> <p>Sous-ensemble des balises de la DTD TEI, la TEILITE. Balises du texte et des images manuellement : conseils sur les étapes du balisage pas à pas⁴⁵.</p> <p>Enregistrement des fichiers balisés</p> <ul style="list-style-type: none"> - Texte intégral : depuis 1992, en sgml, et depuis peu en xml, et les en-têtes TEI sont converties automatiquement au format US-MARC afin de les insérer dans le catalogue général de bibliothèque. - Images : GIF ou JPEG (voir son poids et sa qualité) <p>Gestion des documents sgml et xml</p> <p>Gestion structurée grâce à <i>Near & Far® Designer</i>⁴⁶, architecture basée sur le web pour des performances en temps réel, de la société OpenText.</p> <p>Affichage web</p> <p>Page html, créée à la volée à partir du fichier sgml ou xml, et les images s'affichent en GIF ou JPEG.</p>	<p>Autorisation requise mentionnée par <i>UVA users only</i> mais certaines collections sont en libre accès</p>	<p>La recherche</p> <ul style="list-style-type: none"> - pour les non-abonnés mention <i>Non-UVA users</i> - recherche sur tous les mots du texte ou sur une phrase, dans toute les collections ou seulement sur une particulière. - recherche hypertextuelle en navigant par collections <p>Les résultats</p> <ul style="list-style-type: none"> - présentation des extraits où a été trouvé le mot ou la phrase - possibilité d'accéder au paragraphe ou au chapitre précédent ou suivant, ou encore d'accéder au texte intégral ou à la table des matières

⁴¹ La DTD EAD, Encoded Archival Description, a été conçue par une équipe de chercheurs en bibliothéconomie de l'Université de Californie à Berkeley, dans le but particulier de baliser les renseignements contenus dans des instruments de recherche archivistique. Toutes les informations sur cette DTD sur <http://sunsite.berkeley.edu/ead/>

⁴² Des exemples et des conseils de numérisation de textes puis d'océrisation sont donnés à la page suivante <http://etext.lib.virginia.edu/helpsheets/scantext.html>

⁴³ Ce logiciel présente de nombreux avantages comme gestion de nouveaux caractères et de plusieurs formats de traitement de texte, scanne une partie du document...

⁴⁴ Des exemples et des conseils de numérisation d'images sont donnés à l'adresse suivante : <http://etext.lib.virginia.edu/helpsheets/scanimage.html>

⁴⁵ URL : <http://etext.lib.virginia.edu/tei/uvatei.html>

⁴⁶ Sa principale caractéristique est d'être un système de gestion de contenus pertinent qui permet d'analyser les données de l'entreprise, de développer des DTD et d'automatiser certains processus éditoriaux. Il est complété par d'autres applications concernant la modélisation de documents, les liens entre les documents

Voir http://www.opentext.com/near_and_far/details.html

3. XML et la structuration de thèses, de travaux d'étudiants et de chercheurs

Nom du projet et description	DTD utilisée	Traitement des données / Outils	Accès	Services / Fonctions
<p>CITHER</p> <p>http://csidoc.insa-lyon.fr/these/index.html</p> <p>CITHER, Consultation en texte Intégral des Thèses En Réseau, développé par le centre de documentation de l'INSA de Lyon, Doc'INSA.</p> <p>Objectif Proposer un maximum de nouvelles thèses de l'INSA de Lyon, en texte intégral et en format PDF sur le site web, et ce depuis 1997.</p> <p>Evolution Réflexion, depuis 2001, portée sur l'archivage des thèses au format XML.</p>	<p>DTD de 1^{er} niveau issue d'une DTD publique, et DTD de 2^e niveau créée par l'INSA</p>	<p>Choix d'une DTD DTD de 1^{er} niveau traite de l'organisation du texte, repris dans une DTD type TEI, ISO 12083 ou EDT, DTD de 2^e niveau traite d'éléments tels que Jury, Ecole, Résumé, créée en fonction de la structure des thèses INSA</p> <p>Construction d'une maquette de conversion (interface) du format RTF vers le format XML⁴⁷ : - outil de conversion automatique RTF vers XML. Le(s) fichier(s) d'une thèse en RTF "sont traités avec le script <i>RTF2XML</i> et l'outil <i>Omnimark</i> qui permet de générer des fichiers XML et d'extraire les images insérées."⁴⁸ - lien entre les fichiers xml dans un document xml brut - développement d'une application qui permet d'adapter le fichier XML à l'analyse de la DTD grâce au parsage du document - sortie en XML à l'aide de feuilles de style CSS, appliquée sur la seule balise <paragraphe></p>	<p>Accès gratuit aux thèses en PDF. XML prévu pour l'archivage uniquement ?</p>	<p>Fonctions archivage pérenne des thèses soutenues à l'INSA de Lyon</p>

⁴⁷ NICULETA S., *Conception d'une chaîne de conversion : format RTF vers XML* [On line]. Villeurbanne : INSA-LISI, 2001, 43 p. Rapport final PFE 2000-2001. Consulté le 02/07/01] Available from internet : <url: <http://csidoc.insa-lyon.fr/these/doc/rapport-final2001.pdf>>.

