Rapport de stage septembre 2003

DESS Ingénierie documentaire

Étude de 2 chaînes d'édition numérique XML - Projet de diffusion électronique de la production scientifique de l'INSA.

Frédéric Aliotti

Sous la direction de Madame Monique JOLY
Directrice de Doc'INSA - INSA Lyon.

Dalila Boudia (Responsable des thèses), **Gilles Brochet** (Responsable du service informatique), Doc'INSA, co-responsables du stage.

Emilie Romand-Monnier (Ingénieur d'études), ENSSIB, responsable du stage.

Étude de 2 chaînes d'édition numérique XML - Projet de diffusion électronique de la production scientifique de l'INSA.

Résumé

La diffusion des thèses électroniques est un sujet important pour de nombreux établissements chargés d'en assurer la conservation et la diffusion. Suivant la voie ouverte, en 1987, aux États-Unis par des projets de développement d'outils permettant de convertir les documents produits à l'aide d'éditeurs de texte classiques, en documents SGML, puis XML, les projets de chaîne d'édition numérique (CEN) Cyberdocs et Doc'INSA émergent aujourd'hui.

Ce document présente les intérêts et les lacunes de ces outils, informations destinées à asseoir les réflexions qui conduiront à la mise en place d'une CEN au format XML de la production scientifique des étudiants de l'INSA de Lyon.

Descripteurs : XML, technologie XML, chaîne de traitements, métadonnées, conversion, thèse

Study of two XML processing lines – INSA' scientific production electronic diffusion project

Abstract

The electronic thesis distribution is a main interest purpose for many institutions implied in their preservation and distribution. Following the conversion of edited with common editors to SGML and XML documents tools devellopment projects driven in the United-States since 1987, the Cyberdocs and Doc'INSA processing lines are coming out today.

This report presents the advantages and disadvantages of those tools. Its content will be helpful in the implementation of processing line in XML format for the scientific work of the INSA-Lyon's students.

Keywords: XML, XML technologies, processing line, metadata, conversion, thesis

Remerciements

Un grand merci à Mme Monique JOLY et Mme Nicole Bion pour m'avoir accueilli à Doc'INSA et à toute l'équipe pour sa bonne humeur, son attention et sa gentillesse.

Parmi toutes les personnes qui la compose je remercie plus particulièrement Mme Dalila BOUDIA, M. Gilles BROCHET et Mme Brigitte PRUDHOMME sans lesquels se stage ne m'aurait pas été confié. Je renouvelle mes remerciements à Mme Monique JOLY, Mme Dalila BOUDIA et M. Gilles BROCHET pour les lectures attentives de ce rapport et les remarques dont ils m'ont fait part.

Merci à mes camarades du DESSID, Abderahamane ANNE qui a toujours répondu présent, même le Week-end, et Rosa-Maria GÓMEZ DE REGIL qui m'a supporté tout le reste de la semaine.

Merci à Mme Émilie ROMAND-MONNIER pour ses conseils, sa disponibilité et les re(...)lectures de ce rapport.

Merci à M. Bernard PARIOT pour ses conseils et « english touch » Mme Myriell ELFROTH.

Merci à tous les lecteurs avisés qui n'auront pas jugé utile de parcourir ces lignes et qui seront passés directement au sommaire, et à tous les oubliés qui ne m'en tiendront pas rigueur.

Sommaire

Abréviations

ABES	Agence Bibliographique de l'Enseignement Supérieur.
API	« Application Programming Interface », interface de programmation pour application.
CEN	Chaîne d'Édition Numérique.
CITHER	Consultation en texte Intégral des THèses En Réseau.
CRU	Comité Réseau des Universités [33].
DCMI	« Dublin Core Metadata Initiative », initiative des métadonnées du Dublin Core.
DTD	« Document Type Definition », définition d'un type de document XML.
HTML	« HyperText Markup Language », langage de balisage hypertextuel.
INSA	Institut National des Sciences Appliquées.
OAI	« Open Archives Initiative Protocol », protocole de l'initiative pour les archives ouvertes [25].
PC	« Personal Computer », micro-ordinateur.
PDF	« Portable Document Format », format de document portable.
RDF	« Resource Description Framework », architecture de description des ressources.
SGML	« Standard Generalized Markup Language » (SGML, ISO 8879:1986), langage de balisage standard généralisé.
TEI	« Text Encoding Initiative », initiative d'encodage textuel.
XML	« Extensible Markup Language », langage de balisage extensible.
XSL	« Extensible Stylesheet Language », langage extensible de feuilles de styles.

Introduction

1. Présentation générale de Doc'INSA

La bibliothèque scientifique et technique de l'INSA de Lyon, Doc'INSA, met à la disposition des étudiants, enseignants, chercheurs et autres personnes autorisées, plus de 85 000 ouvrages, 1 740 collections de périodiques et 2 700 microformes ayant trait aux sciences de l'ingénieur. Elle possède en outre plus de 1 800 thèses, qui constituent une partie non négligeable de la production scientifique de l'INSA.

2. Les thèses de l'INSA

Dépositaire officiel des thèses produites dans les laboratoires de l'INSA de Lyon, Doc'INSA reçoit, chaque année, près de 130 nouvelles thèses, dont une partie seulement, en raison des dispositions légales qui en restreignent la distribution électronique, est consultable depuis le site Web de l'établissement. La mise en ligne des thèses de l'INSA fut réalisée dans le cadre du projet CITHER (1996-1998) sous l'impulsion des recommandations du ministère [12]. Les thèses, au format PDF, sont actuellement consultables *via* une interface de recherche performante, basée sur la technologie dtSearch [34] qui permet de localiser des mots clés présents dans ces documents.

3. Organisation de Doc'INSA

Doc'INSA, bibliothèque dans laquelle ce stage a été effectué, est dirigée par Mme Monique JOLY et Mme Nicole BION. Les activités d'accueil des usagers, de prêt entre bibliothèque, d'indexation, de catalogage, de gestion de prêt, de préparation des documents, ainsi que les activités administratives sont partagées (Figure 1) par une trentaine de personnes. Lesquelles sont, en proportions équivalentes, contractuelles ou attachées à la fonction publique.

ACCUEILLIR, RENSEIGNER, INFORMER, FORMER LES USAGERS				**GERER LES COLLECTIONS**			**GERER LE CENTRE**
ACCUEIL - ORIENTATION - PRÊT	RENSEIGNEMENT DOCUMENTAIRE - VEILLE TECHNOLOGIQUE	PRET ENTRE BIBLIOTHEQUES	FORMATION IST - DES ETUDIANTS	TRAITEMENT DES DOCUMENTS SUR SUPPORTS	INFORMATISATION DU FONDS	DOCUMENTATION ELCTRONIQUE	GESTION DU SERVICE
Banque de prêt	Renseignement documentaire	Prêt entre bibliothèques	Cours travaux dirigés	Acquisitions, dépôt des thèses	Administration du catalogue collectif INSA	Ressources Électroniques (Web, Édition électronique,)	Équipe de direction
Magasin	Veille technologique		TD première année	Entrée inventaire	Catalogage		Budget - Comptabilité
			Planning formations	Indexation	Revues		Formation du personnel
				Équipement	Prêt –Relance - Statistiques		Hygiène et sécurité
							Planning - Permanences
							Informatique
							Intradoc
							Comptabilité - Secrétariat

Figure 1 : Diagramme présentant l'organisation fonctionnelle de Doc'INSA.

Mes maîtres de stage, Mme Dalila BOUDIA et M. Gilles BROCHET sont concernés par la mise en place d'une CEN, respectivement en tant que : responsable du traitement des thèses et responsable des ressources informatiques, chargé de la mise en place et de l'administration de la CEN et du serveur qui diffusera les thèses électroniques produites à Doc'INSA.
Ils ont participé au développement d'une chaîne de traitement numérique permettant de convertir les thèses électroniques du format original – Word, $LaTeX$ – vers un format PDF amélioré. Ce projet, nommé CITHER, est actuellement utilisé à Doc'INSA. Ils ont également participé, plus récemment à un projet de mise en place d'une chaîne de traitement numérique destinée à convertir les thèses électroniques déposées à Doc'INSA, du format original – Word, OpenOffice.org, $LaTeX$ – vers le format XML. Ce dernier projet a permis de développer une chaîne d'édition numérique, qui sera désignée dans le reste de ce document par « CEN Doc'INSA ».

4. Définition du stage

4.1. Intitulé du stage

L'intitulé du stage proposé par Mme Monique JOLY est le suivant : « Diffusion électronique de la production scientifique de l'INSA : test de deux chaînes éditoriales, évolution, implémentation et/ou cahier des charges. »

4.2. Contexte opérationnel

4.2.1 Objectifs du stage

Ce stage s'inscrit dans la phase finale d'un projet ayant débuté en 2000, visant à développer la CEN Doc'INSA, outil de conversion des thèses vers le format XML, considéré comme pérenne, particulièrement adapté à l'archivage car consultable et compréhensible en utilisant un simple éditeur de texte, contrairement au format propriétaire des documents Word, qui en l'absence du logiciel à l'aide duquel ils ont été créés, sont difficilement compréhensibles.

Le XML structure les documents et apporte un « ensemble de fonctions utilisant la valeur ajoutée par le balisage » [16] qui permet de les enrichir et de justifier, s'il en était besoin, la conversion des documents vers ce format « pivot », qui servira de matière première à une chaîne de publication de documents dans différents formats permettant leur consultation en ligne.

L'objectif de ce stage est de prendre connaissance des travaux précédents et de décrire l'adéquation entre les outils existants et les besoins actuels. Deux CEN ont été retenues : la CEN Doc'INSA et la CEN Cyberdocs développée dans le cadre du projet Cyberthèses. Il conviendra d'étudier ces deux outils et de présenter les résultats de telle sorte qu'il soit possible de choisir celui qui, le cas échéant, sera le mieux adapté pour mettre en place une plate-forme de publication électronique à Doc'INSA. À défaut de trouver dans les outils étudiés, une réponse satisfaisante aux attentes présentées, une solution alternative pourrait être proposée.

Ce rapport est constitué de quatre parties :

- analyse du projet ;
- comparaison des CEN Doc'INSA et Cyberdocs ;
- mise en place d'un groupe de travail ;
- mise en place d'une CEN.

4.2.2 Environnement informatique

Au cours de ce stage, ont été mises à ma disposition toutes les ressources informatiques souhaitables, bien au delà de ce qui était strictement nécessaire :)

J'ai ainsi bénéficié de trois postes en réseau, reliés à une imprimante :

- un poste de travail personnel :
    - système Microsoft Windows 98 Seconde Édition ;
    - processeur Intel Pentium II cadencé à 350 MHz, 64 M SDRAM, 4 Go d'espace de stockage ;
    - environnement logiciel : messagerie Eudora, suite bureautique Microsoft Office 97, navigateur Internet explorer 6, éditeur XML Spy 4.4 ;
- un poste de travail pour tester la CEN Cyberdocs :
    - système Linux Redhat 7.2 ;

  - processeur Intel Pentium IV cadencé à 1,7 GHz, 512 Mo de mémoire vive SDRAM, 40 Go d'espace de stockage ;
  - environnement logiciel : navigateur Mozilla, SAMBA et LinNeighborhood pour partager des données avec les micro-ordinateurs constituant un réseau Microsoft, suite bureautique OpenOffice.org ;
- un poste de travail pour tester les CEN Doc'INSA et Cyberdocs :
  - système Windows XP ;
  - processeur Intel Pentium IV cadencé à 2.4 GHz, 480 Mo de mémoire vive DDRAM, 80 Go d'espace de stockage ;
  - environnement logiciel : navigateur Internet Explorer 6, suites bureautiques Microsoft Office 2000 et 2003, éditeur XML Spy 4.4.

La mission du stagiaire consiste à :

- évaluer et comparer les CEN Doc'INSA et Cyberdocs ;
- créer un modèle de document pour les thèses de l'INSA compatible avec la chaîne Cyberdocs ;
- proposer une estimation du coût du traitement des thèses.

Les contraintes de la mission :

Les CEN étudiées sont constituées de nombreux logiciels réalisant de façon séquentielle des traitements qui aboutissent à la production des documents escomptés. Chaque modification de l'un des composants d'une telle chaîne de traitement peut en modifier le fonctionnement. Or, au cours de cette étude, nous avons tenté d'utiliser les versions les plus récentes des logiciels composant la chaîne Doc'INSA et avons évalué des versions quotidiennement mises à jour de la chaîne Cyberdocs. Cette dernière n'ayant été distribuée en version « beta » que le 1er septembre 2003. La grande variabilité des objets de cette étude a constitué la principale difficulté rencontrée au cours du stage.

Analyse du projet

1. Étude de l'existant

1.1. Ressources consultées

- Google : la requête « thèse theses thesis +xml +doc +"en ligne" online » permet de recueillir les documents suivants :
    - le rapport de stage DESSID de Carole CLERC [4] ;
    - le rapport de stage DESSID de Jean-Michel MERMET [13] ;
    - un article concernant l'ETD [18], « The Guide to Electronique Thesis and Dissertations » ;
    - des informations concernant la NDLTD [17], « Networked Digital Library of Theses and Dissertations » ;
- CITHER, le projet de consultation en texte intégral des thèses en réseau [10] ;
- l'ABES, Agence Bibliographique de l'Enseignement Supérieur [32].

1.2. Les principaux projets

1.2.1 Projet ETD

Le concept des thèses et rapports électroniques (ETD) [8] fut présenté en 1987 au cours de la réunion de Ann Arbor organisée par l'UMI[i] (University Microfilm, Michigan, États-Unis), à laquelle participaient les représentants de l'université Virginia Tech (Ed Fox du « Computer Science » et Susan Bright du « Computing Center »), l'université du Michigan, et les éditeurs de logiciels SGML SoftQuad et ArborText. Par la suite, l'idée de diffuser des documents électroniques, rédigés par des universitaires, à travers un réseau de bibliothèques numériques voué à la diffusion des thèses et des rapports « Networked Digital Library of Theses and Dissertations » (NDLTD) [14] émergea. Ce réseau regroupe actuellement plus de 160 universités issues de nombreux pays.

Depuis 1996, Virginia Tech développe dans le cadre du projet ETD un ensemble de logiciels destinés au traitement et à la diffusion des thèses et rapports électroniques.

1.2.2 Projet StoneCastle

Ce projet de publication et de diffusion électroniques des thèses de doctorat [15] est issu d'une collaboration entre les universités de Montréal, Lyon 2 et l'Égypte. Ce projet a récupéré presque intégralement la chaîne de traitement développée au cours d'un projet nommé Erudit, et a modifié les scripts OmniMark existants, langage aujourd'hui propriétaire, qui a été abandonné au cours des développements ultérieurs de cette chaîne.

1.2.3 Projet Cyberthèses

Les universités de Lyon et Montréal, associées dans un premier temps à d'autres partenaires représentant les institutions universitaires de l'Égypte, ont mis en place une coopération sur le thème de l'édition et de la diffusion électroniques des thèses s'appuyant sur la norme SGML. Les objectifs de ce projet, nommé Cyberthèses, sont présentés dans les termes suivants :

- s'engager dans une coopération multilatérale sur le thème de l'édition et de la diffusion électroniques sur Internet des thèses universitaires en s'appuyant sur la norme SGML ;
- mettre en service une chaîne de production et de diffusion par la conception et l'élaboration d'un certain nombre de procédures logicielles qui prennent en compte les spécificités de structure de la thèse.
- servir de base à une application de presse électronique universitaire libre et ouverte à toutes les universités francophones et de première pierre à une bibliothèque virtuelle universitaire de la Francophonie.

Ce projet Cyberthèses [7] a été soutenu en 1998, par le fonds francophone [9] des inforoutes. La Bulgarie, le Canada/Québec, la France, l'Égypte, le Maroc et la Tunisie participèrent à cette réalisation. L'Université Lumière Lyon 2 et l'université de Montréal sont les deux établissements initiateurs de ce projet.

---

[i] l'UMI (University Microfilm) est l'organisme qui fait autorité dans le domaine des thèses aux États-Unis. Cet organisme microfiche systématiquement les thèses, gère la base de données « Dissertation Abstract » et commercialise les thèses

1.2.4 Projet CITHER

Le projet CITHER, développé à Doc'INSA, a pour objectif de proposer la consultation en ligne des thèses, déposées à l'INSA de Lyon à partir de janvier 1997, en texte intégral et au format PDF.

1.2.5 Serveur de thèses multidisciplinaire

Le serveur de thèses multidisciplinaire [3], proposé par le Centre National de la Recherche Scientifique [5] (CNRS) et le Centre pour la Communication Scientifique Directe [2] (CCSD) utilise le logiciel eprints.org pour mettre en œuvre l'autoarchivage des thèses qui leur sont confiées.

