Projet CITHER

Rapport de Projet de Fin d'Etude

Rédacteur	: Julien Tognazzi	Projet	: CITHER
Date de rédaction	: 6 septembre 1999	Version	: 1.0
Dernière mise à jour	: 28 septembre 1999	Référence	: Rapport de Projet de Fin d'Etude
Date d'impression	: 13 octobre 1999	Diffusion	: Interne

Remerciements

Je tiens tout d'abord à remercier Madame Monique JOLY, responsable de Doc'INSA, qui m'a accueilli au sein de son service et m'a permis de participer à ce projet.

Je remercie Monsieur Jean–Marie PINON, professeur de l'INSA et enseignant responsable de mon stage, pour son encadrement.

Je remercie particulièrement Jean–Michel MERMET, tuteur de mon stage, pour ses conseils lucides et pertinents.

Je remercie sincèrement tout le personnel de Doc'INSA pour sa sympathie et pour m'avoir fait découvrir l'envers du décor d'une bibliothèque.

A l'issue de trois agréables années au sein du département informatique de l'INSA de Lyon, j'adresse des remerciements particuliers à Monsieur Yves MARTINEZ, directeur du département, pour le dynamisme de ce département d'études, à Jacqueline MARTINEZ et Odile CLEMENT pour leur gentillesse et leur efficacité, et à toute l'équipe enseignante pour la qualité de l'enseignement qui nous a été dispensé.

Sommaire

Ce document présente l'étude réalisée par Julien TOGNAZZI, à Doc'INSA, de Juin à Septembre 1999, lors de son Projet de Fin d'Etude.

1. Objet du Projet

Une première étude à été menée durant l'année 1997/1998 aboutissant à la mise en place d'un serveur de thèses en texte intégral à la bibliothèque Doc'INSA, dépositaire des thèses produites à l'INSA de LYON

Ce projet constitue la suite de cette première étude, par l'extension des fonctionnalités de la chaîne de traitement, pour une montée en charge du serveur (conversion de fichiers sources en Latex, portabilité de la chaîne, etc.) et l'analyse de nouvelles technologies pouvant servir le projet CITHER. Une étude du langage XML a été menée sur ses possibilités en matière d'archivage, de publication et de sécurisation / authentification.

2. Contexte

2.1. Les thèses de Doc'INSA

Doc'INSA, dépositaire officiel de toutes les thèses soutenues au sein des laboratoires de l'INSA de LYON, reçoit chaque année environ 130 thèses. Ces thèses, conservées en deux exemplaires, peuvent être consultées à la bibliothèque. Dans le cadre du prêt entre bibliothèques, des reproductions totales ou partielles de ces thèses (photocopie) sont envoyées aux bibliothèques demandeuses. Il existe, de plus un exemplaire sous forme de microfiche dans toutes les bibliothèques universitaires.

2.2. Intérêt de la publication électronique des thèses

Les thèses sont des documents qui peuvent avoir une durée de vie courte. Par ailleurs, ce sont des documents non commercialisés (ils font partie de la littérature grise) et de ce fait sont peu connus du grand public, et peu accessibles. Il importe donc que ces thèses soient mises à disposition des lecteurs éventuels le plus rapidement possible et qu'un accès international soit proposé.

Un accès électronique à ces thèses (via Internet) offre de nouvelles possibilités, comme la recherche en texte intégrale, le téléchargement et la reproduction partielle d'une thèses suivant les besoins.

D'autre universités de la région Rhône-Alpes devraient rejoindre ce projet, comme Lyon I, et augmenter le nombre de thèses annuelles à traiter. Actuellement, plus de 2000 thèses sont conservées à Doc'INSA, et une vingtaine est d'ores et déjà disponibles en texte intégrale sur le serveur CITHER[1].

[1] http://csidoc.insa-lyon.fr/these

2.3. Déroulement du projet

Le projet s'est déroulé de la manière suivante :

Une première phase comprenant l'étude de l'existant : le projet CITHER dans son ensemble, les réalisations des études précédentes.

Ensuite, l'analyse des besoins, en interviewant les différents acteurs du projet (l'opérateur, le responsable technique, les coordonnateurs)

Un état de l'art sur XML, les possibilités offertes par ce langage dans le cadre du projet.

Puis une phase de développement / maintenance avec l'intégration d'une chaîne de conversion pour les thèses LaTeX à l'application existante (le CEN), la modification du guide opérateur, la correction et l'ajout de nouvelles fonctionnalités et la mise en portabilité de l'ensemble pour tous les systèmes Windows 32 bits.