⁴⁸ *Ibid.*

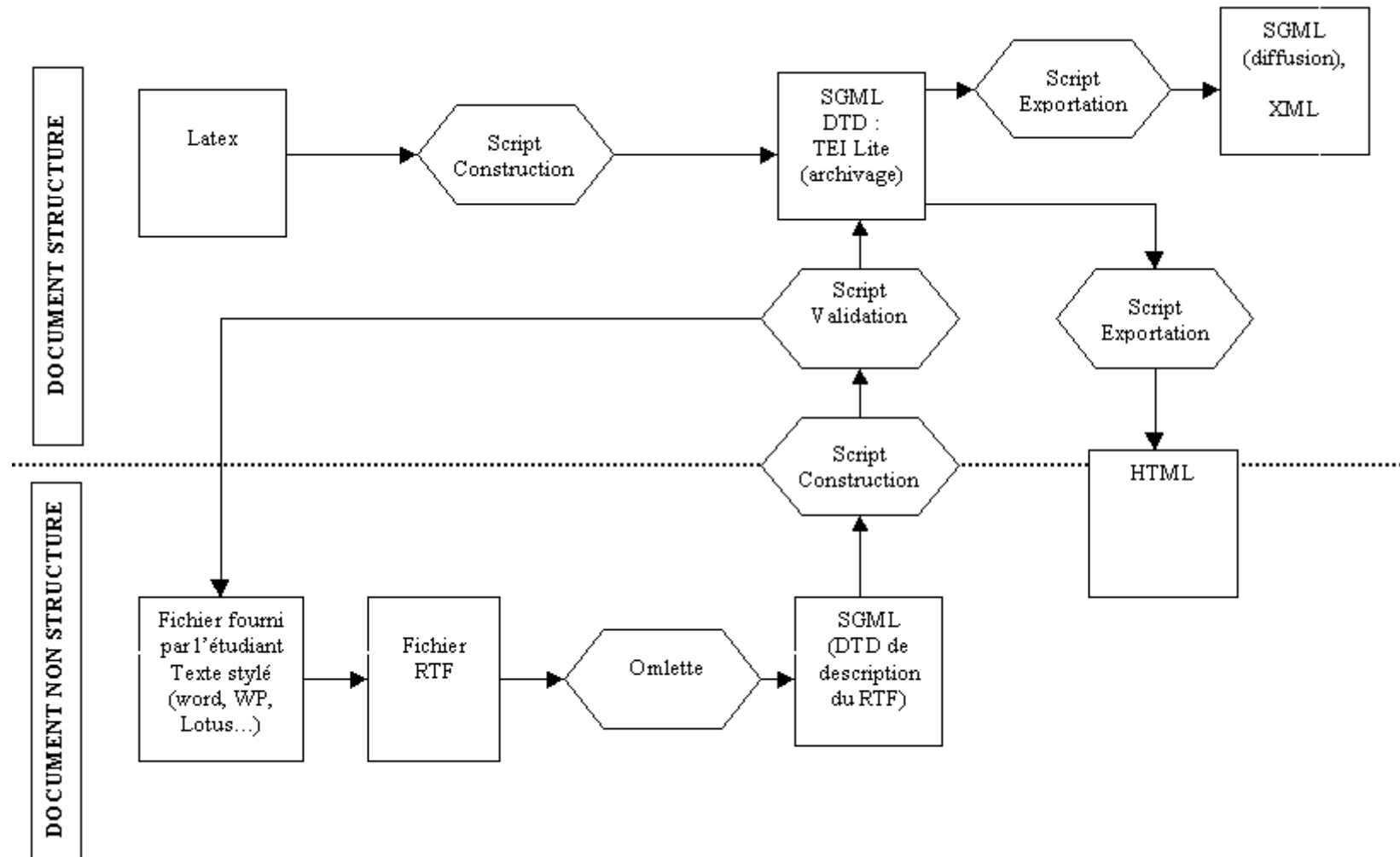


Figure 10 – Chaîne de production utilisée par Cyberthèses

<p>PELLEAS http://www.univ-mlv.fr/bibliotheque/scd/scd.htm</p> <p>Projet en cours de bibliothèque électronique. Réalisé dans le cadre du Polytechnicum de l'université de Marne-la-Vallée, associée à celle de Reims.</p> <p>Objectif Proposer une gestion des documents électroniques, travaux des étudiants et des chercheurs entre autre, aidée par un atelier d'édition électronique.</p> <p>Evolution Etude en 2001 sur le passage du Web vers XML.</p>	<p>DTD ISO 12083 ou DTD TEI, choix indéterminé encore</p>	<p>Les documents primaires subiront l'un de ces 2 traitements :</p> <ol style="list-style-type: none"> 1. Numérisation en mode image avec éventuellement une transformation en mode caractères (par OCR), pour pouvoir mettre à jour le document ou effectuer une indexation automatique, 2. Feuilles de style utilisables avec des logiciels éditeurs standards et d'imposer des feuilles de style aux auteurs de documents : une feuille de style est en cours de définition pour les mémoires de maîtrise, les DEA et les thèses. Puis il y aura un reformatage automatique des documents provenant de logiciels de traitement de texte, dans un format standard SGML/XML (pris comme format pivot de stockage et de conservation). La diffusion sur le Web se fera en XML principalement (mais pas exclusivement). 	<p>Le site Polytechnicum est encore inaccessible</p>	<p>Fonctions Mettre à disposition de la communauté scientifique des documents sous forme textuelle ou multimédia et structurés au format xml</p>
<p>CYBERTHESES et SENTIERS http://theses.univ-lyon2.fr/index.html</p> <p>SENTIER : projet en cours de structuration et de mise en ligne de thèses électroniques développé par le cellule d'édition électronique de l'Université Lyon 2.</p> <p>CYBERTHESES : à l'initiative des Presses de l'Université de Montréal et de l'université, a pour vocation de permettre la recherche et l'indexation de ces thèses au moyen de méta-données (notice) communes. Participe au projet ETD, Electronic Theses and Dissertations de l'Unesco⁴⁹</p>	<p>DTD RTF convertie en DTD TEILite et programmes développés par Omnimark Technologies⁵⁰</p>	<p>Feuilles de style pour la rédaction - Feuilles de style et manuel d'utilisation fournis aux étudiants pour leur rédaction et en fonction de leur traitement de texte : Word 97 ou + pour PC, Word 98 et + pour Mac et StarOffice v.5x. Le tout téléchargeable sur le site</p> <p>Conversion automatique vers sgml et xml⁵¹ - la chaîne de traitement, développée avec les PUM est constituée d'un ensemble de programmes ou scripts rédigés dans le langage Omnimark. Son rôle est de générer automatiquement, à partir de documents de formats divers, principalement RTF, des documents structurés de format homogène, SGML et XML, ou encore PDF (impression) ou HTML (web). Voir figure 10⁵²</p>	<p>Accès au catalogue des thèses électroniques gratuit.</p> <p>Une dizaine de thèses en ligne aux formats SGML, XML et PDF</p>	<p>Fonctions l'archivage pérenne des travaux de recherche effectués au sein de l'université</p> <p>Services la consultation en plein-texte des thèses de doctorat et bientôt puissant outil de recherche grâce à l'utilisation du concept de document structuré, sgml et xml.</p>

⁴⁹ Ce projet, à l'initiative de Virginia Tech (USA), centralise des universités du monde entier pour développer un serveur de thèses et de travaux d'étudiants et de chercheurs. Chacune des universités développent sa propre DTD ETD, aux formats SGML ou XML, et la DTD de l'université Lyon 2 en fait partie, le seul inconvénient est qu'il incombe aux auteurs de fournir leurs travaux déjà formatés en PDF ou XML. Le projet ETD est rattaché au NDLTD, Networked Digital Library of Theses and Dissertations. Toutes les informations sur le site : <http://www.ndltd.org/>

⁵⁰ Omnimark Technologies Inc. est une société qui propose des solutions destinées à la transformation de documents en d'autres formats. Cette société a mis au point un langage de script de transformation de document concurrent de XSLT et de Perl et modulable.

⁵¹ Le site présente un schéma détaillé de la chaîne de production des fichiers <http://www.univ-lyon2.fr/sentiers/edition/theses/ressources/SchemaChaine.gif>

⁵² Cette figure est issue du site Sentiers : <http://www.univ-lyon2.fr/sentiers/edition/theses/ressources/SchemaChaine.gif>

Figure 11 – Chaîne de traitement générale des PUM

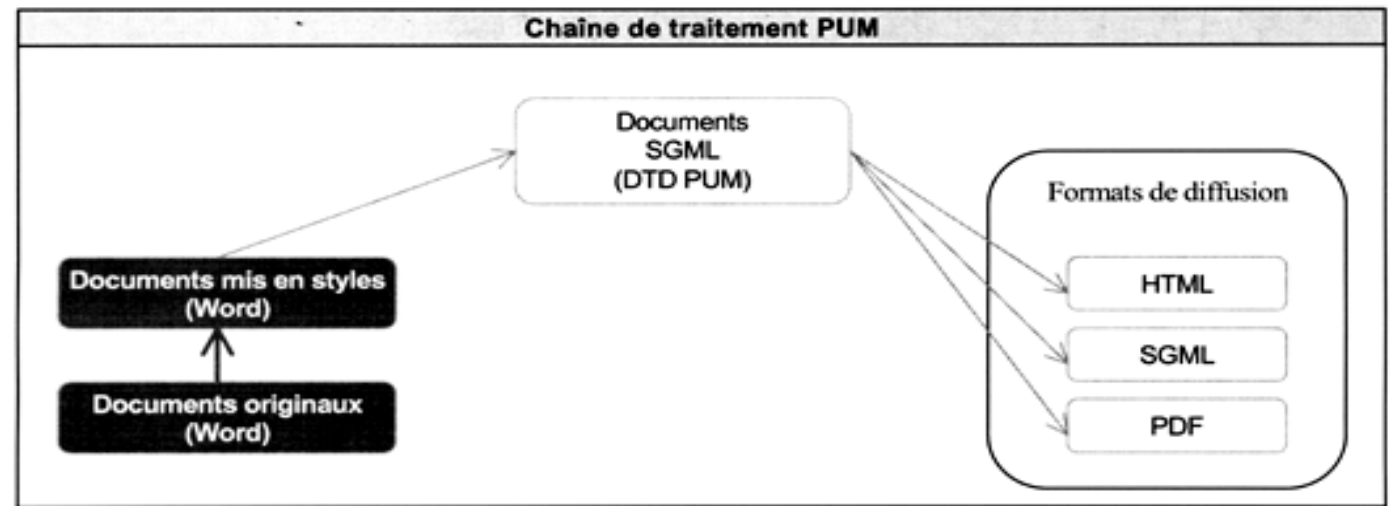
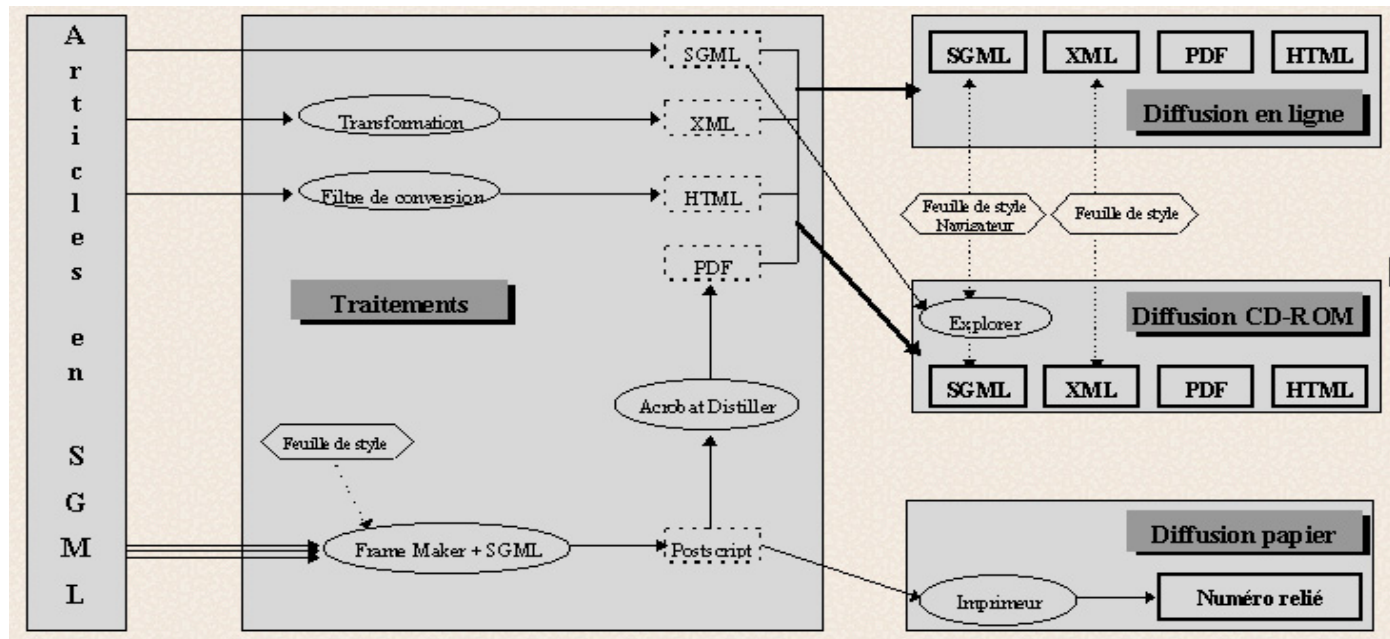


