Rapport de stage Septembre 2003

DESS Ingénierie documentaire

Appropriation d'une plateforme d'édition électronique basée sur XML : Cyberdocs

Abdrahamane ANNE

Sous la direction de M JEAN PAUL DUCASSE
Directeur du service d'Edition, Reproduction
et Archivage du Document
Responsable scientifique du programme Cyberthèses

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

Remerciements

Je tiens à remercier les personnes qui m'ont soutenu et encouragé pendant le déroulement de mon stage :

Messieurs Jean Paul METZGER et Jean Paul DUCASSE qui ont accepté de superviser ce stage.

Mademoiselle Emilie Romand MOUNIER, qui a consacré son temps à la lecture, à la critique et à la correction de ce travail.

Kim, Magalie, Nathalie et tout le personnel de l'ERAD qui m'ont accueilli et accepté dans le cadre de ce stage.

Monsieur et Madame ABDOULAHI qui m'ont supporté durant toute cette année.

Bouchra qui m'a écouté, aidé et conseillé.

Liana, qui a rendu ce travail beaucoup plus lisible.

Et tous ceux qui m'ont aidé, supporté et critiqué.

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

Droits d'auteur réservés.

Appropriation d'une plateforme d'édition électronique basée sur XML : Cyberdocs

Abdrahamane ANNE

Résumé : Après un bref survol du contexte international de l'édition électronique ce rapport présente la plateforme d'édition électronique des thèses conçue par le programme Cyberthèses de l'Université Lumière Lyon 2. Cette plateforme bâtie autour de normes internationales est un ensemble de logiciels libres conçu pour transformer des textes issus de logiciel de traitement de texte en documents structurés et de construire une base documentaire permettant de rechercher et d'afficher des documents en XML.

Mots clés : Edition électronique ; Thèse ; Document structuré ; XML ; Base documentaire.

Title : Cyberdocs : a XML based electronic publishing and information system

Abstract : The goal of this paper is to describe an XML based information system, developed by Cybertheses, a program run by the Université Lumière Lyon 2. Cyberdocs is a set of free softwares intended to produce, index and retrieve XML documents.

Keywords : Electronic publishing, Thesis, Structured document, XML, Information system.

Droits d'auteur réservés.

GLOSSAIRE

Fichier TEI : fichier XML conforme à la DTD TEI Lite

Fichier Word : Fichier informatique au format du logiciel Word de Microsoft

Stylage : Opération consistant à appliquer des styles aux différentes parties d'un document conformément des règles définies dans un modèle de document.

DTD : Document Type Definition, ensemble de règles de construction d'un document XML.

ABRÉVIATIONS

FMPOS : Faculté de Médecine de Pharmacie et d'Odonto-Stomatologie de l'Université de Bamako

OAI: Open Archives Initiative

OO: OpenOffice.org

SDX : Système Documentaire XML. Logiciel permettant la recherche et la consultation de documents XML

TEI : Text Encoding Initiative

XML : Extensible Markup Langage

XSL : Extensible Stylesheet Langage

Droits d'auteur réservés.

SOMMAIRE

INTRODUCTION 7

OBJECTIFS ET CONTEXTES..... 9

1. Objectifs du stage 9

2. Contexte 9
   - 2.1. Contexte international : Internet et la publication scientifique..... 9
   - 2.2. L'Afrique et l'accès à l'information scientifique 11
   - 2.3. Contexte du stage..... 13

3. Problématique 32

LA PRODUCTION DES THÈSES EN LIGNE..... 33

1. Stylage 33

2. La conversion 36
   - 2.1. Les problèmes rencontrés et leurs causes possibles, les solutions envisageables 36
   - 2.2. Les sources des erreurs 38

3. L'indexation 39
   - 3.1. Les erreurs 41

4. La mise en ligne 43

5. L'archivage 43

6. Evaluation : L'interdépendance des étapes 43

INSTALLATION DE LA PLATEFORME 46

1. La machine virtuelle java 46

2. OpenOffice.org 46

3. Tomcat 46

4. SDX 47

5. Installation de Cyberdocs 47

LA PERSONNALISATION DE LA PLATEFORME..... 52

1. Le module de conversion 52
   - 1.1. Les styles..... 52

Droits d'auteur réservés.

2. L'interface de consultation 55
2.1. Ajout de nouvelles institutions 55
2.2. Habillages 56

MODÉLISATION DE LA PUBLICATION D'UNE REVUE ÉLECTRONIQUE SOUS CYBERDOCS 62

1. Intérêt de la diffusion électronique de revues 62

2. Présentation de la revue Mali Médical 62

3. Identification des éléments : création des schémas et modèles de documents 64

4. Adaptation de la plateforme 66
4.1. Adaptation de la conversion 68
4.2. Modification de l'interface de consultation 69
4.3. Accès aux articles 72
4.4. Comment adapter Cyberdocs aux revues 73

CONCLUSION 74

BIBLIOGRAPHIE 76

TABLE DES ANNEXES I

Droits d'auteur réservés.

Introduction

Né dans les années 70, l'Internet est aujourd'hui largement répandu et utilisé dans différentes activités : commerce, éducation, publication, etc. Presque tous les pays du monde y ont accès. Il est devenu un espace de diffusion et d'accès à l'information économique, scientifique, médicale, ludique. C'est également un moyen de communication permettant de transporter, d'échanger, de consulter des documents sonores, graphiques, textuels.

Les avancées des nouvelles technologies de l'information et de la communication (NTIC) permettent aujourd'hui aux universités d'entreprendre la diffusion, entre autres, des thèses en format électronique. Quelles que soient les raisons qui poussent une université à se lancer dans cette entreprise, le résultat incontestable est que ces documents acquièrent une plus large audience. Le stade de l'expérimentation et des initiatives isolées est dépassé. Des projets nationaux et internationaux sont déjà en place ou sont en train d'être mis en place. Ces différents projets sont en train de converger vers la fédération de leurs collections électroniques en se conformant à des standards ou recommandations comme celui de l'Open Archive Initiative (OAI).

Cyberthèses, qui était au début une initiative de l'Université Lyon 2 et l'Université de Montréal, a pour objectif la mise en place d'une plateforme de diffusion électronique des thèses basée sur le document structuré. A la suite d'une série de formations, Cyberthèses s'est ouvert à d'autres institutions en France, en Asie, en Amérique Latine et en Afrique.

Au cours d'un séminaire organisé à Dakar en 2002, le responsable de la Faculté de Médecine de Pharmacie et d'Odonto-Stomatologie de Bamako a été formé aux principes et à la technique de diffusion électronique des thèses. Depuis cette date un système de dépôt électronique des thèses a été instauré à la FMPOS. A ce jour, plus de 300 copies électroniques de thèses attendent d'être mises en ligne.

Droits d'auteur réservés.

La diffusion des thèses sur Internet peut être un moyen alternatif pour les pays en voie de développement de donner une plus large audience aux résultats de travaux de recherche. La part de l'Afrique dans la production scientifique mondiale est relativement peu visible à cause du petit nombre de publications mais aussi à cause des problèmes de distribution. Or, Internet met presque au même pied d'égalité, quant à la facilité de diffusion, les scientifiques du Nord et ceux du Sud.

Le présent rapport essaie dans un premier de temps de situer la publication électronique dans le contexte international de l'édition scientifique et présente les différentes institutions qui pourront être partenaires dans la mise en place d'un système de publication électronique à La Faculté de Médecine de Bamako.

Dans un second temps nous décrirons les différentes tâches que nous avons pu réaliser durant ce stage à savoir le traitement des thèses en vue de leur publication en ligne, l'installation, les tests et la personnalisation de la plateforme Cyberdocs ; la modélisation de la publication électronique d'une revue électronique.

Droits d'auteur réservés.

Objectifs et contextes

1. Objectifs du stage

L'objectif principal du stage était la prise en main d'une plateforme de diffusion des documents structurés en vue de son installation, de son administration et de sa personnalisation dans le contexte de la FMPOS.

Un deuxième objectif non négligeable était d'étudier la possibilité de diffusion d'autres types de documents grâce à la plateforme Cyberdocs.

Finalement le stage devait permettre d'appréhender les problèmes liés à la formation des doctorants, à l'utilisation des modèles de document dans un logiciel de traitement de texte et à la formation des professionnels à la gestion de cette plateforme de production de documents électroniques.

2. Contexte

2.1. Contexte international : Internet et la publication scientifique

La publication électronique des thèses s'inscrit dans la problématique (ou dans la crise) plus large de l'édition scientifique en général. Celle-ci est caractérisée par : l'augmentation du nombre de publications, l'augmentation du prix de l'information, les restrictions budgétaires des centres d'information, le développement et l'expansion de l'informatique et d'Internet, la fracture numérique entre le nord et le sud.

L'augmentation du nombre de publications s'est accompagnée de la concentration de la production de documents scientifiques aux mains de certains groupes commerciaux et de l'augmentation des prix des documents en général et des revues en particulier. Ces deux facteurs couplés aux

Droits d'auteur réservés.

restrictions budgétaires ont conduit les bibliothèques à réduire de plus en plus le volume de leurs acquisitions.

Le développement d'Internet permet la diffusion des résultats de la recherche et facilite la recherche d'information. Il a créé de nouvelles habitudes de recherche, d'accès et de partage de l'information. On peut diffuser toutes sortes de documents : revues, actes de congrès, manuels, cours, bases de données, monographies, etc. L'accès aux ressources en ligne peut être gratuit ou payant. Les acteurs de cette diffusion sont des individus, des universités, des associations professionnelles, des établissements de recherche, des agences gouvernementales, etc. S'en servent comme source d'information les étudiants, les enseignants, les chercheurs, le grand public, différentes catégories professionnelles, les entreprises, les administrations, etc.

Parmi les acteurs de l'Internet figurent également les éditeurs commerciaux. Ils y commercialisent leurs productions. De plus en plus de revues scientifiques « traditionnelles » sont également diffusées sous forme électronique.

Comme moyen de diffusion et d'accès à l'information, l'Internet est en train de devenir un 'terrain de combat' où s'affrontent les entreprises commerciales de l'édition et leur clientèle. Bien que l'informatique ait réduit les coûts de production des documents, le prix d'accès à l'information reste élevé. Les prix élevés des abonnements aux revues électroniques (ou licences) poussent les universités, les bibliothèques et les chercheurs à s'organiser pour résister ou même à contre attaquer. La résistance s'est principalement exprimée par la création de groupes (consortium) de négociation avec les éditeurs. Le but de ces consortiums est de permettre aux bibliothèques d'obtenir des licences d'utilisation des ressources en ligne à des prix raisonnables. La contre - offensive quant à elle s'exprime de différentes façons. L'une des stratégies de cette contre offensive a pour but de redonner aux universités et aux structures de recherche un « rôle éditorial ».

L'expression de ce 'rôle éditorial' se retrouve dans des initiatives comme la mise en place des archives électroniques (preprints, eprints) ou la création de

Droits d'auteur réservés.

revues électroniques à but non commercial. Le but principal de ces initiatives est de mettre en place des collections de documents scientifiques accessibles gratuitement ou à des prix plus modestes. Le point commun de ces initiatives est l'affirmation plus forte du rôle du chercheur comme acteur principal de la nouvelle communication scientifique et du rôle central des institutions universitaires, en particulier les bibliothèques, comme lieu d'échange, de production et de diffusion de la littérature scientifique. On peut citer, entre autres, la PLOS, la BOAI (Initiative de Budapest pour l'Accès Ouvert).

Les thèses électroniques doivent être vues dans la perspective de cette tentative des universités de redevenir éditeurs. Car en définitive les thèses et les articles se trouvent dans une situation similaire de circulation limitée : les thèses à cause des barrières géographiques, les articles à cause des barrières économiques. A cause du nombre limité des exemplaires, les thèses ne sont accessibles que dans quelques endroits, de ce fait elles sont difficilement disponibles à tous ceux qui pourraient en avoir besoin. Quant aux revues, les prix exorbitants des abonnements empêchent les centres d'information d'acquérir tous les documents nécessaires à leurs usagers.

Concernant les thèses électroniques, plusieurs initiatives sont entreprises à travers le monde. Bien que les outils et les pratiques varient selon les pays et les institutions, ces projets s'appuient sur les constats suivants : la majorité des documents est produite à l'aide des outils informatiques ; l'Internet est devenu le premier outil de recherche de l'information scientifique ; la diffusion en ligne permet donc de donner aux thèses une plus large audience et constitue un moyen de divulguer les résultats de la recherche.

2.2. L'Afrique et l'accès à l'information scientifique

L'information est devenue une denrée essentielle qui coûte cher. Or le continent africain est constitué, majoritairement, de pays pauvres. Vu de l'Afrique une thèse ou un article reste à peu près la même chose : un document à diffusion limitée, car les scientifiques africains ont autant de mal

Droits d'auteur réservés.

à accéder aux uns et aux autres. La diffusion et l'accès libre aux documents scientifiques seraient donc un moyen de combler le « retard informationnel » de l'Afrique.

L'informatique et l'Internet représentent un moyen efficace pour l'Afrique de rattraper son retard, pensent certains, sans tenir compte du phénomène de la « fracture numérique ». Celle ci peut être caractérisée par plusieurs aspects, parmi lesquels figurent le sous équipement et la « sous information ». L'Afrique continuant à être consommatrice de la technologie importée, le coût de ces technologies limite la généralisation de son utilisation. La « sous information » est due en partie à des causes économiques, les bibliothèques disposent rarement de budget qui permette d'acheter la documentation scientifique. Certaines bibliothèques continuent de se contenter de la « friperie » (les documents provenant des désherbages des bibliothèques et des dons des pays du Nord).

Cette tendance pourra-t-elle être inversée ? L'Afrique peut-elle être génératrice de contenu ? Probablement, puisque le continent dispose de centres de recherche et de personnels scientifique. Les résultats des travaux de ceux-ci sont publiés et répertoriés. Mais ils sont, le plus souvent, publiés dans les revues du Nord et répertoriés dans les bases de données du Nord. Or, ces sources d'information sont difficilement accessibles aux chercheurs du sud à cause de leurs prix. Les quelques revues scientifiques africaines qui existent sont peu distribuées et ne font pas partie des « must ».

L'instauration de la diffusion électronique des thèses en Afrique doit être vue dans la perspective plus large de rendre visible les résultats de la recherche africaine. S'il est prouvé que les thèses africaines peuvent être mises en ligne, alors la possibilité de diffuser des revues, des monographies, des actes de congrès, des comptes rendus et des rapports de recherche peut être envisagée.

Droits d'auteur réservés.

2.3. Contexte du stage

2.3.1. Le contexte institutionnel

2.3.1.1. Cyberthèses

Le stage s'est déroulé dans le service ERAD de l'université Lumière Lyon 2. Ce service abrite entre autres le programme Cyberthèses. L'originalité de l'approche de Cyberthèses réside dans :

- l'utilisation des normes SGML et XML et de la DTD TEI Lite ;
- le respect des habitudes des auteurs : ceux-ci continuent à travailler avec les mêmes logiciels de traitement de texte ;
- le souci de développer une plateforme logicielle libre et open source ;
- fournir la documentation nécessaire à l'utilisation et créer un réseau collaboratif où les compétences seront partagées.

Origine et historique de Cyberthèses

Partant du constat de la faible diffusion électronique des thèses, les universités de Montréal et de Lyon 2 ont engagé en 1998 une collaboration dont l'objectif est de s'appuyer sur la norme SGML pour diffuser en ligne des thèses. Soutenu par le Fonds Francophone des Inforoutes, le programme Cyberthèses a développé une plateforme logicielle qui repose, en partie, sur les programmes développés par la société Omnimark et qui permet de transformer un document RTF stylé en un document structuré en SGML conforme à la DTD TEI Lite. Le SGML sert de format pivot pour l'archivage et permet de produire d'autres types de documents plus appropriés à la diffusion sur Internet : html et pdf.

Réalisations :

Le projet Cyberthèses a obtenu les résultats suivants :

1. Les modèles de documents pour les logiciels de traitement de texte : La production d'un document structuré à partir de fichiers issus de logiciels de traitement de texte nécessite que

Droits d'auteur réservés.

ces fichiers aient fait l'objet d'une certaine structuration. Un modèle de document a été développé pour les logiciels de traitement de texte Word de Microsoft et Star Office de Sun. Le modèle de document permet de faire un découpage hiérarchique de la thèse et de décrire les blocs de texte qu'elle contient. Globalement, une thèse est composée d'une page de titre, de préliminaires, du corps de la thèse et des post-liminaires. Chaque partie est constituée d'un ensemble d'éléments.

2. La formation des doctorants : régulièrement, des sessions de formation des thèsards à l'utilisation des styles dans le logiciel de traitement de texte sont organisées. L'objectif de cette formation n'est pas d'imposer le modèle de Lyon 2 mais surtout de montrer les avantages que peut apporter l'utilisation d'un modèle de document : la création et la mise à jour automatique des tables de matières et des index. Ensuite les étudiants apprennent à installer et à utiliser le modèle de Lyon 2 sur leur poste de travail.
3. Le didacticiel : En collaboration avec l'université de Genève, un didacticiel a été développé à l'intention des doctorants. Ce didacticiel, disponible en ligne et sous forme de CD-ROM, explique les avantages du respect des normes, démontre l'intérêt du document structuré et explique comment installer et utiliser la feuille de style développée par Lyon2. Il aborde, en outre, les questions juridiques liées à la diffusion électronique des documents.
4. Le dépôt de copie électronique des thèses : Depuis 2000 le dépôt d'une copie électronique de la thèse est obligatoire à l'Université Lyon 2. Un mois environ avant la date de soutenance, l'étudiant dépose un exemplaire en papier et une copie électronique de sa thèse au service ERAD. Les conditions d'acceptation de la copie électronique sont : la

Droits d'auteur réservés.

lisibilité des fichiers et la présence de tous les documents faisant partie de la thèse en format électronique. Si la thèse est déjà structurée, par conséquent transformable rapidement en un fichier d'impression (ps, pdf), l'université prend en charge l'impression de 5 exemplaires de la thèse.

5. Le contrat de diffusion entre l'université et le doctorant : au cours du dépôt le doctorant signe un document qui autorise l'université à diffuser la thèse sur Internet ou Intranet.
6. La plateforme de diffusion : La plateforme logicielle Cyberthèses est basée sur le document structuré et la DTD TEI Lite. Sa première version était basée sur SGML, la nouvelle version est basée sur XML. Cette plateforme permet de convertir un document word pour produire un document conforme à la TEI Lite et à partir de celui-ci de produire des fichiers de diffusion en html et pdf.
7. La diffusion : actuellement, Cyberthèses a produit plus 300 thèses électroniques. Ces thèses sont accessibles sur : http://theses.univ-lyon2.fr/.