3. Documents de référence

Rapports de la première phase du projet menée par Marc-Etienne Huneau de Novembre 1997 à Juin 1998

- *Dossier d'initialisation,*
- *Règles d'édition électronique*
- *Manuel Technique*
- *Manuel Utilisateur*
- *Rapport de Projet de Fin d'Etudes*

Mémoire de stage du DESS en informatique documentaire de Jean-Michel Mermet

- *Coordination et mise en place d'un serveur de thèses en texte intégral à l'INSA de Lyon.*

4. Analyse de l'existant

4.1. Le poste de conversion

Le poste de conversion se compose de l'ensemble Logiciels/Matériels suivant :

- *Un PC sous Windows 95*
- *Un scanner*
- *Un graveur de CD-ROM pour l'archivage*
- *L'application Chaîne d'édition numérique (CEN)*
- *MS Office 97*
- *Adobe Acrobat 3*

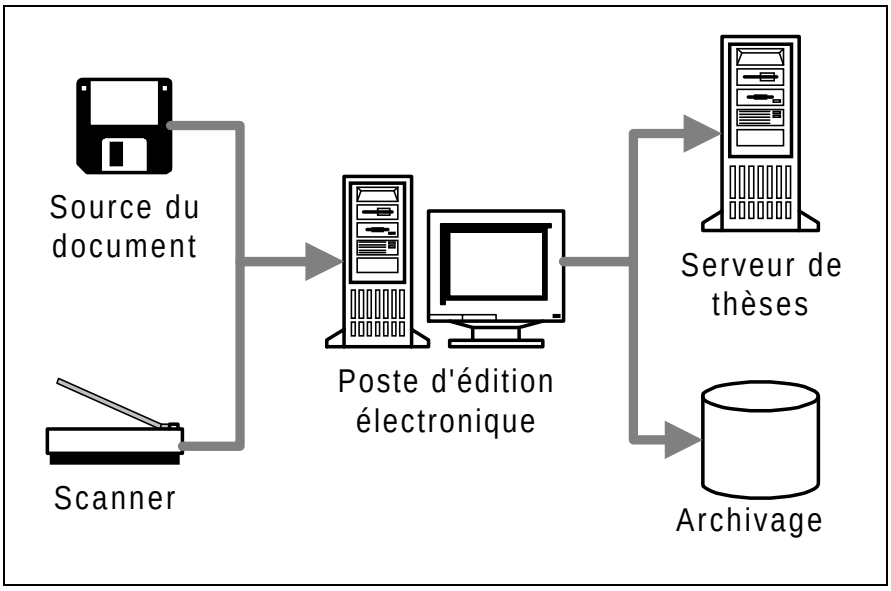

Figure 1 : Vue générale du dispositif

4.2. La chaîne d'édition numérique (CEN)

La chaîne d'édition numérique ou CEN est le logiciel développé lors de la précédente étude. Cette application, programmée sous Delphi 3 dans l'environnement Windows 32 bits, prend en charge le traitement des fichiers électroniques, du fichier source (au format Word 97) jusqu'à la publication sur le serveur.

Elle contrôle les autres applications via plusieurs mécanismes : MS Word et Acrobat Exchange sont pilotés via COM/OLE[2], alors que Acrobat Distiller est contrôlé par des messages Windows[3].

Le format de publication utilisé est le format propriétaire Adobe PDF[4].

[2] **C**ommon **O**bject **M**odel / **O**bject **L**ink **E**mbedding : modèle objet de Windows.
[3] Mécanisme de base de communication entre les entités de Windows
[4] PDF: **P**ortable **D**ocument **F**ormat. Format propriétaire développé par Adobe