1.3. Normes et recommandations

1.3.1 XML

La recommandation « Extensible Markup Language » (XML) définit un langage de structuration des données basé sur un balisage inspiré de la norme SGML (ISO 8879). Cette recommandation a pour objectif de définir un nouveau langage, plus simple que le SGML, mais qui en préserverait les avantages [26].

1.3.2 XSL

« Extensible Stylesheet Language » (XSL) est une famille de recommandations (XSLT, XPath, XSL-FO) qui définissent les instructions de transformation et de présentation des documents XML.

1.3.2.1 XSLT

« XSL Transformations » (XSLT) [27] constitue le langage de transformation du XML.

1.3.2.2 XSL-FO

« XSL Formating Objects » (XSL-FO) [29] définit un vocabulaire de formatage sémantique.

1.3.2.3 XPath

« XML Path Language » (XPath) [28] est un langage contenant des expressions utilisées par XSL pour accéder et se référer à des parties de documents XML.

1.3.3 MathML

« Mathematical Markup Language » (MathML) [30] est une application XML de description des notations mathématiques visant à capturer leur contenu et leur mise en forme.

1.3.4 Dublin Core

Le « Dublin Core Metadata Initiative » (DCMI) [24] est une organisation dédiée à la promotion de l'adoption à large spectre de standards de métadonnées interopérables et au développement d'un vocabulaire de métadonnées ayant pour but de décrire des ressources et de permettre leur recherche par des systèmes plus intelligents.

1.3.5 DocBook

DocBook [20] regroupe un ensemble de balises XML destinées à être utilisées pour rédiger des documents électroniques.

1.3.6 TEI

« Text Encoding Initiative » [21] (TEI) est un standard interdisciplinaire et international, présenté en 1987, qui aide à présenter les textes à intérêt littéraire ou linguistique selon un schéma qui favorise leur exploitation. Les directives de TEI définissent une langue pour décrire la structure des textes, et proposent des désignations pour ses composants. Le consortium TEI, fondé en 2001, société internationale à but non lucratif pour maintenir et pour développer le système TEI, propose une version des directives de TEI entièrement conforme à XML.

2. Étude des besoins

Pour répondre à nos besoins, la CEN doit implémenter les fonctionnalités suivantes :

- production de documents XML à partir de documents Word, OpenOffice.org et $L^{A}T_{\varepsilon}X$;
- production de documents XML conformes aux documents originaux. Dans l'absolu, le document XML doit contenir la même information que le document

Word, OpenOffice.org ou LaTeX à partir duquel il a été produit. Une attention particulière sera portée à la conformité des informations textuelles et graphiques, voire aux informations audiovisuelles dès que leur prise en charge sera possible. L'attente minimale étant de créer des documents XML contenant de façon exhaustive l'information initiale et de limiter les modifications de la mise en page et de la forme du document ;

- production de documents XML conformes à une DTD correspondant à nos besoins (*Cf* chapitre 3.1) ;
- conversion des formules mathématiques contenues dans les thèses en MathML, sous ensemble XML proposé par le W3C comme le format de description et de représentation des équations mathématiques ;
- prise en charge des caractères spéciaux, des en-têtes et pied de pages, des notes de bas de page et de fin de chapitre, des références bibliographiques ;
- prise en charge des documents créés à l'aide de différents modèles de document ;
- production d'un document XML contenant l'ensemble des métadonnées des thèses françaises (Dublin Core et correspondance UNIMARC) proposé par le groupe d'experts « Métadonnées des thèses »- AFNOR CG 46/CN 357 GE 5 ;
- indexation et publication de documents PDF, HTML ou autre, créés à partir du document XML produit par la CEN.

Comparaison de deux chaînes d'édition numérique

1. La CEN Doc'INSA

1.1. Installation

1.1.1 Configuration minimale requise

La configuration suivante est conseillée pour utiliser la CEN Doc'INSA : processeur cadencé à 1,7 GHz, 512 Mo de mémoire vive SDRAM, 10 Go d'espace de stockage, système d'exploitation Windows gérant l'Unicode au niveau du noyau (NT4.0/2000/XP).

1.1.2 Installation de la CEN Doc'INSA

La CEN Doc'INSA fonctionne sous Windows 2000/XP. La séquence des traitements réalisés est définie par trois scripts[ii] qui doivent être exécutés en respectant un ordre déterminé. Pour que les scripts s'exécutent convenablement, il est important de vérifier, au cours de l'installation des logiciels qui constituent la CEN Doc'INSA, que les chemins d'installation sont conformes à ceux décrits dans les paragraphes suivants.

Les fichiers utilisés pour installer la CEN Doc'INSA ne sont pas distribués. Ils sont conservés par le service informatique de Doc'INSA, qui a mis à notre disposition, dans le cadre de cette étude, un CDROM contenant l'ensemble des fichiers accompagnés d'une documentation complète [11].

1.1.3 Installation de MS Word 2000

Les macros fournies avec le SDK de MathType ne fonctionnent pas avec les versions antérieures de ce logiciel. Configuration de MS Word 2000 :

[ii] Un script est un fichier de texte contenant des instructions destinées à être exécutées séquentiellement par un logiciel.

- copier le fichier « Modèle Thèse–Style classique.dot » dans le répertoire c:\Program Files\Microsoft Office\Startup ;
- choisir un niveau de sécurité minimum pour les macros. Pour cela, cocher la case niveau de sécurité bas qui se trouve dans l'onglet Outils>Macro>Sécurité>Niveau de sécurité.

1.1.4 Installation de MathType 5

MathType remplace l'éditeur d'équation Microsoft Equation 3. Aucune difficulté n'a été constatée au cours de l'installation.

1.1.5 Installation du SDK Math Type

Téléchargement de l'archive MTW51_SDK.exe, MathType Software Development Kit [40]. Cette application constitue une API[iii] pour invoquer des fonctions de MathType à partir de macros[iv] développées, par l'auteur [11] de la CEN Doc'INSA, en Visual Basic pour Applications.

1.1.6 Installation de XML Spy 4.4

Téléchargement de l'archive setup44.exe, XML Spy 4.4 [37], puis installation conformément aux instructions de l'éditeur.
XML Spy n'intervient pas dans la chaîne de conversion, il permet de vérifier les documents XML générés.

1.1.7 Installation d'UpCast

Téléchargement de l'archive upCastInstallerVM.exe, Upcast [42]. Ce logiciel a été configuré conformément aux instructions du développeur de la CEN de l'INSA[11]. Pour faciliter les futures installations et configurations de ce logiciel, nous avons sauvegardé dans le fichier upcast.conf (Annexe 2) les paramètres de configuration.
Chemin d'installation : C:\Program Files\infinity-loop\upCast
Le logiciel UpCast permet d'utiliser Word pour créer des documents XML sans se préoccuper de la syntaxe XML. Cette application génère du XML en associant des balises XML à chaque style utilisé dans le document Word. Dans la chaîne de

[iii] Une API ("application programming interface", Interface de programmation pour application) définie une interface qui permet à une application de communiquer avec le système d'exploitation ou d'autres services.
[iv] Une macro est une instruction informatique unique qui induit l'exécution de plusieurs instructions en langage machine, par extension, une macro désigne une (macro)commande qui accomplie plusieurs commandes.

conversion de Doc'INSA, UpCast génère un document XML à partir d'un document Word (*.doc) ou RTF.

1.1.8 Installation de Java (tm) 2

Téléchargement de l'archive j2sdk-1[1].4.2-nb-3.5-bin-windows.exe, J2SE(TM) v 1.4.2 with NetBeans(TM) IDE v 3.5 Cobundle [48]. Après l'installation, nous procédons à la modification des variables d'environnement :

CLASSPATH= C:\jars;C:\Program Files\j2sdk_nb\j2sdk1.4.2\lib

Path=%SystemRoot%\system32;%SystemRoot%;%SystemRoot%\System32\Wbem;C:\Program Files\j2sdk_nb\j2sdk1.4.2\bin

Puis, nous créons un répertoire jars à la racine du lecteur (disque dur ou partition) où se trouve la chaîne de traitement. Ce répertoire jars contient toutes les archives Java nécessaires au fonctionnement de la chaîne.

1.1.9 Installation de Gemini Solo

Téléchargement de l'archive soloDemo.exe, Gemini Solo [41].

Chemin d'installation : C:\Program Files\Iceni\

Gemini Solo est un utilitaire qui extrait des données à partir de documents PDF. Il est utilisé pour générer un document HTML à partir d'un document PDF généré par la chaîne.

1.1.10 Installation d'Acrobat

Ce logiciel est nécessaire au fonctionnement de Gemini Solo. Aucune difficulté n'a été constatée au cours de son installation.

1.1.11 Installation de Xalan

Téléchargement de l'archive xalan-j_2_5_1-bin.exe, Xalan [49].

Xalan-java est un processeur XSLT de transformation de documents XML en HTML, texte, et d'autres types de documents XML.

1.1.12 Installation de Saxon

Téléchargement de l'archive saxon6_5_2.zip, Saxon [43].

Saxon est un processeur XSLT permettant d'extraire et de restructurer des documents XML.

1.1.13 Installation de FOP

Téléchargement de l'archive fop-0.20.5rc3a-bin.tar.gz, FOP [50].

FOP est un processeur de formatage d'objet conçu pour générer des documents à partir d'une source et d'instructions XSL-T. Le principal format de sortie est le PDF. FOP est utilisé pour générer des documents PDF à partir du XML généré par la chaîne Doc'INSA.

1.1.14 Installation de DocBook XSL

Téléchargement de l'archive docbook-xsl-1.61.3.zip, DocBook XSL [45]. Cette archive contient des feuilles de styles qui permettent notamment de convertir un document DocBook en HTML.

1.1.15 Configuration de la CEN

Nous avons modifié les raccourcis contenus dans le répertoire C:\utilisateurs\fred\cen\Chaine\Exécution afin qu'ils correspondent à notre installation.

1.2. Évaluation

1.2.1 Les modèles de document

Des modèles de document Word, créés par l'auteur de la CEN Doc'INSA, sont à notre disposition. Ils sont destinés à faciliter le stylage et la conversion des thèses électroniques par la CEN Doc'INSA et ne permettent pas de styler des thèses en vue de leur traitement par la chaîne Cyberdocs.

Ces modèles de document extrêmement complexes sont associés à plusieurs barres d'outils permettant d'appliquer des styles, mais aussi d'effectuer des traitements « particuliers », tels que l'insertion d'une liste d'équation, des figures, des objets multimédias, d'une table des matières, d'une introduction, ou d'un errata *etc*. Ces traitements « particuliers », sont des macros développées en Visual Basic pour Applications, qui risquent de ne pas fonctionner correctement si elles sont utilisées avec une autre version du logiciel Word que celle utilisée par l'auteur de la CEN Doc'INSA.

1.2.2 La DTD DocBook

Les thèses électroniques traitées par la CEN Doc'INSA sont converties en documents XML, instances de la DTD DocBook XML. La DTD DocBook est conçue pour rédiger de la documentation technique. Elle est cependant utilisée pour créer des documents de types extrêmement variables, tels que : des systèmes d'aide, des sites Web, des livres, des pages de liens, des FAQ (Foires aux question/Frequently Asked Questions), des cours, des articles, des rapports, des spécifications, des guides « HowTo » bien connus dans le monde Unix/Linux, *etc*.
Cette DTD a une structure modulaire qui lui permet de répondre aux exigences d'utilisateurs divers, d'autant que les outils permettant de convertir les documents XML DocBook en différents formats de publication sont nombreux. De plus amples informations sur la DTD DocBook sont disponibles en ligne [23].

1.2.3 Récupération et poursuite du projet de CEN Doc'INSA

Le développement de cette chaîne est interrompu depuis septembre 2002. Un prototype a été réalisé par M. Sébastien Ioannitis, ingénieur de l'INSA de Lyon, dans le cadre de son projet de fin d'étude [11]. Mais aucune version définitive ne fut présentée, en raison de l'absence, au moment du développement de la chaîne, d'outils fiables permettant de réaliser l'ensemble des tâches nécessaires.
Pour évaluer la CEN Doc'INSA, nous avons récupéré la documentation et les exécutables préparés par l'auteur. Le premier problème rencontré fut l'absence d'une installation fonctionnelle de la CEN : celle qui avait été mise en place par l'auteur utilisait des logiciels en version d'évaluation, dont la date limite d'usage avait expiré bien avant le début de notre étude. Nous avons installé la CEN Doc'INSA en nous conformant aux recommandations de l'auteur. Ce qui nous a permis de constater la présence de dysfonctionnements, dont les causes n'ont pu être localisées. Nous avons alors tenté de remplacer les logiciels que l'auteur présentait comme instables, par des versions plus récentes et stables.
Ces modifications n'ont pas permis de faire fonctionner la CEN convenablement, d'autant que les indications laissées par l'auteur de la CEN ne permettaient pas de configurer les nouvelles versions de ces logiciels, dont les interfaces et les fonctionnalités avaient significativement évoluées.

1.2.4 Traitement des documents LaTeX

La CEN Doc'INSA ne traite pas les documents $L^{A}T_{\varepsilon}X$.

1.2.5 Le XML produit par la CEN Doc'INSA

Nous avons malgrès tout réussi à produire des documents XML, qui se sont révélés bien formés et valides par rapport à leur DTD. La CEN Doc'INSA traite correctement les équations mathématiques et produit du MathML bien formé et valide.

1.2.6 Coût de la mise en place de cette chaîne de traitement

Applications	Tarifs (€)
GeminiSolo	159
Acrobat	149
UpCast	100
MathType	129
WindowsXP Professional	299
Office 2000 Standard Edition	399
Total	1235

Tableau 1 : Coût des applications commerciales constituant la CEN Doc'INSA, constatés au 10 septembre 2003.

Si une version fonctionnelle de la CEN Doc'INSA venait à être développée sur la base de la version existante, le coût lié à l'utilisation des logiciels qui la composent serait de 1235 €.

1.2.7 Conclusion

La CEN Doc'INSA utilise de nombreux logiciels, dont certains sont commerciaux. En dehors des coûts liés à l'utilisation de tels logiciels, des limitations importantes sont à prévoir quant aux possibilités de modification de ces logiciels, leur code source n'étant pas disponible. De telles modifications n'auraient pas été nécessaires si une version définitive, parfaitement fonctionnelle de la CEN Doc'INSA avait été produite. Mais, en l'état actuel du projet, il convient de s'interroger sur l'opportunité de poursuivre son développement.

2. La CEN Cyberdocs

2.1. Installation

2.1.1 Configuration minimale requise

La CEN Cyberdocs fonctionne sur un système informatique conçu pour accomplir des tâches bureautiques classiques : processeur cadencé à 500 MHz, 300 Mo de mémoire vive, 10 Go d'espace de stockage, un système d'exploitation Unix/Linux et Windows NT/2000. Une configuration plus performante (15 Go d'espace de stockage, processeur cadencé à 1 GHz) sera nécessaire pour effectuer les mêmes tâches dans un environnement Windows XP. L'utilisation de la plate-forme de publication SDX nécessitera une installation de type serveur, capable de répondre aux sollicitations des utilisateurs.

2.1.2 Choix d'un système d'exploitation

La CEN Cyberdocs est multiplate-forme, elle fonctionne sur des ordinateurs équipés de systèmes d'exploitation Windows NT/2000/XP et Unix/Linux. Ceci est possible grâce à :

- des exécutables codés en Java, langage interprété par une machine virtuelle Java, indépendant du système d'exploitation utilisé ;
- des scripts d'installation et de configuration portant des extensions *.bat et *.sh. Les scripts portant les extensions *.bat sont utilisés pour l'installation dans un environnement Windows. Les scripts portant les extensions *.sh sont utilisés pour l'installation dans un environnement Unix/Linux.

Nous n'avons pas réussi à installer la plate-forme Cyberdocs sur un micro-ordinateur ayant un système d'exploitation Windows 98. En lisant les scripts d'installation de la chaîne Cyberdocs, nous avons trouvé des instructions *pushd* et *popd* qui n'appartiennent pas à la version du langage DOS disponible sur un tel système. À défaut de disposer, au début de cette étude, d'un ordinateur utilisant un système d'exploitation Windows NT/2000/XP ou de trouver des instructions de substitution valides permettant de modifier le script up.bat, nous avons préféré installer la chaîne Cyberdocs dans un environnement Linux.

Ce choix s'est révélé extrêmement intéressant, car il nous a permis de faire remonter des informations aux développeurs, informations qui se sont parfois révélées utiles pour mettre en évidence des problèmes spécifiques à l'utilisation de la chaîne sous Linux. En outre, l'utilisation de la CEN Cyberdocs dans un environnement Linux s'inscrit dans la démarche du projet Cyberthèses, qui consiste à proposer des outils libres de droits, utilisables par un grand nombre d'université à travers le monde. Or un tel objectif ne peut être sérieusement envisagé sans une prise en compte des aspects économiques qu'une telle participation impose. D'où l'intérêt de disposer d'une CEN libre de droit fonctionnant dans un environnement libre de droit.