2.3.1.2. Faculté de Médecine de Bamako

Les connaissances acquises au cours de ce stage seront utilisées dans le futur pour diffuser les thèses de doctorat d'Etat en médecine et en pharmacie soutenues à la FMPOS de Bamako. Créée dans les années 70, la Faculté de Médecine de Pharmacie et d'Odonto-Stomatologie de l'Université de Bamako (ancienne Ecole Nationale de Médecine et de Pharmacie) est la seule école au Mali à former des docteurs en médecine et en pharmacie. En dehors de la formation initiale, la faculté effectue des formations continues et post doctorales : les cours supérieurs d'épidémiologie, les Certificats d'Etudes Spécialisées.

La faculté héberge des laboratoires de recherche en parasitologie et en hématologie. Elle travaille en étroite collaboration avec des institutions

Droits d'auteur réservés.

nationales de la santé : les laboratoires, les centres de recherche, les hôpitaux nationaux.

La FMPOS compte plus de 4500 étudiants. La formation dure 7 ans, au bout desquels chaque étudiant rédige et soutient une thèse pour obtenir le grade de docteur. Chaque année plus de 150 thèses sont soutenues en médecine et en pharmacie.

Les thèses soutenues à la FMPOS sont référencées dans deux bases de données bibliographiques : celle de l'Association pour l'Information et les Bibliothèques de Santé en Afrique et la base de données des thèses de la FMPOS : http://www.keneya.net/fmpos/bd/fmpos.htm.

Actuellement, ces thèses ne sont consultables qu'à la bibliothèque de la faculté. Les meilleures thèses sont recommandées à l'échange avec d'autres facultés africaines de médecine et de pharmacie. Mais ce programme d'échange des meilleures thèses entre les facultés africaines de médecine ne semble plus être pratiqué.

Les thèses sont fréquemment consultées par les chercheurs, les médecins, les pharmaciens et les enseignants. Ceux qui viennent consulter ces thèses expriment souvent le désir d'accéder aux thèses soutenues dans les autres facultés africaines pour comparer les résultats. Dans l'état actuel des choses, l'accès aux simples références bibliographiques de ces thèses reste difficile.

La mise en ligne des thèses de la FMPOS Bamako pourrait les rendre accessibles aux professionnels maliens de la santé travaillant en dehors de la capitale ainsi qu'aux médecins, pharmaciens, chercheurs et enseignants des autres pays africains en premier lieu, mais aussi à toute la communauté scientifique à travers le monde.

La production documentaire de la FMPOS est essentiellement constituée de thèses et de mémoires de Certificats d'Etudes Spécialisées. Néanmoins, les enseignants de la Faculté publient des articles scientifiques. Ces articles sont la plupart du temps publiés dans les revues du Nord, auxquelles la bibliothèque de la FMPOS n'est pas abonnée. Par conséquent, ils sont rarement accessibles à la bibliothèque. Certains articles paraissent dans une

Droits d'auteur réservés.

revue malienne « Mali Médical » qui sort de manière irrégulière et reste peu distribuée.

La bibliothèque de la faculté est une des meilleures bibliothèques universitaires de Bamako. Elle possède un fonds documentaire de plus de 10000 ouvrages, 3000 thèses. De même, elle donne accès à Internet et aux bases de données Pascal, Medline, African Health Anthology. La bibliothèque dispose également de quelques revues. La source principale des acquisitions reste les dons. Les revues proviennent essentiellement de la Conférence Internationale des Doyens des Faculté de Médecine d'Expression française (CIDMEF) par conséquent elles sont rarement à jour. L'abonnement aux bases de données est un don de la Dreyfus Health Foundation.

La FMPOS ne dispose pas actuellement de site web. Si la FMPOS souhaite mettre ses thèses en ligne, il lui faudrait acquérir l'équipement nécessaire, ou le louer à des fournisseurs de services, ou encore chercher des partenaires dont elle pourra utiliser les équipements.

2.3.1.3. Projet de télémédecine Keneya Blown

Origine et historique

« Keneya Blown = Vestibule de la santé » (KB) était un projet pilote de télémédecine initié par un étudiant de la FMPOS dans le cadre de sa thèse. Son objectif principal était de démontrer la faisabilité de la télémédecine à faible coût dans un pays en voie de développement : le Mali. Initialement financé par le Canton de Genève et soutenu par les Hôpitaux Universitaires de Genève, le projet s'est installé à Bamako et à Ségou. Dès le début, un accent particulier fut mis sur la maîtrise des NTIC par les professionnels de la santé, l'accès à l'information médicale, la promotion de l'échange entre les professionnels à l'intérieur du Mali et avec ceux d'autres pays.

Droits d'auteur réservés.

Originalité du projet

L'originalité du projet réside dans deux aspects : il s'agit d'une initiative du Sud et elle n'était pas basée sur une infrastructure sophistiquée. En effet, jusqu'à ce jour les projets pilotes de télémédecine en Afrique étaient l'initiative des organismes internationaux tels que l'Union Internationale des Télécommunications et ils utilisaient pour la plupart du temps des technologies spéciales (des liaisons par satellite ou des lignes spéciales dédiées). Ces projets étaient donc difficilement reproductibles dans d'autres pays. KB quant à lui s'est contenté d'utiliser une liaison à faible bande passante et les moyens du bord pour ses activités (appareils photo numériques, webcams, rétroprojecteurs).

Activités et résultats

Le résultat le plus important de KB a été faire comprendre que la télémédecine est possible au Mali et qu'il n'est pas besoin d'équipement sophistiqué pour la pratiquer. Le projet est parvenu aux résultats suivants :

- La création d'un site web http://www.keneya.net;
- La formation des professionnels de la santé à l'utilisation d'Internet : le courrier électronique, la recherche d'information sur Internet, la création de sites web ;
- L'instauration des séances de télé-enseignement entre des hôpitaux de Genève et du Mali. Ces conférences sont diffusées sur Internet depuis Bamako ou Genève ;
- La réalisation des consultations médicales à distance, l'échange d'images radiologiques entre des hôpitaux du Mali, de Suisse et de France à des fins de diagnostic ou de second avis médical;
- La mise à la disposition des médecins d'une adresse de courrier électronique;
- L'organisation d'une table ronde sur l'utilisation des NTIC dans le secteur de la santé au Mali;

Droits d'auteur réservés.

- La mise à la disposition gratuite des équipements du projet au service des institutions et des professionnels de la santé. Ainsi ses serveurs hébergent les sites de certains congrès et conférences médicaux qui se sont déroulés au Mali, les sites web de certains établissements comme celui du Centre National d'Appui à la Lutte contre la Maladie[1].

Les thèses de la FMPOS pourraient ainsi être hébergées sur les serveurs de Keneya Blown. A l'issue des discussions, KB accepte que la plateforme Cyberthèses soit installée sur ses machines. La faculté ne disposant pas actuellement de serveurs et de site Internet, l'utilisation des appareils de KB permet de résoudre les problèmes de matériel qui auraient pu se poser tout en permettant de faire des économies sur le coût de matériel.

2.3.2. L'environnement logiciel : description de la plateforme

Dés le début, Cyberthèses a mis l'accent sur l'interopérabilité et la portabilité. La prise en compte de ces pré requis a déterminé les choix politiques : se baser sur des normes existantes, développer des logiciels libres et indépendants, autant que faire se peut, des systèmes d'exploitation. Les logiciels devraient être utilisables par les différents partenaires quel que soit leur niveau d'équipement et l'environnement informatique dont ils disposent.

La plateforme logicielle et les outils sont en train de changer, mais la logique et les choix politiques restent les mêmes : mettre en place une plateforme qui garantit l'homogénéité, la pérennité et l'interopérabilité. Le passage de SGML à XML est dicté par plusieurs facteurs. La généralisation de SGML était limitée par l'absence d'outils logiciels (visualiseurs) disponibles à l'ensemble de la communauté sans restriction commerciale. Les programmes informatiques de production des documents en SGML réalisés par la société Omnimark Technologies remettaient en cause la volonté de diffusion ouverte

[1] http://www.keneya.net/cnam/

Droits d'auteur réservés.

et libre des concepteurs du projet, volonté qui est un des fondements du programme financé par les Institutions Francophones. L'évolution de la plateforme vers XML et l'Open Source permet de dépasser ces limites.

2.3.2.1. Les normes et les standards

La plateforme logicielle de Cyberthèses est bâtie autour de normes internationales. L'utilisation des normes assure une certaine garantie de pérennité des documents et une indépendance vis à vis des produits logiciels commerciaux et des systèmes d'exploitation. Parmi les normes les plus significatives citons XML, l'Unicode et les formats des métadonnées :

XML[2] : est un langage de balisage qui permet de décrire la structure logique des documents. C'est une représentation textuelle de données structurées selon une syntaxe normalisée. Développé sous l'égide du World Wide Web Consortium (W3C) depuis fin 1996, XML a été conçu pour permettre d'échanger et de stocker des données indépendamment des programmes ou des processus qui les produisent ou qui les utilisent.

XML permet de représenter différents types de documents : des données documentaires, base de données, une feuille de calcul, l'ensemble des paramètres de configuration d'une application informatique, les flux de données échangés lors de transactions financières, etc. XML a ainsi vocation à devenir le format d'échange universel pour les flux de données structurées qui transiteront sur le World Wide Web.

Plusieurs autres normes, standards et recommandations sont liés à XML et permettent d'exploiter, de transformer les documents XML pour en produire d'autres.

Unicode : C'est un standard définissant des jeux de caractères. Il en définit actuellement plus de 90000. Il existe plusieurs encodages Unicode : utf-8, utf-16, ISO-8859-1. Les thèses en français sont encodées principalement en

[2] http://www.w3.org/TR/REC-xml

Droits d'auteur réservés.

UTF-8. Celui-ci encode les caractères ascii sur un octet et les autres sur 2 à 4 octets.

L'utilisation de l'Unicode apporte la réponse aux questions liées aux multilinguismes : un même document peut contenir des blocs de texte de différentes langues et une même application XML peut contenir des documents écrits dans différentes langues et utilisant différents alphabets : latin, cyrillique, arabe.

XSL : est une recommandation du W3C[3] qui suscite autant d'engouement que le XML. XSLT est un langage de transformation des documents XML pour produire d'autres documents au format XML et en d'autres formats (html par exemple). XSL-FO, quant à lui, permet de décrire la disposition et la mise en page de différents éléments dans les documents dérivés. La principale utilisation actuelle de XSL-FO est la transformation des documents XML en PDF.

Les métadonnées : ont pour rôle de décrire un document. Les métadonnées peuvent être incluses dans le document lui même ou dans un document séparé. Actuellement, à chaque thèse doit être associé un fichier de métadonnées. Cyberthèses utilise actuellement 3 standards de métadonnées identifiés dans le fichier de métadonnées par des espaces de noms. Il s'agit du Dublin Core qualifié et non qualifié, de la norme OAI et des métadonnées propres à Cyberthèses.

La TEI Lite : Ce n'est pas une norme mais une DTD. C'est la version simplifiée de la TEI (Text Encoding Initiative)[4], beaucoup plus complète et complexe. La TEI compte plus de 400 éléments. Initialement développée en SGML, la TEI avait pour objectif de permettre la constitution de corpus de textes électroniques en sciences humaines. Ces documents électroniques doivent être neutres vis à vis des équipements, des applications et des systèmes informatiques. La TEI Lite regroupe un ensemble d'éléments et d'attributs de la TEI les plus utilisés.

---

[3] http://www.w3.org/TR/xslt
[4] http://www.tei-c.org

Droits d'auteur réservés.

Conçue de façon modulaire, la DTD TEI permet de « créer » des sous catégories de documents spécifiques ayant des structures variées. Un document TEI comporte deux parties principales : l'entête et le corps du texte. L'ossature d'un document TEI ressemble à celle-ci[5] :

```
<TEI>
    <teiHeader> [informations contenues dans l'en-tête TEI] </teiHeader>,
    <text>
        <front> [ textes préliminaires...] </front>,
        <body> [ corps du texte... ] </body>
        <back> [annexes... ] </back>
    </text>
</TEI>
```

2.3.2.2. Les logiciels et composantes utilisés par la chaîne Cyberthèses

Cyberthèses utilise actuellement la plateforme baptisée Cyberdocs. Cette plateforme respecte la logique initiale de Cyberthèses à savoir : transformer des fichiers issus d'un logiciel de traitement de texte pour produire des documents consultables sur Internet. Cyberdocs est un ensemble de logiciels coordonnés par des pilotes: OpenOffice.org, une machine virtuelle Java, des parseurs XML, des processeurs XSL, des processeurs XSL-FO, un moteur de servlets (Tomcat), une servlet pour l'indexation et la recherche (SDX). Tous ces logiciels sont accessibles gratuitement sur Internet.

Microsoft Word : c'est un logiciel de traitement faisant partie de la suite bureautique Microsoft office. A vrai dire, Word n'est pas utilisé par la plateforme. Il est utilisé par la majorité des doctorants pour la création de documents, et par les services chargés des thèses électroniques pour le stylage des documents avant de commencer la conversion. Ce qui justifie

[5] André Jacques, Dupoirier Gérard (Coord). Documents numériques : [cédérom]. 3 ed. Paris : Techniques de l'ingénieur, mai 2003.

Droits d'auteur réservés.

l'utilisation de Word, c'est sa popularité parmi la majorité des étudiants et la facilité d'utilisation des feuilles de styles sous forme de menu déroulant.

OpenOffice.org : OpenOffice.org est une suite bureautique libre et gratuite fonctionnant sous Unix, Linux, Windows. Elle est composée d'un traitement de texte, d'un tableur, d'un logiciel de présentation de diapositives, etc. Les fichiers OpenOffice.org sont sauvegardés en XML. Actuellement il est utilisé par le projet Cyberthèses uniquement pour obtenir le fichier XML générique qui sera transformé pour obtenir le fichier TEI Lite. A la longue OpenOffice.org pourrait remplacer Word pour le stylage et même pour la création des thèses.

La machine virtuelle Java : C'est un environnement indispensable pour l'exécution des programmes écrits en Java. Il en existe pour différents systèmes d'exploitation : Windows, Linux, Unix, MacOs, etc., rendant ainsi les programmes écrits en Java indépendants des plateformes. Beaucoup de composantes de Cyberdocs sont des exécutables écrits en Java.

Les parseurs XML : Ce sont des analyseurs syntaxiques qui vérifient la validité des fichiers xml et leur conformité par rapport à une DTD.

Processeurs XSL : Ce sont des applications dont le rôle est de transformer les document XML. Ils appliquent une feuille de style XSL à un document XML pour produire un autre document XML ou dans un autre format. Le processeur utilisé pour la conversion des thèses est *saxon*.

Processeurs fop : Leur rôle essentiel dans la plateforme est de produire des documents en pdf.

Tomcat : C'est un moteur de servlets. Il est utilisé dans Cyberdocs pour la consultation des thèses en ligne.

SDX : C'est une application permettant de développer et de diffuser en ligne des bases de données documentaires contenant des documents XML. Il en est à sa version 2.

Droits d'auteur réservés.

SDX permet d'indexer, de rechercher et d'afficher de documents XML. C'est une servlet Java qui est une extension de Cocoon, utilisant le moteur de recherche Lucene et un processeur XSLT.

SDX contient un ensemble d'objets hiérarchisés sous forme de serveur SDX, d'applications, d'entrepôts. Un serveur est l'ensemble constitué d'un serveur http, d'un moteur de servlets et des applications SDX fonctionnant ensemble. Un serveur SDX contient des applications SDX. Une application SDX est constituée d'une ou plusieurs bases documentaires, chaque base contenant à son tour des dépôts.

Une application est un ensemble de bases documentaires et de groupes d'utilisateurs. Une base est un ensemble de documents partageant des caractéristiques communes et qui peuvent être recherchés. Parmi ces caractéristiques, on peut compter la liste de champs et leurs contenus produits au cours de l'indexation.

Un dépôt représente l'endroit où sont stockés les documents indexés ou leurs adresses. En effet SDX permet d'indexer des documents qui ne sont pas stockés sur le serveur local (par exemple des sites web hébergés ailleurs).

Le processus d'indexation consiste à analyser un document XML, d'en extraire des portions qui servent à remplir le contenu des champs de la base. La recherche s'effectue sur le contenu des champs. La recherche dans SDX peut être effectuée dans une seule base ou dans plusieurs bases appartenant à une même application et à des applications différentes.

2.3.2.3. Description de la plateforme

La diffusion électronique de thèses grâce à la plateforme Cyberdocs peut être décomposée en plusieurs étapes : la construction d'un document word stylé, la transformation du document en un fichier XML conforme à la DTD TEI Lite, la transformation du document TEI Lite en documents diffusables sur Internet, la mise en ligne des thèses et l'archivage. Le fichier word stylé

Droits d'auteur réservés.

sert d'entrée à la chaîne de transformation qui devra produire le fichier pivot à partir duquel seront produits les fichiers de diffusion et d'archivage.