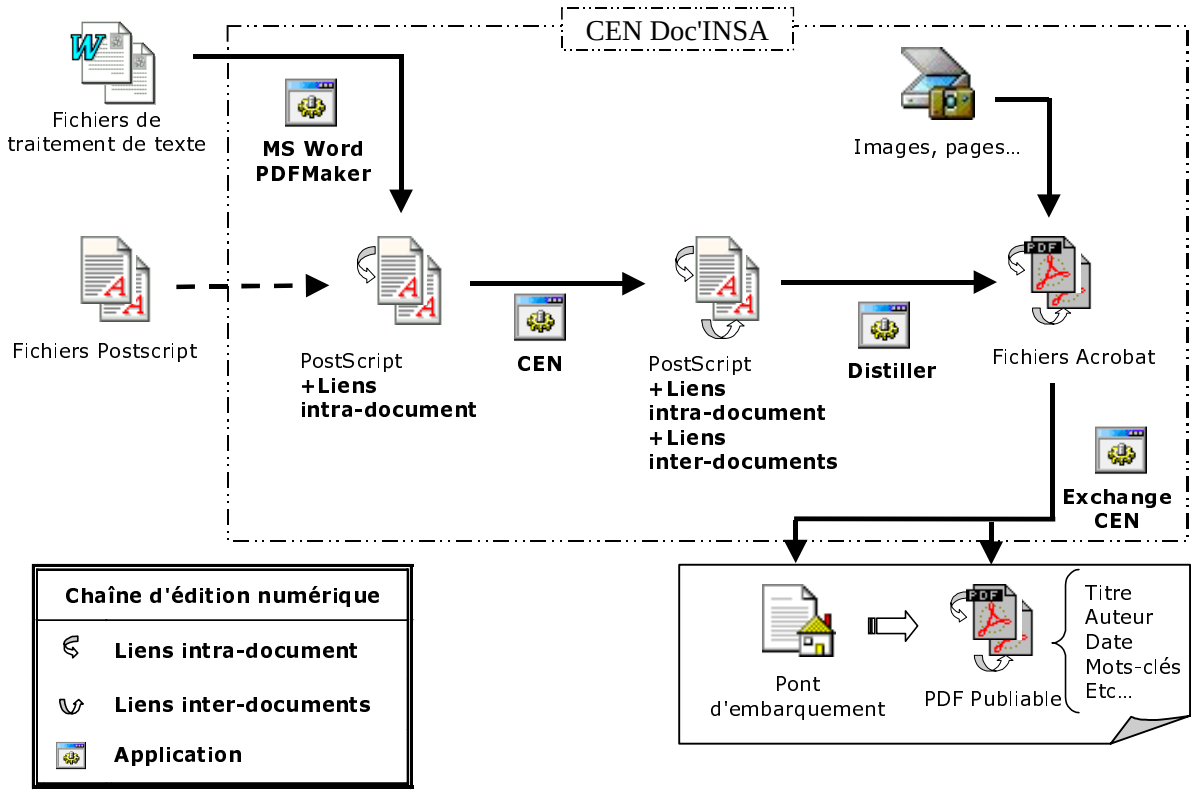

Figure 2 : Opérations de la chaîne d'édition

La conversion se déroule en quatre étapes :

- *Tout d'abord, une macro-commande Word (Adobe PDFMaker [Adobe 98]) crée un fichier PostScript enrichi d'instructions* pdfmark[5] *[Adobe 97] à l'intention d'Acrobat Distiller. Cette macro-commande crée (le cas échéant) des liens à partir des champs 'note', 'table des matières', etc. Elle crée également un repère Acrobat pour chaque titre (Liens intra-document).*
- *Les fichiers PostScript obtenus sont alors directement modifiés par l'application qui y ajoute des repères (toujours via* pdfmark*) désignant les autres fichiers (Liens inter-documents).*
- *Les fichiers PostScript sont ensuite convertis en PDF par Distiller.*
- *Enfin, les fichiers PDF sont 'retraités' à l'aide d'Exchange : leurs champs titre, sujet, auteur, etc. sont renseignés ; les miniatures de pages sont créées et les fichiers optimisés pour une lecture en ligne (opération permettant au serveur d'envoyer le document page à page).*

A ce point, le traitement par lot est terminé, et un rapport de conversion a été généré.

L'application génère en outre un "pont d'embarquement" vers la thèse, page HTML rassemblant la référence bibliographique du document et des liens vers tous les fichiers PDF. Enfin, elle peut préparer les fichiers à un archivage en les rassemblant dans un répertoire.

Le format PDF est un langage de représentation de page, impropre à l'archivage : Ne comprenant pas la notion de structure logique de document (paragraphes, titres, etc.), il ne peut efficacement servir de source à une éventuelle conversion vers un nouveau format. La solution actuelle d'archivage garde donc les fichiers PDF publiables et les documents sources

[5] Opérateur du langage PostScript, destiné à Acrobat Distiller

(fournis par l'auteur et éventuellement retouchés sur le poste d'édition), pour permettre une évolution vers de nouveaux formats (SGML ou XML).

Par ailleurs, un guide de conversion sous forme de liste de contrôles permet à l'opérateur de se repérer dans les différentes phases de la conversion.

5. Analyse des besoins du projet

De nouveaux besoins ont été définis par Doc'INSA avec l'arrivée au sein du projet d'autres universités (notamment Lyon I pour l'année 1999/2000) :

- *Maintenance et évolutions du CEN, pour optimiser le temps de conversion d'un document, et corriger les problèmes existants.*
- *Etude de la portabilité de la chaîne d'édition numérique, pour permettre une installation facile sur de nouveaux postes de conversion.*
- *Extension des types de fichiers sources acceptés en entrée de chaîne, avec plus particulièrement l'intégration des fichiers sources en LaTeX.*
- *Réflexion sur les possibilités offertes par le langage XML comme format d'archivage ou de publication.*

En cours d'étude, une réorientation du projet sur l'intégration des thèses LaTeX a mis en suspens la réflexion sur le langage XML.

5.1. Maintenance de la chaîne d'édition

Plusieurs entretiens avec l'opérateur de conversion ont permis de définir les problèmes ou manques de l'application, notamment au niveau du guide opérateur.

Une mise à jour du guide a été effectuée, tenant compte de l'expérience acquise par l'opérateur.