Système d'exploitation	Version	Installation	Commentaires
Debian GNU/Linux	3 r01 Woody	échec	Problème à l'initialisation d'OpenOffice.org, « java.net.ConnectException: Connection refused »
	2 r01 Potato		
Linux Redhat	7.2	réussite	
	9		
Linux Mandrake	7.1		
Windows	98SE	échec	Problème avec les scripts d'installation
	XP	réussite	

Tableau 2 : Description des installations de la chaîne Cyberdocs effectuées.

L'installation fut aisée sur un micro-ordinateur ayant un système d'exploitation Linux Redhat 7.2. Quelques modifications minimes des scripts d'installation furent réalisées, soumises à l'expertise des développeurs du projet Cyberdocs, et prises en compte.

Le tableau précédent (Tableau 2) regroupe les informations recueillies au cours des différentes installations que nous avons effectuées. L'échec de l'installation de la chaîne sur un système d'exploitation Windows 98 est d'autant moins étonnante que les documents livrés avec la CEN Cyberdocs indiquent clairement que la CEN n'est pas conçue pour fonctionner sur ce système d'exploitation.

Nous avons installé la chaîne Cyberdocs sur une version XP de Windows, système d'exploitation qui possède, à l'instar des version NT et 2000, un DOS plus évolué que celui livré avec Windows 98, qui gère les instructions « *pushd* » et « *popd* ».

La réussite de notre première installation de la chaîne Cyberdocs sur un système d'exploitation Linux Redhat 7.2 nous a encouragé à travailler sous Linux. D'autres installations furent réalisées afin de nous assurer de la portabilité de la chaîne. Nous avons ainsi constaté que, malgrès la réusssite des installations sous Linux

Redhat 7.2, Linux Redhat 9 et Linux Mandrake 7.1, il est impossible d'utiliser la chaîne sur les versions Potato et Woody du système d'exploitation Debian GNU/Linux, en raison d'une erreur «java.net.ConnectException: Connection refused» empêchant l'initialisation du logiciel OpenOffice.org et la conversion des thèses.

2.1.3 Description générale de l'installation de Cyberdocs

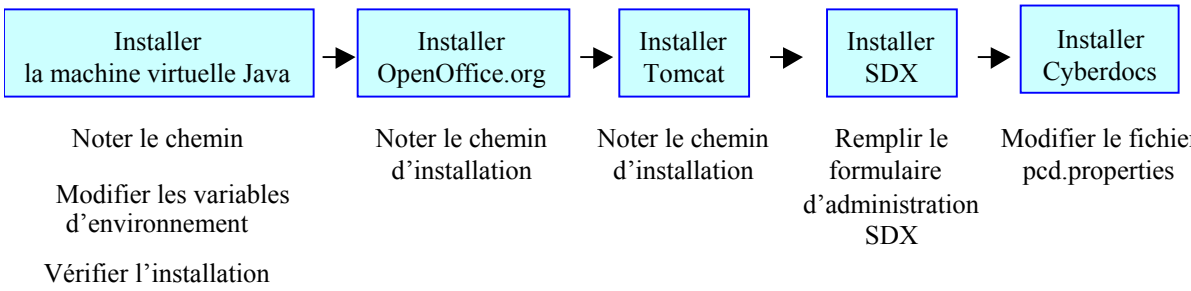


Figure 2 : Description de l'installation de Cyberdocs et des logiciels associés.

L'installation de la CEN Cyberdocs (Figure 2) débute par la mise en place d'une machine virtuelle Java. La suite bureautique OpenOffice.org, Tomcat, SDX, Cyberdocs peuvent ensuite être installés, en se conformant aux instructions fournies par les auteurs de ces logiciels. Une description des installations réalisées sous Linux et Windows XP est présentée dans les chapitres suivants.

2.1.4 Installation sous Linux

2.1.4.1 Installation de Java(tm) 2 SDK

Téléchargement du fichier j2sdk-1.4.1_03-linux.i586-rpm.bin [47].

L'utilisateur root modifie les droits du fichier téléchargé pour le rendre exécutable, puis lance le programme correspondant. Après avoir donné son accord sur les conditions d'utilisation de ce logiciel, un fichier *.rpm est créé dans le répertoire courant. Ce fichier *.rpm permet d'installer Java(tm) 2 SDK, standard Edition 1.4.1_03 :

 # ls –la

*# chmod u+x j2sdk*rpm.bin*

 # ./j2sdk-<>-rpm.bin

 # ls

*# rpm –ivh j2sdk*rpm*

Localisation du répertoire contenant le fichier java que nous venons d'installer :

 # find / -name java

Mise à jour des variables d'environnement :

 # vi /root/.bash_profile

Nous ajoutons les lignes suivantes dans le fichier .bash_profile que nous venons d'ouvrir avec l'éditeur vi :

export JAVA_HOME=/usr/java/j2sdk1.4.1_03

export PATH=${PATH}:${JAVA_HOME}/bin

(Pour que les modifications soient prises en compte, il convient de quitter, puis de réouvrir une session)

Vérification de l'installation de Java(tm) 2 SDK, standard Edition 1.4.1_03 :

 # java –version

Une installation correcte permet de lire le message suivant :

Java (TM) 2 Runtime Environment, Standard Edition (build 1.4.1_03-b02)

Java HotSpot (TM) Client VM (build 1.4.1_03-b02, mixed mode)

2.1.4.2 Installation d'OpenOffice.org

Téléchargement du fichier Ooo_1.0.3.1_LinuxIntel_install.tar.gz [44].

Après décompression de l'archive, nous installons la suite bureautique OpenOffice.org :

*# tar –xvzf Ooo*tar.gz*

 # ls

 # cd install

 # ls

 # ./setup

Au cours de l'installation, nous choisissons un chemin d'installation : /root/OpenOffice.org1.0.3. Il suffit alors de vérifier le bon fonctionnement des logiciels.

2.1.4.3 Installation de la chaîne Cyberdocs

Les sources d'information

L'installation de la chaîne Cyberdocs a été réalisée conformément aux recommandations fournies les développeurs : Martin Sévigny et Myriam Delperier de la société AJLSM. Pour résoudre les problèmes non documentés, nous avons

consulté les archives de la liste cybertheses-users@cru.fr et posé des questions aux développeurs de Cyberdocs qui se sont montrés extrêmement disponibles.

Le téléchargement des sources

La chaîne Cyberdocs étant en développement durant la réalisation de cette étude, nous avons récupéré, plusieurs fois par semaine en utilisant CVS, la dernière mise à jour de la chaîne de conversion des thèses numériques diffusée par le serveur du CRU. Les informations pratiques pour le téléchargement sont disponibles en ligne [6] et retranscrites ici.

Créer puis se déplacer dans un nouveau répertoire :

 # mkdir cybersources

 # cd cybersources

Puis, entrer les instructions suivantes dans un Shell :

 # cvs -d:pserver:cybertheses@sourcesup.cru.fr:/cybertheses login

CVS password: cybertheses (taper :cyberthèses)

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

Les sources se trouvent alors dans le répertoire courant. Nous éditons le fichier pcd.properties (Annexe 1) et personnalisons les variables « dossier.installation.up » et « Openoffice.home » qui reçoivent les valeurs présentées dans le document pcd.properties.

Puis nous modifions les droits des fichiers build-ant.sh et build.sh :

 # chmod u+x build-ant.sh build.sh

Et nous procédons à l'installation de la chaîne Cyberdocs :

 # ./build-ant.sh

 # ./build.sh

Installation et configuration du serveur SDX

Installation de tomcat-4.1.24

Téléchargement du fichier tomcat4.1.24-full.2jpp.noarch.rpm, disponible en ligne [38].

Installation du rpm :

 # rpm –ivh tomcat4.1.24-full.2jpp.noarch.rpm

Ouvrir une session root, utiliser ntsysv pour activer le lancement du service Tomcat 4 au démarrage :

 # ntsysv

Installation de SDX-2.1

Téléchargement du fichier sdx.war, disponible en ligne [46].

Placer le fichier sdx.war dans le répertoire webapps du moteur de servlets[v] Tomcat. Lors du démarrage du serveur, cette archive est automatiquement décompressée, mettant ainsi en place une arborescence de fichier permettant de configurer et d'utiliser SDX.

En se plaçant dans le répertoire qui contient les sources de Cyberdocs, nous exécutons le script build.sh avec les paramètres suivants :

 # ./build.sh installation-sdx

Vérification de l'installation : redémarrer Tomcat, puis envoyer la requête suivante au navigateur Internet :

http://localhost:8080/sdx

Lors du premier accès au serveur SDX, l'application demande le nom et le mot de passe du « super-utilisateur » SDX. Ces paramètres peuvent différer de ceux qui permettent d'ouvrir une session sous Linux. Ces informations pourront être modifiées ultérieurement en complétant le formulaire qui se trouve à l'URL : *{$host}/sdx/sdx/admin/*

Il est alors possible de visualiser l'application sdxtest distribuée avec l'application en envoyant la requête suivante au navigateur Internet :

{$host}/sdx/sdxtest

2.1.5 Installation sous Windows XP

2.1.5.1 Téléchargements

Les sites présentés dans les chapitres décrivants l'installation de la chaîne Cyberdocs sous Linux permettent de télécharger les logiciels destinés au système d'exploitation Windows, sauf indication contraire.

2.1.5.2 Installation de Java(tm) 2 SDK

[v] Un servlet est un morceau de code développé en Java qui ajoute des fonctionnalités à un serveur (généralement un serveur Web). C'est équivalent à un applet (application Java qui ajoute des fonctionnalités au navigateur Web) qui tournerait côté serveur.

L'installation de l'exécutable se fait selon la méthode aviaire classique : un double-clic. Noter le chemin du répertoire d'installation de Java.

La mise à jour des variables d'environnement :

Démarrer>Paramètres>Panneau de configuration>Système>Avancé>Variables d'environnement.

Dans la fenêtre qui s'ouvre, cliquer sur le bouton nouveau situé dans la zone Variables Système. Ajouter les informations suivantes, conformément à l'installation que vous avez réalisée :

Nom de variable : CLASSPATH

Valeur de la variable : C:\Program Files\j2sdk_nb\j2sdk1.4.2\lib

Puis cliquer sur *OK*, et *Nouveau* :

Nom de variable : JAVA_HOME

Valeur de la variable C:\Program Files\j2sdk_nb\j2sdk1.4.2

Puis cliquer sur *OK*, et *Nouveau* :

Nom de variable : PATH

Valeur de la variable {ne rien effacer :o)};C:\Program Files\j2sdk_nb\j2sdk1.4.2

2.1.5.3 Installation d'OpenOffice.org 1.0.x

Seules les versions 1.0.x d'OpenOffice.org sont compatibles avec la CEN Cyberdocs. Noter le chemin du répertoire d'installation d'OpenOffice.org

2.1.5.4 Installation de Tomcat

L'exécutable jakarta-tomcat-4.1.27-LE-jdk14.exe est disponible en ligne [39].

Lancer l'installation de Tomcat en double cliquant sur l'exécutable. Au cours de l'installation, demander à installer Tomcat en tant que service et compléter le formulaire d'information :

HTTP/1.1 Connector Port 8080

Usernameadmin

Password*****

Noter le chemin d'installation de Tomcat.

Vérification de l'installation : ouvrir un navigateur Web est indiquer l'adresse suivante :

http://localhost :8080/

Une page d'accueil doit apparaître.

2.1.5.5 Installation de SDX

Placer l'archive sdx.war dans le répertoire webapps de Tomcat et redémarrer Tomcat.

2.1.5.6 Installation de la chaîne Cyberdocs

Éditer le fichier pcd.properties selon vos besoins. Nous modifions certaines variables en leur attribuant les valeurs suivantes [vi]:

dossier.installation.up=../cyber

openoffice.home=C:\\Program Files\\openoffice.org1.1.0

dossier.installation.consultation=C:/Program Files/Apache Group/Tomca 4.1/webapps/sdx

2.2. Évaluation

2.2.1 Description

2.2.1.1 La DTD TEI-Lite

Les thèses électroniques (documents *.doc) traitées par la chaîne Cyberdocs sont converties en documents XML, instances de la DTD TEI-Lite XML version 1.3.

Cette DTD est conçue pour baliser/coder des textes existants ou en cours d'élaboration, destinés à être traités par des logiciels.

Un texte balisé conformément à la DTD TEI(-Lite) possède la structure suivante [19] :

- un en-tête TEI balisé comme un élément <teiHeader> ;
- la transcription du texte lui-même, ou corps du texte balisé comme un élément <text>.

L'en-tête TEI contient jusqu'à quatre parties, regroupant des informations que l'on trouve sur la page de titre d'un texte imprimé, *i.e.* :

- une description bibliographique du texte électronique ;
- une description de la manière dont il a été codé ;
- une description non-bibliographique du texte, le « profil » du texte ;
- un historique des révisions.

Le corps du texte comporte les éléments suivants :

[vi] La syntaxe des valeurs de ces variables ne fait appel à aucune logique. Elle sont traitées par des programmes Java, et leur syntaxe dépend de ceux-ci.

- <front> regroupe les éléments liminaires, *i.e.* situés avant le début du texte lui-même : en-têtes, page de titre, préfaces, dédicaces, *etc*.
- <group> regroupe plusieurs textes unitaires ou groupes de textes ;
- <body> regroupe le corps entier d'un texte unitaire seul, à l'exclusion de toute pièce liminaire ou annexe ;
- <back> regroupe toutes les annexes qui suivent le texte principal.

De nombreuses balises permettent de représenter la structure de textes littéraires (textes poétiques, dramatiques, ou autres). Ainsi, les balises permettent de décrire les informations suivantes :

- changements des styles de caractères ou des alternances typographiques ;
- numérotation des lignes et des pages ;
- citations et éléments associés ;
- expressions ou mots étrangers ;
- notes, références croisées et liens ;
- interventions éditoriales, omissions, effacements et ajouts ;
- types de données : noms, dates, chiffres et abréviations ;
- citations bibliographiques ;
- formules mathématiques ou chimiques ;
- tables, figures et graphiques ;
- documentation technique ;
- jeux de caractères et signes diacritiques ;

Le choix d'une DTD dépend aussi des outils qui lui sont associés. La DTD TEI-Lite bénéficie d'une gamme de logiciels étendue qui permettent de produire, éditer, transformer et publier des documents. Une liste disponible en ligne est proposé par les éditeurs de la DTD TEI-Lite [22].

2.2.1.2 Cyberdocs, Linux, XML et l'Unicode

En travaillant sous Linux, nous avons été confronté à des difficultés liées à l'encodage des caractères. Tous les documents XML de la CEN Cyberdocs utilisent l'UTF-8, ce qui crée des difficultés d'affichage des documents lorsque le système ou les logiciels utilisés ne supportent pas l'Unicode. Les caractères spéciaux, les accents sont remplacés par 2 caractères, ou sont omis, ce qui en complique la lecture.

L'Unicode définit une table de caractères étendue susceptible de contenir tous les caractères utilisés à travers le monde. Il permet de résoudre les problèmes liés à l'utilisation de tables de caractères codés sur 8 bits, *i.e.* un octet, telle que l'ISO-8859-1, qui ne peut contenir que 256 caractères. Une table constituée de caractères codés sur 8 bits ne permet pas de représenter tous les caractères utilisés dans le monde. L'utilisation de plusieurs tables est donc nécessaire, chacune devant être définie avant d'être utilisée. Bien que les logiciels de traitement de texte modernes soient capables d'utiliser plusieurs tables de caractères, il est impossible de rédiger un document contenant du texte simple associant des caractères provenant de différentes tables de caractères. Cette limitation concerne évidemment les documents XML.

La solution adoptée pour résoudre ce problème est l'Unicode. Mais son utilisation soulève des difficultés techniques liées à la méthode de transport des caractères Unicode *via* des paquets de 8 bits, *i.e.* un octet, taille communément utilisée par les ordinateurs, et les connections réseaux telles que TCP/IP. Quatre solutions existent :

- UTF-8 : 128 caractères codés sur 1 octet, 1920 caractères codés sur 2 octets, 63488 caractères codés sur 3 octets, les autres sur 4, 5, 6 octets.
- UCS-2 : tous les caractères sont codés sur 2 octets, une limite étant fixée à 65536 caractères.
- UTF-16 : 65536 caractères codés sur 2 octets, les autres, jusqu'à 11144112 sur 4.
- UCS-4 : tous les caractères sont codés sur 4 octets.

L'Unicode UTF-8 présente l'avantage de ne pas pénaliser les pays européens, utilisant les tables de caractères ASCII US et ISO-8859-1. De plus, son utilisation ne nécessite aucune modification des programmes existants.