Figure 12 – Chaîne de traitement détaillée des PUM



4. XML et la structuration de revues et journaux électroniques

Nom du projet et description	DTD utilisée	Traitement des données / Outils	Accès	Services / Fonctions
<p>ERUDIT http://www.erudit.org/erudit/index.html</p> <p>ERUDIT, Edition de Revues Universitaires et Diffusion sur Inforoutes, développé par les Presses Universitaires de Montréal, PUM.</p> <p>Objectif Donner aux revues savantes les moyens de s'engager dans un processus de conversion vers l'électronique en leur proposant des solutions techniques.</p> <p>Formatage en sgml et xml sur 6 revue en sciences humaines et sociales⁵³, dans une cellule spécialisée dans le traitement éditorial de ces revues</p>	<p>Adaptation d'autres DTDs pour la création d'une DTD PUM</p>	<p>Création de la DTD PUM Noyau de la DTD ISO : 12083, auquel ont été greffées des DTDs normalisées telles que CALS (Continuous Acquisition and Lyfe Cycle Support), TEI, celle de la revue Earth Interactions.</p> <p>Chaîne de traitement automatisée, basée sur la norme SGML⁵⁴ (voir figures 11 et 12 ci-dessus) - document primaire en Word et sortie papier grâce à FrameMaker+SGML - pour d'autres sorties, document repris à l'aide d'une feuille de style sous Word et enregistré en RTF : uniformisation et normalisation des textes et tableaux. 1 feuille de style par revues avec 30-50 styles différents chacune. - première conversion automatique des données en SGML grâce au convertisseur OmniMark de OmniMark. Ce document est conforme à une DTD décrivant la structure et le contenu d'un document RTF. - seconde conversion en sgml conforme à la DTD-PUM, axée sur le formatage et sur la structure sémantique.</p> <p>Formats de sortie - html, sgml, et pdf (1 fichier pour 1 revue). - le format xml est prévu à moyen terme. La sortie se fait en fonction d'une feuille de style propre à chacune des revues.</p> <p>Insertion d'images Images toujours de très bonne qualité (haute résolution et en couleur) qui seront ensuite traitées (formats, résolution, couleur ou NB) en fonction de leur support de diffusion</p>	<p>Accès gratuit aux revues aux formats HTML, PDF et pour certaines SGML</p>	<p>Services offerts aux revues : - publication et diffusion sur supports électronique et papier - indexation - numérisation rétrospective - souci de conciliation des objectifs d'accès, de pérennité et de qualité graphique.</p> <p>Outil de recherche : moteur de recherche, <i>Harvest</i>, en texte intégral qui indexe les documents en sgml</p> <p>Inconvénient : il faut un viewer sgml pour visionner les documents ainsi formatés</p>

⁵³ Géographie Physique et Quaternaire, Meta, Relations Industrielles/Industrial Relations, Revue d'histoire de l'Amérique française, Sociologie et sociétés et Surfaces.

⁵⁴ Les détails de la chaîne de traitement sont donnés sur le site www.erudit.org/erudit/rapport/index.htm et dans SEVIGNY M. *Un modèle de traitement pour l'édition électronique de revues savantes* [on line]. Montréal : PUM, 1997. [Consulté le 03/07/01]. Available from internet : <url : www.pum.umontreal.ca/pum/publ_electr/rapports/sev01/partie_11.html>

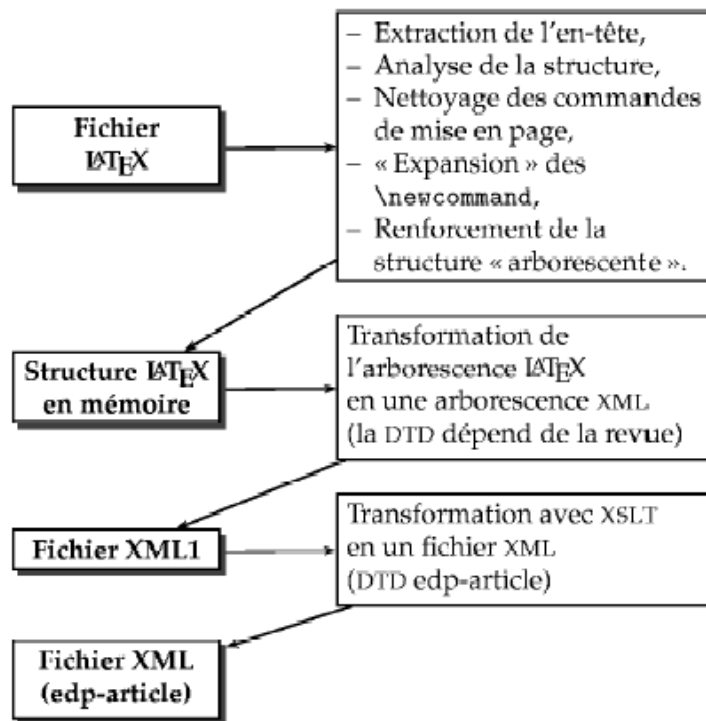


Figure 13 - Principe de la transformation de LaTeX en XML
chez EDP Sciences

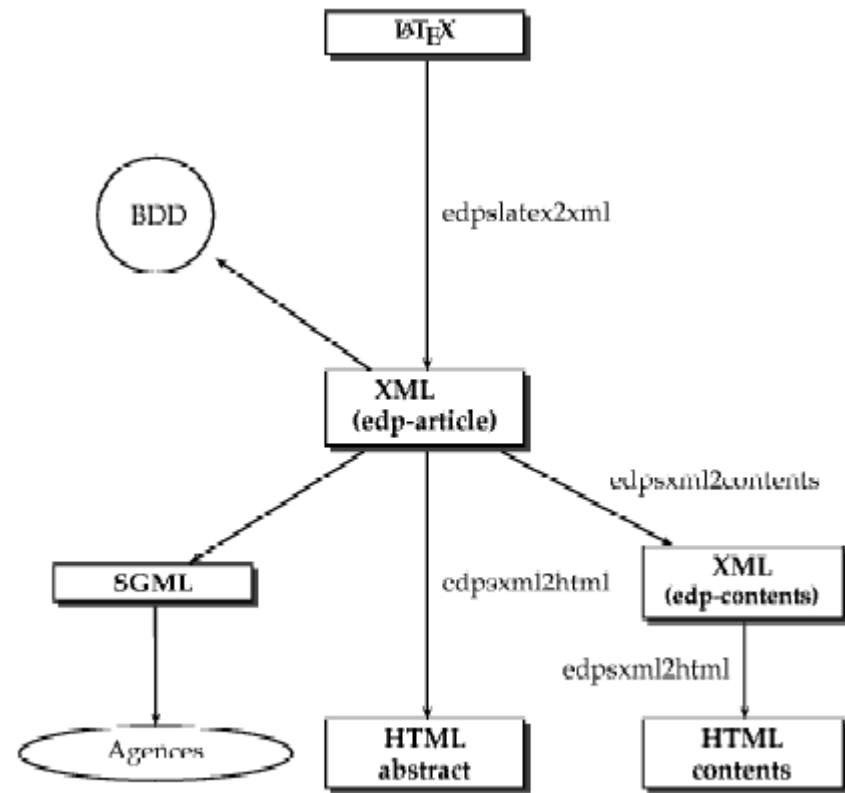


Figure 14 - Utilisation des fichiers XML
chez EDP Sciences

<p>Projet d'EDP Sciences http://www.edpsciences.org/bienvenue.html</p> <p>EDP Sciences : un des premiers éditeurs français dans les domaines de la physique fondamentale et appliquée mais aussi en chimie, astronomie, mathématique, et sciences de la vie. Production de 45 000 pages, sur support papier et électronique, par an depuis 2001.</p> <p>Objectifs Edition électronique de différentes revues dans divers formats tels que html, PostScript et pdf. Politique en place depuis 1995 en vue de rationaliser la production.</p> <p>Evolution Migration des formats vers xml depuis janvier 2000 en vue de stocker et de manipuler les principales informations de chacune des revues publiées. Exemple de traitement sur <i>Astronomy and Astrophysics</i>⁵⁵. Elle représente environ 380 pages hebdomadaires soient 19 000 pages par an.</p>	<p>Une DTD par revue créée en interne, et une DTD finale pour l'article; DTD edp-article</p>	<p>Transformation à partir de LateX d'un fichier XML Transformations exécutées par des outils développés en interne (voir figures 13 et 14). 1. Actions sur les fichiers originaux formatés en LateX : extraction de l'en-tête, analyse de la structure, nettoyage des commandes de mise en pages, renforcement de la structure arborescente, et mise en mémoire de cette structure LateX 2. Transformation de l'arborescence LateX en une arborescence XML en fonction de la DTD de la revue pour créer un premier fichier XML 3. Transformation automatique du 1^{er} fichier XML par XSLT pour produire un 2^e fichier XML conforme à la DTD edp-article Informations contenues dans le fichier XML final - les informations essentielles sur la revue telles le titre, le numéro de volume, l'ISSN - l'article c'est-à-dire le texte, les auteurs, les résumés anglais et / ou français Utilisation des fichiers XML Production de sommaires et de résumés en HTML grâce à un interpréteur XSLT, Saxon, passé sur le document XML. Le résumé est obtenu à partir du document XML final alors que le sommaire est généré dans un fichier à part en xml puis en html. Pour les formules, utilisation de LateX2HTML qui les transforme en HTML ou GIF, elles sont ensuite réintroduites dans le fichier HTML. Alimentation de leur bases de données où les documents sont stockés en XML Problèmes rencontrés - correction humaine quand les macros utilisées sont trop compliquées ou quand les règles de composition n'ont pas été respectées. - question en suspens : formules mathématiques et chimiques en xml</p>	<p>Accès à la version électronique compris dans l'abonnement papier.</p> <p>Non-abonnés : accès aux résumés et mots-clés de l'article</p> <p>Les tarifs d'abonnement dépendent de la revue à laquelle on veut s'abonner.</p> <p>Possibilités de recherche sur toute la base pour tous :</p> <ul style="list-style-type: none"> - à partir d'un index des revues - à partir d'un index des sujets traités par les revues - recherche plein texte sur les résumés et les mots-clés mais ne supporte aucun caractères spéciaux, - à partir du nom de la revue (boîte déroulante listée), du n° de volume ou de l'année, du n° de page <p>Services de la version en ligne des revues :</p> <ul style="list-style-type: none"> - articles multi-supports - liens avec la base d'objets stellaires du CDS (Centre de données astronomiques de Strasbourg) et la base bibliographique de la NASA. Ces liens sont visibles dans le résumé et le corps de l'article - liens avec les références réciproques d'autres revues avec d'autres éditeurs
--	--	--	--

⁵⁵ La chaîne de production des revues éditées par EDP Sciences a fait l'objet d'un article : JORDA J.P., CHAIX M. L., MAHBOUB A. LateX et XML dans la chaîne éditoriale d'EDP Sciences. *Cahiers GUTenberg*. [On line]. Mai 2001, n°39-40, 13 p.[visité le 12/07/01]. Available from internet : <url: www.gutenberg.eu.org/pub/GUTenberg/publicationsPDF/39-jorda.pdf>