La correction et l'ajout de plusieurs fonctionnalités ont été implémentées :

- *Fonction d'impression du rapport de conversion*
- *Définition de l'URL[6] en fonction du nom de l'auteur et de la date de soutenance*
- *Ajout automatique d'un nouveau lien dans les fichiers PDF pour revenir au pont d'embarquement.*

5.2. Portablilité de l'application

L'application CEN a été développée sous Delphi 3, en environnement Windows 95. Mais, jamais aucun test n'avait été effectué quant à sa portabilité sur d'autres machines, ou sur d'autres systèmes Windows 32 bits (Windows NT/98).

Une installation sur un poste Windows NT, et sur un nouveau poste de conversion équipé de Windows 98, mît en évidence certains problèmes :

- *Clés manquantes dans la base de registre Windows pour l'interface COM/OLE des produits Acrobat.*
- *Fonctionnement perturbé par le déplacement des répertoires de travail*

[6] URL à laquelle les fichiers seront transférés sur le serveur.

Une fois ces problèmes détectés, ils ont été résolus en modifiant la procédure d'installation et en corrigeant le code correspondant de l'application CEN.

6. Intégration de LaTeX

La part de thèses rédigées en LaTeX sur l'INSA est faible mais non négligeable[7], et avec l'arrivée de Lyon I dans le projet, elle va augmenter fortement.

6.1. Présentation de LaTeX

LaTeX est un traitement de texte particulièrement adapté à la rédaction de documents scientifiques et mathématiques, mais il sert aussi à écrire toutes sortes de documents, de la simple lettre, à des livres complets. Il est utilisé par beaucoup d'étudiants, de chercheurs et d'éditeurs à travers le monde.

LaTeX faisant partie du monde des logiciels libres, il est disponible sur la plupart des plates-formes informatiques, du PC au Mac, en passant par les systèmes Unix et VMS.

Dans le cadre du projet, il a été décidé d'étudier l'intégration du traitement des fichiers LaTeX à la chaîne d'édition numérique à partir d'une distribution Windows 32 bits.

6.2. Choix de la distribition

Plusieurs distributions existent pour Windows, proposant toutes un environnement complet (Miktex, Fptex, etc.). Fptex [Fptex 99] a été choisi pour les tests, pour son suivi des programmes en cours de développement, (notamment PdfTex, un programme de conversion de fichiers latex en PDF) et ses mises à jour régulières.

[7] 7% des thèses recensées lors d'une enquête de Novembre 1996 à Novembre 1997

6.3. Une nouvelle chaîne de traitement

Une étude des différents programmes de conversion présents sous LaTeX à mis en évidence deux chaînes de traitement possibles :

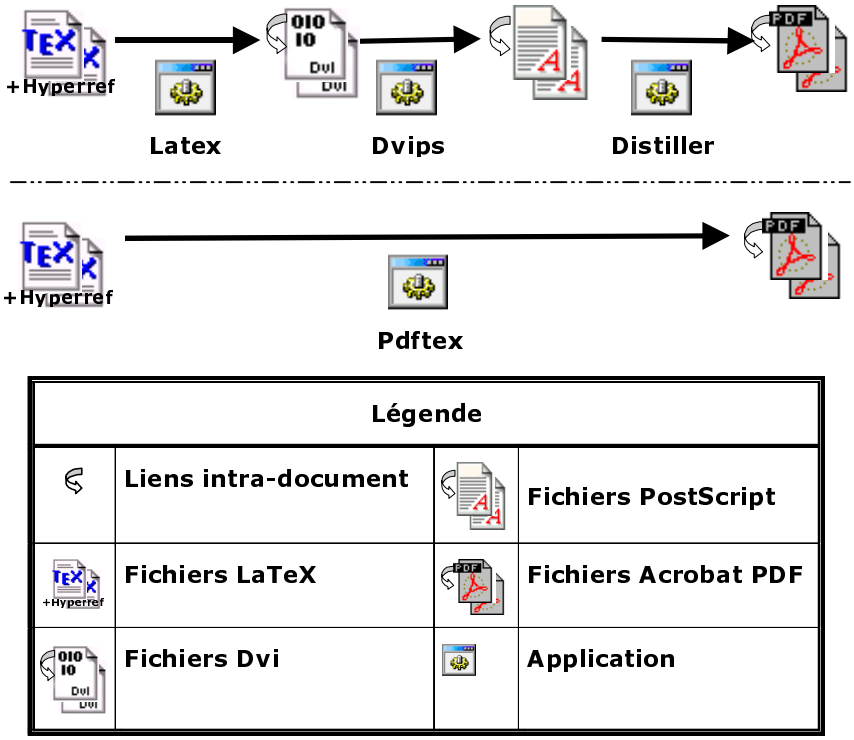

Figure 3 : Chaînes de traitement LaTeX

La première chaîne utilise le format de sortie traditionnel de LaTeX : le fichier dvi[8]. Ensuite, un premier programme, Dvips, convertit le fichier Dvi en fichier PostScript, et enfin le programme Distiller d'Adobe Acrobat, transforme le fichier PostScript en fichier PDF.