2.3.2.4. Chaîne de traitement

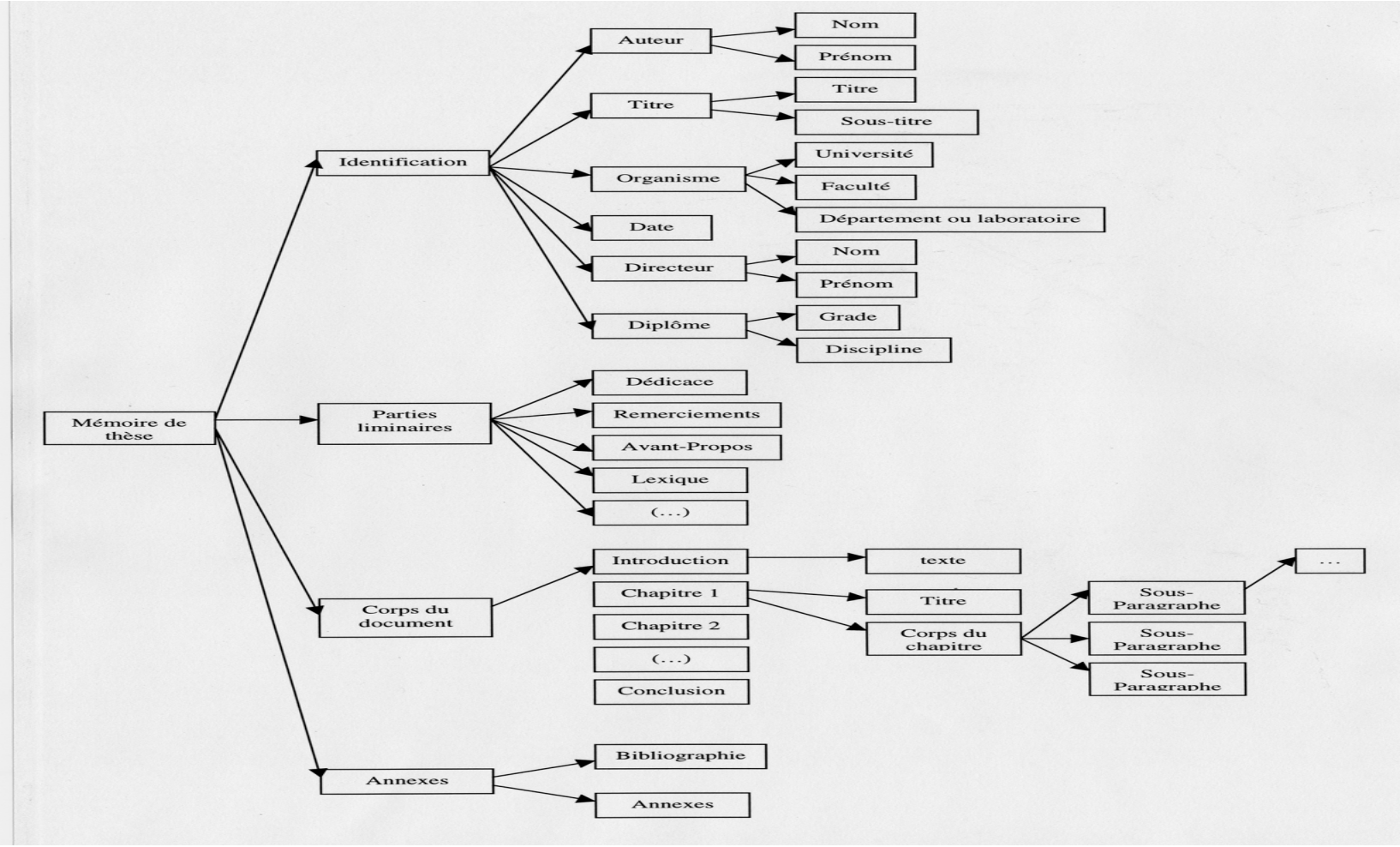
1- Stylage dans le traitement de texte

Le stylage du fichier word est effectué par un être humain. Il a pour but de produire un document structuré. Le modèle de document utilisé est celui développé par Lyon 2. Bien que les doctorants suivent une formation à l'utilisation des feuilles de styles, les fichiers qu'ils déposent sont rarement utilisables tels quels. Des personnes sont chargées au sein d'ERAD de styler ces fichiers. Le stylage d'une thèse prend du temps, variable en fonction de la complexité des documents ou de l'utilisation d'une feuille de style quelconque par le doctorant.

Le modèle de document de Lyon 2 contient une cinquantaine de styles de plusieurs niveaux. L'utilisation des styles est largement facilitée par une barre de menus déroulants. Ces styles permettent de marquer les différentes parties et informations d'un document. Ces parties sont : la page de titre, les préliminaires, les titres des chapitres, les paragraphes, les légendes des illustrations et des tableaux, les illustrations, les citations, les annexes, la bibliographie. Les différentes informations contenues dans une thèse sont hiérarchisées, comme le montre la figure 1.

Figure 1 : hiérarchie des parties d'une thèse

Droits d'auteur réservés.



Droits d'auteur réservés.

La page de titre contient le titre, l'université, la faculté, l'auteur, la date de la soutenance, le directeur de la thèse, les membres du jury. Le corps de la thèse est divisé en parties, chapitres, sous-chapitres. Ceux-ci contiennent des paragraphes, des illustrations, des citations. Les post-liminaires contiennent les annexes, les bibliographies.

Le développement de la nouvelle plateforme allège le stylage d'un document. Il n'est plus nécessaire de détacher chaque image pour la sauvegarder dans un fichier séparé. En plus, il n'est plus besoin de sauvegarder les fichiers word en RTF, la conversion étant possible à partir des fichiers word.

2- Cyberdocs

Cyberdocs est la nouvelle plateforme logicielle qui sera utilisée pour la conversion et la diffusion des thèses. Elle peut être divisée en deux principaux modules : celui de la conversion et celui de la consultation. Le module de consultation est une application SDX permettant d'indexer, de rechercher et d'afficher les thèses. Un troisième module de pilotage et d'administration de la plateforme est disponible.

Le module de conversion

Les fichiers de départ sont des fichiers word, les fichiers des images, et éventuellement le fichier des métadonnées. La conversion d'une thèse s'effectue en plusieurs étapes, les principales étant la création d'un fichier XML conforme à la DTD TEI Lite, la conversion du fichier TEI Lite en html et pdf. La durée de la conversion d'une thèse dépend de la taille de celle-ci et des performances de la machine.

Déroulement de la conversion

Un script shell ou batch est exécuté. La conversion exécute OpenOffice.org qui ouvre le fichier source et le sauvegarde sous son propre format. La décompression du fichier OO produit un fichier en XML (content.xml). C'est à partir de ce fichier que s'effectueront toutes les transformations ultérieures. Les différentes transformations intermédiaires ont pour but d'épurer le fichier

Droits d'auteur réservés.

XML issu d'OO. Le dernier fichier intermédiaire est transformé en un fichier conforme à la DTD TEI Lite. A partir de ce fichier seront créés les fichiers pour la diffusion et l'impression en html et en pdf.

Les différents fichiers d'une thèses sont organisés en sous répertoires, un par étape et type de fichiers : ***sources, oo, prod, xml, html, pdf***.

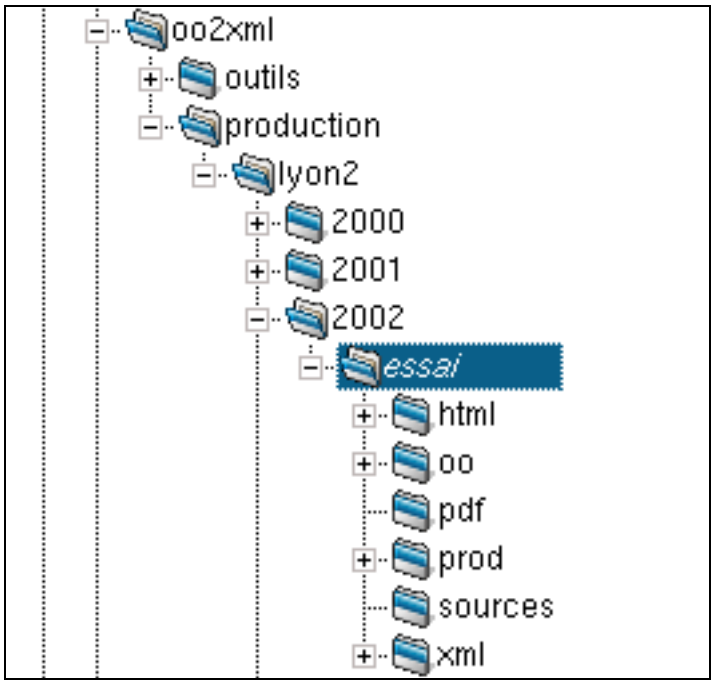


Figure 1 : arborescence des fichiers du document essai

Le répertoire ***sources*** contient le fichier word, le fichier des métadonnées, les images ; ***oo*** contient les fichiers OpenOffice.org (l'original et les fichiers issus de la décompression de celui-ci), ***prod*** contient les fichiers XML intermédiaires, ***xml*** contient le fichier TEI, le fichier des métadonnées, les fichiers des notes, le répertoire des images, ***pdf*** et ***html*** contiennent les fichiers de diffusion au format pdf et html.

Au cours de la conversion, le système affiche des messages informatifs, des avertissements, ou des erreurs. Le message à titre informatif renseigne sur l'étape et les opérations de la conversion. Un avertissement indique qu'une

Droits d'auteur réservés.

anomalie a été rencontrée. Les avertissements peuvent être provoqués par la présence de styles non déclarés dans le fichier de styles, l'absence de certains éléments obligatoires (le titre ou l'université par exemple), ou l'absence d'une ressource que devrait contenir le document (une image par exemple).

Le module de consultation

Celui-ci peut à son tour se subdiviser en 3 trois parties : l'indexation, la recherche et la consultation des documents entiers.

L'indexation

Sont indexés le fichier TEI en XML et celui des métadonnées. L'application contient deux bases : la base des utilisateurs et la base de documents. La base des documents est constituée d'un seul dépôt. Elle contient 46 champs dont 40 sont indexés. Certains champs sont utilisés à des fins de gestion, d'autres à des fins d'affichage, les autres sont destinés à la recherche, à la construction des index et au tri des résultats. Les éléments du fichier XML qui sont indexés sont : l'auteur, le directeur de recherche, les membres du jury, l'université, la faculté, le département, l'école doctorale, l'année, la ville de soutenance, les titres, l'identifiant, les titres des chapitres, les tableaux, les illustrations, le résumé, le « texte intégral » de la thèse, les références bibliographiques.

Lors de l'indexation, chaque thèse est découpée en plusieurs unités documentaires (« enregistrement SDX ») : une pour la totalité de la thèse et une par chapitre. Une thèse produira un sous document « principal » qui contient la page des titres et des informations sur les chapitres de la thèse et un sous document par chapitre. Chaque partie a un identifiant, et contient également l'identifiant de la thèse dont elle est issue.

La recherche

Il existe deux interfaces de recherche : la recherche simple et la recherche avancée. La recherche peut être effectuée par champs ou en texte intégral. Tous les champs indexés peuvent être interrogeables.

Droits d'auteur réservés.

La recherche simple : Par défaut la recherche simple s'effectue sur le texte intégral. L'opérateur par défaut est le « OU » booléen. Cependant l'utilisateur peut utiliser les opérateurs booléens (AND, OR, NOT) pour combiner les termes. Il est également possible de définir les champs dans lesquels s'effectuera la recherche en utilisant une syntaxe spéciale, dont l'apprentissage par les futurs visiteurs des sites Cyberthèses semble peu probable, par : *+auteur :|Dupont, Jean|* permet de retrouver le document dont l'auteur est Dupont, Jean.

La recherche par champs : La manière la plus facile d'effectuer une recherche par champs est d'utiliser l'interface de la recherche avancée. Dans cette interface, l'utilisateur a la possibilité de choisir les champs dans lesquels il souhaite effectuer la recherche. L'interface de recherche donne accès aux champs suivants : l'auteur, le titre, le directeur de recherche, les membres du jury, les contributeurs, le code de l'institution, la discipline, l'école doctorale, le résumé, les tableaux, les illustrations, les références bibliographiques, les titres des chapitres, les citations, le texte intégral. La recherche peut être effectuée dans plusieurs champs en choisissant les opérateurs pour combiner les termes.

Les résultats de la recherche : une recherche fournit non pas la liste des thèses qui correspondent à la requête, mais la liste des parties (chapitres, annexes, bibliographies) de la thèse qui contient les termes de la recherche. A partir de la page des résultats l'utilisateur peut accéder au chapitre de la thèse ou à la fiche descriptive de celle-ci. Cette fiche provient du fichier des métadonnées, par conséquent elle ne fournit que les informations contenues dans celui-ci.

La consultation des documents

La consultation du texte intégral des thèses est basée sur le découpage de la thèse en morceaux. Une thèse est, en effet, un document volumineux dont la consultation à l'écran peut être fastidieuse. L'accès aux différentes parties d'une thèse est effectué grâce à une table de matières dynamique. L'interface de consultation est découpée en quatre parties (frames html) : le cadre de

Droits d'auteur réservés.

titre, le cadre de la table des matières, le cadre du contenu, le cadres des notes. La table des matières est subdivisée en onglets : un pour les titres des chapitres, un pour les illustrations, un pour les tableaux. Lors de la consultation d'une thèse, les termes de la requête sont mis en surbrillance.

Le module de gestion

C'est une application informatique, écrit en PHP, permettant de piloter la plateforme à distance à partir d'un poste client grâce à une interface Web qui remplace le travail en ligne de commande. Toutes les fonctions nécessaires à la conversion et à la production, à l'archivage et ensuite à la diffusion sont accessibles de manière transparente, permettant la prise en main des opérations par des non informaticiens. Différents niveaux de prise en main sont disponibles : super administrateur, administrateur d'un domaine institutionnel, utilisateur, styleur. Chacun possède des droits différents en fonction de ses compétences. On peut créer des espaces par institution, ensuite par utilisateurs, qui peuvent gérer l'espace nécessaire à la conversion d'une thèse, créer les différents répertoires nécessaires à la phase de conversion, remplir le bordereau de metadonnées, l'éditer, entreprendre la conversion, en une seule fois ou par étape, vérifier les résultats obtenus, puis ensuite faire le lien avec la plate forme d'indexation et d'interrogation SDX. Ce module de gestion permet à plusieurs utilisateurs de travailler simultanément sur une plate forme unique et ensuite de transférer les documents sur des plate formes de diffusion différentes.

Les archives ouvertes

Avec Cyberdocs, Cyberthèses ne sera plus un serveur de thèses d'une institution isolée mais plutôt un serveur faisant partie d'un réseau de serveurs de thèses grâce au protocole Open Archives Initiative (OAI). L'idée est de mettre en place un système d'archives reparties à travers la planète qu'un utilisateur peut interroger simultanément.

Droits d'auteur réservés.

3. Problématique

Cyberdocs est une plateforme open source utilisant XML et les technologies associées à celui-ci. Il est conçu pour tourner sur différents systèmes d'exploitation. L'idée est d'en faire une plateforme générique permettant de publier des documents en XML quel que soit le type de document et la DTD utilisés.

Dans l'état actuel de son développement, la plateforme permet de traiter des thèses, stylées selon la feuille de styles word de Lyon 2, en les convertissant en TEI Lite XML. L'interface de consultation par défaut est celle de Lyon 2. Sont déjà pris en compte la possibilité de publier sur un même serveur les thèses de plusieurs universités. La problématique de la généralisation de la plateforme peut se formuler au travers des questions suivantes :

- Est-il possible de convertir et publier des documents structurés avec un modèle de document différent celui de Lyon 2 ?
- La plateforme permet-elle de publier des documents conformes à la DTD TEI sans pour autant être des thèses ?
- Peut-on publier avec la plateforme des documents conformes à une DTD différente de la TEI ?

La réponse à ces questions permettra de faire de Cyberdocs une plateforme utilisable par plusieurs établissements pour traiter des documents construits selon différents modèles.

Droits d'auteur réservés.

La production des thèses en ligne

Pour nous familiariser avec la plateforme, nous avons principalement travaillé sur des thèses soutenues à la Faculté de Médecine de Bamako. Ceci nous a permis d'appréhender certains problèmes qui pourraient être rencontrés lors de mise en marche de Cyberthèses à Bamako.

La production d'une thèse en ligne est constituée de plusieurs étapes : la création du document par son auteur, son traitement et son stylage par le service chargé des thèses électroniques, sa conversion, sa diffusion et sa consultation.

1. Stylage

Le stylage a été effectué en utilisant le modèle de document de Lyon2 avec logiciel de traitement de texte Word de Microsoft. L'ensemble des styles est accessible grâce à un menu déroulant accessible dans la barre des outils. Les styles sont regroupés dans ce menu par parties ou types de styles : page des titres, les pages liminaire, corps de la thèse, listes, citations, les illustrations, les annexes.

Barre de menu de modèle de Lyon 2

Le stylage des thèses a permis de relever les problèmes éventuels qui pourraient être rencontrés. Certains problèmes sont génériques, d'autres sont spécifiques aux thèses de bamako.

Des éléments non prévus dans la feuille de style : presque toutes les thèses de Bamako contiennent dans la page de titre: la dénomination du pays, le ministère auquel est rattachée l'université.

Droits d'auteur réservés.

Les éléments non prévues par la plateforme

Texte de l'élément	Description
République du Mali	Dénomination du pays
Un peuple Un but Une Foi	Devise du Mali
Ministère de l'Education Nationale	Ministère auquel est rattaché l'université

La prise en compte de ces éléments nécessite de créer de nouveaux styles ou d'utiliser des styles prévus pour d'autres éléments. Ces informations auraient pu être stylées comme un simple texte, cependant, la page de titre ne peut pas contenir d'éléments texte. L'ajout de nouveaux styles entraîne quelques modifications dans la chaîne de conversion.

La hiérarchisation des chapitres et des sections

Le problème principal qui se pose est la numérotation automatique. En effet les éléments ne sont pas numérotés de manière univoque par les auteurs : ils utilisent des chiffres romains, latins, des lettres majuscules ou minuscules. Dans une même thèse des sections de même niveau, situées dans deux chapitres peuvent être numérotées différemment par l'auteur. Par exemple, un sous-chapitre de niveau 2 peut être numéroté avec des lettres majuscules dans premier chapitre, et avec des chiffres latins dans un autre chapitre. La définition d'une règle de numérotation automatique affectera le texte de l'auteur. Or il faut rester conforme le plus possible au choix de l'auteur.

Les découpages trop fins

Il arrive que la thèse soit découpée par son auteur de manière trop fine, certaines sections ne comportant qu'un seul court paragraphe. Faut-il styler ces sections comme des titre de chapitres ou juste comme du texte mis en évidence. L'utilisation d'un style chapitre alourdirait la table des matières, alors qu'utiliser un style de texte normal pourrait compromettre la fidélité que nous devons garder par rapport aux intentions de l'auteur.

Droits d'auteur réservés.

Les tableaux

Les auteurs regroupent dans une cellule des informations sur plusieurs lignes séparées par des retours à la ligne, alors qu'une cellule de tableau ne doit pas contenir ce signe. Elle ne peut que contenir que du texte.

Exemple de tableau d'une thèse :

Sexe	***Age moyen***
Féminin	33,15
Masculin	45,05

Cette pratique oblige la reprise de plusieurs tableaux pour obtenir quelque chose qui ressemble à ceci. C'est un travail fastidieux.

Tableau corrigé

Sexe	***Age moyen***
Féminin	33,15
Masculin	45,05

En définitive le stylage n'est pas un travail mécanique, mais un travail intellectuel qui demande d'entrer dans la logique de l'auteur. Il nécessite la plupart du temps des choix et des prises de décisions. Les seuls indices dont dispose la personne chargée du stylage sont la table des matières, les aspects typographiques du texte ou les numéros.

Droits d'auteur réservés.