Pour de plus amples explications, on consultera la page de manuel Linux UTF-8 (7) [36].

2.2.1.3 Prise en compte des métadonnées des thèses françaises

Les métadonnées des thèses françaises sont définies par le groupe d'experts AFNOR CG 46/CN 357 GE 5. Elles seront disponibles prochainement.

Pour qu'elles soient prises en compte au moment de l'indexation, un fichier contenant les métadonnées doit être créé dans le même dossier que le document XML. Ce fichier que l'on éditait manuellement, avant la sortie de la version *beta* de la CEN Cyberdocs, porte le nom du document XML, auquel on ajoute le suffixe -md.xml. Depuis, un module de gestion est associé à la CEN. Il facilite la saisie les métadonnées depuis un formulaire Web.

2.2.1.4 Formats des documents traités par la CEN Cyberdocs

Les formats de documents suivants sont traités par la CEN Cyberdocs :

- format OpenOffice.org (*.sxw) ;
- format Word (*.doc) créés avec Word ;
- format Word (*.doc) créés avec OpenOffice.org ;
- format RTF (*.rtf) créés avec Word ;
- format RTF (*.rtf) créés avec OpenOffice.org.

Au cours de la conversion des documents Word par la chaîne, un document intermédiaire au format *.sxw est généré, à partir duquel, un document XML est produit. Le format *.sxw est un format pivot dans le traitement réalisé par la CEN.

OpenOffice.org utilise le XML comme format d'échange. Les fichiers *.sxw sont des archives compressées au format ZIP. Après décompression, on peut observer un ensemble de fichiers XML contenant les données du document sauvegardé :

- content.xml contient le texte du document ;
- meta.xml regroupe des informations sur le document ;
- settings.xml contient d'autres informations sur les paramètres du document (historique des modifications, source de données connectée, *etc.*) ;
- styles.xml contient les styles définis pour le document ;
- meta-inf/manifest.xml décrit la structure du fichier XML ;
- le sous-répertoire Pictures regroupe les images associées au document.

Nous avons considéré que les documents créés à partir d'OpenOffice.org au format *.sxw sont traités par la CEN de façon équivalente à ceux créés à l'aide de Word, afin de ne réaliser les essais de la CEN qu'à partir de documents Word. Cette démarche se justifie par le fait que 80% des thèses rédigées et soutenues par les étudiants de l'INSA sont au format Word, et que le traitement de ces documents constitue actuellement une priorité.

2.2.1.5 Description de la structure de fichiers

Le fonctionnement de la chaîne de conversion est conditionné par la mise en place d'une structure de fichiers (arborescence) permettant, aux applications qui la constituent, de récupérer et de sauvegarder des fichiers au cours du traitement.

Le nom du dossier racine est défini lors de l'installation de la chaîne : il correspond à la variable dossier.installation.up déclarée dans le fichier pcd.properties.

Les deux sous-répertoires production et outils contiennent respectivement les fichiers traités par la CEN et les fichiers qui réalisent les traitements.

L'arborescence (Figure 3) constituée par les dossiers numérotés de 1 à 4, permet d'accéder à un répertoire source, dans lequel se trouve une thèse que l'on souhaite convertir et un fichier nom-thèse-md.xml contenant les métadonnées de thèses que l'on souhaite associer au document *.xml. Après avoir mis en place une telle arborescence, il faut créer un fichier nom–thèse.sh que l'on place dans le répertoire 2. Le fichier nom–thèse.sh, dans l'exemple de la Figure 3 contient les instructions suivantes :

```
#!/bin/sh
# Exemple de fichier de commandes pour un document en particulier
initial=$PWD
cd ../..
./up.sh tout thèse.doc institution nom-thèse code_style fr année
cd $initial
```

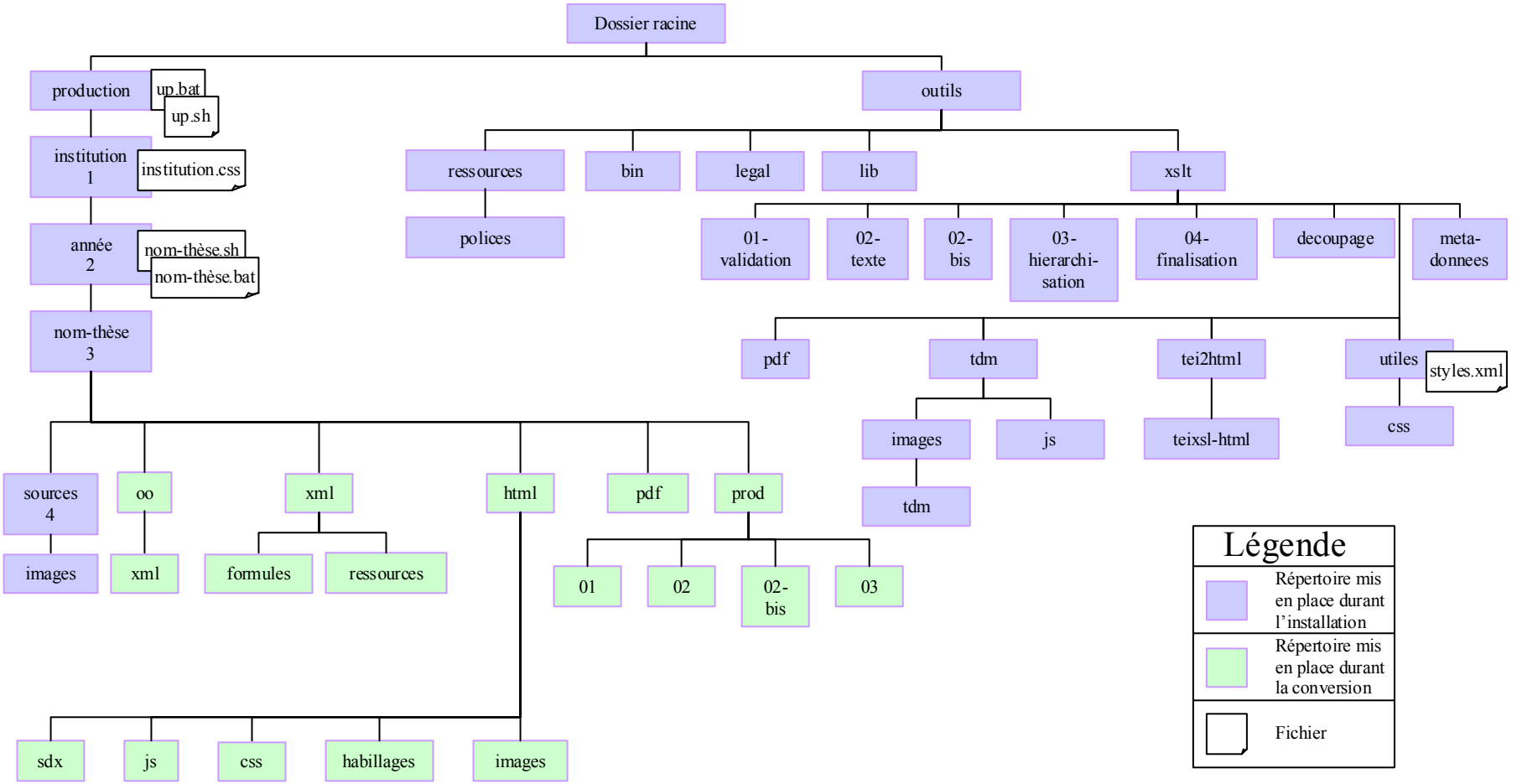


Figure 3 : Arborescence de la CEN Cyberdocs après installation

Ce script appelle le script up.sh qui se trouve dans le répertoire production, et lui passe les paramètres suivants :

tout indique à la chaîne d'effectuer tous les traitements possibles.

thèse(n).doc nom de la thèse à convertir.

institutionnom du répertoire 1, définit le code institution qui apparaît dans le document XML.

nom-thèsenom du répertoire 3.

code_style nom du style utilisé pour mettre en forme la thèse à traiter.

fr code pays.

année(n) nom du répertoire 2.

Le traitement d'une thèse débute par l'exécution du fichier nom-thèse.sh (nom-thèse.bat sous Windows). Si le traitement arrive à son terme, les répertoires créés sont :

oo contient le résultat de la conversion du document effectuée par OpenOffice.org ;

xml contient le document final, résultat de la conversion, au format xml ;

html contient une version HTML statique du document, la page de départ étant index-frames.html. Le sous-répertoire sdx contient des versions HTML imprimables du document et de ses parties ;

pdf contient une version PDF du document et de ses parties, en plus des fichiers XSL-FO qui ont servi à les produire ;

prod contient les fichiers produits au cours des conversions et classés en fonction des étapes de traitement.

2.2.2 Méthodes d'évaluation

Pour évaluer la CEN Cyberdocs, nous avons procédé à son installation puis à des essais. Nous disposions de deux micro-ordinateurs pour tester la CEN Cyberdocs :

- micro-ordinateur 1 : processeur Intel Pentium II cadencé à 350 MHz, mémoire vive 64 Mo SDRAM, espace de stockage 4 Go, Windows 98, Office 97 ;
- micro-ordinateur 2 :Processeur Intel Pentium IV cadencé à 1,7 GHz, mémoire vive 512 Mo SDRAM, espace de stockage 40 Go, Linux Redhat 7.2, OpenOffice.org.

Il convenait d'utiliser ces ressources pour mettre en place notre environnement de travail. Nous avons donc installé une distribution Linux Redhat 7.2 sur le micro-ordinateur 2, ainsi que l'ensemble des outils nécessaires au fonctionnement de la chaîne Cyberdocs. Le micro-ordinateur 1, sur lequel était installée la suite

bureautique Microsoft Office 97 permettait d'éditer des documents Word. La principale contrainte liée à cet environnement de travail était de transférer les documents Word du poste 1 vers le poste 2, sur lequel était installée la CEN Cyberdocs.

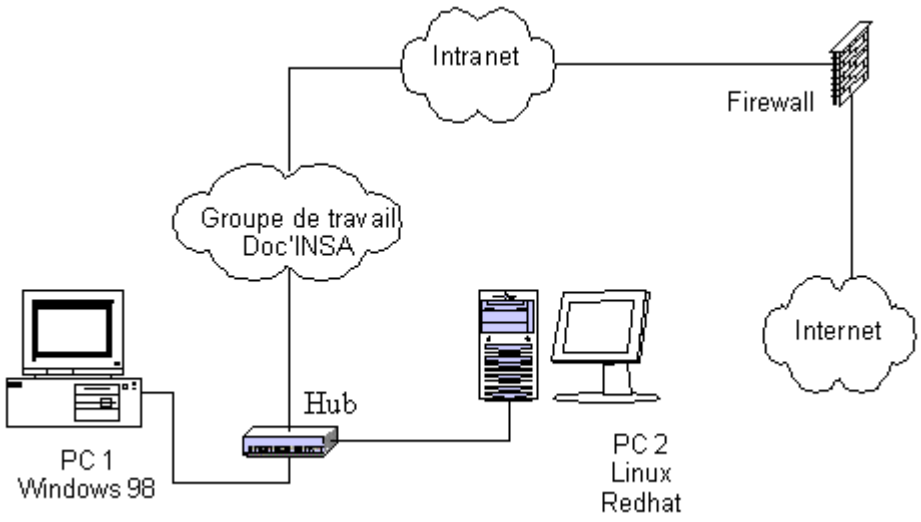


Figure 4 : Schéma descriptif de l'installation mise en place pour tester de la chaîne Cyberdocs

L'installation de SAMBA et du logiciel LinNeighborhood nous a permis d'accéder aux données du micro-ordinateur 1 depuis le micro-ordinateur 2. Pour cela, nous avons associé au groupe de travail Doc'INSA constitué exclusivement de micro-ordinateurs Windows, le micro-ordinateur 2 utilisant un système d'exploitation Linux. Il devenait possible de créer, à partir de cette installation (Figure 4), un document Word sur le micro-ordinateur 1, puis de le récupérer *via* le réseau sur le micro-ordinateur 2 et de le convertir en document XML en utilisant la CEN Cyberdocs.

Une autre solution avait été envisagée. Elle consistait à créer une partition Windows sur le disque dur du micro-ordinateur 2 et d'utiliser le logiciel Wine pour créer des documents Word depuis Linux. Cette solution aurait permis de travailler sur un seul poste de travail, et par conséquent de ne plus être tributaire du bon fonctionnement du réseau. La production et le traitement des thèses électroniques aurait ainsi été entièrement effectués sous Linux.

2.2.2.1 Description des documents d'évalutation

Les essais ont été réalisés à partir de trois fichiers Word destinés à être convertis par la CEN, que nous nommerons « documents d'évaluation ». Ces documents devaient respecter les contraintes de contenu et de mise en forme imposées par la chaîne.

Les fichiers Beleca.doc et Dieng.doc sont des documents d'évaluation créés à partir du modèle de document lyon2. Izbel.doc est basé sur un modèle de document insa, créé pour l'occasion.

Le style lyon2, style proposé aux doctorants inscrits à l'université Lyon 2, est pris en charge par défaut par la chaîne Cyberdocs. Cette prise en charge dépend d'un fichier styles.xml distribué avec les sources du projet Cyberdocs. Le document Izbel.doc, basé sur le modèle insa nous a permis de valider la possibilité de traiter *via* la CEN Cyberdocs, des documents créés à partir de modèles de documents différents du modèle lyon2.

La création d'un modèle de document n'est pas une procédure triviale, même si le principe général ne présente aucune difficulté théorique. Le stylage d'un document Word permet d'associer, à la mise en forme d'un paragraphe ou d'une suite de caractères, une balise XML décrite dans la DTD TEI-Lite. Les modalités de mise en oeuvre d'un tel processus sont toutefois complexes. La chaîne Cyberdocs utilise un fichier styles.xml pour associer les styles du modèle Word à des styles intermédiaires reconnus par OpenOffice.org, et finalement convertis en balises appartenant à la DTD TEI-lite. Ce traitement séquentiel est suffisamment complexe et obscur pour considérer, en l'état de nos connaissances actuelles, qu'il est impossible d'ajouter ou de supprimer les styles définis dans le modèle lyon2. Par conséquent, nous avons créé un modèle insa faisant correspondre à chaque style défini dans le modèle lyon2, un style personnalisé. Et avons ainsi soigneusement évité de créer ou de supprimer de nouveaux styles.

2.2.2.2 Stylage des thèses

Cette étape est essentielle, les erreurs ayant systématiquement des conséquences néfastes dans les phases ultérieures du traitement réalisé par la CEN.

Les essais et erreurs, ainsi que l'étude des documents et des messages générés par la chaîne au cours du traitement des documents d'évaluation nous permettent de proposer les conseils suivants pour réaliser un stylage convenable :

La page de titre doit impérativement contenir les informations suivantes :

- un paragraphe contenant le titre de la thèse(style code titre-these) ;
- un paragraphe contenant le sous titre de la thèse, ou un paragraphe vide (style code sous-titre) ;
- au moins un paragraphe contentant le nom de l'université (style code universite) ;
- un paragraphe discipline(style code discipline) ;
- un paragraphe nom du directeur de thèse (style code directeur) ;
- au moins un paragraphe intitulé du grade (style code grade) ;
- un paragraphe nom de l'école doctorale (style code ecole-doct);
- un paragraphe nom de l'auteur (style code auteur) ;
- un paragraphe date de dépôt (style code depot) ;
- un paragraphe composition du jury(style code jury) ;
- un paragraphe numéro d'ordre de la thèse(style code no-officiel) ;
- un paragraphe mention du droit d'auteur(style code copyright) .

Les parties liminaires contiennent :

- la page de dédicaces constituée de :
    - d'un paragraphe contenant les dédicaces (style code dedicace).
- d'autres parties liminaires :
    - un paragraphe de titre (style code titre-front)
    - un ou plusieurs paragraphes proprement dits (style code text). Il est possible d'insérer des images, *etc*.

Les parties du document :

- La première partie du document est l'introduction. Elle contient :
    - un paragraphe de titre (style code intro) ;
    - un ou plusieurs paragraphes (style code text), des images, tableaux, *etc*.
- Les parties suivantes du document contiennent : aucun ou un paragraphe de titre (style code partie). Si un paragraphe portant un style code partie est créé,

il doit impérativement être suivi d'un paragraphe (style code text). Sinon, le premier paragraphe sera de niveau 1 (style code heading1).

- La dernière partie du document contient :
    - un paragraphe de titre (style code conclu) ;
    - un ou plusieurs paragraphes (style code text), des images, tableaux, *etc*.