<p>Earth Interactions http://earthinteractions.org/</p> <p>Initiative de 5 sociétés scientifiques savantes: American Geophysical Union, AGU, American Meteorological Society, AMS⁵⁶, Association of American Geographers, AAG, Ecological Society of America, ESA et The Oceanography Society, TOS et la NASA depuis 1992.</p> <p>Journal électronique traitant des interactions entre la lithosphère, l'hydrosphère, l'atmosphère, et la biosphère dans le contexte d'enjeux globaux ou de changement global. Archives en ligne et disponibles aux formats html et pdf, si abonné.</p> <p>Objectif : codage des articles en sgml plutôt qu'en html.</p>	<p>DTD propre à la revue.</p>	<p>Rédaction des articles</p> <ul style="list-style-type: none"> - utilisation d'un modèle de rédaction stylé et mise à disposition de leur guide d'utilisation. Document évolutif. - styles de base à respecter⁵⁷ : le titre de la section, le résumé, le texte, les tables, les figures (possibilité de comprendre des animations ou de supporter des ensembles de données), les remerciements, les références, et les annexes - les citations aux documents ou aux ensembles de données accessibles Internet peuvent être faites sous forme d'URLs écrit en texte. <p>Transformation automatique d'un format traitement de texte vers SGML</p> <ul style="list-style-type: none"> - Si l'auteur soumet son article comme document html, ce document sera utilisé directement dans le processus de conversion. - Si l'auteur soumet un document en PostScript ou pdf, il sera invité à fournir le fichier de traitement de texte employé pour créer ce document. - Si le manuscrit soumis est un fichier HTML produit par un outil de type traitement de texte, la partie automatisée de la conversion en SGML se produira mieux si le fichier de traitement de texte initial peut être utilisé, ainsi des auteurs seront invités à fournir ce fichier. <p>Les figures statiques</p> <p>Incluses comme images GIF. Toutes les figures doivent être disponibles en tant que fichiers séparés pouvant être utilisés dans la création de l'article. Les images non formatées en .GIF seront formatées par l'équipe chargée du processus de production.</p>	<p>Accès aux articles réservés à ceux ayant souscrit un abonnement</p> <p>Accès pour tous aux résumés</p> <p>Recherches sur :</p> <ul style="list-style-type: none"> - un index des auteurs et un - index des mots-clés de chaque numéro archivé - sur un des champs de l'article : titre, résumé, plein texte. Choix d'un des 3 opérateurs booléens, exprimés par « tous les mots », « certains mots et « phrase exacte » <p>Recherches approfondies</p> <ul style="list-style-type: none"> - troncature à droite et à gauche sur un seul () ou plusieurs caractères (%), et au milieu d'un terme () - liste de mots et caractères réservés - recherche sur des mots de la même famille avec le \$ placé avant le terme - recherche sur des homonymes en anglais : ! placé avant le terme <p>Inconvénient : posséder un viewer SGML, Panoram Viewer</p>
--	-------------------------------	---	---

⁵⁶ En octobre 1998, l'American Meteorological Society a gagné un prix d'InfoWorld (site dédié à l'informatique et ses nouvelles tendances : <http://www.infoworld.com/>) pour son innovation en technologie: "Au lieu de regarder les cieux, l'AMS a observé ses journaux se déplacer du seulement-imprimable (printable-only) des versions dans des créations multimédia avec une petite aide de sgml. Pendant que les manuscrits sont entrés dans le système, le SGML est automatiquement mis en application pour préparer le texte multimédia. Puisque l'Ams veut que chacun ait accès facilement à son information, il compile ses journaux sur CD-ROMs à la fin de chaque année pour une consultation plus facile. L'AMS a également créé une publication sur Internet seulement pour supporter les articles qui ne peuvent pas être réduits à une forme papier, telle que des articles avec des clips de bruit ou de film".

Nom du projet et description	DTD utilisée	Traitement des données / Outils	Accès	Services / Fonctions
<p>ATLA Serials Project http://rosetta.atla-certr.org/CERTR/ATLAS/ http://www.atla.com/</p> <p>ATLA : American Theological Library Association créée en 1999, par le CERTR⁵⁸ Couvre 50 000 articles retenus parmi 6 disciplines : Archéologie, Bible, Oecuménisme, Ethique, Missions, Ministère pastoral, Philosophie, Religion, Société, et Théologie</p> <p>Objectif Numériser la valeur de 50 ans de 50 journaux qui traitent de l'étude des religions et les rendre accessibles sur le web. Les revues les plus exceptionnelles ont été sélectionnées par un groupe de travail.</p>	<p>DTD EBind⁵⁹ et DTD TEI</p>	<p>La numérisation s'est faite de 2 façons :</p> <ul style="list-style-type: none"> - images encapsulées - encodage plein texte <p>Images encapsulées Des images de page scannée sont enveloppées dans une métadonnée « enveloppe » xml, conformément à la DTD EBind</p> <p>Encodage XML plein texte Encodage des documents se fait grâce à la DTD TEI. L'affichage sur le web se fait en xml. Si le navigateur web de l'utilisateur ne permet de visionner un document xml, alors conversion à la volée d'une page html à partir du fichier xml.</p> <p>Moteur de recherche Lui-même basé sur la base de donnée ATLA Religion, interroge un index commun, apparemment connu et estimé.</p>	<p>Accès à la base de données ATLA Serials en xml, soumis à abonnement</p>	<p>Recherche sur images</p> <ul style="list-style-type: none"> - sur des articles écrits par des auteurs particuliers - sur des articles qui contiennent certains mots du titre - sur des articles qui traitent de sujets particuliers ou qui traitent de passages sélectionnés dans divers textes religieux <p>Recherche plein texte</p> <ul style="list-style-type: none"> - sur tous les champs du document - simplement sur du plein texte - en utilisant les opérateurs booléens : and, or, not - de façon approximative : le mot A a 10 mots du mot B

⁵⁷ Les détails sur les styles à utiliser sont disponibles à l'adresse suivante http://www.agu.org/ei/ei_style.html#2.

⁵⁸ Centre for electronic resources in theology and religion basé à Stone Mountain en Georgie (USA).

⁵⁹ Cette DTD est le travail de chercheurs de l'Université de Californie à Berkeley, elle est destinée à lier "électroniquement" des images de pages numérisées et à fournir des mécanismes de navigation aisée. Téléchargeable sur le site : <http://sunsite.berkeley.edu/Ebind/>.

5. XML et la structuration de textes littéraires

Nom du projet et description	DTD utilisée	Traitement des données / Outils	Accès	Services / Fonctions
<p>Edition savante électronique aux PUL</p> <p>Objectif : édition électronique de plusieurs types de documents réalisée par les Presses Universitaires de Laval (Canada) :</p> <ul style="list-style-type: none"> - périodiques électroniques au format LaTeX, html, pdf et sgml - deux publications électroniques des PUL aux formats html et sgml. <p>Le traitement des données n'est expliqué que pour les 2 publications éditées par les PUL et qui ont subi le traitement</p>	DTD ISO 12083 ⁶⁰	<p>Formatage traitement de texte</p> <ul style="list-style-type: none"> - document source composé sous PageMaker pour Mac et converti pour une sortie en Word Perfect 5.1 DOS. - document Word Perfect non style mais composé de constantes facilement 'recherchables' dans le texte : elles permettent l'automatisation du processus en remplaçant les balises de formatage par des balises de structure tout en respectant l'ordre déclaré dans la DTD⁶¹. <p>Automatisation du processus grâce aux macros Word Perfect</p> <p>Elles pointent sur les éléments à remplacer :</p> <ul style="list-style-type: none"> - macros pour le traitement du corps du livre : balise les titres et auteurs de chaque chapitre, les sections et sous sections, les citations - macros de traitement des hyperliens : crée les hyperliens entre le texte et les notes en fin de document, entre une citation et la référence bibliographique - autres macros : la macro exposant et la macro tableau <p>Conversion automatique vers sgml</p> <p>Conversion faite à l'aide de feuille de style issue de Panorama, viewer sgml, pour le web.</p> <p>Problèmes rencontrés au cours du traitement</p> <ul style="list-style-type: none"> - la taille du fichier sgml, - la représentation des caractères sur le web, - des références bibliographiques et des graphiques 	<p>Accès réservé pour les périodiques aux membres du campus</p> <p>Accès gratuit sur le web aux 2 publications électroniques</p>	<p>Services</p> <p>Formats html et sgml pour visionner les documents en texte intégral.</p> <p>Inconvénient : posséder un viewer sgml</p>

⁶⁰ DTD issue de l'American Association of Publishers (AAP) avec son projet de manuscrit électronique. La norme ISO 12083 comprend en fait 4 DTD : DTD article pour décrire des articles, DTD serial pour décrire des collections d'articles, DTD book pour décrire des monographies et des thèses, DTD maths pour décrire des formules mathématiques. Les informations sur ces DTDs sur le site <http://www.xmlxperts.com/12083xml.htm>

⁶¹ Ce processus peut être complètement automatisé grâce à l'outil *SGML Author for Word*. Il permet de convertir facilement un fichier Word en SGML ; pour cela, il associe, au moyen d'un tableau d'équivalences, un style Word à une balise SGML. Le seul problème est qu'il n'est pas compatible avec les nouvelles versions de Word. Site : <http://www.microsoft.com/catalog/display.asp?subid=22&site=723&pg=1>