2. La conversion

2.1. Les problèmes rencontrés et leurs causes possibles, les solutions envisageables

Après le stylage, la conversion d'une thèse peut commencer. Avant de lancer la conversion, les fichiers à convertir doivent être organisés à l'intérieur du répertoire du module de conversion. Les fichiers de toutes les thèses sont regroupés dans le répertoire ***oo2xml/production***. Ce répertoire contient un sous répertoire par institution. Le répertoire de chaque institution contient un sous répertoire par année. Le répertoire d'une année contient un répertoire par thèse. Le fichier word à convertir doit être copié dans le répertoire ***sources*** de répertoire de la thèse. Ainsi la thèse de Jean Pierre Dupont, soutenue en 2OOO à l'Université de Lyon 2, aura l'arborescence suivante :

oo2xml/production/lyon2/2000/dupont_jp/sources/dupont_jp.doc

Les noms des répertoires ***lyon2*** et ***dupont_jp*** correspondent respectivement au code de l'université et au code de la thèse. A Lyon 2, le code de la thèse correspond généralement au nom de l'auteur suivis des initiales de ses prénoms.

Le répertoire ***sources*** d'une thèse peut contenir également les images organisées dans un sous répertoire ou un fichier compressé, un fichier xml de métadonnées (voir la figure 1).

Le programme qui lance la conversion prend 7 arguments : l'étape de la conversion à effectuer, le code de l'institution, le code de la thèse, le code de la feuille de styles, la langue de la thèse, et l'année de soutenance. En ligne de commande, le programme de conversion est appelé par un script shell, qui par commodité est nommé par le code de la thèse. Ce script est situé dans le répertoire de l'année de soutenance de la thèse. Le script permettant de convertir la thèse de JP Dupont s'appellera, par exemple, `dupont_jp.sh`.

Droits d'auteur réservés.

Il contient les lignes suivantes :

cd ../..

./up.sh tout dupont_jp.doc lyon2 essai lyon2 fr 2002.

Ce fichier appelle exécute le fichier up.sh qui est situé deux répertoires plus haut en lui passant la liste des arguments (paramètres) qu'il attend:

Liste des arguments

Arguments	**Exemples**
Étape de la conversion	tout
Nom du fichier à convertir	dupont_jp.doc
Code de l'université	lyon2
Code de la thèse	essai
Code de la feuille de styles	lyon2
Langue de la thèse	fr
Année de soutenance	2002

Dans cet exemple, toutes les conversions (xml intermédiaires, TEI, html, pdf) seront effectuées.

Le programme affiche des messages sur le déroulement de la conversion : étapes de conversion, pilotes utilisés pour chaque conversion, les fichiers créés, les erreurs rencontrées.

L'échec de la conversion peut provenir d'une mauvaise installation des composantes de la plateforme ou d'une mauvaise organisation des fichiers de la thèse.

Droits d'auteur réservés.

2.2. Les sources des erreurs

La conversion se poursuit généralement jusqu'au bout, même si des erreurs se sont produites. Les messages permettent d'identifier l'origine des problèmes. Les erreurs rencontrées provenaient généralement des styles ou des fichiers de la thèse.

2.2.1. Les erreurs liées aux styles

Tous les styles utilisés dans le fichier word doivent être déclarés dans le fichier ***oo2xml/outils/xslt/utiles/styles.xml***. Certains de ces styles sont considérés comme obligatoires. Les éléments obligatoires proviennent de la page de titre. Ce sont : l'université, l'école doctorale, la faculté, le grade, le directeur de la recherche, l'auteur, le titre, le sous-titre, le date de soutenance. Les problèmes liés au style sont de deux sortes : l'absence des styles considérés obligatoires, la présence dans la thèse des styles non déclarés.

L'absence d'un style obligatoire

Elle provoque l'affichage d'un avertissement indiquant les styles manquants. Ce message est ensuite inclu dans le fichier TEI sous forme d'un élément avertissement apparaissant après l'élément racine TEI.2 (Voir annexe 1).

La présence de styles non déclarés

Quand la thèse comporte des styles qui ne sont pas déclarés dans les fichiers ***styles.xml***, un message le signale au cours de la conversion. Un texte décrit avec un style non déclaré pourrait disparaître dans le document final. Le cas le plus souvent rencontré au cours de ce stage a été celui du style légende. Ce style existe dans le modèle de document pour Word de Lyon 2. Il sert à styler les légendes des tableaux et des images. Au cours de la conversion, il devient WW-Légende. Or le style WW-Légende n'était pas déclaré dans la plateforme. Il en résultait que le texte des légendes disparaissait dans le fichier TEI XML et par conséquence des fichiers dérivés de celui-ci.

Droits d'auteur réservés.

2.2.2. Les ressources manquantes

Ce sont des parties de la thèse qui se présentent généralement sous forme de fichiers placés dans le répertoire "sources". C'est le cas des images quand elles sont détachées du document word. Une image est sauvegardée dans un fichier, le nom de ce fichier remplace l'image dans le fichier image et stylé avec le style figure. A certain moment de la conversion, des liens sont créés vers les fichiers des images. Si le nom de l'image dans le fichier XML ne correspond pas à un nom de fichier, un message avertit que ce fichier n'existe pas.

3. L'indexation

L'indexation a pour but d'incorporer une thèse dans la base documentaire SDX de la plateforme. Elle s'effectue dans une interface web. Une thèse ne peut être ajoutée que par un administrateur de la base. Les conditions préalables à l'indexation d'une thèse sont : la présence de tous les fichiers de la thèse nécessaires à l'indexation, le respect de l'arborescence des fichiers de la plateforme.

L'arborescence est la même que celle du répertoire de production du module de conversion à savoir code de l'institution, année de soutenance, code de la thèse. Le répertoire concerné par l'indexation est le sous-répertoire ***xml*** de la thèse. Il doit contenir le fichier TEI de la thèse, le fichier des métadonnées, le fichier des notes, le répertoire contenant les formules, le fichier de la table des matières, le répertoire ressources contenant les images.

Droits d'auteur réservés.

Contenu du répertoire xml de la thèse de dupont_jp

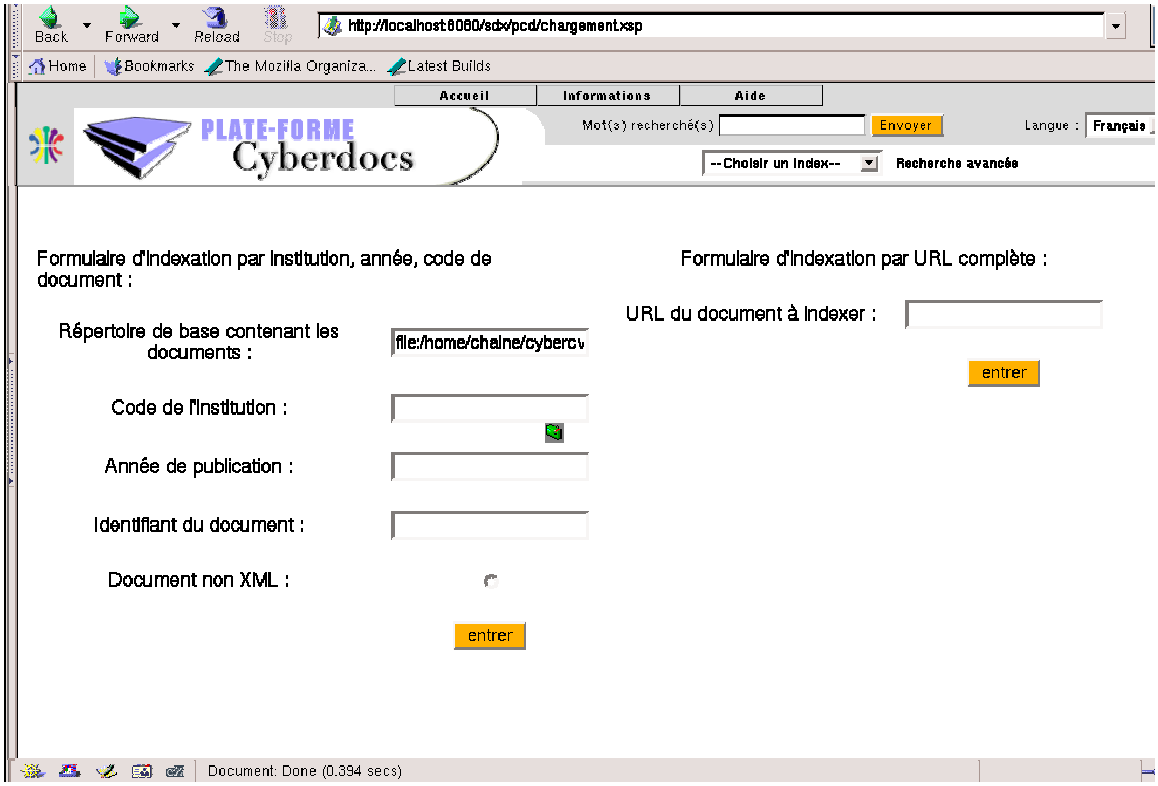
Fichier ou répertoire	*Description*
dupont_jp.xml	Fichier TEI obtenu après la conversion du fichier word
dupont_jp-md.xml	Fichier des métadonnées conformes aux normes utilisées par Cyberthèses
notes.xml	Les notes de bas de page de la thèse
tocTab.js	Fichier javascript contenant la table des matières
teixlite.dtd	Fichier contenant la DTD TEI Lite
Formules	Répertoire contenant les formules codées conformément à la MathML
Ressources	Répertoire contenant les images

Les fichiers indispensables sont : le fichier TEI, la table des matières, les images auxquelles il est fait référence dans la thèse. L'indexation ne se fera pas correctement si un de ces fichiers manque.

L'indexation d'une thèse s'effectue grâce à un formulaire qui permet de saisir le nom et le chemin du répertoire contenant les thèses, le code de l'institution, l'année de soutenance, le code de la thèse.

Figure 2 : Formulaire d'indexation

Droits d'auteur réservés.



Les valeurs des champs correspondent aux noms de répertoires du dossier contenant les documents. Elles permettront de reconstruire l'arborescence des fichiers et de localiser les fichiers de la thèse.

Après la validation du formulaire, le résultat de l'indexation est indiqué par un message. Si le chemin indiqué est correct et si tous les fichiers sont présents, l'indexation s'effectue généralement sans incident. Le résultat de l'indexation est indiqué par un message adéquat.

3.1. Les erreurs

Les problèmes peuvent provenir d'un mauvais chemin, de fichiers manquants, ou des problèmes de codage des fichiers xml.

Droits d'auteur réservés.

3.1.1. La reconstruction de l'arborescence

La reconstruction de l'arborescence des fichiers peut échouer si les informations saisies dans le formulaire sont incorrectes (ou mal orthographiées) : répertoire contenant les documents, le code de l'institution, code de la thèse, année. Dans ce cas un message d'erreur est affiché (voir annexe 2-1).

Pour résoudre ce problème, il suffit de revenir pour saisir les bonnes informations

3.1.2. Fichiers manquants

L'absence de certains fichiers dans le repertoire xml de la thèse peut empêcher l'indexation de se dérouler. Les fichiers dont l'absence peut causer cet incident sont : la table des matières, les images, etc. Le nom de la ressource manquante peut être dans les fichiers ***logs*** de SDX (voir annexe 2)

Remarque : Le fichier de métadonnées joue un rôle important dans l'indexation d'une thèse car il fournit le contenu de plusieurs champs SDX de la base documentaire. Cependant son absence n'empêche pas une thèse d'être indexée. On ne se rend compte de son absence que lors de l'affichage de la fiche de la thèse. Vu son importance, il serait préférable qu'il soit réclamé lors de l'indexation.

Les erreurs liées au codage du document : Les documents XML sont encodés en unicode. Le module de conversion produit des fichiers encodés en UTF-8. La modification d'un fichier (le fichier TEI par exemple) avec un éditeur de texte qui ne supporte pas Unicode rend ce dernier impossible à indexer.

Un encodage incorrect provoque l'affichage d'un message indiquant que le fichier n'a pas pu être parsé (voir annexe 2-3).

Même si le fichier des métadonnées existe, il ne sera pas indexé s'il n'est pas bien encodé. Cependant aucun message ne l'indiquera pendant la phase d'indexation.

Droits d'auteur réservés.

4. La mise en ligne

La mise en ligne d'une thèse est effectuée dès qu'elle indexée. Elle peut être retrouvée et affichée. L'emplacement de la page consultation est indiqué dans l'élément <dc:identifier> du fichier des métadonnées. Le contenu de ce champs servira à alimenter le contenu du champs SDX url, qui sera utilisé pour la création d'un lien vers cette page. Cependant, le fait d'indexer une thèse ne recopie pas automatiquement les fichiers à cet emplacement.

Dans l'état actuel des choses, la page de consultation d'une thèse est identique au répertoire contenant les documents à indexer. Le chemin de ce répertoire est stocké dans le champ *documentUrl*.

5. L'archivage

La plateforme Cyberdocs n'offre pas des fonctionnalités spécifiques d'archivage des documents. Cependant, à l'Université Lyon 2, il existe des procédures d'archivage des copies électroniques des thèses. Les documents électroniques archivés sont : les fichiers originaux déposés par les doctorants avant et après la soutenance, les documents structurés issus de la transformation des fichiers originaux, les fichiers d'impression et publication issus de la transformation des document structurés. Les documents sont archivés sous forme de cédéroms en plusieurs exemplaires.

6. Evaluation : L'interdépendance des étapes

Le but de ce chapitre est de montrer l'interdépendance de toutes les opérations réalisées et quelles sont les incidences de chaque opération sur la finalité de la plateforme à savoir la recherche et l'affichage des documents.

L'étape qui prend le plus de temps est le stylage. Son but est de fournir un document conforme au modèle de document utilisé par une institution. Ce document servira de matière première aux étapes ultérieures. La qualité des

Droits d'auteur réservés.

résultats des futures tâches dépend de la qualité de ce document. Ce travail est réalisé actuellement à Lyon 2 par le service chargé de la diffusion électronique des thèses. L'idéal est que ce travail soit réalisé en grande partie par le doctorant lui même, le service des thèses électroniques ne faisant qu'apporter quelques modifications. La maîtrise de l'utilisation des modèles de documents et des styles par les étudiants pose le problème de la formation de ceux ci à l'utilisation d'un logiciel de traitement de texte. Doit-elle être assurée par les services chargés de thèses électroniques ou faire partie de leur cursus ? L'utilisation de manière cohérente d'un modèle de document (peu importe lequel) par le doctorant facilite énormément le travail de stylage.

La conversion se sert du document issu de l'étape de stylage. Son but est de produire un fichier conforme à la TEI Lite et des fichiers de diffusion au format PDF et HTML dérivés de celui-ci. Pour que tous les blocs de texte décrits dans le fichier source soient correctement convertis, il est nécessaire que les styles utilisés dans le modèle de document du logiciel de traitement de texte soit déclarés et pris en compte. Sans quoi des fragments de la thèse disparaîtront du fichier résultat ou prendront alors une autre signification, si ils sont décrits par les balises qui ne conviennent pas.

L'indexation quant à elle vise à rendre le document accessible sur Internet. Elle se base sur les résultats de l'étape de conversion et du fichier des métadonnées. Elle utilise une arborescence de fichiers déterminée que l'étape de conversion permet de construire. La modification de cette arborescence implique une modification des programmes chargés de l'indexation.

L'indexation découpe la thèse en plusieurs morceaux. Ce découpage est basé sur la structure du fichier TEI. Pour qu'il soit correctement effectué, il faut que le fichier ait été correctement construit.

La recherche s'effectue sur le contenu des champs indexés. Le fichier des métadonnées sert à alimenter certains champs qui permettent de regrouper les résultats de certains types de recherche. Ce groupage fort utile (index par université) n'est possible que si les données sont présentées d'une manière uniforme. Par exemple l'index des universités et des facultés fournira

Droits d'auteur réservés.

plusieurs entrées pour une même université à cause de l'utilisation d'une casse différente ou des appellations différentes. Le contenu du fichier des métadonnées influent énormément sur la performance de la recherche. Une recherche sur « Claude Bertrand » en tant que directeur de recherche fournira des résultat incomplets si son nom n'est pas saisi de la même manière dans les fichiers de métadonnées de toutes les thèses qu'il a dirigées : « Bertrand Claude », « Bertrand, Claude », « Bertrand C », etc.

La conversion dépend du stylage, l'indexation dépend de la conversion, la recherche dépend de l'indexation. Mais au tout début se trouve le producteur du document (le doctorant). La qualité de son produit a une grande incidence sur la performance de l'ensemble de la chaîne

Droits d'auteur réservés.

Installation de la plateforme

Comme expliqué plus haut, Cyberthèses nécessite la présence des composantes suivantes : Java, OpenOffice.org, Tomcat, SDX et Cyberdocs

Nous avons effectué nos différentes installation dans l'ordre suivant : JAVA, OpenOffice.org, Tomcat, SDX, Cyberdocs.

1. La machine virtuelle java

Sources : http://java.sun.com

C'est le premier logiciel à installer, car il est nécessaire au fonctionnement des autres outils utilisés par Cyberdocs. Après son installation il faut initialiser la variable d'environnement JAVA_HOME qui indique l'emplacement du répertoire d'installation de la machine virtuelle java.

2. OpenOffice.org

Sources : http://www.OpenOffice.org

Après son installation, il vaut mieux l'exécuter une première fois afin de procéder à l'enregistrement ou désactiver ce message.

3. Tomcat

Sources : http://jakarta.apache.org

Par défaut Tomcat écoute sur le port 8080. Pour lui affecter un autre port, il faudrait changer ce port dans le fichier de configuration ***server.xml***. Il existe deux fichiers exécutables permettant de démarrer et de stopper le serveur. Une fois le serveur démarré, il devient accessible à l'adresse

Droits d'auteur réservés.

http://nomduserveur:port/ (http:/localhost:8080/, par exemple, pour installation par défaut).

Tomcat est un moteur de servlets, mais il peut également jouer le rôle de serveur web. Cependant, il peut être configuré pour fonctionner avec un serveur web (Apache par exemple). Dans une telle configuration toutes les requêtes sont adressées au serveur web. C'est ce dernier qui invoquera Tomcat en cas de besoin. Pour faire coopérer Apache et Tomcat par exemple, il faut ajouter le module mod_jk à Apache, et effectuer les configurations nécessaires.

4. SDX

Sources : http://savannah.nongnu.org/files/?group=SDX

Les fichiers d'installation de SDX sont disponibles sous forme de fichiers compilables ou d'un fichier binaire déployable. Ce dernier est facile à installer. Il suffit de copier le fichier ***sdx.war*** dans le répertoire ***webapps*** du moteur de servlet et de redémarrer celui-ci. Au redémarrage ***sdx.war*** est décompressé et les fichiers de SDX sont déployés. L'accès à SDX se fait depuis un navigateur à l'adresse http://nomduserveur/SDX/ (par exemple http://localhost:8080/sdx/). Au premier accès SDX demande de créer un compte de super-utilisateur qui celui sera l'administrateur du serveur SDX.