La deuxième chaîne est basée sur un nouveau programme, encore en cours de développement, PdfTex [Pdftex 99]. Pdftex remplace la compilation traditionnelle Latex, pour donner directement un fichier de sortie au format PDF, et non plus un fichier Dvi.

Dans les deux cas, l'intégration des liens intra-document, s'effectue par l'ajout du module Hyperref [Hyper 99]dans le préambule (en-tête) du fichier source Latex.

Ce module permet de définir au moyen de commandes Pdfmark les renvois aux notes, la table des matières dynamiques, etc., de la même manière que la macro-commande PDFMaker pour les fichiers Word. Il permet de plus une gestion des "back references" pour la bibliographie, en indiquant après chaque référence bibliographique les pages où elles ont été citées.

Ces commandes sont intégrées, pour la première chaîne, au fichier Dvi et PostScript puis interprétées par Distiller lors de la conversion au format PDF.

[8] DVI : Device Independent (indépendant du périphérique de sortie)

Pour la deuxième chaîne, ces commandes sont directement intégrées lors de la création du fichier PDF.

6.4. Comparaison des différentes chaînes

I^{ère} chaîne	**2ème chaîne**
Utilisation de 3 programmes Latex, Dvips, Distiller	Un seul programme Pdftex
Programmes stabilisés, offrant un comportement sûr	Programme encore en phase de développement
Compatible avec tous les formats d'image utilisés sous LaTeX	Les fichiers d'image Eps ne sont encore pas reconnus

La chaîne de traitement basée sur Pdftex, permet une conversion plus simple et plus rapide, mais l'absence de reconnaissance du format Eps est un inconvénient majeur, ce type de fichier étant très utilisé par les utilisateurs Unix/Linux, principaux rédacteurs sous LaTeX.

Notre choix s'est donc porté sur la première chaîne de traitement présentée, comprenant l'utilisation successive des programmes latex, dvips, distiller.

Remarque à propos des références croisées

Pour une bonne gestion des références croisées sous Latex, il est nécessaire d'effectuer plusieurs passes (généralement deux). De plus, dans le cas d'un document contenant une bibliographie, comme c'est le cas pour une thèse, On doit faire appel à un autre programme, Bibtex, pour les références à la bibliographie.

On obtient donc une chaîne de conversion faisant appel à 4 programmes différents (Latex, Bibtex, Dvips, Distiler) dont certain doivent être lancés plusieurs fois successivement (Latex, Bibtex).

L'utilisation d'un script Perl[9] Latexmk, résout ce problème, en s'assurant lui-même du bon enchaînement des programmes Latex, Bibtex et Dvips. Il ne reste plus qu'à lancer Distiller pour obtenir le fichier PDF.

[9] Perl : Practical Extraction and Report Language. Langage de script très puissant développé par Larry Wall

6.5. Intégration au CEN

La chaîne de conversion LaTeX choisie, utilisant le programme Acrobat Distiler, comme la chaîne de conversion Word, elle peut être intégrée facilement à la chaîne existante.