6. XML et les plate-formes documentaires

Nom du projet et description	DTD utilisée	Traitement des données / Outils	Accès	Services / Fonctions
<p>OPERA http://www.inrialpes.fr/opera.html Sous la tutelle de l'INRIA, Institut de Recherche en Informatique et Automatique, s'intéresse aux documents électroniques, hypertextes et multimédia.</p> <p>Etudie les modèles de documents qui rendent compte à la fois de leur organisation logique, de leur présentation graphique et des contenus multimédia.</p> <p>Met également au point des techniques d'édition qui s'appuient sur ces modèles.</p>	Outils développés en interne	<p>Les outils développés le sont en fonction des axes de recherche du projet :</p> <ul style="list-style-type: none"> - structure de documents : techniques de transformations de structure pour les documents XML et méthodes de comparaison automatique et langages d'expression des transformations. Ces techniques sont expérimentées dans les éditeurs Thot⁶² et Amaya⁶³. - documents multimédias : environnements d'édition de documents multimédia et environnements de présentation offrant des services efficaces et de haut niveau comme la navigation temporelle, la synchronisation, l'accès réseau. Outil MADEUS, éditeur de documents multimédias - coopération sur le web : édition coopérative de documents sur le web grâce à l'outil BYZANCE. 	Accès gratuits aux outils comme Thot ou Amaya.	<p>Fonctions</p> <ul style="list-style-type: none"> - permet la structuration de documents - permet la création de documents multimédias utilisant des standards - permet l'édition coopérative de documents sur le web
<p>DILIB http://www.loria.fr/projets/DILIB/ Projet du LRIA, Laboratoire lorrain de recherche en informatique et ses applications.</p> <p>Plate-forme pour l'ingénierie du document et de l'information scientifique et technique qui propose des solutions telles que la prise en compte des aspects volume, temps de réponse, données approximatives (clusters), la structuration et le formatage de documents en Unimarc, SGML et XML.</p>	Outils développés par le centre de recherche LORIA	<p>DILIB propose des outils tels que :</p> <ul style="list-style-type: none"> - Une boîte à outils SGML : bibliothèque de fonctions, commandes de manipulations de documents sgml, outils de conversion - Des modules pour construire des Systèmes de Recherche d'Information : outils de construction de système, d'interrogation, interfaces web - Des modules pour l'analyse infométrique des informations : fonctions et modules destinés à analyser les fonds d'information pour réaliser des études ou des systèmes de fouille de données, de recherches ou de veille 	Accès gratuit à certaines démonstrations, et sur demande, un mot de passe peut être fourni pour une exploration ponctuelle	<p>Fonctions</p> <ul style="list-style-type: none"> - l'investigation documentaire, - la construction de systèmes de recherche d'information, - les outils pour les bibliothèques électroniques

⁶² Ce logiciel est en open source et téléchargeable sur le site <http://www.inrialpes.fr/opera/Thot.fr.html>

⁶³ Ce logiciel est désormais l'éditeur et le navigateur de document xml recommandé par le W3C, il est téléchargeable sur leur site <http://www.w3.org/Amaya/>

7. Commentaires

En comparant les processus éditoriaux du *BBF* électronique et ceux précédemment exposés, on peut faire plusieurs remarques.

Tout d'abord, l'utilisation d'une DTD reconnue telle la DTD TEI. Une DTD normalisée présente de nombreux avantages : elle bénéficie d'un suivi, de mises à jour et d'un réseau d'experts, elle est aussi adaptable aux besoins (ajout/suppression d'éléments par exemple). En vue de l'édition électronique des étudiants, le service informatique devait connaître les réalisations à ce sujet. Il apparaît dans ce cas que la DTD TEILite répondrait à leurs exigences. Elle est dédiée à la structuration de textes électroniques, est largement répandue et bénéficie d'un réseau international en raison de la reconnaissance de ses recommandations.

Dans le cadre d'un prochain projet, le service informatique devrait donc s'orienter vers ce type de DTD se déchargeant ainsi des évolutions futures en terme d'écriture ou de révision de DTDs, ce choix lui permettrait de se rapprocher d'autres initiatives utilisant la même structuration de documents. En effet, la plupart de ces projets participent à des initiatives communes plus importantes favorisant l'entraide et la mise en place de système d'information stable et semblable.

Par ailleurs, les politiques d'édition électronique datent bien souvent de plusieurs années et certains projets ont acquis une première expérience du document structuré en SGML, ce qui n'est pas le cas pour l'équipe du *BBF* électronique, dans leur cas la conversion et/ou le balisage des documents se fait automatiquement en XML.

Lors d'un balisage manuel, les solutions adoptées sont bien souvent les mêmes que pour le *BBF* (feuilles de style et/ou création de macros...). Rappelons que ce projet est récent et qu'il représente une première expérience pour le service informatique. Les méthodes de traitement et les applications sont perfectibles et le transfert de compétences d'Araxe

dont il bénéficie lui permettra d'asseoir ses connaissances et de les approfondir.

D'autre part, il convient d'ajouter que la revue électronique est fortement liée à sa version papier, contrairement à d'autres projets où le support papier devient un sous-produit de la forme électronique.

Le *BBF* électronique, lui, en est directement issu et ses contraintes éditoriales (organisation du contenu, graphisme et typographie) sont déterminantes.

En se référant à l'article⁶⁴ G. Beaudry et G. Boismenu, le traitement de publication du *BBF* se situe typiquement dans l'option de "Mise en page vers XML". Cette option est avant tout dédiée aux documents issus de logiciel de PAO (Quark XPress) et migrant vers un format numérique.

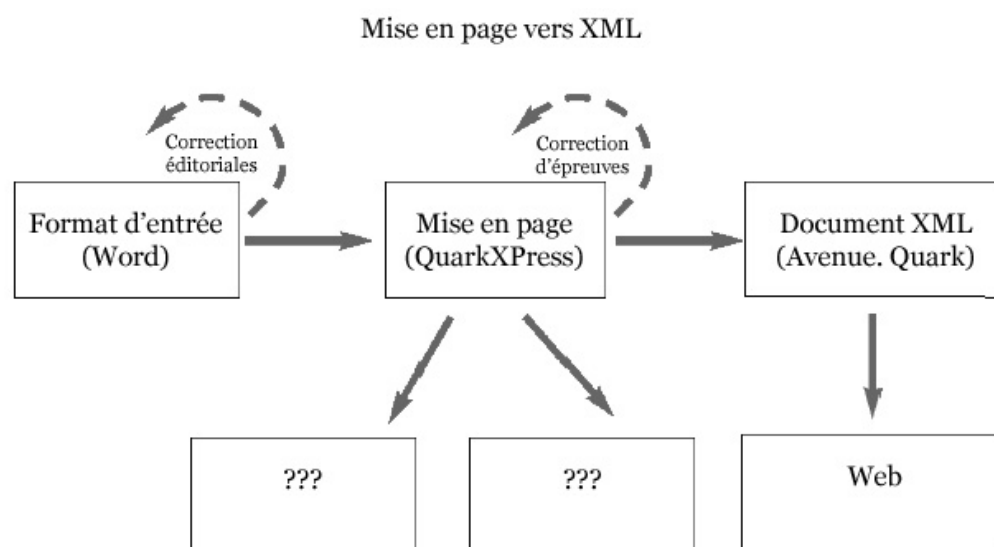


Figure 15 – Option de mise en page vers XML

⁶⁴ *Ibid.*

Elle présente deux inconvénients :

- *"(...) dépendance pour un format propriétaire, tel que celui de Quark"*
- *"(...) cette option fait du numérique un produit d'un processus conçu pour le papier. La chaîne de production est en fonction du papier, une fois la revue prête à imprimer réalisée, un fichier est à la fois acheminé à l'imprimeur et utilisé pour les opérations de conversion vers XML"*

Les réflexions de ces deux auteurs confortent le service informatique dans ses choix au sujet de la chaîne de production du *BBF* électronique puisque *"compte-tenu de l'importance du papier, de l'expertise et de ces façons de faire éprouvées par les équipes des revues, l'option "Mise en page vers XML" est certainement à considérer et plus facile à implanter qu'une chaîne de traitement XML-intégrée"*⁶⁵. En effet, elle ne nécessite pas l'achat d'applications souvent basées sur des environnements XML-natifs qui s'élèvent à plusieurs milliers de francs. Dans le cas du *BBF* ce type d'investissement n'était pas envisageable.

Il est bon de noter que *"(...) ce modèle ["Mise en page vers XML"] offre exactement les mêmes avantages que le modèle XML-intégré. Les variables touchant les possibilités de recherche, la génération automatique des métadonnées et l'indice de pérennité élevé sont du même ordre pour les deux modèles"*⁶⁶. L'intérêt d'une solution XML-intégrée repose surtout sur l'automatisation des tâches puisque d'un document XML natif dépend plusieurs supports de sortie (papier, CD-ROM, fichiers XML, PDF, HTML ...).