SDX est installé avec une importante documentation en français et en anglais et une application de démonstration, qui constitue une bonne entrée en matière de SDX.

5. Installation de Cyberdocs

Le stage s'est déroulé pendant une période de développement de Cyberdocs. Les fichiers d'installation étaient modifiés presque tous les jours et parfois plusieurs fois par jour. Les fichiers à jour étaient accessibles sur un serveur CVS de Comité Réseau des Universités (CRU). L'installation de Cyberdocs à

Droits d'auteur réservés.

partir des CVS comporte plusieurs étapes : le téléchargement des fichiers sources, la configuration de l'installation, la compilation des librairies, l'installation du module de conversion, l'installation du module d'indexation et de consultation, l'installation des fichiers des institutions.

Le téléchargement :

Les fichiers sources de Cyberdocs sont disponibles sur internet à l'adresse http://sourcesup.cru.fr/cybertheses/fr/installation/download.html. Ils peuvent être également obtenus par cvs en exécutant les commandes suivantes :

```
cvs -d:pserver:cybertheses@sourcesup.cru.fr:/cybertheses login CVS password:cybertheses
```

```
cvs -d:pserver:cybertheses@sourcesup.cru.fr:/cybertheses co .
```

La configuration :

La configuration consiste à modifier le fichier ***pcd.properties***. Il faut indiquer dans celui-ci : le chemin d'installation du module de conversion, le chemin d'installation de SDX, le nom du répertoire du module de consultation, le nom de l'application SDX, le type d'habillage

Droits d'auteur réservés.

Description d'un exemple de fichier de configuration :

Instructions de Configuration	***Description***
dossier.installation.up=../oo2xml	Indique où sera installé le module de conversion.
OpenOffice.org.home=/home/admin/bin/OpenOffice.org1.0.3	Indique l'emplacement de OpenOffice.org
sdx.application.path=theses	Indique le nom du répertoire qui contiendra l'application Cyberdocs. Ce répertoire sera créé sous le répertoire de SDX, lui même localisé dans le répertoire ***webapps*** du moteur de servlets.
habillage=pcd	Indique l'habillage qui sera utilisé par l'application.
application.sdx.identifiant=org.cyberdocs.theses	Indique le nom que portera l'application SDX. Ce nom permet d'identifier l'application sur un serveur SDX
dossier.installation.consultation=/usr/tomcat-4.1.24/webapps/sdx	Indique l'emplacement où SDX est installé
sdx.application.open=true	Paramètre d'installation de l'application : ouverte ou non

Avec ce fichier, Cyberdocs s'installera sous deux répertoire : ***oo2xml*** (pour la conversion) situé dans le répertoire parent du répertoire contenant les fichiers sources, ***theses*** (pour la consultation) situé sous le répertoire d'installation de SDX. La conversion utilisera une installation de OpenOffice.org située ***/home/admin/bin/OpenOffice.org.org1.0.3.*** Le module de consultation est une application SDX dont l'identifiant est ***org.cyberdocs.theses***. L'apparence

Droits d'auteur réservés.

des pages de l'interface est déterminée par l'ensemble des règles définies grâce à ***pcd***.

La compilation de la librairie ANT :

L'objectif est de créer une archive java exécutable. La compilation est réalisée grâce à la commande :

```
./build-ant.sh.
```

L'installation du module de conversion :

Pour installer le module de conversion, il faut exécuter la commande

```
./build.sh
```

Ceci crée le répertoire ***oo2xml*** et y copie les répertoires et les fichiers indispensables à la conversion. Ce répertoire contient deux sous-répertoires : ***outils*** et ***production***.

Le répertoire ***production*** contient le fichier ***up.sh*** et un répertoire par institution. ***up.sh*** est un script shell qui sert au lancement d'une conversion. La plateforme est distribuée avec quelques thèses de Lyon 2 qui permettent de la tester.

Le répertoire ***outils*** contient les programmes nécessaires pour la conversion : des fichiers de configuration ou de paramètres, des feuilles de styles XSL, des pilotes.

L'installation du module d'indexation

L'installation s'effectue grâce à la commande

```
./installation-web.sh
```

ou

```
./build.sh installation-sdx
```

L'objectif est de créer sous SDX le répertoire qui contient l'arborescence de l'application SDX de consultation. Le répertoire de Cyberdocs contiennent 6 répertoires regroupant l'ensemble des fichiers de l'application.

La copie des institutions :

Droits d'auteur réservés.

Pour installer les institutions, il faut exécuter la commande

./copie-institutions.sh.

Cyberdocs est prévu pour abriter sur un même serveur les thèses provenant de plusieurs institutions. Les paramètres de chaque institution seront regroupés dans un sous répertoire portant le nom de l'institution (plus précisément le code de l'institution) dans le répertoire institutions de Cyberdocs-sdx. La plateforme est distribuée avec les fichiers de paramètres de consultation des thèses de Lyon 2. L'Université Lumière Lyon 2 est identifiée par le code lyon2. L'exécution de la commande précédente créera un répertoire :

{$emplacement-sdx}/{$repertoire-cyberdocs}/institutions/lyon2

Ex : ***/usr/tomcat-4.1.24/webapps/sdx/pcd/institutions/lyon2***.

Résumé des commandes pour installer Cyberdocs :

commande*	***Description***
build-ant.sh	Compiler la librairie ANT
build.sh	Installer le module de conversion
installation-web.sh	Installer le module de consultation
copie-institutions.sh	Copie les fichiers de paramètre des institutions
*NB: Sous Windows, les fichiers correspondants ont une extension .bat	

Droits d'auteur réservés.

La personnalisation de la plateforme

Cyberdocs est conçu comme un logiciel libre et open source. Il est distribué selon les termes de la GPL (GNU General Public License) publiée par la Free Software Foundation. Ce n'est donc pas une boite noire. Cette partie vise à montrer comment adapter Cyberdocs aux besoins d'une institution. Les modifications dont ils est question ne concerne qu'une partie des technologies utilisées par SDX et Cyberdocs.

1. Le module de conversion

1.1. Les styles

Modification de la feuille de style Lyon 2 : ajouter de nouveaux éléments

Les styles utilisés dans le document word sont déclarés dans un fichier oo2xml/outils/xslt/utiles/styles.xml (annexe 3 contient un extrait de ce fichier). Ce document contient une déclaration de chaque modèle de document Word et la description de chaque style que la plateforme est capable de traiter. Le modèle de document est généralement identifié par le code de l'institution qui l'a créé. Le modèle de document de l'Université Lyon 2 par exemple sera déclaré par

```
<institution code="lyon2"/>
```

Un style est identifié par un attribut code et peut avoir plusieurs dénominations (nom), un même style pouvant être appelé différemment selon le modèle de document utilisé. L'attribut code de l'élément nom sert à indiquer le modèle de document.

Un nouveau nom de style dans le document word peut correspondre à un style non prévu ou tout simplement une nouvelle appellation d'un style déjà existant. Dans ce dernier cas, il suffit de rajouter le nouveau nom au style

Droits d'auteur réservés.

qu'il désigne. Ce que nous avons fait pour le style WW-Légénde. Il peut correspondre à une légende de figure ou une légende de tableau. Nous l'avons donc rajouté dans le nom des styles correspondants (Les ajouts sont gras).

```
<style code="legende-fig">
```

<nom code="lyon2" xml:lang="fr">LegendeFig</nom>

<nom code="lyon2" xml:lang="fr">WW-Légende</nom>

</style>

<style code="legende-tab">

<nom code="lyon2" xml:lang="fr">LegendeTab</nom>

<nom code="lyon2" xml:lang="fr">WW-Légende</nom>

</style>

Si par contre le style n'est pas prévu, comme c'est le cas du style pays pour les thèses de la FMPOS, alors il faudrait ajouter un élément style dont est code est pays :

```
<style code="pays">
```

<nom code="lyon2" xml:lang="fr">1|pays</nom>

</style>

Utiliser un nouveau modèle de document

L'utilisation d'un nouveau modèle de document (nouvelle feuille de style pour Word) implique la déclaration du modèle (l'élément institution) et la déclaration pour chaque élément style d'un nom ayant comme code le nom donné à la feuille de style. Le modèle de document utilisé par une thèse est indiqué par un des paramètres du script lançant la conversion d'une thèse. Si la FMPOS développe un modèle de document qui appelle autrement les éléments, la modification du fichier ***styles.xml*** consistera à déclarer le modèle de la FMPOS et à indiquer par quel nom chaque style est désigné :

Droits d'auteur réservés.

Exemple d'ajout d'une nouvelle feuille de style fmpos

```
<institutions>
        <institution code="fmpos"/>
</institutions>
```

Ce passage déclare le modèle de document de la FMPOS

```
<style code="auteur">
        <nom code="lyon2" xml:lang="fr">1|Auteur</nom>
        <nom code="fmpos" xml:lang="fr">_auteur</nom>
</style>
```

Celui-ci indique comment le modèle de la FMPOS a nommé le style auteur.

Remarque : Cet extrait est inspiré des essais réalisés par M. Aliotti, étudiant DESSID, stagiaire à l'INSA.

Le script de conversion indiquera 'fmpos' comme code du modèle de document.

Les styles obligatoires :

La liste des styles considérés comme étant obligatoire est fournie dans oo2xml/outils/xslt/01-validation/corps.xsl. Cette feuille de style **xsl** recherche dans le fichier xml de OpenOffice.org le nom de chaque style énuméré dans la liste. Si ce style n'est pas présent dans le document un message est émis à l'écran et ce message est repris en tant qu'avertissement dans le document TEI XML résultant de la conversion.

Droits d'auteur réservés.

2. L'interface de consultation

2.1. Ajout de nouvelles institutions

Cyberdocs est prévue pour publier sur un même serveur SDX des documents provenant de plusieurs institutions. La plateforme est livrée avec la configuration de l'Université de Lyon 2. Les institutions sont regroupées (un répertoire par institution) dans le répertoire ***sdx/{$cyberdocs}/institutions***. Celui de l'université Lyon 2, par exemple, correspond à ***sdx/theses/institutions/lyon2***. Le répertoire de chaque institution contient : un fichier ***config.xml***, un fichiers CSS et un répertoire ***logos*** qui contiennent les images des logos.

config.xml est un fichier en XML qui définit les étiquettes à utiliser pour afficher les métadonnées associées à une thèse, les masques des adresses IP de l'intranet, la dénomination de l'université, l'emplacement du logo de l'institution. Les adresses IP de l'intranet servent à limiter l'accès aux thèses à diffusion restreinte. Une machine ne faisant pas partie de l'intranet ne pourra pas consulter une thèse dont la diffusion est restreinte à l'intranet. Le logo de l'institution sera affiché dans le cadre supérieur de l'écran de consultation de chaque thèse de celle-ci.

Nous avons ajouté une nouvelle institution: la FMPOS de Bamako identifiée par le code de « ***bamako*** ». La manière la plus simple consiste à copier le répertoire de lyon2 dans le nouveau répertoire (Par exemple sous Unix : `cp -R lyon2 bamako`), et à modifier les différents fichiers.

La dénomination de l'institution est indiquée dans l'élément *ins:nom*, les masques du sous-réseau de l'intranet dans l'élément *ins:intranet*. Les étiquettes dans les sous-éléments dans l'élément *ins:métadonnées*. L'emplacement du fichier du logos est indiqué dans *ins:logo*.

Droits d'auteur réservés.

2.2. Habillages

L'aspect des pages, les menus, les étiquettes des champs des formulaires sont définis dans des fichiers regroupés au sein du répertoire d'installation de Cyberdocs. Il existe deux habillages : ***cybertheses*** et ***pcd***. Le code de l'habillage par défaut est pcd. Toutes les modifications ont été faites après l'installation. Ce répertoire est alors disponible à ***{$emplacement-sdx}/{$cyberdocs}/habillages***. Il est également possible d'effectuer des modifications dans les fichiers sources, dans ce cas les dossiers de l'habillage sont dans le répertoire ***src/web/habillages.***

2.2.1. Ajout de nouvelles pages statiques

Les pages statiques sont des documents XHTML qui doivent être encodés en unicode (comme tout document xml). Ces fichiers sont situés dans le sous répertoire ***statique*** du répertoire de l'habillage choisi au moment de l'installation. Par conséquent les pages statiques se trouveront à ***{$habillage}/statique*** (par exemple : ***/usr/tomcat-4.1.24/webapps/sdx/theses/habillages/pcd/statique***). A l'intérieur de ce répertoire les pages sont regroupées par langues (***fr*** pour le français, ***en*** pour l'anglais). L'accès depuis un navigateur à une page statique se fait par l'url http://{$host}/sdx/{$cyberdocs}/{$nomfichier}.shtm. http://localhost:8080/sdx/theses/aide.shtm, par exemple, permet d'accéder à la page : ***/usr/tomcat-4.1.24/webapps/sdx/theses/habillages/pcd/statique/{$langue}/aide.xhtml.*** {$langue}correspondant à la langue que l'utilisateur a choisie d'utiliser (anglais ou français). Les versions anglaises et françaises d'un même fichier auront le même nom, chacun étant située dans le répertoire d'une langue.

Nous avons ajouté une nouvelle page ***test.xhtml*** en français. Ce fichier est situé à :

Droits d'auteur réservés.

/usr/tomcat-4.1.24/webapps/sdx/theses/habillages/pcd/statique/fr/test.xhtml et accessible par internet à l'adresse http://localhost:8080/sdx/theses/test.shtm

Quand on essaie d'accéder à cette page depuis une interface en anglais un message d'erreur indiquant que le fichier n'existe pas s'affiche.

2.2.2. Ajout de nouveaux menus

Les menus sont disposés sur la partie supérieure des pages du site. L'organisation et la gestion des menus est réalisée dans le fichiers {$habillage}/navigation/general.xml et {$habillage}/messages/menu-general.xml. Par défaut il existe 3 menus principaux : *accueil, informations, aide*. Le menu *informations* contient deux sous menus : *projet* et *technique*. ***general.xml*** contient les identifiants des pages, les noms des fichiers et indiquent comment les menus sont hiérarchisés. Les étiquettes des menus (le texte du menu ou sous menus) sont indiquées dans le fichier ***habillages/messages/menu-general.xml. (Voir annexe 4)***

Ajouter un nouveau menu signifie donc faire des ajouts dans les deux fichiers gérant les menus. L'ajout de menu a été fait en ajoutant quelques lignes dans general.xml et menu-general.xml. Nous avons rajouté un menu ayant deux sous menus (voir annexe 4).

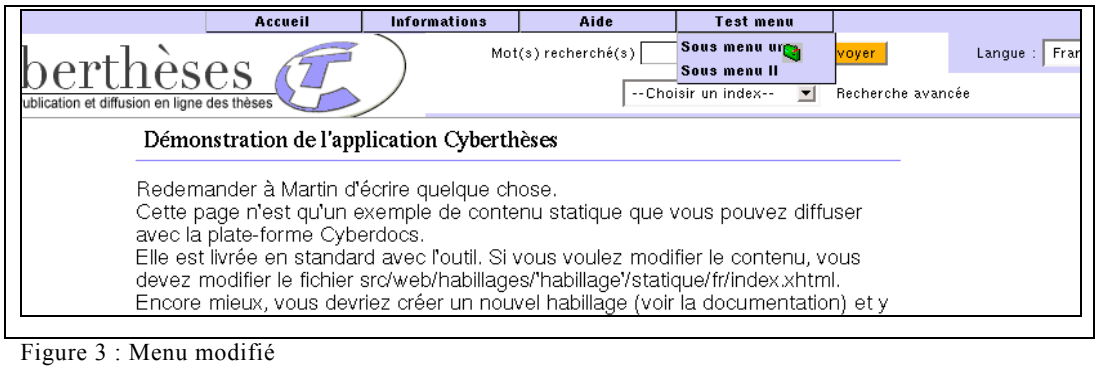


Figure 3 : Menu modifié

2.2.3. Ajout de nouveaux index

Les documents contenus dans la base documentaire sont accessibles par recherche ou par navigation. La recherche consiste à saisir les termes à

Droits d'auteur réservés.

rechercher et à afficher les documents correspondant à la requête. La navigation par index consiste à suivre un parcours par le biais de liens. Les liens sont construits à partir des termes extraits des champs de la base. Par défaut il existe un index : celui des universités et des facultés. Il regroupe les thèses par université de soutenance et à l'intérieur de chaque université par faculté. En cliquant sur le nom d'une faculté, l'utilisateur a accès à toutes les thèses de la faculté.

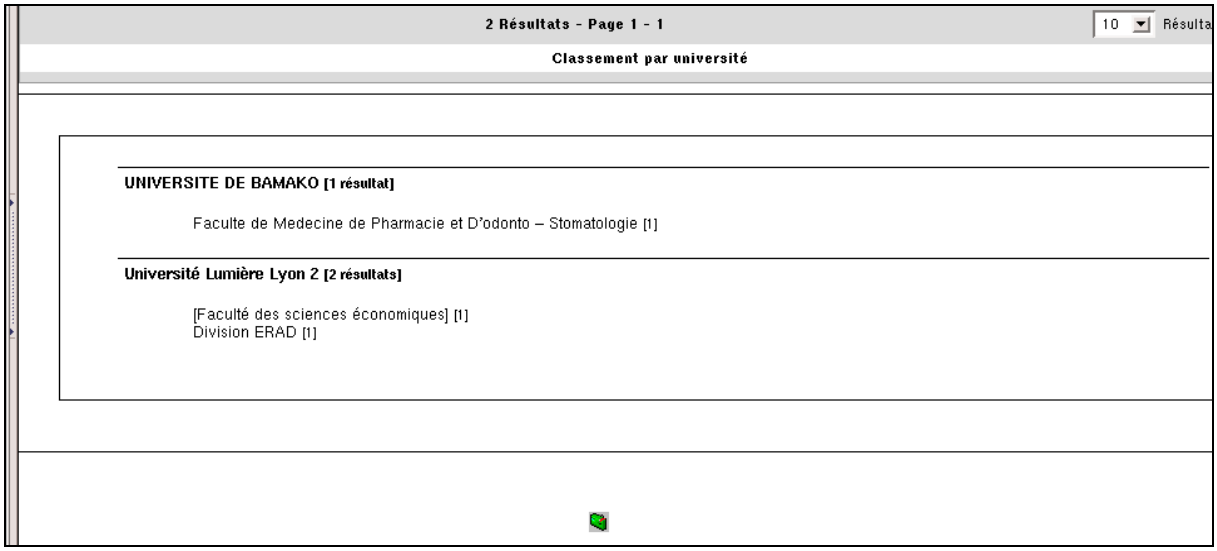


Figure 4 : Accès par université et faculté

Les index, pouvant être utilisés pour un parcours par navigation, dépendent des métadonnées définies dans chaque thèse.