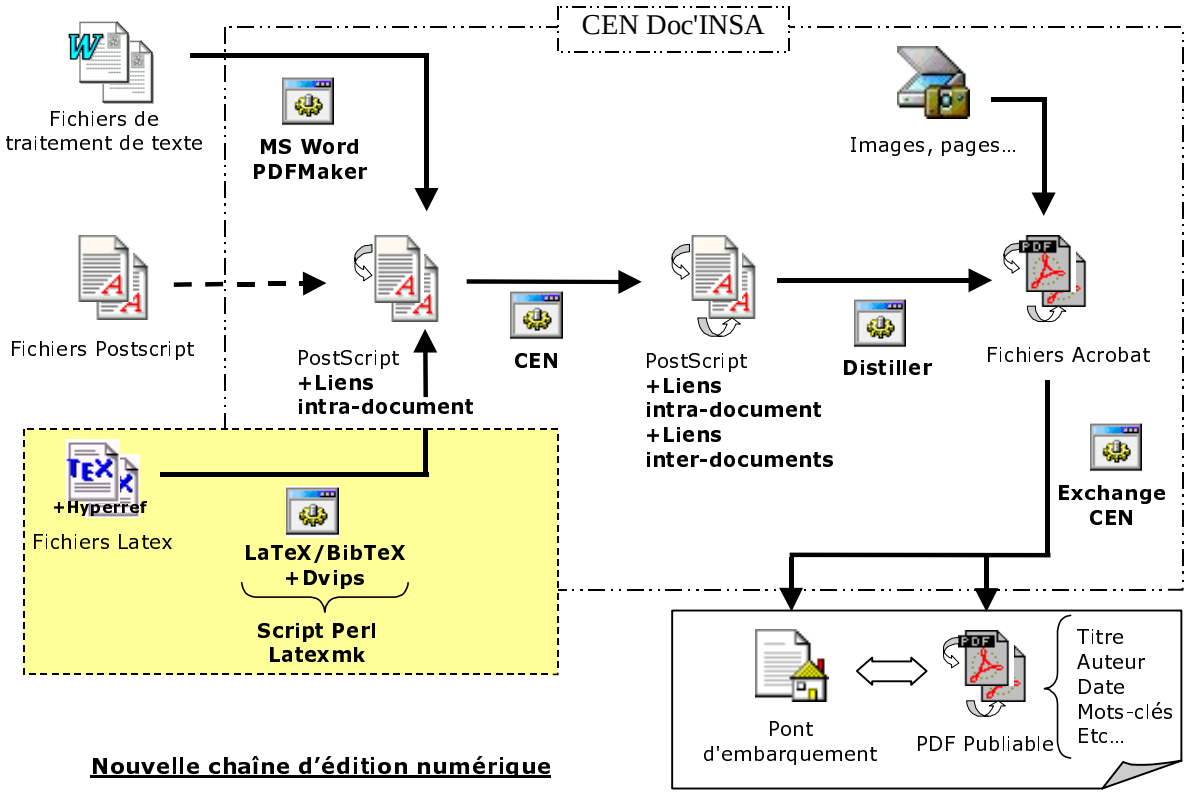

Figure 4 : Intégration à la chaîne existante

L'application CEN lance le script Perl qui s'occupe de la conversion des fichiers Latex en fichiers Postscript contenant les liens intra-document. A partir de là, on rejoint la chaîne existante qui poursuit la conversion par la création des liens inter-documents (dans le cas de plusieurs fichiers à traiter) puis des fichiers PDF par le Distiller et enfin, un retraitement et l'optimisation des fichiers avec Acrobat Exchange.

Remarque :

Le CEN prévoit le cas de thèses mixtes, où une partie du document serait développée sous LaTeX, et une autre sous MS Word (par exemple, la page de titre et certaines annexes en Word, et la thèse en Latex).

7. Evolution vers XML ?

7.1. Le langage XML

Le langage XML est un langage de balisage de document, comme HTML. Une recommandation du W3C[10] du 10 février 1998 définit XML dans sa version 1.0 [XML 98].

XML semble un format prometteur puisque tous les grands acteurs du monde informatique le soutiennent (Oracle, IBM, SUN, Microsoft, etc.).

Ses principales caractéristiques sont :

- *Un ensemble extensible de balises (contrairement à HTML)*
- *Une séparation entre la présentation et les données.*
- *Un codage de caractère en UNICODE – ISO 10646*
- *Une bonne adaptation à la diffusion sur internet.*

Un document XML s'accompagne généralement d'une DTD (Document Type Définition) qui permet de définir la structure du document XML, et de le valider.

7.2. Les développements liés

D'autre développements sont en cours, liés au langage XML.

XSL [XSL 99] est en langage pour définir des feuilles de style, il est composé de deux parties :

- *un langage de transformation de documents XML permettant, à partir d'un document source XML, de produire un document cible XML composé de nouveaux éléments et / ou d'éléments présents dans le document source.*
- *un langage permettant de spécifier de manière très précise la présentation des données.*

Une feuille de style XSL spécifie la présentation d'une classe de documents XML en décrivant comment une instance de cette classe est transformée en un autre document XML utilisant le langage de présentation des données.

Il est donc possible, à partir d'un fichier pivot, de dériver plusieurs versions adaptées au périphériques de sortie, en définissant les feuilles de style appropriées. (par exemple : impression sur papier A4, affichage à l'écran, etc.)

Enfin, XLL [XLL 99] (XML Linking Language) définit les liens hypertextes au sein du document XML. Une distinction est faite entre les liens externes, et les liens internes pointant sur des documents XML. Un lien est une relation explicite entre au moins deux données ou ensemble de données.

[10] W3C :World Wide Web Consortium

7.3. CITHER et XML

Une première étape consiste à vérifier l'existence de DTD permettant la définition de longs documents structurés. Plusieurs sont disponibles sur internet :

- *The Book DTD – ISO 12083*
- *La DTD du TEI (Text Encoding Initiative) ou sa version simplifiée TEI lite [TEI 99]*

Ces DTD ont été reprises de SGML, elles sont très complètes, et permettent la définition d'une thèse.

Il semble donc possible de définir un fichier au format XML, ne contenant que les données, la structure logique (Titre, auteur, texte, citation) et d'utiliser ce fichier pour l'archivage et la génération des différents formats pour la publication (HTML, PDF, XML).

Le problème réside dans l'obtention d'un tel fichier à partir des documents sources fournis par les doctorants (fichiers Word ou Latex)

Les outils de conversion (pour la génération du fichier pivot XML et ensuite sa dérivation en plusieurs autre formats de sortie) n'étant pas encore complètement disponibles, il a été jugé que l'on n'obtiendrait pas la qualité offerte par les fichiers PDF et qu'il valait mieux attendre l'aboutissement de toutes les normes liées à XML (XSL, XLL, etc.) et l'arrivée d'outils de conversion et de visualisation.