Les tableaux contiennent :

- un paragraphe qui précède le tableau (style code legende-tab ou caption)
- un paragraphe contenant un tableau (style code text). La première ligne est mise en valeur automatiquement, elle correspond aux titres des colonnes du tableau ;

Les équations : apparaissent dans un paragraphe seul (style code text), sans légende ;

Les images : la chaîne ne traite que les formats *.jpg, *.tif, *.bmp, *.gif et *.png. Deux méthodes de stylage peuvent être utilisées : la méthode classique Cyberthèses consiste à extraire les images du document Word original en prenant soin d'insérer une référence, le nom de l'image, à la place où celle-ci résidait. L'autre méthode, propre à Cyberdocs, consiste à associer un code style legende-fig ou caption au paragraphe qui précède l'image. Cette dernière recevant le style figure ou image-ligne.

Les images créées dans Word, au format WMF (Microsoft Windows Metafile Format) constituent un ensemble de données hétérogène, contenant des images vectorielles et point par point, qui n'est pas traité par la CEN Cyberdocs. Il convient de grouper l'ensemble des éléments constituant l'image, de le couper, puis de sélectionner Collage spécial dans le menu Edition > Image en mode point, en s'assurant que bouton « Dissocier du texte » n'est pas sélectionné. Si l'option Image en mode point n'apparaît pas, sélectionnez Image, puis recommencez et sélectionnez Image en mode point.

2.2.2.3 Les styles traités par la CEN Cyberdocs

Nous avons tenté de récupérer des documents décrivant la prise en charge et le traitement des styles par Cyberdocs. N'ayant trouvé aucun document *ad hoc*, nous

avons eu recours au forum de discussion des utilisateurs de Cyberdocs pour récupérer quelques informations. Il apparait que :

- le fichier /outils/xslt/utiles/styles.xml associe un code institution à des styles. Une modification de ce fichier est donc nécessaire pour pour que les styles d'un modèle de document soient reconnus par la chaîne. L'édition du fichier styles.xml consiste à ajouter un code institution, puis à associer à chaque code style existant, le nom d'un style issu de notre modèle de document Word ;
- certains traitements effectués sur les styles sont codés « en dur » dans la plate-forme Cyberdocs. Des modifications des programmes développés en Java seraient par conséquent nécessaires pour que des styles ne correspondant à aucun style existant soient complètement pris en charge par la chaîne. En effet, une telle modification revient à créer une nouvelle balise XML, ce qui implique de modifier la DTD, les traitements XSL, et les exécutables Java qui constituent la CEN Cyberdocs.
- certains styles sont obligatoires. En leur absence, des traitements ne sont pas effectués correctement.
- le modèle de document lyon2 ne contient pas de style permettant de styler les informations contenues dans la page résumé. La chaîne évalue les titres des parties liminaires, et reconnaît les valeurs résumé et abstract. Les données correspondantes font l'objet d'une procédure de traitement particulière, qui baliser automatiquement les données considérées comme des résumés français et anglais dans le document XML.

2.2.2.4 Vérification de la conformité du document XML par rapport au document source

Le document XML créé par la chaîne est bien formé et valide par rapport à sa DTD (TEI-Lite). Les formules mathématiques générées par la chaîne lors de la conversion du document Word en document XML sont stockées dans des fichiers *.mml. Ces fichiers MathML se trouvent dans le répertoire /xml/formules. Ils sont bien formés et valides par rapport à la DTD MathML [31]. Mais des difficultés persistent dans les étapes ultérieures du traitement : les documents HTML et PDF produits ne permettent pas d'afficher les formules mathématiques correctement avec les outils disponibles actuellement.

Les développeurs de la chaîne Cyberdocs n'ayant pas trouvé de solution optimale permettant d'afficher les formules mathématiques dans des documents HTML ou PDF, n'autorisaient qu'un affichage des descriptions des formules mathématiques dans la version *alpha* de la CEN Cyberdocs, téléchargeable depuis un serveur CVS.

La version *beta* de la CEN Cyberdocs, disponible depuis le premier septembre 2003, gére correctement les équations mathématiques créées avec l'éditeur d'équation de Word, à condition que celle-ci soient éditables avec l'éditeur d'équation d'OpenOffice.org. Elle produit des documents HTML et PDF dans lesquels les formules mathématiques sont présentes. Des problèmes d'affichage sont cependant observés, qui sont suffisament important pour remettre en question la possibilité de diffuser ces documents.

2.2.2.5 Indexation des thèses par l'application SDX

Les informations suivantes sont issues de la documentation SDX, disponible en ligne [46]. Pour plus d'informations sur l'utilisation de SDX avec l'application Cyberdocs, on se reportera au rapport de stage de M. Anne Abderahamane [1].

Historique

SDX est un puissant outil de recherche XML, placé sous licence GPL. Il permet de créer rapidement des sites Web dynamiques à partir de collections de documents XML, quelle que soit leur structure. Cette plate-forme est développée, depuis la fin de l'année 2000, sous l'impulsion de la mission de la recherche et de la technologie du ministère de la culture et de la communication.

Outil de recherche XML

SDX est un outil de recherche XML qui permet d'indexer des documents XML dans le but de permettre leur recherche ultérieure et leur consultation. Les documents XML originaux sont stockés par SDX dans des entrepôts, *i.e.* dans des répertoires qui peuvent être internes ou externes à l'application.

SDX offre un moteur de recherche hautement configurable (possibilité de définir un ou plusieurs schémas, avec calcul de pertinence et maquette de présentation). Il permet récupérer des informations à partir d'URLs[vii];.

Concevoir une application SDX consiste à écrire des pages serveur Cocoon (XSP), des transformations XSLT, des documents XML et des fichiers de configuration (application.xconf, sitemap.xmap). Ce langage permet à un développeur de créer des applications SDX capables de stocker et trouver des documents, de les indexer pour la recherche, d'offrir des requêtes complexes, de gérer des droits d'utilisateurs, *etc*.

Les applications SDX n'imposent pas de restriction sur les documents indexés au sein d'une base. Il est tout à fait possible d'indexer des documents XML respectant différents schémas ou DTD au sein d'une même base.

Outil de consultation

SDX est un outil de recherche pour documents XML qui constitue une plate-forme intéressante pour afficher des documents XML ou d'autres types de documents. Une base de données, nommée entrepôt, contient toutes les informations recueillies au cours des processus indexations de documents que l'on souhaite publier *via* SDX. Ainsi, l'application SDX retrouve les documents (pas seulement des documents XML) tels qu'ils lui ont été soumis pour indexation. L'entrepôt peut stocker les documents indexés ou tout simplement en garder l'adresse. Il existe différents types d'entrepôts qui choisissent l'une ou l'autre de ces options tout en offrant une certaine souplesse au moment de la configuration.

2.2.3 Les coûts

2.2.3.1 Coût de la mise en place de la CEN Cyberdocs

La plate-forme Cyberdocs est constituée de logiciels libres, protégés par la licence GPL [35]. Les sources sont distribués gratuitement, et peuvent être modifiés conformément aux conditions fixées par cette licence, ce qui implique qu'après modification, ils demeurent protégés par la même licence et distribués dans les mêmes conditions.

[vii] Uniform Resource Locator

Ainsi, la mise en place de la chaîne Cyberdocs ne nécessite aucun autre investissement financier que le matériel nécessaire à son fonctionnement. Plusieurs forums de discussion permettent de récupérer les informations nécessaires pour la mise en place et l'utilisation de la chaîne d'édition numérique. Cette ressource, comparable à un service après vente de qualité, est très utile et cependant entièrement gratuite.

2.2.3.2 Coût du traitement d'une thèse

Sous les termes « traitement d'une thèse », nous distinguons :

- la réception du document et de l'ensemble des données nécessaires à son traitement ;
- la vérification de la conformité du document par rapport aux indications qui ont été données à l'auteur ;
- la vérification du stylage du document ;
- le traitement par la CEN Cyberdocs et la vérification de la conformité du document XML produit par rapport au document original ;
- l'indexation et la publication de la thèse.

À l'usage, ces étapes feront l'objet d'une normalisation afin d'optimiser et de standardiser la procédure. La durée du traitement dépend de la qualité des documents reçus : le traitement d'une thèse convenablement stylée est beaucoup plus rapide que celui d'un document dont le stylage nécessite une mise en conformité. La vérification du document XML dépend de l'appréciation et de l'expertise de l'opérateur, du niveau d'exigence imposé.

Le temps de traitement d'une thèse ne devrait pas dépasser le temps qui lui est actuellement consacré dans le cadre du projet CITHER, soient 1 à 2 jours.

2.2.4 Conclusion

La version *beta* de la CEN Cyberdocs, développée dans le cadre du projet Cyberthèses, est disponible en téléchargement depuis le 1er septembre 2003. L'étude a essentiellement porté sur la version *alpha* de la CEN Cyberdocs, qui était téléchargeable, du fait des modifications quotidiennes, sur un serveur CVS. Cette application n'arrivera à maturité que dans les prochains mois. Cependant, les conversions réalisées au cours du stage, sont prometteuses :

- le XML, ainsi que le MathML produits sont valides ;

- les images, les tableaux, les titres et parties du document sont pris en charge convenablement ;
- les métadonnées des thèses françaises sont prises en compte au moment de l'indexation qui est réalisée par l'application SDX.

Nous avons cependant constaté que :

- les documents HTML et PDF produits par la chaîne ne sont pas dépourvus d'erreurs et ne permettent pas d'envisager de les diffuser en ligne actuellement.
- les formules mathématiques produites sont au format MathML, format qui reste illisible sur la plupart des navigateurs utilisés actuellement.
- les documents $L^AT_\varepsilon X$ ne sont pas traités. La liste de discussion du forum Cyberthèses (cybertheses-latex@cru.fr, Groupe de travail – $L^AT_\varepsilon X$) ne permet pas d'entrevoir de développement à court ou moyen terme, car les discussions sont peu nombreuses et destinées à proposer des idées qui pourraient être reprises dans un cahier des charges.

3. Conclusion générale sur les CEN

3.1. Les DTD

Les DTD TEI-Lite et DocBook présentées dans ce document sont actuellement largement utilisées. Chaque nouvelle version répond à des exigences nouvelles et élargissent leur domaine d'application. Ainsi, bien que la DTD TEI-Lite ait été conçue pour répondre aux attentes d'auteurs de documents littéraires, et que la DTD DocBook ait été conçue pour rédiger des documents techniques, ces DTD peuvent être utilisées pour créer des documents très éloignés des domaines de d'application définis initialement.

La DTD TEI-Lite est ainsi utilisée par l'Université du Michigan Ann Arbor (Sciences de la Vie) et l'Université Lumière Lyon 2 (Sciences humaines). L'une de ses variantes contenant des extensions pour les sciences naturelles est utilisée par l'Université des Sciences Agricoles d'Uppsala, Suède.

3.2. Traitement des documents LaTeX

Aucune des CEN étudiées ne permet de convertir les document $L^AT_\varepsilon X$ en XML.

On remarquera cependant que les développeurs de la CEN Cyberthèses évoquent la possibilité de développer un outil qui réaliserait ce traitement.

3.3. Le XML obtenu

Les documents XML produits par les CEN Doc'INSA et Cyberdocs sont valides par rapport à leur DTD.

3.4. La publication

La CEN Cyberdocs est associée au puissant outil de stockage, indexation, publication qu'est SDX. Aucun outil de publication n'est associé à la CEN Doc'INSA, mais il est possible de créer des applications SDX pour publier les thèses qui pourraient être produites à partir de cette chaîne d'édition numérique.

3.5. Les coûts

Il n'a pas été possible d'indiquer la durée de traitement que la CEN Doc'INSA nécessiterait. Nous limiterons la comparaison aux coûts liés à leur mise en place.

La configuration matérielle minimale requise est équivalente. Le coût des applications nécessaires au fonctionnement de la CEN Doc'INSA est approximativement de 1235 €, auxquels il convient d'ajouter des frais pour les développements qui s'imposent, contre 0 € pour une installation de la CEN Cyberdocs sous Linux.

Quelle que soit la plate-forme retenue, de nombreuses difficultées liées au mauvais stylage des thèses pourraient être évitées par une formation adéquate des auteurs. La simple prise en compte de recommandations et d'une feuille de style correctement élaborée devrait assurer une adéquation entre le document papier original et la version electronique qu'il sera possible de produire.

Mise en place d'un groupe de travail

Des réunions de travail informelles entre les étudiants l'ENSSIB ont abouti à la mise en place, au cours de l'été 2003, d'un groupe de travail regroupant des personnes travaillant dans des institutions lyonnaises (ENSSIB, Doc'INSA) intéressées par le projet Cyberdocs. Ce groupe de travail a permis d'échanger des informations techniques et générales sur l'évolution du projet Cyberthèses.

Les participants (Mme Émilie ROMAND-MONNIER, M. Richard GRENIER, Mme Liliane Miremont de l'ENSSIB, M. abderahamane ANNE stagiaire à l'université Lyon 2, M. Gilles BROCHET, Mme Dalila Boudia de Doc'INSA, M. Frédéric ALIOTTI, stagiaire à Doc'INSA) ont échangé, au cours de deux réunions, des informations sur les thèmes suivants :

- état des projets de mise en place de chaînes d'édition numérique ;
- difficultés rencontrées au cours des études réalisées et solutions envisagées ;
- choix d'une DTD, possibilité d'utiliser avec Cyberdocs une DTD existante, ou d'en créer de nouvelles ;
- prise en compte des modèles de documents existants par la chaîne Cyberdocs ;
- intérêt de la plate-forme de publication SDX.

Ces discussions ont permis de faire ressortir des centres d'intérêt communs et de replacer les enjeux du développement des chaînes d'édition numérique dans un contexte plus large, que celui de la conversion des thèses électroniques. Il apparaît, en outre, que si l'intérêt de produire des documents XML est communément accepté, le choix d'outils permettant de convertir les documents n'est actuellement pas arrêté.

Mise en place d'une CEN

Les paragraphes suivants constituent une étude de faisabilité destinée à faciliter une éventuelle mise en place de la CEN Cyberdocs à Doc'INSA.

1. Création d'un modèle de document INSA-Lyon

1.1. Étude de la structure des thèses soutenues en 2002

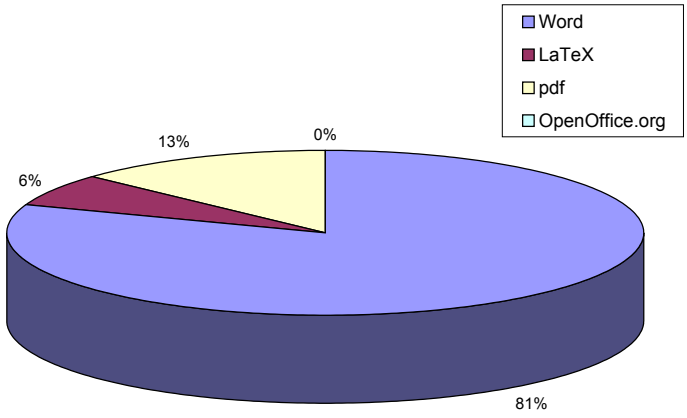


Figure 5 : Formats des thèses traitées en 2002. Étude réalisée sur 108 documents.

Plus de 80% des thèses traitées en 2002 (Figure 5) étaient des documents Word. Les documents $L^{A}T_{\varepsilon}X$ ne constituaient que 6% des 108 documents recueillis. Aucun document OpenOffice.org n'a été trouvé.

Nous avons retenu 4 indicateurs pour mettre en évidence les principales caractéristiques du stylage des thèses électroniques déposées à doc'INSA sous forme de documents Word et soutenues au cours de l'année 2002 : la présence d'un sommaire automatique, de tables des illustrations automatique, de numérotation des pages, et d'utilisation du modèle de document proposé par Doc'INSA.

La présence d'un sommaire automatique implique une structuration cohérente du document et l'utilisation des différents styles de paragraphe.
La présence d'une table des illustrations implique une maîtrise avancée du logiciel de traitement de texte.
La présence d'une numérotation automatique nécessite une maîtrise de base du logiciel de traitement de texte.
L'utilisation du modèle de document proposé par Doc'INSA peut correspondre à une nécessité (se conformer à un modèle), une envie (utiliser un modèle dont l'aspect est attrayant et/ou bénéficier des outils inclus dans le modèle).