La recherche par index est intéressante sur les champs de type 'field', car leur index contient l'intégralité du champ. Les résultats des index des champs de type 'word' semblent difficilement exploitables, car l'index du champ fragmente le contenu du champ en mot.

La liste des index pour la navigation est définie dans l'élément 'listes' du fichier ***$habillages/messages/global.xml.*** Nous avons rajouté un accès par discipline dans la liste des index (vor annexe 5).

Droits d'auteur réservés.

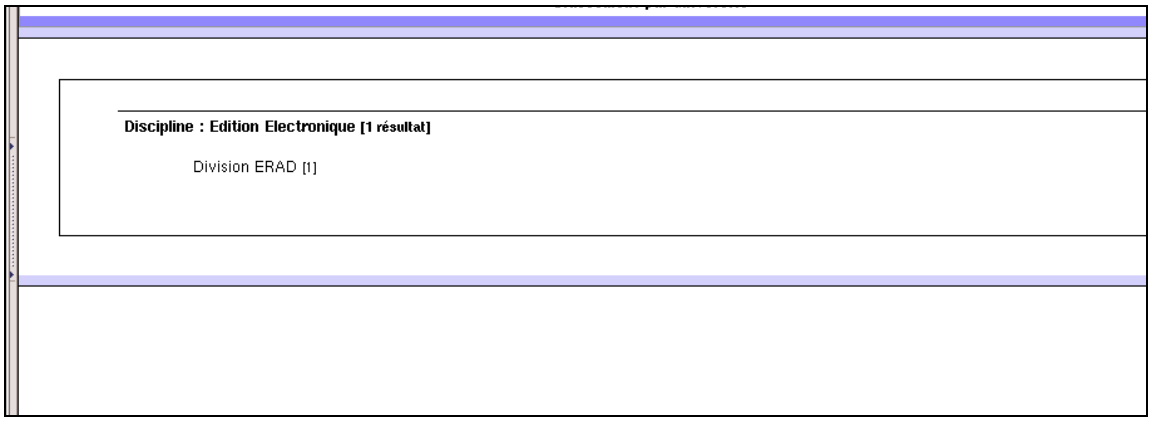


Figure 5 : Accès par discipline et faculté

2.2.4. Ajout de nouvelles langues

Nous n'avons pas réalisé ce travail, mais il peut s'inspirer de la logique de la construction de l'application autour de l'anglais et du français. Globalement nous pensons qu'il existe 3 étapes principales pour ajouter une nouvelle langue dans l'interface de consultation :

1. Déclarer la langue dans l'élément langues du fichier ***sdx/cyberdocs/habillages/pcd/message/global.xml***. Pour chaque langue il existe deux éléments : *titre* correspondant à une étiquette visible dans l'interface web, *langue* correspondant à une entrée dans le menu de choix des langues. Les xml:lang de titre et code de langue correspondent au code de la langue selon la norme iso (*fr* pour français, *en* pour l'anglais, *ru* pour russe). Les lignes suivantes permettent d'ajouter l'italien.

```
<langues>
...
            <titre xml:lang="it">Lingua:</titre>
           <langue code="it">Italiano</langue>
</langues>
```

2. Créer un sous répertoire code de la langue dans le répertoire statique (ex : statique/it). Ce répertoire contiendra la traduction des pages statiques.

Droits d'auteur réservés.

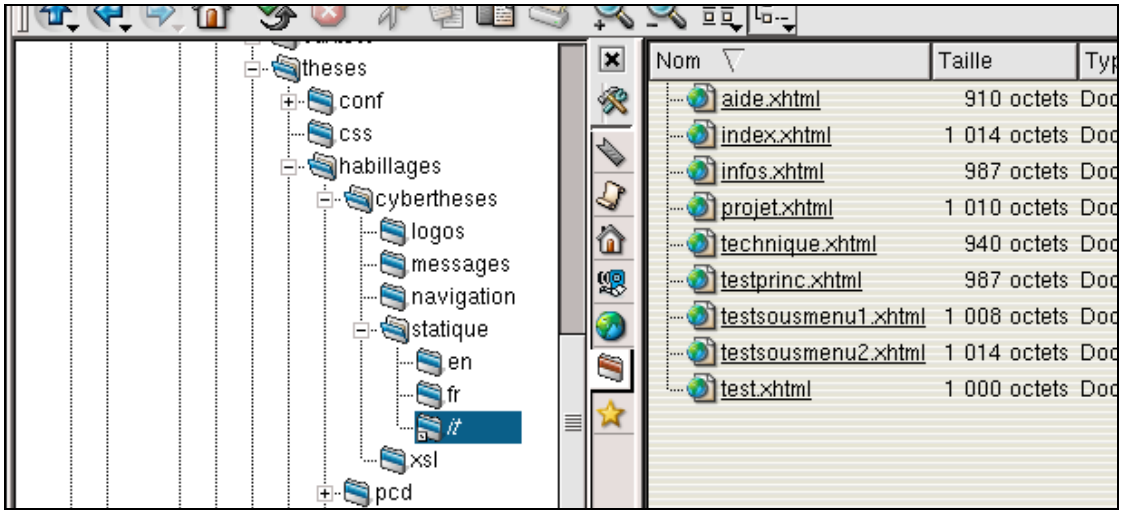


Figure 6 : Repertoire des pages xhtml statiques

3. Traduire tous les fragments de texte visibles dans l'interface web : messages, étiquettes, boutons, etc. Ces fragments sont répartis dans différents fichiers xml. Chaque fragment de texte est contenu dans un élément xml ayant un attribut xml:langue. Il faut donc créer une nouvelle entrée de chaque élément avec l'attribut xml:lang='codedelangue' (<element xml:lang='it'>...</element> pour l'italien).

Extrait de global.xml

```
<boutons>
      <suppr xml:lang="it">[intitulé en italien]</suppr>
</boutons>
```

Ces lignes d'un intitulé en italien pour le bouton 'supprimer'

2.2.5. Utilisation d'une nouvelle DTD

Pour que la plateforme soit générique, elle doit permettre de traiter des documents construits conformément à des DTD autres que la TEI Lite.

Droits d'auteur réservés.

L'utilisation d'une nouvelle DTD implique des ajustements à deux endroits : au niveau de la conversion et au niveau de l'indexation.

Au niveau de la conversion il doit être possible de générer le nouveau type de document à partir du fichier XML générique produit par OpenOffice.org. Pour que cela puisse être fait l'écriture de nouveaux pilotes de conversion (scripts xsl) ou la modification de ceux qui existent est nécessaire. Le choix du pilote à utiliser dépendra alors du type de document qu'on l'on désire produire. L'annexe 6 contient un extrait montrant comment le pilote de conversion est choisi.

Au niveau de l'indexation, les scripts xsl d'indexation devront permettre d'extraire le contenu des nouveaux types de document pour alimenter les champs SDX de la base. Pour cela, il faudra écrire de nouveaux scripts xsl ou modifiés ceux qui existent déjà. Il doit être possible de déterminer le type de document à indexer (la DTD utilisée) et par conséquent choisir le modèle d'indexation adéquat.

A priori, la prise en compte de nouveaux DTD ne semble avoir d'incidence sur la logique de la plateforme. Par contre, il va falloir introduire de nouveaux paramètres qui serviront à déterminer le type de document à produire, à indexer ou à afficher.

Droits d'auteur réservés.

Modélisation de la publication d'une revue électronique sous Cyberdocs

1. Intérêt de la diffusion électronique de revues

Le surnombre, les tarifs élevés, les critères de classement des revues scientifiques ajoutés aux problèmes de distribution constituent des facteurs discriminants pour celles qui sont publiées au Sud.

Selon les initiateurs de la bibliothèque en ligne "Scielo", une initiative sud américaine de publication scientifique en ligne, les « problèmes de distribution auxquels sont confrontées les revues scientifiques du Sud limitent l'accès et l'exploitation de l'information scientifique produite localement ». Ils voient la diffusion en ligne comme un « moyen efficace permettant d'assurer la visibilité et l'accessibilité universelles à la littérature scientifique » des pays du sud et de « combattre le phénomène de science perdue »[6].

La présence sur Internet d'une revue revêt certes un intérêt stratégique, mais nous allons nous pencher uniquement sur l'aspect technique : comment avec la plateforme Cyberdocs publier une revue électronique? La démarche devra aboutir à l'identification des modifications à apporter à la plateforme.

2. Présentation de la revue Mali Médical

Diffusée en 500 exemplaires, la revue « Mali Médical » est publiée par les professeurs de la faculté de médecine de Bamako. C'est une revue trimestrielle qui est distribuée au Mali. Les auteurs des articles sont des spécialistes maliens et d'autres pays de la sous-région.

[6] Scientific Library Online : [en ligne]. Disponible sur <http:/www.scielo.org>.

Droits d'auteur réservés.

Les tests ont été réalisés à partir du fichier word des numéros 3 et 4 de 2002. Une étude du document permet de faire le découpage suivant d'un numéro de cette revue : la page de couverture, les informations générales, le sommaire du numéro courant, les articles.

La page de couverture : elle contient le titre de la revue, une image de couverture, l'année de publication, le numéro du tome et le numéro du fascicule au sein du tome.

Les informations communes : ce sont les recommandations aux auteurs, la fiche d'abonnement, les informations administratives (le siège social, l'adresse, le compte bancaire, etc.) ; les comités de rédaction et de lecture, les tarifs de publicité et d'abonnement.

Le sommaire : à l'intérieur du sommaire les articles sont regroupés en rubriques. Le numéro 3&4 contient trois rubriques : « articles originaux », « cas clinique », « lettre à l'éditeur ». Pour chaque article, le sommaire indique : le titre de l'article, les auteurs, le numéro de la première page.

Les articles : chaque article débute sur une nouvelle page. Un article comprend les éléments suivants : le titre, la liste des auteurs, les services de rattachement des auteurs, le résumé, les mots clés, le corps de l'article. Le corps de l'article est découpé en sections. Une section comporte un intitulé, des paragraphes, des tables, des listes à puces, des illustrations etc. Le texte de l'article est en double colonne.

La mise en page : la page d'un article est divisée horizontalement en trois parties : une entête, une partie centrale, un pied de page. L'entête reprend l'intitulé de la rubrique à laquelle l'article appartient et les premiers mots du titre de l'article. La partie centrale contient le texte de l'article en deux colonnes. Le pied de page reprend le titre de la revue, l'année, les numéros du tome et du fascicule. Sur la première page le titre, les auteurs, les adresses des auteurs, le résumé, les mots clés occupent toute la largeur de la page, tandis que le corps de l'article est présenté en deux colonnes.

Droits d'auteur réservés.

3. Identification des éléments : création des schémas et modèles de documents

Le but de cette partie est de dégager une structure logique sommaire mais suffisante pour le déroulement des tests.

Pour adapter Cyberdocs à la publication d'une revue électronique, il existe plusieurs approches. Une première approche consiste à considérer un fascicule de revue comme document principal. Dans ce cas, chaque fascicule sera traité comme une thèse, chacun des articles sera considéré comme un chapitre. Une autre approche possible consiste à prendre en compte trois documents distincts mais liés : la revue, le fascicule (ou numéro) et l'article. C'est celle ci qui va être appliquée. Nous n'avons pris en compte que quelques éléments qui nous ont semblé indispensables et significatifs pour tester notre approche et les modèles qui en découlent.

Une revue est un ensemble de volumes (ou tomes), chaque volume regroupant un ensemble de fascicules ayant chacun un numéro. Un fascicule est constitué de plusieurs articles pouvant être groupés en rubriques. C'est à partir de cette supposition que nous avons fait trois schémas : celui de la revue, celui du fascicule et celui de l'article.

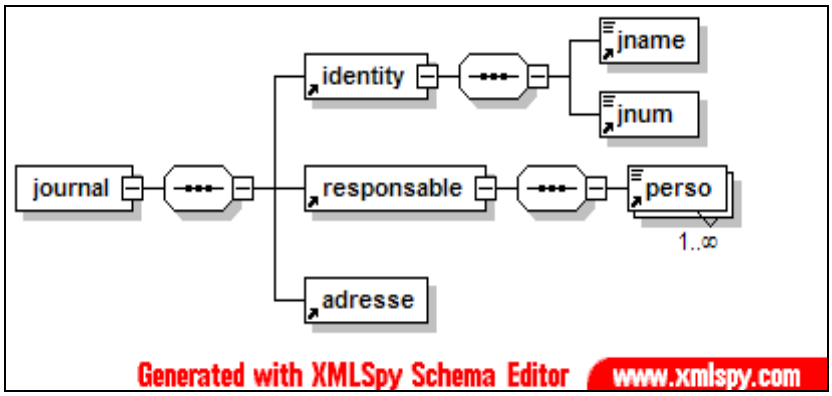


Figure 7 : Schéma d'une revue

Droits d'auteur réservés.

Ce schéma présente les éléments significatifs permettant d'identifier et de décrire une revue : le titre, l'identifiant, les responsables et l'adresse.

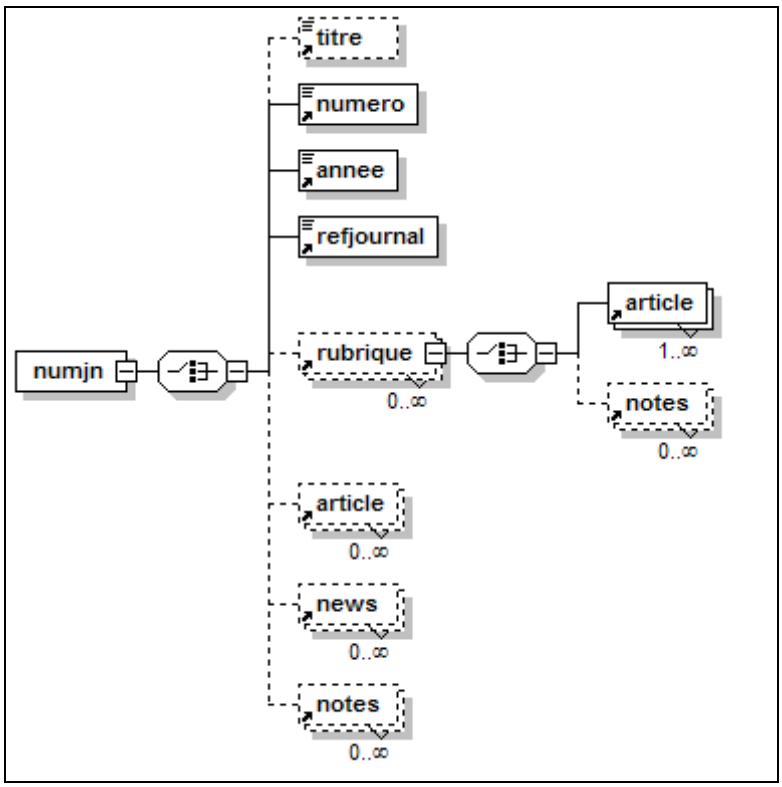


Figure 8 : Schéma d'un numéro de revue

Ce schéma s'applique à un fascicule d'une revue. Les éléments qui le composent peuvent être regroupés en trois catégories : les éléments qui permettent d'identifier le fascicule; ceux permettant de le relier à la revue et ceux relatifs aux différents documents qui le composent.

Droits d'auteur réservés.

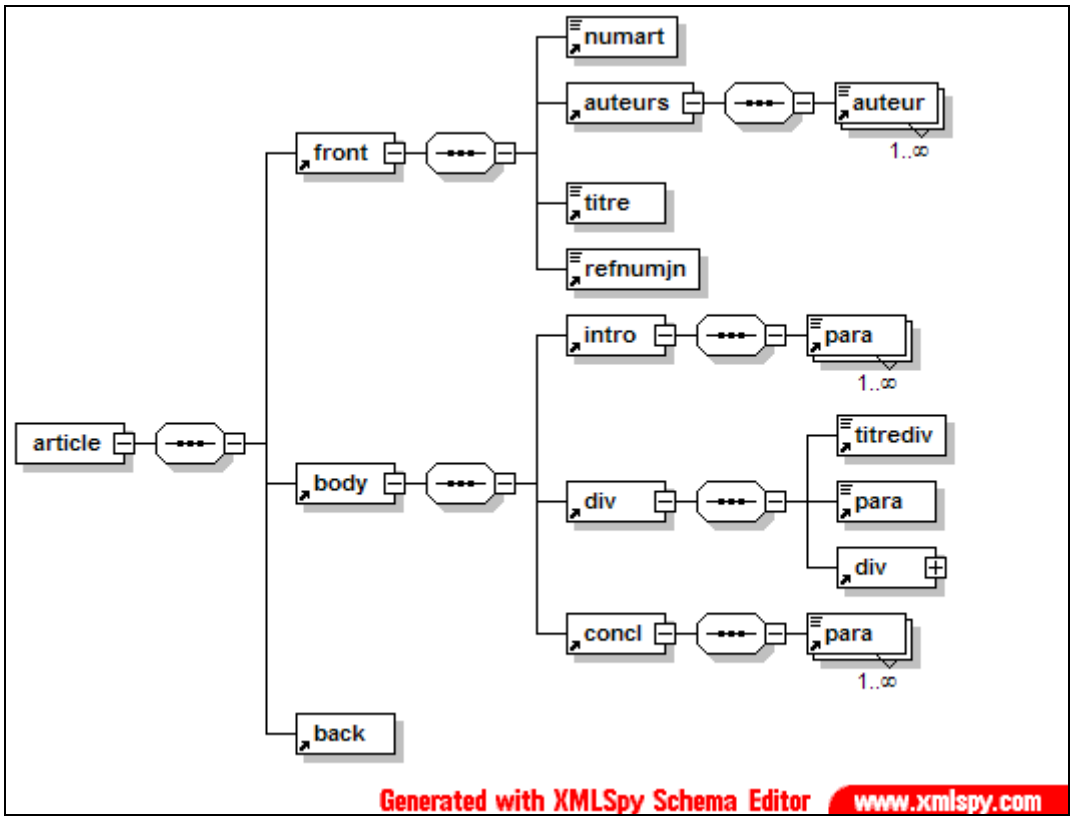


Figure 9 : Schéma d'un article

L'article est construit selon le modèle de la TEI : il comporte une entête; le corps du document et les annexes. L'entête (front) contient les éléments descriptifs de l'article et ceux permettant de le rattacher à un fascicule d'une revue

4. Adaptation de la plateforme

Pour adapter, il fallait comparer les thèses (documents publiés actuellement avec Cyberdocs) et les revues pour relever les similitudes. L'objectif de cette comparaison est de voir comment les informations similaires pouvaient être traités.