⁶⁵ *Ibid.*

⁶⁶ *Ibid.*

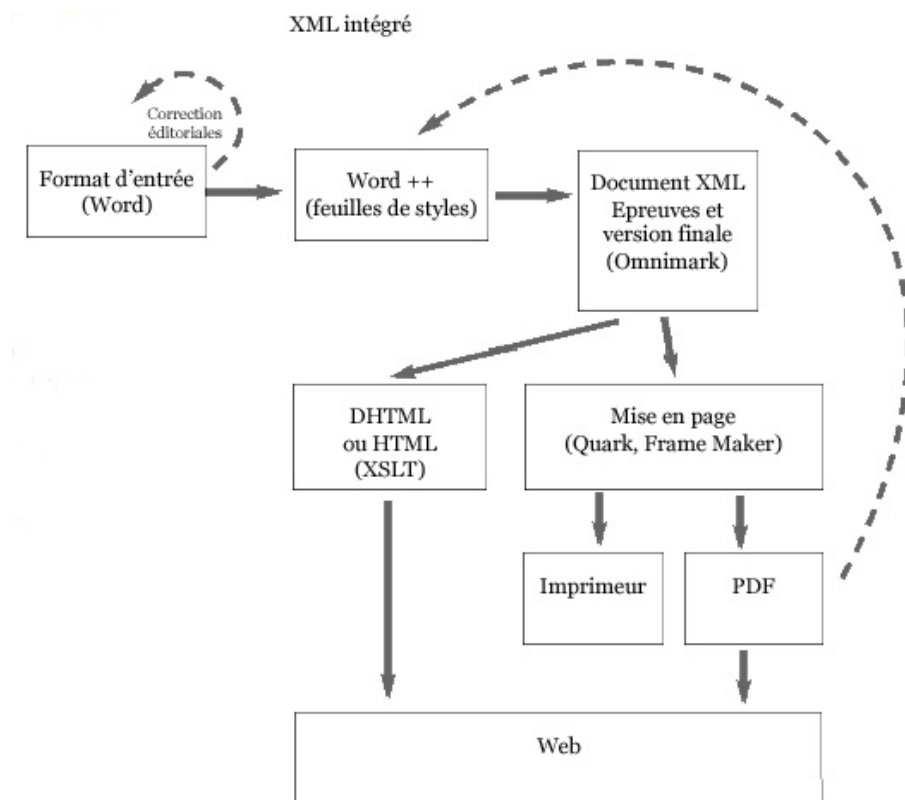


Figure 16 – Option XML intégrée

Par rapport aux services proposés par le *BBF*, la gazette électronique est originale tout comme les possibilités de recherche sur des informations précises comme les illustrations et les citations. Quant à la souscription d'un abonnement pour avoir accès à la version électronique, ce choix est partagé suivant les initiatives..

Bilan sur le *BBF* électronique et son traitement informatique

1. Observations au terme de ces quatre mois

1.1. Un choix éditorial "prédéterminé" dès le départ

Puisqu'elle ne s'affranchit pas de la version papier de la revue et de ses contraintes de présentation, graphiques et typographiques, la revue électronique n'est pas un produit ex-nihilo. Or habituellement dans ce genre de projet, la revue papier devient un sous-produit de la revue électronique en raison des nombreux avantages que présente le format XML : à savoir la sortie multi-support, électronique ou papier, d'un document XML.

Dès la mise en place du projet, l'une des conditions était de garder la signature de la version papier. Le formatage en PDF en est d'ailleurs la preuve.

Les feuilles de style XSL-FO permettaient, en effet, de traiter automatiquement un fichier XML en PDF, avec une présentation épurée. Mais il a été décidé de continuer de traiter les fichiers PDF via XPress, copie conforme de la revue papier.

1.2. Un choix technique judicieux

1.2.1 Réalisations du SI

A la lecture de différents articles, il apparaît que le stockage et la gestion de documents XML dans un SGBDR est un choix judicieux qui a fait ses

preuves. Selon M. Varandat, dans un article⁶⁷ de *01 informatique*, "*le SGBDR peut prétendre au stockage de documents XML, ce qui n'est pas sans remettre en cause la raison d'être des bases natives XML*". Entre la base de données relationnelle et la base de données native XML, on se ne sait pas en fait laquelle possède les meilleurs résultats en ce qui concerne le traitement de documents XML.

La seule certitude est que le SGBDR est capable de gérer la richesse des documents XML. Il propose en plus des outils de mapping, outils qui traduisent le contenu d'un document XML en fonction de son arborescence dans des tables relationnelles.

A la demande de J. Lavandier, j'ai également dû positionner l'orientation technique prise pour le *BBF* électronique par rapport à d'autres produits présents sur le marché.

1.2.2 Autres outils disponibles

Le principal concurrent du SGBDR est, on l'a vu, la base de données native XML. Voici un tableau récapitulant les principaux outils de stockage des documents XML⁶⁸ (Microsoft SQLServer 2000 ne sera pas traité ici) :

	Produits / Fournisseurs	Caractéristiques principales
SGBDR	DB2 UDB 7.1 / IBM	XML Extender est un outil de mapping permettant d'éclater un document XML dans des tables relationnelles puis de les restituer. Génère des documents XML à partir de requêtes SQL.
	Internet Foundation 2000 / Informix	Extension XML (Datablade) et restitution de documents XML à partir de requêtes SQL. Mapping entre les tables et les documents XML.
	Oracle 9i / Oracle	Extension XML (cartridge). Restitution de documents XML à partir de requêtes SQL. Mapping entre les tables et les documents

⁶⁷ VARANDAT M. Quand le SGBDR devient un document XML. *01 informatique*, 18 mai 2001, n°1634, pp. 44.

⁶⁸ *Ibid.*

	XML Suite / e-XMLMedia	Référentiel pour SGBDR : e-XML Repository. Chargement de document XML dans une base relationnelle et restitution par requête. Mapping entre tables et schémas avec e-XML Mediator...
Serveur natif XML	dbXML Core 4 / dbXML Group	Commercialisé sous licence GNU ⁶⁹ . Services Corba, outils d'administration ...
	Tamino / Software AG	Moteur (X-Machine) combinée à une base hiérarchiquement organisée. Module X-Node pour le mapping de tables en XML. Implémentation de XQL, XPATH. Moteur SQL.
	EXcelon / EXcelon	Serveur XML s'appuyant sur un SGBD objet. Stockage de documents sous forme d'arbres DOM accessibles par requêtes XQL. Importation de données relationnelles, outil de conception de schémas.

J'ai recensé de nombreux éditeurs et parseurs XML, qui concurrencent à XMLSpy, ses qualités étant reconnues, je ne m'attarderai pas davantage.

Les outils de conversion⁷⁰ peuvent toutefois s'avérer utiles. Ils permettent de générer automatiquement un document XML à partir de Microsoft Word, plus précisément, à partir de RTF. Parmi eux, on peut citer Majix⁷¹, RTF2XML 0.9, RTF4XML, UpCast 2.0. Peut-être faudrait-il tester ces produits pour alléger le balisage des documents XML.

Il est intéressant également d'appréhender les solutions qu'offrent les spécialistes de la GED⁷², puisque leurs outils s'adaptent à XML. Comme le montre le tableau récapitulatif⁷³ suivant :

⁶⁹ GNU, traduction française de GPL., General Public License. Cette licence est souvent utilisée pour les logiciels libres, et non gratuits.

⁷⁰ *Ibid.*

⁷¹ Une version "light" est à télécharger sur leur site www.tetrasix.com

⁷² Gestion Electronique de Documents

⁷³ Le tableau est issu de l'article de PARPINELLI S. La migration des documents est en marche. *01 informatique*, 18 mai 2001, n°1634, 45 p.

Editeur / Produit	Prise en compte de XML
Arbortext Epic	Création de contenu XML structuré. Conversion de documents existants (Word, SGML) enXML. Diffusion du contenu sur le web, WAP...
Documentum 4i	Création et gestion de contenus : importation de documents XML, y compris avec les éditeurs d'Arbortext et de SoftQuad, validation de DTDs et de schémas. Génération de documents XML composites pour un affichage sur divers canaux.
FileNet Panagon 5.2	Workflow en XML. Requête nécessaire à la generation d'un document est décrite en XML.
IBM Content Manager 7.1	Incorporation de contenu XML dans le référentiel de Content Manager. Fiche d'indexation peut être décrite en XML
Intranet Solutions Xpedio	Conversion de documents Word, Excel, Powerpoint en contenu XML. Utilisation de feuille de style XSL. Métadonnées du référentiel incorporées aux documents XML générés.
OpenText LiveLink 9	Parseur et moteur XSLT intégrés. Contenu et informations sur le document exportés en XML. Prise en compte par le moteur de toute source externe pouvant être représentée en XML.

Enfin, il existe des entreprises proposant de développer des solutions sur mesure, c'est le cas de la société AJLSM, chargée par le ministère de la culture et de la communication, de développer BiblioML pour l'échange de notices bibliographiques. Le groupe Berger-Levrault⁷⁴ et la société Eurodoc-Sofilog⁷⁵ proposent également les services d'experts sur la chaîne de traitement du document au format XML.

Ces outils offrent de multiples potentialités mais celles-ci ne sont pas forcément appropriées au traitement du *BBF* électronique. Les bases natives XML restent un investissement trop coûteux et qui n'est pas justifié pour le moment dans le cas du *BBF*, en raison de son trop faible volume de documents.

⁷⁴ <http://www.xml-ais.com/>

⁷⁵ <http://www.eurodoc-sofilog.com/fr/tec/index.html>

En ce qui concerne les applications de GED, elles sont davantage tournées vers le travail collaboratif et le commerce-électronique qui demandent de plus à plus une diffusion multi-support du document (Web, Wap, assistant personnel, papier ...).

1.3. Une expertise sur un système d'information liant la technologie XML

Cette première expérience de gestion de projet d'édition XML représente une réelle expertise pour le service informatique. Il possède désormais un savoir-faire sur la création et la gestion de documents XML et une maîtrise de la chaîne de traitement de tels documents.

Fort de cette expérience d'édition électronique au format XML, le service informatique pourra dorénavant choisir un modèle de traitement, de gestion et les technologies qui sont les plus adaptées.

E. Romand-Monnier représente en effet la personne ressource à l'enssib pour ce genre d'applications.

2. Pour un balisage plus humain

J'entends par là faciliter le plus possible cette tâche, en réduisant les gestes répétitifs et le temps consacré au balisage. Par ailleurs, les macros se sont avérées utiles pour le baliseur puisqu'il traite un numéro dans sa totalité, formatage en PDF et balisage inclus, en 10 jours.

2.1. Aide au balisage

La gazette électronique est une rubrique purement électronique. Elle est entièrement réalisée par la cyberbibliothécaire qui est la seule à maîtriser la logique des documents dont elle est l'auteur.

Peut-être faudrait-il envisager qu'elle écrive et balise simultanément puis indexe les documents de la *Gazette électronique*.

Comme pour le baliseur, E. Romand-Monnier pourrait installer le modèle contenant les macros dans Word, sur son poste et lui expliquer la structure de la DTD dédiée à la *Gazette électronique*.