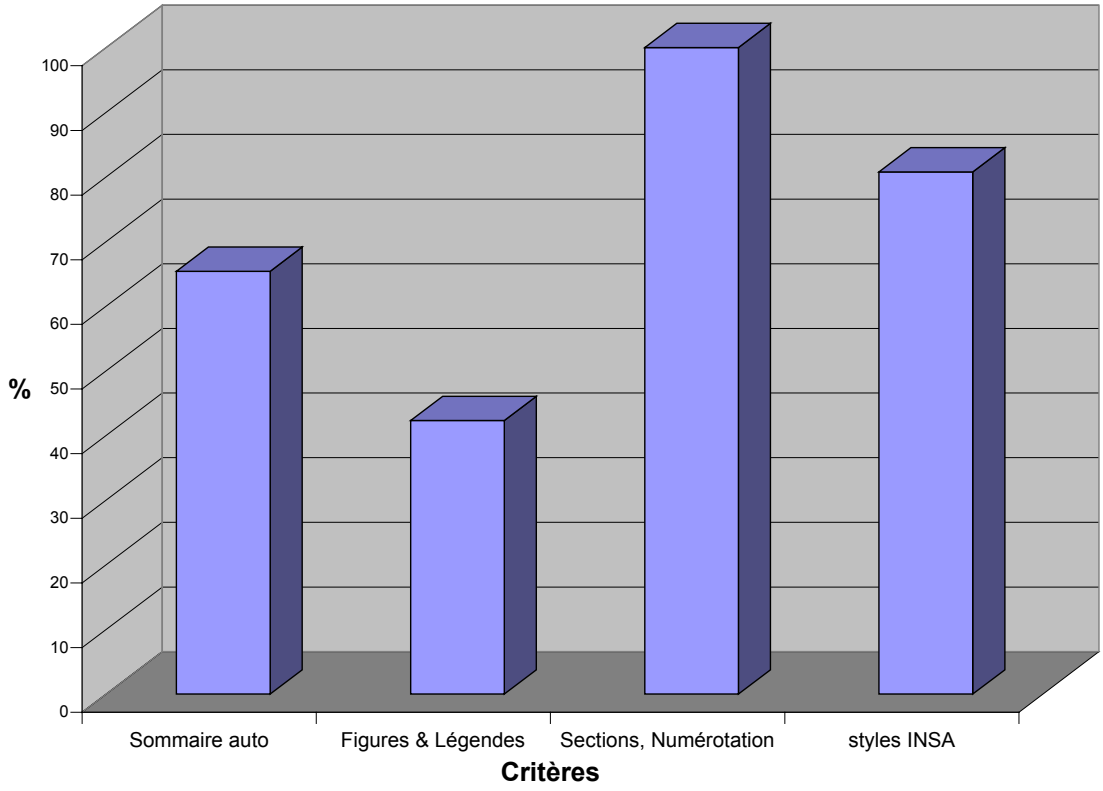


Figure 6 : Caractéristiques du stylage des thèses électroniques déposées à Doc'INSA, sous forme de documents Word, et soutenues au cours de l'année 2002. Étude de 53 documents parmi les 87 disponibles.

L'étude de ces données (Figure 6) indique que tous les doctorants maîtrisent les fonctions de base du logiciel de traitement de texte, et que 40% d'entre eux maîtrisent les fonctions avancées. On remarquera que 65% des doctorants structurent leur document et génèrent un sommaire automatique. Ce résultat est

minoré, en raison de difficultés d'interprétation. Comme nous l'indiquons dans le paragraphe suivant, une proportion non négligeable des thèses sont scindés en plusieurs parties, ce qui incite les auteurs à créer un sommaire de document principal en collant des parties de sommaires créées automatiquement. Il est difficile de distinguer les sommaires recomposés de ceux qui ont été rédigés manuellement par les auteurs, ce qui nous à conduit à sous-estimer l'utilisation des sommaires automatiques.

Plus 75% des documents Word déposés à Doc'INSA en 2002 sont scindés en plusieurs parties. Il serait nécessaire de s'entretenir avec les doctorant pour comprendre les raisons qui justifient cette démarche. Et éventuellementcréer un modèle de document permettant de traiter aisément les documents composites reçus avec la CEN Cyberdocs.

1.2. Présentation des styles

Code de style	Elément xml TEI-Lite	Nom lyon2	Nom insa
la page de titre			
auteur	<docAuthor>	1|Auteur	_auteur
copyright	<titlePart type="copyright">	1|Copyright	_copyright
depot	<docDate>	1|Depot	_depot
dept	<titlePart type="dept">	1|Dept	_departement
directeur	<titlePart type="directeur">	1|Directeur	_directeur
discipline	<titlePart type="discipline">	1|Discipline	_discipline
ecole-doct	<titlePart type="ED">	1|EcoleDoct	_ecoledoctorale
faculte	<titlePart type="faculte">	1|Faculte	_laboratoire
grade	<titlePart type="grade">	1|Grade	_grade
jury	<titlePart type="jury">	1|Jury	_jury
no-officiel	<titlePart type="Reference">	1|NoOfficiel	_noofficiel
sous-titre	<docTitle>/<titlePart type="sub" lang="fr">	1|Sous-titre	_soustitre
titre-these	<docTitle>/<titlePart type="main" lang="fr">	1|TitreThese	_titrethese
universite	<titlePart type="univ">	1|Universite	_insa
les liminaires			
dedicace		1|Dedicace	_dedicace
titre-front		1|TitreFront	_titrefront
les annexes			
ann-titre	----------	3|Ann_titre	_annexetitre
ann-titre1**9		3|Ann_titre1**9	_annexetitre1**9
la bibliographie			
bibli-item	<bibl>	3|Bibli_item	_biblioitem
bibli-tit	----------	3|Bibli_tit	_bibliotitre
bibli-tit1**4		3|Bibli_tit1**4	_bibliotitre1**4
les citations			
citation	<q rend="block">	Citation, WW-Citation	Citation, WW-Citation
citation-bloc1**2	<q rend="block">	CitatioBloc1**2	_citationbloc1**2
les parties			
intro		Intro	_introduction
conclu		Conclu	_conclusion
partie		Partie	_partie
le closer			

closer	<closer>	closer	_closer
les notes			
source	<note place="interlinear">	Source	Source
les listes			
entree	<list type="gloss">/<item>	Entree	Entree
liste-num	----------	ListeNum	_listenum
liste-num1	<list type="ordered">/<item>	ListeNum1	_listenum1
liste-num2	<list type="ordered">/<item>	ListeNum2	_listenum2
liste-puce	----------	ListePuce	_listepuce
liste-puce1**8	<list type="bulleted">/<item>	ListePuce1**8	_listepuce1**8
liste-simple	<list type="simple">/<item>	ListeSimple	_liste
liste-titre	<list>/<head>	ListeTitre	_listetitre
le texte			
les titres			
heading1**9	/<head>	heading 1**9	Heading 1**9
les images			
figure	<figure>	Figure	_figure
image-ligne	<figure>	ImageLigne	ImageLigne
image-tab	<figure>	ImageTab	ImageTab
les légendes			
legende-fig	<* id="fig">/<head>	LegendeFig	_legendefigure
legende-tab	<* id="tab">/<head>	LegendeTab	_legendetableau
caption	<head>	Caption	Caption
les épigraphes			
epigraphe	<epigraph>	1|Epigraphe	_epigraphe
ce qui n'est pas traité			

Tableau 3 : présentation des codes (définis dans le fichier outils/xslt/utiles/styles.xml), structures XML produites par la présence de ce style, ainsi que des noms de styles correspondant dans la feuille de styles lyon2 et insa.

Le tableau précédent (Tableau 3) présente le nom des styles insa et les correspondances avec les éléments XML définis dans la DTD TEI-Lite, tels qu'ils apparaissent dans le document XML produit lors du traitement par la CEN Cyberdocs.

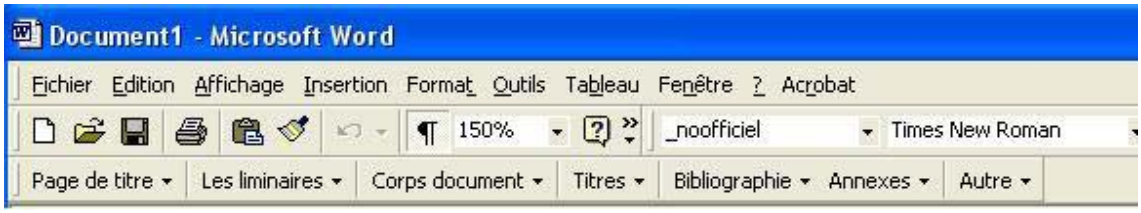
Comme nous l'indiquions précédemment (*Cf* 2.2.2.2), certains styles sont obligatoires. Or, contrairement au modèle de document utilisé par l'université Lyon 2, le modèle de l'INSA ne nécessite pas de style « 1|Faculté ». Nous avons utilisé le code style « faculte » pour créer un style _laboratoire, information qui doit figurer dans la page de titre des thèses déposées à l'INSA, mais qui ne correspondait à aucun code style existant.

Pour des raisons de compatibilité avec la CEN Cyberdocs, les noms de styles définis dans le modèle de document Word ne doivent contenir aucun caractère accentué, aucune espace. Les caractères de soulignement et de séparation « | » (correspondant à une séquence Alt-Gr-6 sur un clavier classique) peuvent être

utilisés.Les noms de style Word réservés et qui ne respectent pas les recommandations précédentes, tels que le style nommé « Légende » doivent être préfixés par un WW- lorsqu'ils sont déclarés dans le fichier styles.xml.

1.3. Présentation de la barre d'outils

La conception de la barre d'outils est délibérément très simple. Le critère de simplicité a été choisi en se basant sur les observations des modèles de documents proposés par l'université Lyon 2, l'ENSSIB et Doc'INSA.



1.4. Les macros

La présence de macros dans les documents traités par la CEN Cyberdocs entraine l'apparition d'un message d'avertissement indiquant que la chaîne ne traite pas ces scripts, à la suite duquel le traitement se poursuit normalement. Il est donc possible de créer un modèle de style contenant des macros, bien que cette solution ne soit pas sans risque. Les macros sont susceptibles de ne pas fonctionner sur toutes les vesions de Word existantes. Ce problème de compatibilité pourrait freiner l'utilisation du modèle de document Word par les utilisateurs et induirait un effet contraire à celui escompté.

En outre, la présence de macros dans les modèles de document Word limite la compatibilité de ces document avec le logiciel de traitement de texte OpenOffice.org Writer. Il n'est donc pas souhaitable d'incorporer de macros dans les modèles de documents, car leur présence est de nature à promouvoir l'utilisation du logiciel Microsoft Word.

2. Perspectives

La formation des doctorants, qui produisent les thèses que nous devons transformer en documents XML, est une étape essentielle qu'il convient de ne pas négliger. L'étude des documents Word traîtés au cours de l'année 2002 indique que cette

population maîtrise ce logiciel de traitement de texte. Il serait intéressant de réaliser une enquête qui porterait sur les doctorants et leur connaissance des enjeux de la publication des thèses électroniques. Une telle étude permettrait de proposer des formations adaptées à leurs besoins et de poursuivre efficacement les travaux entrepris dans ce domaine, travaux qui donnent déjà des résultats satisfaisants.

Pour conclure, si l'on souhaite produire et diffuser des thèses électroniques conformes aux documents papiers, nous devrons tenir compte des limites des outils disponibles actuellement. Une solution pourrait consister à produire des documents papier ou Word plus proches du document électronique qu'ils serviront à produire. Il s'agit d'envisager d'établir une charte, un contrat, selon des modalités qui doivent être définies, entre les doctorants et l'institution qui réalisera la production d'une version électronique à partir du document original qui leur a été confié. Ce contrat doit clairement expliquer les limites des procédés utilisés pour produire les documents électroniques : homogénisation et standardisation de la forme des documents, respect de conventions stylistiques et typographiques qui constituent des contraintes importantes. Il indiquera également tous les avantages que l'on peut espérer du respect des recommandations : le document électronique sera conforme à la version papier originale et prendra sa place dans un corpus virtuel à forte valeur ajoutée, disponible en ligne et qui constitue une vitrine pour le travail des auteurs et l'école qui les a formés.

Le CCSD propose une solution de diffusion des thèses électroniques qui, à défaut de se préoccuper de leur conservation, se révèle très efficace : les thèses électroniques sont déposées en lignes par les étudiants qui réalisent la conversion de leur document vers un format de diffusion PostScript ou PDF. Ce qui réduit considérablement le coût du traitement des documents réalisé par le CCSD.

L'orientation des projets de CEN Doc'INSA et Cyberthèses se caractérise par des contraintes techniques et des coûts importants liés à des cahiers des charges particulièrement contraignants. Les retards constatés dans la mise en œuvre des solutions développées dans le cadre de ces projets, s'opposent naturellement à la rapidité qui caractérise le développement et la mise en place du serveur de thèses multidisciplinaire du CCSD. Alors, diffusion ou conservation ?

Conclusion

Au cours de mon stage à Doc'INSA j'ai bénéficié de 16h de formation, organisées par l'INSA de Lyon, à destination de son personnel. Nous avons également participé à deux réunions du groupe de travail sur les métadonnées des thèses - AFNOR CG 46/CN 357 GE 5, qui se tenaient à Paris, dans les locaux de la Bibliothèque Nationale de France.

Les formations « Linux utilisateur autonome » et « Logiciel d'édition $L^AT_\varepsilon X$ » m'ont permis de valider des acquis et d'acquérir de nouvelles compétences, qui se sont rapidement révélées utiles dans le cadre du stage. Les réunions du groupe de travail sur les métadonnées des thèses, auxquelles des représentants de Doc'INSA participent depuis sa création, m'ont permis d'entrevoir l'intérêt, les enjeux et la complexité du travail de production des normes destinées aux professionnels de la gestion de l'information.

Au cours de cette étude, nous avons été attentifs aux informations qui concernaient la suite bureautique Microsoft Office 2003 dont la commercialisation est annoncée pour la fin octobre 2003. La version d'évaluation que nous avons installée ne nous à pas permis de convertir des documents Word en documents XML conformes à une DTD de notre choix. Il semblerait que les fonctions d'exportation des documents Word en format XML ne soient pas incluses dans les versions standards de la nouvelle suite bureautique.

Outre les limitations imposées à l'utilisation des fonctions de production de documents dans le format non propriétaire qu'est le XML, des problèmes de compatibilité avec les les outils de bureautique alternatifs tels que Star Office de Sun Microsystems et OpenOffice.org sont à craindre. S'il devenait impossible d'éditer ou de consulter des documents créés à partir de Word 2003 avec OpenOffice.org writer, les projets de CEN XML présentés dans cette étude rencontreraient des difficultés évidentes.

La généralisation de l'utilisation de la suite bureautique OpenOffice.org, initiée à l'INSA de Lyon, pourrait être une solution.

Bibliographie

Diffusion des thèses électroniques

1. **ANNE A.** Appropriation d'une plateforme d'édition électronique basée sur XML : Cyberdocs. Rapport de stage DESS en Ingénierie Documentaire: ENSSIB/Université Claude Bernard - Lyon 1, 2003, p. 98.
2. **CCSD**. Centre pour la Communication Scientifique Directe **[en ligne]** 2003. Disponible sur: < http://www.ccsd.cnrs.fr/ >. (Consulté le 17 06 2003).
3. **CENTRE POUR LA COMMUNICATION SCIENTIFIQUE DIRECTE**. Serveur de thèses multidisciplinaire **[en ligne]** 2003. Disponible sur: < http://tel.ccsd.cnrs.fr/ >. (Consulté le 17 06 2003).
4. **CLERC C.** Contribution au développement d'un serveur de thèses électroniques. Rapport de stage DESS en Informatique Documentaire: ENSSIB/Université Claude Bernard - Lyon 1, 1999, p. 72.
5. **CNRS**. Centre national de la recherche scientifique **[en ligne]** 2003. Disponible sur: < http://www.cnrs.fr/ >.
6. **CRU**. Informations téléchargement des sources Cyberdocs **[en ligne]** 2003. Disponible sur: < http://www.ajlsm.com/projets/cybertheses/site-dev/en/download.html >. (Consulté le 17 06 2003).
7. **CYBERTHÈSES**. Projet Cyberthèses **[en ligne]** 1998. Disponible sur: < http://www.cybertheses.org et http://sophia.univ-lyon2.fr/CyberTheses/index.php >. (Consulté le 02 07 2003).
8. **ETD**. Projet ETD **[en ligne]**: ETD, 1997. Disponible sur: < http://etd.vt.edu/ >. (Consulté le 02 07 2003).
9. **FRANCOPHONIE.ORG**. Projet Cyberthèses **[en ligne]** 1998. Disponible sur: < http://www.francophonie.org/fonds/projets/ProjetRetenu98.htm >. (Consulté le 02 07 2003).
10. **INSA DE LYON**. CITHER, consultation en texte intégral des thèses en réseau **[en ligne]**: Doc'INSA, 1999. Disponible sur: < http://csidoc.insa-lyon.fr/these/ >. (Consulté le 17 06 2003).
11. **IOANNITIS S.** Elaboration d'un cahier des charges pour les thèses numériques de Doc'INSA (cither 2002) - Explication de la chaîne de traitements. Projet de fin d'étude INSA de Lyon: INSA de Lyon, 2002, p. 72.
12. **JOLLY C.** Rapport sur la diffusion électronique des thèses établi par un groupe de travail **[en ligne]**: ABES, 2003. Disponible sur: < http://www.abes.fr/abes/DesktopDefault.aspx?tabid=213 >. (Consulté le 17 06 2003).
13. **MERMET J.-M.** Coordination et mise en place d'un serveur de thèses en texte intégral à l'INSA de lyon. Conception du frontoffice. Rapport de stage DESS en Informatique Documentaire: ENSSIB/Université Claude Bernard - Lyon 1, 1998, p. 67.
14. **NDLTD**. Networked digital library of theses and dissertations **[en ligne]**: NTLD, Networked Digital Library of Theses and Dissertations, 1998.