Après une étude et une analyse de la plateforme, la mise en oeuvre sera basée sur les constats suivants : une revue peut être considérée comme une institution ; un tome peut être assimilé à une faculté; un numéro peut être considéré comme une thèse, un article comme un chapitre de la thèse. Cette

Droits d'auteur réservés.

comparaison a permis d'identifier les éléments d'hiérarchisation et d'organiser la navigation.

Cette comparaison a cependant des limites : les nom des champs utilisées par l'application des thèses (institution; école; faculté) ne conviennent pas à la revue, l'organisation des fichiers doit être également changé. Il faut trouver des dénominations de champs et une organisation des fichiers à indexer plus adéquates.

L'adaptation de la plateforme soulève certaines questions organisationnelles et techniques.

Faut-il convertir chaque article individuellement ou les rassembler tous dans un seul fichier correspond à l'intégralité du fascicule ?

Comment sera organisée l'arborescence des fichiers : faudra-il rassembler tous les articles d'un fascicule dans un seul répertoire ou faut-il créer un répertoire par article ?

Chaque article devra-t-il contenir l'intitulé de la rubrique à laquelle il est rattaché au sein du fascicule ou bien cette information devra être mentionnée dans le sommaire ?

Les informations générales (comité de rédaction, tarif d'abonnement, recommandations aux auteurs) concernent-elles le fascicule ou bien la revue en général ? Si elles sont rattachées au fascicule, l'historique des différents changements sera conservé. Comment seront-elles traitées pendant l'indexation : seront elles pertinentes pour la recherche ou seront-elles justes conservées comme des documents attachés qui seront consultables en cas de besoin ? Par contre si elles sont rattachées à la revue, seules les versions à jour seront disponibles.

L'indexation devra-t-elle être effectuée en une seule opération (indexation du fascicule) ou en plusieurs opérations (indexation individuelle de chaque article) ?

Un constat s'impose : la taille d'un article est relativement modeste pour l'instant. De ce fait il semble peu pertinent de le découper en plusieurs

Droits d'auteur réservés.

morceaux que ce soit au moment de la conversion ou au moment de l'indexation. Il sera donc plus pertinent de créer un seul fichier pdf et un fichier html par article.

4.1. Adaptation de la conversion

Chaque article a été enregistré dans un fichier word stylé avec le modèle de document de Cyberthèses. Les auteurs d'un article appartiennent à des services différents. Le nom de chaque auteur est suivi du numéro de l'institution à laquelle il appartient. Chaque institution a été stylée avec le style 'département' et est précédée d'un numéro. Ce numéro sert de signet vers lequel pointe le renvoi placé après chaque auteur. Malheureusement au cours de la conversion, le lien entre les auteurs et les institutions n'est pas conservé. On aurait voulu obtenir quelque chose ressemblant à :

```
<docAuthor service='1>Nom Auteur</docAuthor>
<titlePart type='dept' id='1'>Adresse de l'auteur</titlePart>.
```

Peu de modifications ont été apportées à la chaîne de conversion. Notre seule préoccupation a été de rassembler tous les articles dans un répertoire. Tous les fichiers word sont rassemblés dans le sous répertoire ***sources*** et tous les fichiers pdf, xml, html dans le répertoire correspondant. Les fichiers produits au cours de l'article porteront le même nom que le fichier sources au lieu du nom répertoire Pour y parvenir nous avons apporté des modification des scripts lançant et gérant la conversion : ***document.sh*** et ***oo-vers-tei.xml*** (voir annexe 7).

Il suffisait de changer le nom du fichier de sortie de chaque étape de la conversion : au lieu du nom du code du document, il porte le nom du fichier.

Nous aurions voulu empêcher le découpage d'un article en plusieurs fichiers html et pdf (un par chapitre). Une autre modification à apporter serait la mise

Droits d'auteur réservés.

en page du document pdf afin que le texte soit disposé sur deux colonnes et qu'il n'y ait pas de saut de page en début de chapitre.

L'intitulé de la rubrique à laquelle appartient chaque article a été ajouté manuellement au fichier XML de celui-ci.

Le sommaire de la revue sera produit dynamiquement par l'interface de consultation en extrayant tous les articles appartenant au fascicule dont le code a été recherché.

4.2. Modification de l'interface de consultation

Les tests sur l'interface de consultation ont été faits sur une installation dédiée exclusivement aux revues. C'est presque une autre application SDX différente de celle des thèses. Pour que Cyberdocs puisse accueillir les thèses et les revues (ou d'autres types de documents) il faut soit créer des nouvelles bases (une pour les thèses et une pour les revues), soit ajouter de nouveaux champs (titre de la revue, numéro de fascicule par exemple) qui ne sont pas pris en compte actuellement par l'application existante.

Comme pour les thèses, nous avons prévu deux modes de consultation : la navigation et la recherche par requête. Le but de la navigation est de permettre d'avoir la liste des revues, pour chaque revue la liste des tomes et des fascicules et pour chaque fascicule le sommaire. La recherche par requête peut être une recherche simple ou une recherche avancée.

4.2.1. L'unité documentaire

Chaque thèse donne lieu à plusieurs unités documentaires : une, principale, pour toute la page des titres et le sommaire de la thèse, une par chapitre de premier niveau. De façon similaire, un numéro de revue donnera plusieurs unités documentaires : une, principale, qui décrit tout le numéro et une unité documentaire par article. Les documents communs à l'ensemble de la revue

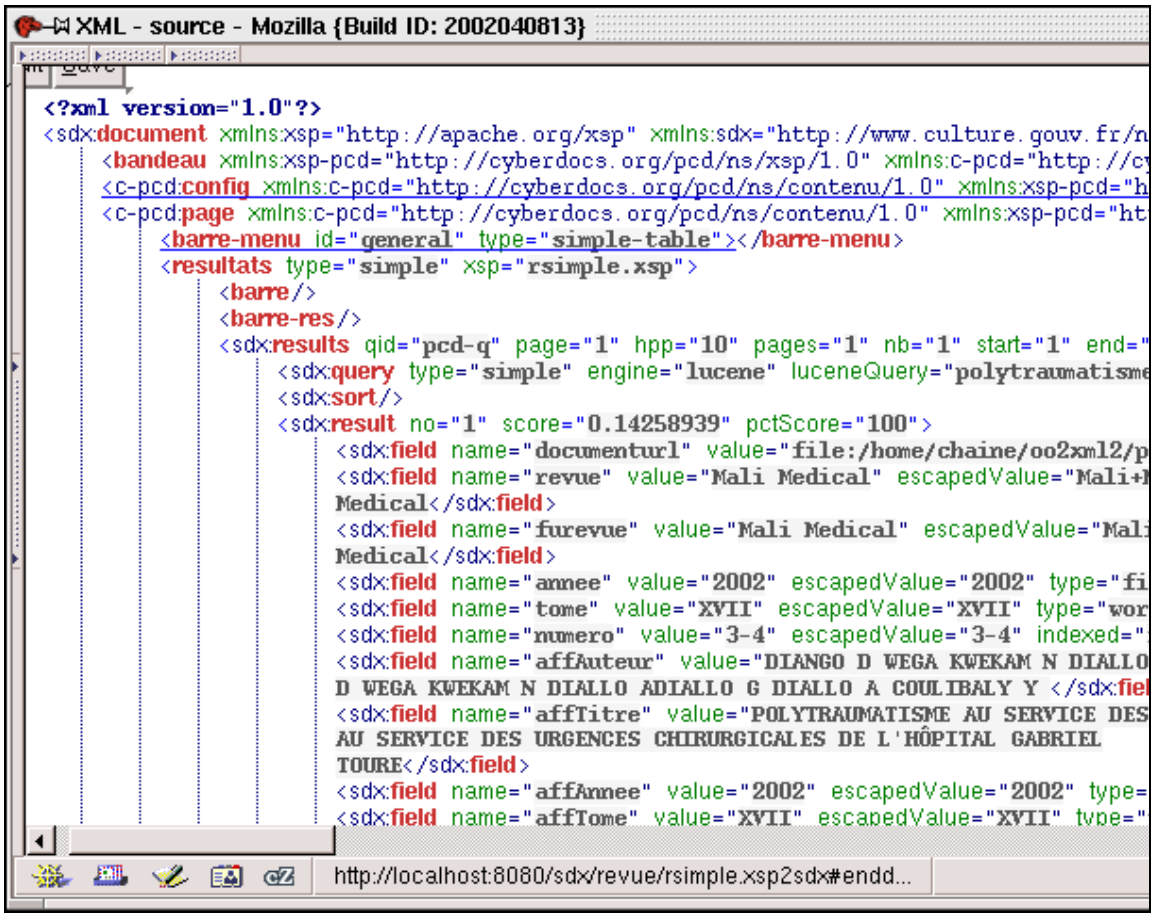
Droits d'auteur réservés.

seront encodés sous forme de pages XHTML statiques dans le répertoire de la revue. Les unités principales serviront à la navigation; les unités des articles serviront à la recherche par requête. Nous aurons ainsi deux types d'enregistrements SDX : un pour le sommaire d'un fascicule (numéro) de revue, un autre pour les articles.

Les champs (annexe 8) ont différentes fonctions : pour la gestion; pour la recherche; l'affichage; la navigation. La différence avec l'application de thèses n'est pas grande. Certains champs présents dans la description d'une thèse (université, école, faculté, jury, directeur de recherche) ne sont pas présents ici et inversement, des champs pertinents pour cette application (titre de la revue; tome de la revue; numéro de fascicule) ne sont pas nécessaires pour une thèse.

Figure 10 : Enregistrement SDX d'un article de revue

Droits d'auteur réservés.



Lors de l'indexation d'un article, un enregistrement contenant les métadonnées de la thèse est créé, en même temps son contenu est copié dans un document xml attaché à l'enregistrement. L'enregistrement permet de retrouver le document lors des recherches et le fichier attaché permet de visualiser le texte intégral. Ce document a la structure suivante :

```
<sous-document id="xxxx" id-parent="yyyy" titre-doc="titre"
type='article' auteur='liste des auteurs>
    <sources>
        <nomrevue>Mali Medical</nomrevue>
        <coderevue>mlmed</coderevue>
        <anneepub>2002</anneepub>
        <tomepub>XVII</tomepub>
```


Droits d'auteur réservés.

```
        <numeropub>3-4</numeropub>
        <pages>xx-yy</pages>
  </sources>
    [texte inégral de l'article]
 </sous-document>
```

L'indexation est réalisée par un ensemble de scripts ***xsl*** que nous avons écrits, qui sont inspirés de ceux utilisés par Cyberdocs. Si Cyberdocs doit faire cohabiter des thèses et des revues dans le futur :

- chaque enregistrement de la base devra contenir une indication sur le type de document indexé ;
- il serait envisageable d'écrire un script xsl ou un modèle (« xsl:template ») par type de document. Le modèle adéquat sera déterminé en fonction du type de document qu'on souhaite indexer.

4.3. Accès aux articles

Il existe deux modes d'accès : la recherche et la navigation. La recherche exécute une requête dont le résultat est une liste des articles contenant les termes de la recherche. Pour chaque article retrouvé sont affichés les auteurs, le titre de l'article, le titre, l'année, le tome, le numéro de fascicule et les pages de l'article.

La navigation a pour but d'accéder à la liste des fascicules d'une revue puis de reconstruire le contenu (sommaire) d'un fascicule. Le parcours idéal serait : la liste des revues disponibles sur le serveur, la liste des fascicules d'une revue regroupés par année, le sommaire d'un fascicule. A partir du sommaire, il devra être possible d'accéder au texte intégral des articles.

Droits d'auteur réservés.

4.4. Comment adapter Cyberdocs aux revues

Cette partie a pour but de dégager certaines réflexions sur les démarches possibles pour adapter Cyberdocs aux revues électroniques en particulier et à d'autres types de document.

- Pour la conversion : prévoir la possibilité de convertir un document en des documents xml ayant des structures différentes. Pour ce faire le pilote de conversion devra être en mesure de connaître la structure du document résultat et par conséquent utiliser les scripts xsl adéquats.
- Pour l'indexation
  - Ajouter des nouveaux champs dans la base documentaire ;
  - développer des scripts xsl pour chaque type de document xml à indexer ;
  - prévoir des scripts permettant d'afficher chaque type de documents.

Droits d'auteur réservés.

Conclusion

A l'issue de ce travail dont le but était d'analyser la version 2.0 de la plateforme cyberdocs pour l'indexation et la diffusion des documents XML, nous pouvons souligner certains aspects:

- Le module de conversion accepte des documents Word ayant été structurés avec des modèles de documents différents de celui de Lyon 2. Pour ce faire, la déclaration du modèle de document et des styles utilisés est nécessaire dans le fichier des styles.

- L'application SDX qui permet d'indexer et de consulter les documents a été conçue en fonction des thèses. A priori, la plateforme est en mesure d'indexer des documents conformes à la TEI types mais d'une structure différente de la thèse. Cela suppose une évolution des fonctionnalités de la plateforme.

La modélisation de la publication d'une revue électronique a permis de démontrer que la plateforme peut prendre en compte de nouveaux types de documents au prix de quelques modifications à apporter. Ce travail a mis en évidence les parties de l'application qu'il faut modifier. Les modifications apportées concernent : la structure de la base, les scripts XSL d'indexation, d'affichage des résultats et du texte intégral.

Globalement le stage a permis de réfléchir aux démarches pour la mise en place d'une plateforme de publication de thèses électroniques au Mali :

- Le modèle de document de Lyon 2 peut être utilisé par les étudiants de la FMPOS pour la rédaction de leur thèse ;
- Le système qui sera mis en place nécessite la collaboration de plusieurs institutions : le programme Cyberthèses pour le développement de la plateforme support et l'appui à la mise en place organisationnelle, la Faculté de Médecine en tant que producteur de contenu et responsable de la plateforme, Keneya

Droits d'auteur réservés.

Blown pour les équipements informatiques et l'espace de diffusion sur Internet ;

- La mise en place d'un politique d'édition électronique nécessite la formation des auteurs pour l'utilisation des modèles des documents, le personnel de la FMPOS aux traitements des documents en vue de leur publication en ligne.

Au delà des thèses : vers des serveurs de documentation électronique (portails). La plateforme Cyberdocs pourrait permettre la mise en place d'un serveur de documentation électronique. Pour que cela soit possible, elle doit être capable de traiter, d'indexer et de restituer des documents de différentes catégories (thèses, revues, livres, sites web), quelques soient les modèles des documents et DTD utilisés.

Droits d'auteur réservés.

BIBLIOGRAPHIE

THÈSES ÉLECTRONIQUES

1. **CLERC, Carole**. Contribution au développement d'un serveur de thèses électronique : mémoire de stage. DESS Ingénierie Documentaire. Villeurbanne : ENSSIB, 1999. 44 p.

2. **BOULETREAU, Viviane ; GAUVIN, Jean François ; DUCASSE, Jean Paul.** La publication électronique des thèses : un exemple de collaboration franco-québecoise. *Documentaliste Sciences de l'Information*, nov 1999, Vol 36 n°6, pp.337-344.

3. **BOULETREAU, Viviane ; DUCASSE, Jean Paul.** La production de documents électroniques structurés à grande échelle. *Cahiers Gutenberg*, mai 2000, n°35-36, pp.25-35.

4. **BOULETREAU, Viviane ; DUCASSE, Jean Paul.** La diffusion électronique des thèses : un enjeu politique et non pas technique. In : **GUICHARD, Eric** (dir.). Comprendre les usages de l'internet. Paris : Editions Rue d'Ulm, 2001, pp.118-122.

5. **BOULETREAU, Viviane ; DUCASSE, Jean Paul, GILLIERON-GRABER, Marie-Pierre.** Cyberthèses en Europe. Bulletin des Bibliothèques de France, 2001, tome 46, n°6 pp.122-129.

Droits d'auteur réservés.

PUBLICATION SCIENTIFIQUE

6. **GUEDON, Jean-Claude.** Numériser les revues savantes : d'un commerce à un autre. *La recherche*, 2000, n°335, pp.78-85.

7. **GUEDON, Jean-Claude.** A l'ombre d'Oldenburg : bibliothécaires, chercheurs scientifiques, maisons d'édition et le contrôle des publications scientifiques. 138ème meeting de l'Association of Research Libraries (ARL). [en ligne]. Disponible sur <http://www.arl.org/> Version française disponible sur : <http://doc-iep.univ-lyon2.fr/Edelec/> (Consulté le 12.07.2003)

TEI

8. **Text Encoding Initiative** [site web] accessible.

Disponible sur: http://www.tei-c.org/ (consulté le 8 sept. 2003)

9. **Cahiers Gutenberg**, 1996, n° 24 : Text Encoding Initiative.

10. **ANDRE, Jacques** ; **DUPOIRIER, Gérard** (Coord). Documents numériques : [cédérom]. 3 ed. Paris : Techniques de l'ingénieur, mai 2003.

XML

11. **YOUNG, Michael**. Formation à XML. Les Ulis (Essonne) : Microsoft Press, 2000.

Droits d'auteur réservés.

12. **World Wide Web Consortium**. Extensible Markup Language (XML) [en ligne]. Disponible sur <http://www.w3.org/XML/> (consulté le 08.09.2003)

12. **World Wide Web Consortium**. Extensible Markup Language (XML) 1.0 : W3C recommandation : [en ligne]. Disponible sur http://www.w3.org/TR/REC-xml (consulté le 10.09. 2003)

XSLT

13. **HOLTZNER, Steven**. XSLT par la pratique. Paris : Eyrolles, 2002.

14. **World Wide Web Consortium**. The Extensible Stylesheet Language Family (XSL) : [en ligne]. Disponible sur <http://www.w3.org/Style/XSL/> (consulté le 08.09.2003).

SDX

15. **Documentation SDX-2** : [en ligne]. Disponible sur <http://www.nongnu.org/sdx/docs/html/doc-sdx2/fr/index.html> (consulté le 08.09.2003)

REVUES ÉLECTRONIQUES

16. **Scientific Electronic Library Online** : [en ligne]. Disponible sur : <http://www.scielo.org/ >. (consulté le 08.09.2003).

Droits d'auteur réservés.