Remarque

Plusieurs d'universités, (l'université Laval au Québec, Les Presses universitaires de Montréal, l'université Lyon II, etc.) se sont associées pour le développent d'une chaîne de conversion de thèses autour des langages SGML[11]/XML [LyonII 99]. Dans le cadre de la réflexion sur l'évolution du projet CITHER, plusieurs réunions ont été tenues avec ce groupe d'universités pour la présentation de la chaîne, et son test avec une thèse scientifique de l'INSA. Les résultats sont encore insuffisants dans l'état actuel du projet, mais une collaboration est envisageable.

8. Conclusion

Le projet CITHER offre maintenant une chaîne de traitement plus complète pour la publication électronique des thèses de l'INSA.

Plus d'une vingtaine de thèses ont déjà été converties, et la chaîne semble maintenant prête à une montée en charge.

Mais le projet ne s'arrête pas là, un groupe de travail à été formé sous l'impulsion de Doc'INSA, pour étudier la mise en place de feuilles de style propres aux thèses pour aider les étudiants dans leur rédaction et pour faciliter les traitements de conversion. Une formation à la rédaction de longs documents structurés va être instaurée.

[11] Langage de balisage définit en 1986, à l'origine d'HTML et d'XML.

Un rapprochement avec d'autres projets est aussi envisagé :

- *Pour le test de nouvelles chaînes de conversion,*
- *Le partage de connaissances,*
- *La mise en commun des méta-données*[12]*,*
- *La constitution d'un catalogue des thèses électroniques.*

Avec l'arrivée de MS Office 2000 sur le marché et la sortie de la version 4.0 d'Adobe Acrobat, les prochains développements du projet vont consister à étudier leur intégration au CEN. La veille technologique autour de XML va être maintenue pour permettre l'évolution du projet vers ce langage dès que cela sera possible.

9. Références bibliographiques

[Hyper 99], *Hypertext marks in LaTeX: the hyperref package*, [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet :
<URL:http://tug.org/applications/hyperref/manual.html>

[TIE 99], *Text Encoding Initiative* [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet :
<URL:http://www-tei.uic.edu/orgs/tei/>

[Pdftex 99] *PDFTeX support* [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : <URL:http://www.tug.org/applications/pdftex/>

[LaTeX 99] *Une courte (?) introduction à LaTeX* [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet :
<URL:ftp://ctan.tug.org/tex-archive/info/lshort/french/flshort-3.3.pdf>

[Fptex 99] fpTeX 0.3 User's manual [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet :
<URL:ftp://ftp.loria.fr/tex-archive/systems/win32/fptex/fptex.pdf>

[XML 98] Extensible Markup Language (XML) 1.0 [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet :
<URL:http://www.w3.org/TR/1998/REC-xml-19980210>

[XSL 99] Extensible Stylesheet Language (XSL) working draft [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet :
<URL:http://www.w3.org/TR/WD-xsl/>

[XLL 98] XML Linking Language (XLink) working draft [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet :
<URL:http://www.w3.org/TR/1998/WD-xlink-19980303>

[Huneau 98] Huneau M.E., "*Serveur de thèses en texte intégral : Rapport de Projet de Fin d'Etudes"* [On-line]. Villeurbanne (Fr.) : INSA – IF, 1998, 29 p. Available from internet : <URL: http://csidoc.insa-lyon.fr/these/doc/rapport_pfe.pdf>

[12] Les meta-données sont les informations portant sur le document (nom de l'auteur, année, laboratoire, résumé, abstract, mots-clés, etc.)

[Adobe 97] **Adobe Developer Support**, *Acrobat Distiller Control Interface Specification*. Adobe, July 1997, Technical Note #5158

[Adobe 98] *Adobe PDFMaker 1.0 for Microsoft Word 97* [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : <URL:http://www.adobe.com/supportservice/custsupport/LIBRARY/4d9e.htm>

[LyonII 99] Service des Nouvelles Technologies pour l'Information Et la Réalisation de Serveurs (SENTIERS) [On-line]. Septembre 1999 [Visité le 13 Septembre 1999] Available from internet : <URL:http://phebus.univ-lyon2.fr/sentiers/>

ANNEXES

Projet Cither

Intégration de LaTeX

Rédacteur	: Julien Tognazzi	Projet	: CITHER
Date de rédaction	: 6 septembre 1999	Version	: 1.0
Dernière mise à jour	: 28 septembre 1999	Référence	: Rapport de Projet de Fin d'Etude
Date d'impression	: 13 octobre 1999	Diffusion	: Interne

Sommaire

10. Introduction

Ce document présente les modification apportées au logiciel de conversion "Chaîne d'Edition Numérique" pour l'intégration des thèses rédigées sous LaTeX.