Cette manipulation n'est par guère envisageable pour les auteurs des articles. Par contre, l'utilisation par ces derniers d'une feuille de style Word pour la rédaction des articles et d'une feuille de style dans X-Press par l'équipe de Paris, pour la mise en page préparerait le balisage. Chaque feuille de style reprendrait les éléments de la DTD ce qui faciliterait le balisage final de l'article.

2.2. Révision des DTDs

La révision des DTDs pourraient être envisageable dans le cadre de la rétro-conversion des anciens numéros (fonds 1995-1999). Il faudrait en effet s'assurer de l'utilité de DTDs si complexes, qui rendent laborieuses, les tâches de conversion automatique d'un format natif type traitement de texte vers un format XML.

Il existe également les W3C Schémas qui offrent des structures arborescentes et qui ont été créés pour pallier les imperfections des DTDs. L'inconvénient est qu'ils n'ont pas encore la même reconnaissance et ne sont pas aussi répandus que l'utilisation que d'une DTD.

Le balisage se faisant manuellement, il est donc important de savoir si l'on met en place une procédure de rétro-conversion à partir de ces DTDs ou non. Et l'enssib semble de toute façon intéressée par l'enrichissement de la base XML du *BBF* et la mise en place d'autres projets d'édition dans ce même format.

Conclusion

L'intérêt du stage a été multiple pour moi. La découverte d'une nouvelle technologie, XML, et de ses multiples potentialités, m'a tout d'abord fait connaître un aspect de l'édition numérique qui m'était jusqu'alors inconnu. Il m'apparaît désormais évident d'appréhender préalablement, dans sa globalité la chaîne de production d'un document destiné à la publication, aussi bien électronique que papier.

Le stage, se déroulant au moment même où la base de donnée se constituait, m'a révélé la difficulté de mener à bien un projet. La nécessité de constituer des documents de référence, récapitulatifs des différentes attentes sur ce projet, est indispensable.

L'intérêt était également de faire face à des imprévus et des urgences inhérents à la mise en place d'un nouveau projet et d'assister à toutes les étapes de développements, de révisions et de balisages me permettant ainsi d'avoir une vue globale du *BBF* électronique.

Abréviations

Cette liste présente les abréviations et pour certaines, leurs principales caractéristiques. Elle est principalement issue de l'article : Après le Big Bang, la galaxie XML poursuit son expansion. *01 informatique*, 18 mai 2001, n°1634, pp. 50.

ASP : Active Server Pages.

DTD : Document Type Definition. Elle décrit la structure d'un document XML.

HTML : HyperText Markup Language. Format d'échanges de données sur le web et qui définit la présentation graphique de pages web.

PDF : Portable Document Format. Format propriétaire d'Adobe. Principalement destiné à l'impression, il a l'avantage de préserver l'apparence originale du document.

RTF : Rich Text Format. Format d'échange proposé par Microsoft, pour des environnements différents. Il est de plus en plus utilisé pour la conversion automatique vers XML.

SGML : Standardized General Markup Language. Il est moins contraignant que le XML car ne nécessite pas la fermeture des balises mais le balayage d'un document SGML se fait en boucle, il n'y a donc jamais de fin, ce qui posait des problèmes de gestion des documents qui atteignait un niveau de complexité trop important. Il est facile de passer automatiquement d'un format SGML à XML. Le SGML est donc plus souple mais moins propre que le XML.

XML : eXtensible Markup Language. Il a de nombreux caractéristiques et avantages développés dans le rapport. L'avantage par rapport au XML est que le document XML a un jeu de balises prédéfinies au départ qui donnent une cohérence au document et une fin.

XSL : XML StyleSheet. Feuilles de style permettant la transformation (XSLT) et le formatage (XSL-FO) de fichiers XML.

XSL-FO : XML StyleSheet – Formatting Object.

XSLT : XML StyleSheet Transformation.

Références bibliographiques

1. BIBLIOGRAPHIE

BERANGER V. CHAMONARD M. et alii. *Indexation de la version du BBF électronique au format XML*. Villeurbanne : ENSSIB, 2001. Etude de cas informatique, DCB.

MICHARD A. *XML, Langages et applications*. Paris : 1998, Ed ; Eyrolles, 361 p.

PARPINELLI S. La migration des documents est en marche. *01 informatique*, 18 mai 2001, n°1634, 45 p.

VARANDAT M. Quand le SGBDR devient un document XML. *01 informatique*, 18 mai 2001, n°1634, pp. 44.

2. WEBLIOGRAPHIE

ANDRIES P., CUNY S. et alii. Langage de balisage extensible (XML) 1.0. Traduction des recommandations du W3C, 10 février 1998 [on line]. [Consulté le 03/07/01]. Available from internet : <url : http://babel.alis.com/web_ml/xml/REC-xml-en-tete.fr.html>

ATTAR J., CHATEL B. Etat des recommandations XML dans le domaine documentaire. *Cahiers GUTenberg*. [On line]. Décembre 2000, n°37-38, 32 p.[visité le 12/07/01]. Available from internet : <url: www.gutenberg.eu.org/pub/GUTenberg/publicationsPDF/37-attar.pdf>

BOISMENU G., SEVIGNY M., VEZINA M.H., BEAUDRY G. *Le projet ERUDIT : un laboratoire québécois pour la publication et la*

diffusion électroniques des revues universitaires [On line]. Montréal : Presses Universitaires de Montréal, juin 1999, 260 p. Available from internet <url: <http://www.erudit.org/erudit/rapport/rapport.pdf>>

BOISMENU G., BEAUDRY G. *Expertise technique et organisationnelle* [On-line] Villeurbanne, [consulté le 02/07/01]. Available from internet : <url: <http://revues.enssib.fr/Index/indextecnic.htm>>

NICULETA S. *Conception d'une chaîne de conversion : format RTF vers XML* [On line]. Villeurbanne : INSA-LISI, 2001, 43 p. Rapport final PFE 2000-2001. [Consulté le 02/07/01] Available from internet : <url: <http://csidoc.insa-lyon.fr/these/doc/rapport-final2001.pdf>>

ROMAND-MONNIER E., *Migration d'une revue professionnelle vers un modèle structuré en ligne* [On line]. Villeurbanne : ENSSIB-LYON1, 2000, 92 p. Available from internet : <url: www.enssib.fr/bibliotheque/documents/dessid/rsromandmonnier.pdf>

SEVIGNY M. *Un modèle de traitement pour l'édition électronique de revues savantes* [on line]. Montréal : PUM, 1997. [Consulté le 03/07/01]. Available from internet : <url: www.pum.umontreal.ca/pum/publ_electr/rapports/sev01/modele_de_traitement.html>

Sites web de référence

1. Portails XML

Francophones

<http://www.abf.asso.fr-enrichi-xmldoc.htm>

<http://www.mutu-xml.org/index.html>

<http://www.chez.com/xml/>

<http://xmlfr.org/>

Anglophones

<http://www.oasis-open.org/cover/sgml-xml.html>

<http://www.xmlconference.com/>

<http://xml.apache.org/>

<http://www.xmlsoftware.org>

<http://www.xmlxperts.com/>

2. DTDs normalisées

EAD : <http://sunsite.berkeley.edu/ead/>

EBIND : <http://sunsite.berkeley.edu/Ebind/>

ISO 12083 : <http://www.xmlxperts.com/12083xml.htm>

MASTER :

<http://www.hcu.ox.ac.uk/TEI/Master/Reference/DTD/masterx.dtd>

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

ANNEXES

Table des annexes

annexes 1 : DTDgazette	I
annexe 2 : Liste détaillée des concepts	III
annexe 3 : Structure de la base de donnée	VI

annexes 1 : DTDgazette

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- edited with XML Spy v3.5 (http://www.xmlspy.com) by EMILIE ROMAND MONNIER
(ENSSIB DE LYON) -->
<!--structure de la racine 'document'-->
<!ENTITY % Panorama "Panorama du Web">
<!ENTITY % Actualité "Actualité des textes officiels">
<!ENTITY % Débats "Débats et publications électroniques">
<!ELEMENT document (num-BBF, titre, theme?, chapo?, article)>
<!ATTLIST document
  rubrique CDATA #REQUIRED>
<!--déclaration de l' élément identifiant le numéro du BBFauquel le document appartient-->
<!ELEMENT num-BBF (#PCDATA)>
<!ATTLIST num-BBF
  annee CDATA #REQUIRED
  tome CDATA #REQUIRED
  numero CDATA #REQUIRED
  img_couv CDATA #IMPLIED>
<!ELEMENT titre (#PCDATA)>
<!ATTLIST titre
  titre-rubrique CDATA #REQUIRED>
<!ELEMENT theme (#PCDATA)>
<!ELEMENT chapo (para+)>
<!--Structure de l'élément article-->
<!--déclaration des éléments définissant le contenu de l'article, i-e le texte -->
<!ELEMENT article (section-niveau1)*>
<!--déclaration des éléments définissant la structure de l'article: partie de niveau1, sous
partie de niveau2 et sous-sous parties de niveau3-->
<!ELEMENT section-niveau1 (titre-section1, para*, (texte* | section-niveau2* | BR*)*,
encyclo?, vedette?)*>
<!ELEMENT titre-section1 (#PCDATA)>
<!ELEMENT section-niveau2 (titre-section2, para*, (((p+ | p-appel-note* | p-
illustration?)+)? | texte* | section-niveau3* | BR*)*, encyclo?, vedette?)+>
<!ELEMENT titre-section2 (#PCDATA)>
<!ELEMENT section-niveau3 (titre-section3 | para* | texte* | BR* | encyclo? | vedette?)*>
<!ELEMENT titre-section3 (#PCDATA)>
<!ELEMENT texte (para+ | reference+ | note*)+>
<!ELEMENT BR (#PCDATA)>
<!ATTLIST texte
  auteur CDATA #IMPLIED
  titre CDATA #IMPLIED
  editeur CDATA #IMPLIED
  date-docu CDATA #IMPLIED>
<!--déclaration des éléments définissant le contenu-->
<!ELEMENT para ((p+ | p-appel-note* | p-illustration?)+)>
<!ELEMENT p (#PCDATA)>
<!ELEMENT p-illustration (#PCDATA)>
<!ATTLIST p-illustration
  legende CDATA #IMPLIED>
<!ELEMENT p-appel-note (#PCDATA)>
<!ATTLIST p-appel-note
  num-appel IDREF #REQUIRED>
<!--déclaration des éléments définissant les notes-->
<!ELEMENT note (p-appel-note, (p? | p-URL?)+)>
<!ATTLIST note
  num-note ID #REQUIRED>
<!--description des references aux site-->
<!ELEMENT reference (p-URL, date-consultation)>
<!ATTLIST reference
  logo-pdf CDATA #FIXED "chemin logo"
  type-site (meta-site | site | page) #IMPLIED>
<!ELEMENT p-URL (#PCDATA)>
<!ELEMENT date-consultation (#PCDATA)>
<!--Concepts encyclopédiques-->
```