Disponible sur: < http://www.ndltd.org/info/index.htm >. (Consulté le 02 07 2003).

15. **PROJET DE PUBLICATION ET DE DIFFUSION ÉLECTRONIQUES DES THÈSES DE DOCTORAT**.
Projet de publication et de diffusion électroniques des thèses de doctorat **[en ligne]** 2000. Disponible sur: < http://www.pum.umontreal.ca/theses/RapportThesesUdeM.pdf >. (Consulté le 02 07 2003).
16. **ROMAND-MONNIER E.** Migration d'une revue professionnelle vers un modèle structuré en ligne. Rapport de stage DESS en Ingénierie Documentaire: ENSSIB/Université Claude Bernard - Lyon 1, 2000, p. 108.
17. **SULEMAN H., ATKINS A., GONÇALVES M.A., *et al.*** Networked digital library of theses and dissertations: Bridging the gaps for global access - part 1: Mission and progress. D-Lib Magazine, 2001.
18. **UNESCO**. The guide to electronic theses & dissertations **[en ligne]** 2002. Disponible sur: < http://etdguide.org/ >. (Consulté le 17 06 2003).
19. **BURNARD L., SPERBERG-MCQUEEN C.M.** La TEI simplifiée : Une introduction au codage des textes électroniques en vue de leur échange **[en ligne]** 1996. Disponible sur: < http://www.univ-rennes1.fr/pub/GUTenberg/publicationsPS/24-teilite.ps.gz >. (Consulté le 17 06 2003).

Les DTDs

20. **COMITÉ TECHNIQUE DOCBOOK D'OASIS**. Page d'accueil de DocBook **[en ligne]**: Comité technique DocBook d'OASIS, 2003. Disponible sur: < http://www.oasis-open.org/docbook/ >. (Consulté le 02072003).
21. **TEI**. Text encoding initiative **[en ligne]** 1987. Disponible sur: < http://www.tei-c.org/Publicity/p4release-fr.html >. (Consulté le 02 07 2003).
22. **TEI CONSORTIUM**. Outils TEI-Lite **[en ligne]** 2003. Disponible sur: < http://www.tei-c.org/Software/ >. (Consulté le 17 06 2003).
23. **TRAVAIL COLLABORATIF**. The DocBook wiki **[en ligne]** 2003. Disponible sur: < http://www.docbook.org/wiki/moin.cgi/ >.

Les métadonnées

24. **DCMI**. Dublin Core Metadata Initiative **[en ligne]**:
Dublin Core Metadata Initiative, 2003. Disponible sur: < http://dublincore.org/about/ >. (Consulté le 02 07 2003).
25. **OAI**. OAI, the open archives initiative protocol for metadata harvesting **[en ligne]** 2003. Disponible sur: < http://www.openarchives.org/OAI/openarchivesprotocol.html >. (Consulté le 17 06 2003).

Les standards et recommandations

26. **BRAY T., PAOLI J., SPERBERG-MCQUEEN C.M., *et al.*** Extensible markup language (XML) 1.0 (second edition) W3C recommendation **[en ligne]**: W3C, 2000. Disponible sur: < http://www.w3c.org >. (Consulté le 02 07 2003).
27. **W3C**. XSL Transformations (XSLT) version 1.0 w3c recommendation **[en ligne]** 1999. Disponible sur: < http://www.w3.org/TR/1999/REC-xslt-19991116 >. (Consulté le 02 07 2003).
28. **W3C**. XML Path Language (XPATH) version 1.0 W3C recommendation **[en ligne]** 1999. Disponible sur: < http://www.w3.org/TR/1999/REC-xpath-19991116 >. (Consulté le 02 07 2003).
29. **W3C**. Extensible Stylesheet Language (XSL) version 1.0 w3c recommendation **[en ligne]** 2001. Disponible sur: < http://www.w3.org/TR/2001/REC-xsl-20011015/ >. (Consulté le 02 07 2003).
30. **W3C**. Mathematical Markup Language (MathML) version 2.0 w3c recommendation **[en ligne]** 2001. Disponible sur: < http://www.w3.org/TR/2001/REC-MathML2-20010221 >. (Consulté le 02 07 2003).
31. **W3C**. La DTD MathML **[en ligne]** 2003. Disponible sur: < http://www.w3.org/TR/MathML2/appendixa.html#parsing_dtd >. (Consulté le 17 06 2003).

Autres sources d'informations

32. **ABES**. L'ABES, agence bibliographique de l'enseignement supérieur **[en ligne]**: ABES, 2003. Disponible sur: < http://www.abes.fr/ >. (Consulté le 17 06 2003).
33. **CRU**. Comité réseau des universités **[en ligne]** 2003. Disponible sur: < http://www.cru.fr/ >. (Consulté le 17 06 2003).
34. **DTSEARCH CORPORATION**. Dtsearch - the smart choice for text retrieval since 1991 **[en ligne]**: dtSearch Corporation, 2003. Disponible sur: < http://www.dtsearch.com >. (Consulté le 17 06 2003).
35. **GPL L.** GNU Public Licence **[en ligne]** 2003. Disponible sur: < http://www.gnu.org/copyleft/gpl.html >.
36. **KUHN M.** UTF-8 - an ASCII compatible multibyte unicode encoding **[en ligne]** 2001. Disponible sur: < http://www.icewalkers.com/Linux/ManPages/utf8-7.html >. (Consulté le 17 06 2003).

Logiciels présentés

37. **ALTOVA GMBH**. XML Spy. [Programme informatique] version 4.4. Altova GmbH, Informations en ligne sur: < http://www.altova.com/download_archive.html >. (Consulté en 2000).
38. **APACHE JAKARTA PROJECT**. Tomcat binaire pour Linux. [Programme informatique] version 4.1.24. Apache, Informations en ligne sur: < http://jakarta.apache.org/builds/jakarta-tomcat-4.0/release/v4.1.24/rpms/ >. (Consulté en 2003).

39. **APACHE JAKARTA PROJECT**. Tomcat binaire pour Windows. [Programme informatique] version 4.1.27. Apache, Informations en ligne sur: < http://www.apache.org/jakarta/tomcat-4/binaries/ >. (Consulté en 2003).
40. **DESIGN SCIENCE**. Mathtype software development kit. [Programme informatique] version 5.1. Design Science, Informations en ligne sur: < http://www.dessci.com/en/reference/sdk/ >. (Consulté en 2003).
41. **ICENI TECHNOLOGY**. Gemini Solo. [Programme informatique] version 1.6. Iceni technology, Informations en ligne sur: < http://www.iceni.com/downloadSet.html >. (Consulté en 2003).
42. **INFINITY-LOOP**. UpCast. Infinity-loop, Informations en ligne sur: < http://www.infinity-loop.de/products/upcast/downloads.html >. (Consulté en 2003).
43. **KAY M.** Saxon. [Programme informatique] version 6.5.2 et 7.6. Michael Kay, Informations en ligne sur: < http://saxon.sourceforge.net/ >. (Consulté en 2003).
44. **OPENOFFICE.ORG**. Suite bureautique OpenOffice.org. [Programme informatique] version 1.0.3.1. OpenOffice.org, Informations en ligne sur: < http://www.openoffice.org/dev-docs/source/1.0.3/ >. (Consulté en 2003).
45. **ORGANIZATION FOR THE ADVANCEMENT OF STRUCTURED INFORMATION STANDARDS**. DocBook XSL, ensemble de balises pour écrire des documents structurés. [Programme informatique] version 1.61.3. Oasis, Organization for the Advancement of Structured Information Standards, Informations en ligne sur: < http://sourceforge.net/project/showfiles.php?group_id=21935 >. (Consulté en 2003).
46. **SÉVIGNY M., BOTIN M.** SDX. [Programme informatique] version 2.0. AJLSM, Informations en ligne sur: < http://sdx.culture.fr/sdx/ >. (Consulté en 2003).
47. **SUN**. Java(tm) 2 SDK, standard edition 1.4.1_03. [Programme informatique] version 1.4.1_03. sun, Informations en ligne sur: < http://java.sun.com >. (Consulté en 2003).
48. **SUN**. J2se(tm) v 1.4.2 with NetBeans(tm) ide v 3.5 Cobundle. [Programme informatique] version 1.4.2. sun, Informations en ligne sur: < http://java.sun.com/j2se/1.4.2/download.html >. (Consulté en 2003).
49. **THE APACHE XML PROJECT**. Xalan. [Programme informatique] version Xalan-J 2.5.1. The Apache XML project, Informations en ligne sur: < http://xml.apache.org/xalan-j/downloads.html >. (Consulté en 2003).
50. **THE APACHE XML PROJECT**. FOP, the formatting objects processor. [Programme informatique] version 0.20.5. The Apache XML Project, Informations en ligne sur: < http://xml.apache.org/dist/fop/ >. (Consulté en 2003).

Table des annexes

Annexe 1

Fichiers nécessaires à l'installation et à l'utilisation de la CEN Cyberdocs

Fichier pcd.properties

dossier.installation.up=../20030701

openoffice.home=/root/OpenOffice.org1.0.3

sdx.application.path=pcd

habillage=pcd

application.sdx.identifiant=org.cyberdocs.demo

dossier.installation.consultation=/usr/local/tomcat5/webapps/sdx

sdx.application.open=true

Fichier styles.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<!--
Plate-forme Cyberdocs
Copyright (C) 2003
Projet Cyberthèses
(..)
-->
<styles>
<institutions>
 <institution code="lyon2"/>
</institutions>
<institutions>
 <institution code="insa"/>
</institutions>
<!--la page de titre-->
<style code="auteur">
 <nom code="lyon2" xml:lang="fr">1|Auteur</nom>
 <nom code="insa" xml:lang="fr">_auteur</nom>
</style>
<style code="copyright">
 <nom code="lyon2" xml:lang="fr">1|Copyright</nom>
 <nom code="insa" xml:lang="fr">_copyright</nom>
</style>
<style code="depot">
 <nom code="lyon2" xml:lang="fr">1|Depot</nom>
 <nom code="insa" xml:lang="fr">_depot</nom>
</style>
<style code="dept">
 <nom code="lyon2" xml:lang="fr">1|Dept</nom>
 <nom code="insa" xml:lang="fr">_departement</nom>
</style>
<style code="directeur">
 <nom code="lyon2" xml:lang="fr">1|Directeur</nom>
 <nom code="insa" xml:lang="fr">_directeur</nom>
</style>
<style code="discipline">
 <nom code="lyon2" xml:lang="fr">1|Discipline</nom>
 <nom code="insa" xml:lang="fr">_discipline</nom>
</style>
<style code="ecole-doct">
 <nom code="lyon2" xml:lang="fr">1|EcoleDoct</nom>
 <nom code="insa" xml:lang="fr">_ecoledoctorale</nom>
</style>
<style code="faculte">
 <nom code="lyon2" xml:lang="fr">1|Faculte</nom>
 <nom code="insa" xml:lang="fr">_insa</nom>
</style>
<style code="grade">
```

```
 <nom code="lyon2" xml:lang="fr">1|Grade</nom>
 <nom code="insa" xml:lang="fr">_grade</nom>
</style>
<style code="jury">
 <nom code="lyon2" xml:lang="fr">1|Jury</nom>
 <nom code="insa" xml:lang="fr">_jury</nom>
</style>
<style code="no-officiel">
 <nom code="lyon2" xml:lang="fr">1|NoOfficiel</nom>
 <nom code="insa" xml:lang="fr">_noofficiel</nom>
</style>
<style code="sous-titre">
 <nom code="lyon2" xml:lang="fr">1|Sous-titre</nom>
 <nom code="insa" xml:lang="fr">_soustitre</nom>
</style>
<style code="titre-these">
 <nom code="lyon2" xml:lang="fr">1|TitreThese</nom>
 <nom code="insa" xml:lang="fr">_titrethese</nom>
</style>
<style code="universite">
 <nom code="lyon2" xml:lang="fr">1|Universite</nom>
 <nom code="insa" xml:lang="fr">_insa</nom>
</style>
<!--les liminaires-->
<style code="dedicace">
 <nom code="lyon2" xml:lang="fr">1|Dedicace</nom>
 <nom code="insa" xml:lang="fr">_dedicace</nom>
</style>
<style code="titre-front">
 <nom code="lyon2" xml:lang="fr">1|TitreFront</nom>
 <nom code="insa" xml:lang="fr">_titrefront</nom>
</style>
<!--les annexes-->
<style code="ann-titre">
 <nom code="lyon2" xml:lang="fr">3|Ann_titre</nom>
 <nom code="insa" xml:lang="fr">_annexetitre</nom>
</style>
<style code="ann-titre1">
 <nom code="lyon2" xml:lang="fr">3|Ann_titre1</nom>
 <nom code="insa" xml:lang="fr">_annexetitre1</nom>
</style>
<style code="ann-titre2">
 <nom code="lyon2" xml:lang="fr">3|Ann_titre2</nom>
 <nom code="insa" xml:lang="fr">_annexetitre2</nom>
</style>
<style code="ann-titre3">
 <nom code="lyon2" xml:lang="fr">3|Ann_titre3</nom>
 <nom code="insa" xml:lang="fr">_annexetitre3</nom>
</style>
<style code="ann-titre4">
 <nom code="lyon2" xml:lang="fr">3|Ann_titre4</nom>
 <nom code="insa" xml:lang="fr">_annexetitre4</nom>
</style>
<style code="ann-titre5">
 <nom code="lyon2" xml:lang="fr">3|Ann_titre5</nom>
 <nom code="insa" xml:lang="fr">_annexetitre5</nom>
</style>
<style code="ann-titre6">
 <nom code="lyon2" xml:lang="fr">3|Ann_titre6</nom>
 <nom code="insa" xml:lang="fr">_annexetitre6</nom>
</style>
<style code="ann-titre7">
 <nom code="lyon2" xml:lang="fr">3|Ann_titre7</nom>
 <nom code="insa" xml:lang="fr">_annexetitre7</nom>
</style>
<style code="ann-titre8">
 <nom code="lyon2" xml:lang="fr">3|Ann_titre8</nom>
 <nom code="insa" xml:lang="fr">_annexetitre8</nom>
</style>
<style code="ann-titre9">
 <nom code="lyon2" xml:lang="fr">3|Ann_titre9</nom>
 <nom code="insa" xml:lang="fr">_annexetitre9</nom>
</style>
<!--la bibliographie-->
<style code="bibli-item">
 <nom code="lyon2" xml:lang="fr">3|Bibli_item</nom>
 <nom code="insa" xml:lang="fr">_biblioitem</nom>
</style>
<style code="bibli-tit">
 <nom code="lyon2" xml:lang="fr">3|Bibli_tit</nom>
 <nom code="insa" xml:lang="fr">_bibliotitre</nom>
</style>
<style code="bibli-tit1">
 <nom code="lyon2" xml:lang="fr">3|Bibli_tit1</nom>
 <nom code="insa" xml:lang="fr">_bibliotitre1</nom>
</style>
<style code="bibli-tit2">
 <nom code="lyon2" xml:lang="fr">3|Bibli_tit2</nom>
 <nom code="insa" xml:lang="fr">_bibliotitre2</nom>
</style>
<style code="bibli-tit3">
 <nom code="lyon2" xml:lang="fr">3|Bibli_tit3</nom>
 <nom code="insa" xml:lang="fr">_bibliotitre3</nom>
</style>
<style code="bibli-tit4">
 <nom code="lyon2" xml:lang="fr">3|Bibli_tit4</nom>
 <nom code="insa" xml:lang="fr">_bibliotitre4</nom>
</style>
<!--les citations-->
<style code="citation">
 <nom code="lyon2" xml:lang="fr">Citation</nom>
 <nom code="lyon2" xml:lang="fr">Citation (user)</nom>
 <nom code="lyon2" xml:lang="fr">WW-Citation</nom>
 <nom code="insa" xml:lang="fr">Citation</nom>
 <nom code="insa" xml:lang="fr">WW-Citation</nom>
 <nom code="insa" xml:lang="fr">_citationdoctorant</nom>
</style>
<style code="citation-bloc1">
 <nom code="lyon2" xml:lang="fr">CitatioBloc1</nom>
 <nom code="insa" xml:lang="fr">_citationbloc1</nom>
</style>
<style code="citation-bloc2">
 <nom code="lyon2" xml:lang="fr">CitatioBloc2</nom>
 <nom code="insa" xml:lang="fr">_citationbloc2</nom>
</style>
<!--les parties-->
<style code="intro">
 <nom code="lyon2" xml:lang="fr">Intro</nom>
 <nom code="insa" xml:lang="fr">_introduction</nom>
</style>
<style code="conclu">
 <nom code="lyon2" xml:lang="fr">Conclu</nom>
 <nom code="insa" xml:lang="fr">_conclusion</nom>
</style>
<style code="partie">
 <nom code="lyon2" xml:lang="fr">Partie</nom>
 <nom code="insa" xml:lang="fr">_partie</nom>
</style>
<!--le closer-->
<style code="closer">
 <nom code="lyon2" xml:lang="fr">closer</nom>
 <nom code="insa" xml:lang="fr">_closer</nom>
</style>
<!--les notes-->
<style code="source">
 <nom code="lyon2" xml:lang="fr">Source</nom>
 <nom code="insa" xml:lang="fr">Source</nom>
</style>
<style code="endnote">
 <nom code="lyon2" xml:lang="fr">endnote text</nom>
 <nom code="insa" xml:lang="fr">_endnotetxt</nom>
</style>
<style code="footnote">
 <nom code="lyon2" xml:lang="fr">footnote text</nom>
 <nom code="lyon2" xml:lang="fr">Footnote</nom>
 <nom code="lyon2" xml:lang="fr">Footnote Symbol</nom>
 <nom code="insa" xml:lang="fr">_piedpagetxt</nom>
 <nom code="insa" xml:lang="fr">_piedpage</nom>
 <nom code="insa" xml:lang="fr">_piedpagesymbol</nom>
</style>
<!--les listes-->
<style code="entree">
 <nom code="lyon2" xml:lang="fr">Entree</nom>
 <nom code="insa" xml:lang="fr">Entree</nom>
</style>
<style code="liste-num">
 <nom code="lyon2" xml:lang="fr">ListeNum</nom>
 <nom code="insa" xml:lang="fr">_listenum</nom>
</style>
<style code="liste-num1">
 <nom code="lyon2" xml:lang="fr">ListeNum1</nom>
 <nom code="insa" xml:lang="fr">_listenum1</nom>
</style>
<style code="liste-num2">
 <nom code="lyon2" xml:lang="fr">ListeNum2</nom>
 <nom code="insa" xml:lang="fr">_listenum2</nom>
```