Table des annexes

ANNEXE 1 : MESSAGES DE L'ÉTAPE DE CONVERSION II

ANNEXE 2 : INDEXATION.....III

Annexe 2-1 : fichier non trouvé III

Annexe 2-2 : Ressources manquantes III

Annexe 2-3 : erreurs d'encodage IV

ANNEXE 3 : LES STYLES..... V

DTD du fichier des styles V

Exemple de fichier de style V

ANNEXE 4 : MENUS..... VII

ANNEXE 5 : INDEX.....XI

ANNEXE 6 : PILOTES DE CONVERSION XII

ANNEXE 7 : CONVERSION DES ARTICLES XIII

ANNEXE 8 : LISTE DES CHAMPS DE LA BASE DES REVUES XVI

ANNEXE 10 : PETIT GUIDE D'INSTALLATION DE CYBERDOCS SOUS LINUX.....XVIII

Droits d'auteur réservés.

Annexe 1 : messages de l'étape de conversion

Lors de la conversion, l'absence d'un style obligatoire produit l'affichage du mesage suivant :

```
[<a href="#" title="Effectue une transformation XSLT">style</a>]<a href="#" title="Un avertissement, ce qui signifie probablement qu'un traitement ne s'effectuera pas correctement.">AVERTISSEMENT</a>:<a href="#" title="Il s'agit d'un style obligatoire, veuillez vérifier le stylage du document">400:Il manque le style 1|EcoleDoct, 1|Faculte, 1|Grade, 1|Sous-titre,</a>
```

Ce message est repris sous la forme d'un élément <avertissement> dans le fichier TEI de la thèse :

```
<TEI.2 id="lyon2.2000.dieng_sa">
   <avertissement Cyberdocs="Il manque le style 1|EcoleDoct, 1|Faculte, 1|Grade, 1|Sous-titre,"/>
```


Droits d'auteur réservés.

Annexe 2 : indexation

Annexe 2-1 : fichier non trouvé

Ce message affiché lors de l'indexation indique que le fichier n'a pas été trouvé. La cause peut être une valeur mal orthographiée dans le formulaire d'indexation

```
Un problème est survenu lors de l'indexation file:/home/chaine/cybercvs/../oo2xml/production/lyon2/2002/essai/xml/essai.xml : SDX - Document - XML : erreur dans le document à file:/home/chaine/cybercvs/../oo2xml/production/lyon2/2002/essai/xml/essai.xml : no more input
```

Annexe 2-2 : Ressources manquantes

Message affiché :

```
Un problème est survenu lors de l'indexation
lyon2.2002.essai : SDX - Document - source : Impossible datteindre une source pour le document {0}.
```

Signification :

Ce message indique que un des fichiers auxquels il est fait référence dans le fichier TEI n'existe pas (table des matières, une image). La résolution de ce problème est un peu plus difficile, car le message n'indique pas le nom du fichier manquant. Pour identifier le fichier manquant, il faut lire les fichiers

Droits d'auteur réservés.

de logs de sdx. Ce fichiers se trouve dans le répertoire : ***sdx/WEB-INF/logs***. C'est le fichier sdx.log qui renseigne sur le nom de la ressource manquante. Généralement les lignes signalant les noms de ces fichiers se trouvent parmi les lignes commençant par le mot 'ERROR'.

Annexe 2-3 : erreurs d'encodage

Ce message affiché lors de l'indexation d'un document indique que le fichier n'est pas correctement encodé en Unicode

Un problème est survenu lors de l'indexation file:/home/chaine/cybercvs/../oo2xml2/production/lyon2/2002/essai/xml/essai.xml : SDX - Document - XML : erreur dans le document à file:/home/chaine/cybercvs/../oo2xml2/production/lyon2/2002/essai/xml/essai.xml : Fatal error parsing file:/home/chaine/cybercvs/../oo2xml2/production/lyon2/2002/essai/xml/essai.xml (line -1 col. 869): bad continuation of multi-byte UTF-8 sequence (code: 0x72) .

Droits d'auteur réservés.

Annexe 3 : Les styles

DTD du fichier des styles

```
<!DOCTYPE                                                styles[
<!ELEMENT              styles(institutions,              style+)>
<!ELEMENT                              institutions(institution+)>
<!ELEMENT                                      institution(EMPTY)>
<!ATTLIST                                                institution
      code                    #CDATA                     required>
<!ELEMENT                                               style(nom+)>
<!ATTLIST        style        code        #CDATA         required>
<!ELEMENT                                              nom(#PCDATA)>
<!ATTLIST                                                        nom
      code                    #CDATA                     required
      xml:lang                  #CDATA                   required>
<!ELEMENT                                                        >
]>
```

Exemple de fichier de style

```
<styles>
    <institutions>
        <institution code="lyon2"/>
    </institutions>
    <style code="auteur">
        <nom code="lyon2" xml:lang="fr">1|Auteur</nom>
    </style>
```


Droits d'auteur réservés.

```
        <style code="copyright">
                <nom code="lyon2" xml:lang="fr">1|Copyright</nom>
        </style>
    <style code="liste-puce3">
                <nom code="lyon2" xml:lang="fr">ListePuce3</nom>
                <nom code="lyon2" xml:lang="fr">List Bullet 3</nom>
                <nom code="lyon2" xml:lang="fr">WW-Liste Ã  puces 3</nom>
     </style>
...
</styles>
```


Droits d'auteur réservés.

Annexe 4 : menus

Les menus visibles sur les pages de l'interface de consultation sont gérés par deux fichiers :

/installation-sdx/cyberdocs/habillages/pcd/navigation/general.xml

/installation-sdx/cyberdocs/habillages/pcd/messages/menu-general.xml

Extrait de general.xml : exemple de menu sans sous menu

```
<menu link="aide.shtm" id="aide">
      <item id="aide"/>
 </menu>
```

Extrait de general.xml : exemple de menu avec sous menu

```
<menu link="infos.shtm" id="infos">
       <item id="infos"/>
       <sous-menu>
           <item id="projet" link="projet.shtm"/>
           <item id="technique" link="technique.shtm"/>
       </sous-menu>
</menu>
```

link = le nom du fichier à appelé

Droits d'auteur réservés.

id = identifiant de la page

Extrait de menu-general.xml

```
<nav:item id="index">
   <nav:label xml:lang="fr" status="on">Accueil</nav:label>
   <nav:label xml:lang="fr" status="off">Accueil</nav:label>
   <nav:link-title xml:lang="fr">Page d'accueil de l'application</nav:link-title>
</nav:item>
```

nav:label = étiquette du menu

xml:lang = langue de l'interface de consultation

Ajouts de general.xml

```
<menu link="testprinc.shtm" id="testmenu">
 <item id="testmenu"/>
   <sous-menu>
      <item id="testsousmenu1" link="testsousmenu1.shtm"/>
      <item id="testsousmenu2" link="testsousmenu2.shtm"/>
   </sous-menu>
</menu>
```

ajouts dans menu-general.xml

```
<nav:item id="testmenu">
      <nav:label xml:lang="fr" status="on">Test Menu</nav:label>
      <nav:label xml:lang="en" status="on">Menu test</nav:label>

      <nav:label xml:lang="fr" status="off">Test menu</nav:label>
```


Droits d'auteur réservés.

```
        <nav:label xml:lang="en" status="off">Menu test</nav:label>

        <nav:link-title xml:lang="fr">Test menu</nav:link-title>
        <nav:link-title xml:lang="en">Menu test</nav:link-title>
</nav:item>
<nav:item id="testsousmenu1">
        <nav:label xml:lang="fr" status="on">Sous menu un</nav:label>
        <nav:label xml:lang="en" status="on">Sub menu One</nav:label>

        <nav:label xml:lang="fr" status="off">Sous menu un</nav:label>
        <nav:label xml:lang="en" status="off">Sub menu One</nav:label>

        <nav:link-title xml:lang="fr">Sous menu un</nav:link-title>
        <nav:link-title xml:lang="en">Sub menu One</nav:link-title>
</nav:item>
<nav:item id="testsousmenu2">
        <nav:label xml:lang="fr" status="on"> Sous menu II </nav:label>
        <nav:label xml:lang="en" status="on"> Sub Menu two </nav:label>

        <nav:label xml:lang="fr" status="off">Sous menu II</nav:label>
        <nav:label xml:lang="en" status="off">Sub Menu two</nav:label>

        <nav:link-title xml:lang="fr">Sous menu II</nav:link-title>
        <nav:link-title xml:lang="en">Sub Menu two</nav:link-title>
</nav:item>
```


Droits d'auteur réservés.

Ces ajouts aux deux fichiers créent un menu ayant deux sous-menu.

Droits d'auteur réservés.

Annexe 5 : index

Extrait de global.xml modifié

```
<listes action="termes.xsp">
  <liste xml:lang="fr" champ="funiversite">UniversitÃ©s et facultÃ©s</liste>
  <liste xml:lang="en" champ="funiversite">Universities and faculties</liste>
  <liste xml:lang="fr" champ="fdiscipline">Discipline et tests</liste>
  <liste xml:lang="en" champ="fdiscipline">Fields of studies -tests</liste>
</listes>
```


Droits d'auteur réservés.

Annexe 6 : pilotes de conversion

Actuellement la plateforme contient des pilotes pour convertir les documents en TEI Lite. Ces pilote sont regroupés au sein des répertoire correspondant aux étapes des conversion : ***01-validation, 02-texte, 03-hierarchisation, 04-finalisation, métadonnées, découpage, pdf et tei2html***. L'invocation des pilotes est gérée par le fichier ***oo2xml/outils/bin/oo-vers-tei.xml.***

```
<target name="etape_02"
     depends="initialisation-etape_02"
     description="ExÃ©cute la deuxiÃšme Ã©tape de conversion aprÃšs OpenOffice.org"
        >
                <!-- La XSLT Ã|  utiliser -->
     <property    name="xslt.02"    value="${dossier.xslt}/02-texte/pilote.xsl"/>
           <!--  On convertit le dossier oÃ¹ se trouve le fichier XML en URL -->
                ...
```

target correspond à l'étape de conversion

<property name="xslt.02" value="${dossier.xslt}/02-texte/pilote.xsl"/> indique le pilote qui sera utilisé durant cette transformation.

Droits d'auteur réservés.

Annexe 7 : conversion des articles

./up.sh tout essai.doc lyon2 essai lyon2 fr 2002

Cette ligne permet de lancer la conversion de document essai.doc : essai.doc correspond au nom du fichier à convertir et essai le nom du répertoire où seront stockés les fichier produits. Avec l'ancienne version de oo-vers-tei.xml et tous les fichiers dérivés de la conversion porteront le nom du répertoire essai. Nous obtenions ainsi :

essai/prod/01/essai.xml

...

essai/xml/essai.xml

essai/html/essai.html

Pour convertir un article ce fichier a été modifié de la manière suivante :

./up.sh tout journArt02 mlmed articles lyon2 fr 2002

journArt02 correspond au nom du fichier à convertir et ***articles*** correspond le répertoire où seront stockés les fichiers produits. Sans une modification de ***oo-vers-tei.xml,*** la conversion d'un article écraserait ceux de l'article précédemment converti parce que le nom des fichiers est basée sur le nom du répertoire (articles en l'occurrence).

Les extraits suivants mettent en évidence les modifications apportées à oo2xml/outils/bin/oo-vers-tei.xml (les parties modifiés sont en gras souligné).

Extrait du fichier original

```
<target name="etape_oo"
```


Droits d'auteur réservés.

```
    depends="initialisation"
    description="ExÃ©cute la conversion de Word vers
OpenOffice.org"
 >
                <!-- Le document Ã  convertir -->
  <property                                    name="fichier.doc"
value="${dossier.sources}/${nom.fichier}"/>
                <!-- Le fichier OpenOffice.org Ã  produire -->
  <property                                     name="fichier.oo"
value="${dossier.oo}/${code.doc}.sxw"/>
  ...
</target>
```

extrait du fichier modifié

```
<!-- Conversion de Word vers OpenOffice.org -->
<target name="etape_oo"
  depends="initialisation"
  description="ExÃ©cute la conversion de Word vers OpenOffice.org"
 >
          <!-- Le document Ã  convertir -->
  <property                                          name="fichier.doc"
value="${dossier.sources}/${nom.fichier}.doc"/>
          <!-- Le fichier OpenOffice.org Ã  produire -->
  <property name="fichier.oo" value="${dossier.oo}/${nom.fichier}.sxw"/>
  ...
</target>
```

extrait du fichier original

```
 <!-- PremiÃšre Ã©tape de conversion aprÃšs OpenOffice.org -->
```


Droits d'auteur réservés.

```
<target name="initialisation-etape_01" depends="initialisation">
        <!-- Le document source -->
  <property name="fichier.oo.xml" value="${dossier.oo.xml}/content.xml"/>
        <!-- Le fichier de sortie -->
  <property name="fichier.sortie.01"
      value="${dossier.prod.01}/${code.doc}.xml"/>
</target>
```

extrait du fichier modifié

```
<!-- PremiÃšre Ã©tape de conversion aprÃšs OpenOffice.org -->
<target name="initialisation-etape_01" depends="initialisation">
        <!-- Le document source -->
  <property name="fichier.oo.xml" value="${dossier.oo.xml}/content.xml"/>
        <!-- Le fichier de sortie -->
  <property name="fichier.sortie.01"
      value="${dossier.prod.01}/${nom.fichier}.xml"/>
</target>
```


Droits d'auteur réservés.

Annexe 8 : liste des champs de la base des revues

```
<SDX:field name="gcoderevue" type="field"/>
<SDX:field name="gannee" type="field"/>
<SDX:field name="gtome" type="field"/>
<SDX:field name="gnumerevue" type="field"/>
<SDX:field name="gnumarticle" type="field"/>
<SDX:field name="uauteur" type="unindexed" brief="true"/>
<SDX:field name="utitre" type="unindexed" brief="true"/>
<SDX:field name="affRevue" type="field" brief="true"/>
<SDX:field name="affAnnee" type="unindexed" brief="true"/>
<SDX:field name="affTome" type="unindexed" brief="true"/>
<SDX:field name="affNuremo" type="unindexed" brief="true"/>
 <SDX:field name="documenturl" type="field" brief="true"/>
<SDX:field name="auteur" type="word"/>
<SDX:field name="fuauteur" type="word" brief="true"/>
<SDX:field name="futitre" type="word" brief="true"/>
<SDX:field name="titre" type="word"/>
```


Droits d'auteur réservés.

<SDX:field name="titres" type="word"/>

<SDX:field name="revue" type="word" brief="true"/>

<SDX:field name="furevue" type="field" brief="true"/>

<SDX:field name="tome" type="word" brief="true"/>

<SDX:field name="annee" type="field" brief="true"/>

<SDX:field name="texte" type="word" default="true"/>

<SDX:field name="numerevue" type="word" brief="true"/>

<SDX:field name="id-numerevue" type="field" brief="true"/>

<SDX:field name="tome" type="word" brief="true"/>

<SDX:field name="resume" type="word"/>

<SDX:field name="sujet" type="word"/>

<SDX:field name="biblio" type="word"/>

<SDX:field name="id-numerevue" type="field" brief="true"/>

<SDX:field name="coderevue" type="word" brief="true"/>

<SDX:field name="fcoderevue" type="unindexed" brief="true"/>

Droits d'auteur réservés.

Annexe 10 : petit guide d'installation de Cyberdocs sous Linux

Installation de la plateforme Cyberthèses

Composants :

La plateforme cyberthèses a besoin de 5 composants :

Un environnement Java

OpenOffice.org

Apache Tomcat

Sdx

PCD

Etapes d'installation :

1.Créer un répertoire de travail

ce repertoire contient les fichiers d'installation de tous les composants.

Ex : mkdir chaine

Astuce : *créer également un répertoire temporaire dans lequel on copie les fichiers source de chaque composant après son installation.*

2. Installation des différents composants

2.1 Installation de java

Droits d'auteur réservés.

L'installation peut être faite à partir d'une archive compressée (j2sdk.x.x.tar.gz) ou partir d'un package RPM

Installation d'un package rpm

```
rpm –ivh j2sdk.xxx.rpm
```

installation d'une archive compressée

```
tar -zxvf j2sdk.x.x.tar.gz
```

Initialisation de la variable d'environnement JAVA_HOME

Ex :

```
JAVA_HOME=/usr/java/j2sdk1.0.4_03
export JAVA_HOME
```

Note *: la variable d'environnement JAVA_HOME doit également initialisé dans le fichier .bash_profile de l'utilisateur.*

```
cat >>.bash_profile
JAVA_HOME=/usr/java/j2sdk1.0.4_03
export JAVA_HOME
```

2.2 Installation de openoffice.org

Les fichiers d'installation sont disponibles sous forme d'archive compressée OpenOffice.org.x.x.tar.gz

Decompresser l'archive

tar –zxvf OpenOffice.org.x.x.tar.gz

La décompression crée un repertoire install

Droits d'auteur réservés.

cd install/

./setup

- accepter la licence

- indiquer le chemin d'installation. /home/chaine/OpenOffice.Org1.0

- Indiquer l'installation de java à utiliser : soit en choix parmi ceux qui ont été détecté, soit indiquer le chemin d'installation de java

2.3 Installation de apache tomcat

Les fichiers d'installation sont disponibles sous forme d'archive compressée jakarta-tomcat.x.x.tar.gz

- Décompresser l'archive

tar –zxvf jakarta-tomcat.x.x.tar.gr

- ***Changer éventuellement le port par défaut 8080 par une autre valeur***, si un autre serveur utilise déjà le port 8080

2.4 Installation de sdx

Les fichiers d'installation sont disponibles sous forme d'une archive : sdx.war

- Copier le fichier sdx.war dans le repertoire des documents de apache tomcat :

cp sdx.war /chemin/installation/apache-tomcat.x.x/webbapps

Droits d'auteur réservés.

- Lancer le serveur web : apache tomcat

cd /chemin/installation/apache-tomcat.x.x/bin

./startup.sh

- Accéder à sdx à partir d'un navigateur. Au premier accès le fichier est décompressé pour créer l'arborescence de sdx.

1.5 Installation de PCD

Les fichiers d'installation sont disponibles sous forme d'un fichier compressé.

- Decomprésser le fichier

unzip pcd.zip

la décompression crée un repertoire pcd

- modifier le fichier de paramètres

cd pcd

vi pcd.properties

- Indiquer l'emplacement exact de openoffice et de sdx

- Enregistrer et quitter vi

:wq

- compiler pcd

. /build.sh

recompiler pcd avec sdx

. /build.sh installation-sdx

Lancer un navigateur et acceder à la page d'acceuil de cybertheses

http://localhost:8080/sdx/pcd

Droits d'auteur réservés.