```

<!--liste détaillée des sujets des 8 concepts de l'encyclopédie: 1 ou 2 sujets dans chaque
concept"-->
<!--Sujets du concept OBJET: au moins-->
<!ELEMENT encyclo (objet, forme, (objet? | etablissement? | support? | public? |
contenus? | lieu? | histoire?)*)>
<!ELEMENT objet EMPTY>
<!ATTLIST objet
    choix-objet (Batiments | Catalogue-bibliographie-classification-indexation | Collections |
Collections-patri-conser | Cooperation | Generalites-divers | Gestion-evaluation | Histoire-
philosophie-bibliothèques | Juridique | Metiers-bibl-doc | Normes-normalisation |
Aspects-politiques-economiques | Production-diffusion-documents | Publics | Savoirs |
Services-non-documentaires | Formation-utilisateurs-mediation | Services-spec-
documentaires | Technologies | Technologies-imprime | Bibliotheque-numerique)
#REQUIRED>
<!--Sujets du concept FORME-->
<!ELEMENT forme (#PCDATA)>
<!ATTLIST forme
    forme4 (Panorama-Web | Actualites-textes-officiels | Debats-publications-
electroniques) #REQUIRED>
<!--Sujets du concept ETABLISSEMENT: 1 ou 2 facultatif-->
<!ELEMENT etablissement EMPTY>
<!ATTLIST etablissement
    choix-etablissement (B-general | Administrations-centrales | Archives | B-academique-
pluri-BU | B-specialisee | B-associative-privees | B-scolaire | B-tiers-reseau | BM-lecture-
publique | BDP-reseau-departemental | BN | Centre-doc | Espace-hors-bibl | Maison-
ecrivain-musee-litt | Maison-edition-librairie | Musees) #REQUIRED>
<!--Sujets du concept SUPPORT: 1 ou 2 facultatif-->
<!ELEMENT support EMPTY>
<!ATTLIST support
    choix-support (Doc-numerique-internet | Estampe-image | Phonogrammes |
Monographie-imprimees | periodiques-imprimes | periodiques-electroniques)
#REQUIRED>
<!--Sujets du concept PUBLICS: 1 ou 2 facultatif-->
<!ELEMENT public EMPTY>
<!ATTLIST public
    choix-public (Chomeurs | Differenciation-sexuelle | Enseignants-chercheurs | Etudiants
| Handicapes | Illettres | Jeunes | Ados | Enfants | Personnes-agees | Professionnels)
#REQUIRED>
<!--Sujets du concept PUBLICS: 1 ou 2 facultatif-->
<!ELEMENT contenus EMPTY>
<!ATTLIST contenus
    choix-contenu (Art | Musique | Droit-Economie-Gestion | Litt-jeunesse | Sciences-
humaines-sociales | Romans | Science-pures-ingenieurs | IST-informatique | STM-
sciences-sante) #REQUIRED>
<!--Sujets du concept LIEU: 1 ou 2 facultatif-->
<!--Principe: France: identification a la région, Europe au pays, au-delà au continent (3
parties pour l'Asie)-->
<!ELEMENT lieu EMPTY>
<!ATTLIST lieu
    choix-lieu (France | Alsace | Aquitaine | Auvergne | Basse-Normandie | Bourgogne |
Bretagne | Centre | Champagne-Ardenne | DOM-TOM | Franche-Comté | Haute-
Normandie | Ile-France | Languedoc-Roussillon | Limousin | Lorraine | Midi-Pyrénées |
Nord-Pas-de-Calais | Pays-de-la-Loire | Picardie | Poitou-Charentes | PACA_Corse |
Rhone-Alpes | Europe | Allemagne | Autriche | Belgique | Espagne | Europe-
centrale_Balkans | Grèce | Italie | Luxembourg | Pays-Bas | Pays-scandinaves | Portugal |
Royaume-Uni_Irlande | Russie_Europe-de-Est | Suisse | Afrique | Amerique-Nord |
Amerique-latine | Asie-centrale | Extreme-Orient | Proche-Orient | Oceanie-Australie)
#IMPLIED>
<!--Sujets du concept HISTOIRE: 1 ou 2 facultatif-->
<!ELEMENT histoire EMPTY>
<!ATTLIST histoire
    choix-histoire (Apres-1990 | Moyen-Age-aujourd'hui | 1945-90 | 1918-1945 | 1870-1918
| 1789-1870 | 18eme | 16eme-17eme | Moyen-Age | Antiquite) #REQUIRED>
<!--indexation RAMEAU-->
<!ELEMENT vedette (sujet*)>
<!ATTLIST vedette
    ram_geo CDATA #IMPLIED
    ram_chrono CDATA #IMPLIED
    ram_forme CDATA #IMPLIED>
<!ELEMENT sujet (#PCDATA)>

```

annexe 2 : Liste détaillée des concepts

Rappel : chaque fichier peut être associé à 1 ou 2 sujets dans chaque concept. Seul le concept Objet est obligatoire, et bien sûr la Forme –point 8- du fichier
Pour chaque ensemble conceptuel, les concepts-fils sont signalés en retrait, sous leur concept-père.

1) Objet

- Bâtiments (équipements, bibliobus, etc)
- Catalogues et bibliographies (y compris bibliographies commerciales)+classifications et indexations
- Collections (et leur traitement)
 - Collections patrimoniales et conservation
- Coopération (réseaux, ...)
- Généralités, divers
- Gestion- évaluation (management, direction, budgets, knowledge management)
- Histoire et philosophie des bibliothèques
- Juridique
- Métiers dans la bibl.et la doc. (compétences, formation professionnelle,...)
- Normes et normalisation
- Aspects politiques et économiques(notamment culturelles et éducatives ; inclut la censure)
- Production et diffusion des documents (édition, librairie,...)
- Publics (étude sociologique, lecture, ...)
- Savoirs (disciplines, recherche, IST, ...)
- Services non-documentaires (accueil, animations, ...)
 - Formation des utilisateurs et médiation
- Services spécifiquement documentaires (prêt, référence, fourniture à distance, recherche et analyse documentaires)
- Technologies (systèmes d'information; inclut logiciels et bibl. électronique en général)
 - Technologies de l'imprimé (papier, imprimerie,...)

2) Etablissement (par défaut : bibliothèques en général)

- Administrations centrales
- Archives
- Bibliothèque (*à n'utiliser que si l'on veut signaler 2 types d'établissements différents*)
- B. académique pluridisciplinaire et BU en général
 - B. Spécialisée
- B. associatives (et bibl. privées en général)
- B. Scolaires
- B. tiers réseau
- BM (et bibl. de lecture publique en général)
 - BDP et réseaux de type départemental
- BN
- Centres de doc
- Espaces hors bibliothèques (en général)
 - Maisons d'écrivain et musées littéraires
 - Maisons d'édition, librairies
 - Musées

3) Support

- Docs numériques et Internet
- Estampes (et images en général)
- Monographies imprimées (et imprimé en général)
- Périodiques imprimés (et périodiques en général)
 - Périodique électronique

4) Publics

- Chômeurs
- différenciation sexuelle
- Enseignants et chercheurs
- Etudiants
- Handicapés
- Illettrés et personnes en diffic. de lecture/écriture
- Jeunes
 - ados
 - enfants
- personnes âgées
- Professionnels

5) Contenus

- Arts (en général)
- Musique
- Droit-Economie-Gestion
- Littérature jeunesse
- Sciences humaines et sociales (inclut Littérature)
 - Romans
- Sciences pures et ingénieurs
 - IST et informatique
- STM (Sciences santé)

6) Lieu

Principe : en France, identification à la région, en Europe au pays, au-delà au continent –distingué en 3 parties pour l'Asie)

- Bourgogne
- Bretagne
- DOM-TOM
- Ile de France
- Languedoc-Roussillon
- Midi-Pyrénées
- PACA
- Picardie
- Rhône-Alpes
- Europe (en général) (*inclut la France dans les requêtes, mais en fin de liste, après les résultats de 1-Europe, 2-Pays dans ordre alphabétique, 3-France*)
 - *Allemagne*
 - *Belgique*
 - Espagne
 - Pays-Bas
 - Royaume-Uni
 - Suède
 - Suisse
- Afrique
- Amérique du Nord
- Amérique latine
- Asie centrale

- Extrême Orient
- Proche-Orient
- Océanie et Australie

7) Histoire

- par défaut : après 1990
- Du Moyen-Age à nos jours (pour aspects historiques larges)
- 1945-1990
- 1918-1945
- 1870-1918
- 1789-1870
- 18^{ème}
- 16^{ème}-17^{ème}
- Moyen-Age
- Antiquité

8) Forme

- Art = Article
- TH = Compte-rendu (Tour d'horizon ou chronique)
- BB = Analyse bibliographique (Bibliothèque du bibliothécaire)

annexe 3 : Structure de la base de donnée