```
</style>
<style code="liste-puce">
 <nom code="lyon2" xml:lang="fr">ListePuce</nom>
 <nom code="lyon2" xml:lang="fr">List Bullet</nom>
 <nom code="lyon2" xml:lang="fr">WW-Liste à puces</nom>
 <nom code="insa" xml:lang="fr">WW-Liste à puces</nom>
 <nom code="insa" xml:lang="fr">_listepuce</nom>
</style>
<style code="liste-puce1">
 <nom code="lyon2" xml:lang="fr">ListePuce1</nom>
 <nom code="lyon2" xml:lang="fr">List Bullet</nom>
 <nom code="lyon2" xml:lang="fr">WW-Liste à puces</nom>
 <nom code="insa" xml:lang="fr">WW-Liste à puces</nom>
 <nom code="insa" xml:lang="fr">_listepuce1</nom>
</style>
<style code="liste-puce2">
 <nom code="lyon2" xml:lang="fr">ListePuce2</nom>
 <nom code="lyon2" xml:lang="fr">List Bullet 2</nom>
 <nom code="lyon2" xml:lang="fr">WW-Liste à puces 2</nom>
 <nom code="insa" xml:lang="fr">WW-Liste à puces 2</nom>
 <nom code="insa" xml:lang="fr">_listepuce1</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 à puces 3</nom>
 <nom code="insa" xml:lang="fr">WW-Liste à puces 3</nom>
 <nom code="insa" xml:lang="fr">_listepuce3</nom>
</style>
<style code="liste-puce4">
 <nom code="lyon2" xml:lang="fr">ListePuce4</nom>
 <nom code="lyon2" xml:lang="fr">List Bullet 4</nom>
 <nom code="lyon2" xml:lang="fr">WW-Liste à puces 4</nom>
 <nom code="insa" xml:lang="fr">WW-Liste à puces 4</nom>
 <nom code="insa" xml:lang="fr">_listepuce4</nom>
</style>
<style code="liste-puce5">
 <nom code="lyon2" xml:lang="fr">ListePuce5</nom>
 <nom code="lyon2" xml:lang="fr">List Bullet 5</nom>
 <nom code="lyon2" xml:lang="fr">WW-Liste à puces 5</nom>
 <nom code="insa" xml:lang="fr">WW-Liste à puces 5</nom>
 <nom code="insa" xml:lang="fr">_listepuce5</nom>
</style>
<style code="liste-puce6">
 <nom code="lyon2" xml:lang="fr">ListePuce6</nom>
 <nom code="lyon2" xml:lang="fr">List Bullet 6</nom>
 <nom code="lyon2" xml:lang="fr">WW-Liste à puces 6</nom>
 <nom code="insa" xml:lang="fr">WW-Liste à puces 6</nom>
 <nom code="insa" xml:lang="fr">_listepuce6</nom>
</style>
<style code="liste-puce7">
 <nom code="lyon2" xml:lang="fr">ListePuce7</nom>
 <nom code="lyon2" xml:lang="fr">List Bullet 7</nom>
 <nom code="lyon2" xml:lang="fr">WW-Liste à puces 7</nom>
 <nom code="insa" xml:lang="fr">WW-Liste à puces 7</nom>
 <nom code="insa" xml:lang="fr">_listepuce7</nom>
</style>
<style code="liste-puce8">
 <nom code="lyon2" xml:lang="fr">ListePuce8</nom>
 <nom code="lyon2" xml:lang="fr">List Bullet 8</nom>
 <nom code="lyon2" xml:lang="fr">WW-Liste à puces 8</nom>
 <nom code="insa" xml:lang="fr">WW-Liste à puces 8</nom>
 <nom code="insa" xml:lang="fr">_listepuce8</nom>
</style>
<style code="liste-simple">
 <nom code="lyon2" xml:lang="fr">ListeSimple</nom>
 <nom code="insa" xml:lang="fr">_liste</nom>
</style>
<style code="liste-titre">
 <nom code="lyon2" xml:lang="fr">ListeTitre</nom>
 <nom code="insa" xml:lang="fr">_listetitre</nom>
</style>
<!--le texte-->
<style code="text">
 <nom code="lyon2" xml:lang="fr">Text</nom>
 <nom code="lyon2" xml:lang="fr">Texte</nom>
 <nom code="lyon2" xml:lang="fr">Standard</nom>
 <nom code="insa" xml:lang="fr">_txt</nom>
 <nom code="insa" xml:lang="fr">_txte</nom>
 <nom code="insa" xml:lang="fr">Standard</nom>
</style>
<!--les titres-->
<style code="heading1">
 <nom code="lyon2" xml:lang="fr">heading 1</nom>
 <nom code="lyon2" xml:lang="fr">Heading 1</nom>
 <nom code="insa" xml:lang="fr">heading 1</nom>
 <nom code="insa" xml:lang="fr">Heading 1</nom>
</style>
<style code="heading2">
 <nom code="lyon2" xml:lang="fr">heading 2</nom>
 <nom code="lyon2" xml:lang="fr">Heading 2</nom>
 <nom code="insa" xml:lang="fr">heading 2</nom>
 <nom code="insa" xml:lang="fr">Heading 2</nom>
</style>
<style code="heading3">
 <nom code="lyon2" xml:lang="fr">heading 3</nom>
 <nom code="lyon2" xml:lang="fr">Heading 3</nom>
 <nom code="insa" xml:lang="fr">heading 3</nom>
 <nom code="insa" xml:lang="fr">Heading 3</nom>
</style>
<style code="heading4">
 <nom code="lyon2" xml:lang="fr">heading 4</nom>
 <nom code="lyon2" xml:lang="fr">Heading 4</nom>
 <nom code="insa" xml:lang="fr">heading 4</nom>
 <nom code="insa" xml:lang="fr">Heading 4</nom>
</style>
<style code="heading5">
 <nom code="lyon2" xml:lang="fr">heading 5</nom>
 <nom code="lyon2" xml:lang="fr">Heading 5</nom>
 <nom code="insa" xml:lang="fr">heading 5</nom>
 <nom code="insa" xml:lang="fr">Heading 5</nom>
</style>
<style code="heading6">
 <nom code="lyon2" xml:lang="fr">heading 6</nom>
 <nom code="lyon2" xml:lang="fr">Heading 6</nom>
 <nom code="insa" xml:lang="fr">heading 6</nom>
 <nom code="insa" xml:lang="fr">Heading 6</nom>
</style>
<style code="heading7">
 <nom code="lyon2" xml:lang="fr">heading 7</nom>
 <nom code="lyon2" xml:lang="fr">Heading 7</nom>
 <nom code="insa" xml:lang="fr">heading 7</nom>
 <nom code="insa" xml:lang="fr">Heading 7</nom>
</style>
<style code="heading8">
 <nom code="lyon2" xml:lang="fr">heading 8</nom>
 <nom code="lyon2" xml:lang="fr">Heading 8</nom>
 <nom code="insa" xml:lang="fr">heading 8</nom>
 <nom code="insa" xml:lang="fr">Heading 8</nom>
</style>
<style code="heading9">
 <nom code="lyon2" xml:lang="fr">heading 9</nom>
 <nom code="lyon2" xml:lang="fr">Heading 9</nom>
 <nom code="insa" xml:lang="fr">heading 9</nom>
 <nom code="insa" xml:lang="fr">Heading 9</nom>
</style>
<!--les images-->
<style code="figure">
 <nom code="lyon2" xml:lang="fr">Figure</nom>
 <nom code="insa" xml:lang="fr">_figure</nom>
</style>
<style code="image-ligne">
 <nom code="lyon2" xml:lang="fr">ImageLigne</nom>
 <nom code="insa" xml:lang="fr">ImageLigne</nom>
</style>
<style code="image-tab">
 <nom code="lyon2" xml:lang="fr">ImageTab</nom>
 <nom code="insa" xml:lang="fr">ImageTab</nom>
</style>
<!--les légendes-->
```

```
<style code="legende-fig">
 <nom code="lyon2" xml:lang="fr">LegendeFig</nom>
 <nom code="insa" xml:lang="fr">_legendefigure</nom>
</style>
<style code="legende-tab">
 <nom code="lyon2" xml:lang="fr">LegendeTab</nom>
 <nom code="insa" xml:lang="fr">_legendetableau</nom>
</style>
<style code="caption">
 <nom code="lyon2" xml:lang="fr">Caption</nom>
 <nom code="lyon2" xml:lang="fr">caption</nom>
 <nom code="insa" xml:lang="fr">Caption</nom>
 <nom code="insa" xml:lang="fr">caption</nom>
</style>
<!--les epigraphes-->
<style code="epigraphe">
 <nom code="lyon2" xml:lang="fr">1|Epigraphe</nom>
 <nom code="insa" xml:lang="fr">_epigraphe</nom>
</style>
<!--ce qui n'est pas traité-->
<style code="table-liste">
 <nom code="lyon2" xml:lang="fr">1|TableListe</nom>
 <nom code="insa" xml:lang="fr">_tableliste</nom>
</style>
<style code="jalon">
 <nom code="lyon2" xml:lang="fr">Jalon</nom>
 <nom code="insa" xml:lang="fr">_jalon</nom>
</style>
</styles>
```

Annexe 2

Fichier nécessaire à l'utilisation de la CEN Doc'INSA

Fichier UpCast.conf

```
<?xml version="1.0" encoding="UTF-8"?>
<plist version="1.0">
 <dict>
  <key>RTFImportFilter</key>
  <dict>
   <key>1</key>
   <dict>
    <key>DocDestResolution</key>
    <integer>96</integer>
    <key>DocbaseImageNaming</key>
    <true/>
    <key>EMFDestFormat</key>
    <string>UseWMFSubstitute</string>
    <key>ImageResolution</key>
    <integer>96</integer>
    <key>IncludeImages</key>
    <true/>
    <key>JPEGDestFormat</key>
    <string>unchanged</string>
    <key>JPEGDest</key>
    <dict>
     <key>JPEG</key>
     <dict>
      <key>Quality</key>
      <integer>100</integer>
     </dict>
     <key>PNG</key>
     <dict>
      <key>CmprType</key>
      <string>default</string>
     </dict>
    </dict>
    <key>LiteralCharStyle</key>
    <string>LITERALCHAR</string>
    <key>LiteralParStyle</key>
    <string>LITERAL</string>
    <key>LiteralProcessing</key>
    <false/>
    <key>OrigNumbering</key>
    <true/>
    <key>PICTDestFormat</key>
    <string>JPEG</string>
    <key>PICTDest</key>
    <dict>
     <key>JPEG</key>
     <dict>
      <key>Quality</key>
      <integer>100</integer>
     </dict>
     <key>PNG</key>
     <dict>
      <key>CmprType</key>
      <string>default</string>
     </dict>
    </dict>
    <key>PNGDestFormat</key>
    <string>unchanged</string>
    <key>PNGDest</key>
    <dict>
     <key>JPEG</key>
     <dict>
      <key>Quality</key>
      <integer>100</integer>
     </dict>
     <key>PNG</key>
     <dict>
      <key>CmprType</key>
      <string>default</string>
     </dict>
    </dict>
    <key>WMFDestFormat</key>
    <string>JPEG</string>
    <key>WMFDest</key>
    <dict>
     <key>JPEG</key>
     <dict>
      <key>Quality</key>
      <integer>100</integer>
     </dict>
     <key>PNG</key>
     <dict>
      <key>CmprType</key>
      <string>default</string>
     </dict>
    </dict>
   </dict>
```

```
  </dict>
  <key>XML</key>
  <dict>
   <key>5</key>
   <dict>
    <key>CSSUnitMap</key>
    <string>upcast:default-map</string>
    <key>CombineWithLogicalStyle</key>
    <false/>
    <key>CustomStylesheetPI</key>
    <string>*</string>
    <key>DOCTYPEDecl</key>
    <string>*</string>
    <key>DTDType</key>
    <string>DTD</string>
    <key>DeleteEmpties</key>
    <false/>
    <key>Extension</key>
    <string>.xml</string>
    <key>FilterName</key>
    <string>XML (upCast DTD)</string>
    <key>IncludeHiddenContents</key>
    <true/>
    <key>IncludeVisual</key>
    <true/>
    <key>NamespacePrefix</key>
    <string/>
    <key>OutputEncoding</key>
    <string>UTF-8</string>
    <key>TableModel</key>
    <string>CALS</string>
    <key>UnicodeTranslationMap</key>
    <string>upcast:xml-map</string>
    <key>UseNamespace</key>
    <false/>
    <key>Validate</key>
    <false/>
    <key>WriteDTD</key>
    <true/>
    <key>WriteTOC</key>
    <false/>
   </dict>
  </dict>
  <key>application</key>
  <dict>
   <key>name</key>
   <string>upCast</string>
   <key>stats</key>
   <dict>
    <key>CharacterCounter</key>
    <integer>0</integer>
    <key>ConversionCounter</key>
    <integer>0</integer>
    <key>StartDate</key>
    <string>Thu, 03 Jul 2003 09:16:02 CEST</string>
   </dict>
   <key>version</key>
   <integer>1024</integer>
  </dict>
  <key>converter</key>
  <dict>
   <key>Dflt</key>
   <dict>
    <key>SwingPLAF</key>

 <string>com.sun.java.swing.plaf.windows.WindowsL
ookAndFeel</string>
    <key>imageDir</key>

 <string>C:\utilisateurs\fred\cen\Chaine\Fichiers\Corps
\images\</string>
    <key>inFilterClass</key>

 <string>de.infinityloop.upcast.importfilters.RTFImpor
tFilter</string>
    <key>inFilterInstID</key>
    <string>1</string>
    <key>inputFile</key>

 <string>C:\utilisateurs\fred\cen\Chaine\Fichiers\Corps
\RTF.rtf</string>
    <key>instanceID</key>
    <integer>6</integer>
    <key>outFilterClass0</key>

 <string>de.infinityloop.upcast.exportfilters.ExportFilt
er3</string>
    <key>outFilterClass1</key>

 <string>de.infinityloop.upcast.exportfilters.ExportFilt
er6</string>
    <key>outFilterCount</key>
    <integer>1</integer>
    <key>outFilterInstID0</key>
    <string>5</string>
    <key>outFilterInstID1</key>
    <string>4</string>
    <key>outputDir</key>

 <string>C:\utilisateurs\fred\cen\Chaine\Fichiers\Corps
\rtf_gen\</string>
   </dict>
   <key>UserInstanceID</key>
   <string>Dflt</string>
  </dict>
  <key>global</key>
  <dict>
   <key>TimedExecutionSwitch</key>
   <false/>
  </dict>
 </dict>
</plist>
```