11. Documents de référence

Les documents cités ci-dessous sont disponible au format PDF dans le répertoire "C:\DocInsa\Latex\Documents de référence\".

- [lshort.pdf] Tobias Oetiker, Hubert Partl, Irene Hyna and Elisabeth SchleglThe. "The Not So Short Introduction to LaTeX 2ε", Version 3.7, 14. April, 1999
- [flshort.pdf] Tobias Oetiker, Hubert Partl, Irene Hyna et Elisabeth Schlegl, traduit en français par Matthieu Herrb, "Une courte (?) introduction à LaTeX 2ε", Version 3.3, Février 1999
- [guide1998.pdf] J.-M. Hufflen, D. Roegel, K. Tombre LORIA, *"Guide local (L A)T E X du LORIA Millésime 1998"*, Septembre 1998
- [Hyperref's manuel.pdf] Sebastian Rahtz, *"Hypertext marks in LaTeX: the **hyperref** package"*, June 1998

12. Conversion de documents avec LaTeX

Un document LaTeX est un document texte normal contenant des commandes typographique (ou autres) qui sont interprétées par le compilateur LaTeX pour générer un fichier affichable et/ou imprimable. Le fichier de sortie d'une compilation LaTeX est un fichier DVI.

12.1. Structure du Fichier source

Quand LaTeX analyse un Fichier source, il s'attend à y trouver une certaine structure. C'est pourquoi chaque fichier source doit commencer par la commande :

```
\documentclass{…}
```

Elle indique quel type de document vous voulez écrire. Après cela vous pouvez insérer des commandes qui vont influencer le style du document ou vous pouvez charger des extensions qui ajoutent de nouvelles fonctions au système LaTeX. Pour charger une extension, utilisez la commande :

```
\usepackage{…}
```

Quand tout le travail de préparation est fait, vous pouvez commencer le corps du texte avec la commande :

```
\begin{document}
```

Maintenant vous pouvez saisir votre texte et y insérer des commandes LaTeX. A la fin de votre document, utilisez la commande

```
\end{document}
```

12.2. La chaîne de conversion

Une compilation complète (pour obtenir le fichier DVI) s'effectue de la manière suivante :

- 1ere passe Latex
- lancer BibTeX si un fichier bibliographie est utilisé.
- 2eme passe Latex
- Si Latex signale des référence indéfinies (Undefined References), une 3eme passe est nénessaire

Quand Latex ne signale plus de références indéfinies, on a un fichier Dvi qu'il va être nécessaire de transformer en fichier PostScript à l'aide du programme Dvips.

La ligne de commande est la suivante :

```
Dvips –o fichier.ps fichier.dvi
```

L'option –o permet de définir le nom du fichier de sortie postscript, la sortie par défaut est 'lpr', qui sous Unix désigne le driver d'impression, provoque une erreur sous Windows.

Enfin, le fichier PostScript obtenu est converti en fichier PDF, par le distiller d'Acrobat. Lancer le distiller, ouvrir le fichier '.ps', on obtient alors un fichier PDF.

Normalement, il n'est pas nécessaire de savoir tout cela, un fichier script Latexmk, s'occupant du bon enchaînement des programmes jusqu'à la production du fichier PostScript.

Le fichier PostScript est transformé en fichier PDF par Acrobat Distiller, de la même manière que lors du traitement d'un fichier Word.

12.3. L'extension Hyperref

Pour permettre la gestion des liens hypertextes sur les références croisées et les références bibliographique, il est nécessaire d'utiliser l'extension Hyperref. La version utilisée actuellement est la `6.65d`.

La commande à rajouter dans les fichiers sources LaTeX est :

```
\usepackage[option(s)]{hyperref}
```

12.3.1.Options générales

Nom	Type	Défaut	Description
draft	boolean	false	all hypertext options are turned off

debug	boolean	false	extra diagnostic messages are printed in the log file
a4paper	boolean	true	sets paper size to 210mm x 297mm
a5paper	boolean	false	sets paper size to 148mm x 210mm
b5paper	boolean	false	sets paper size to 176mm x 250mm
letterpaper	boolean	false	sets paper size to 8.5in x 11in
legalpaper	boolean	false	sets paper size to 8.5in x 14in
executivepaper	boolean	false	sets paper size to 7.25in x 10.5in

12.3.2.Options de configuration

Nom	**Type**	**Défaut**	**Description**
Raiselinks	boolean	true	In the hypertex driver, the height of links is normally calculcated by the driver as simply the base line of contained text; this options forces \special commands to reflect the real height of the link (which could contain a graphic)
Breaklinks	boolean	false	Allows link text to break across lines; since this cannot be accomodated in PDF, it is only set true by default if the pdftex driver is used. This makes links on multiple lines into different PDF links to the same target.
Pageanchor	boolean	true	Determines whether every page is given an implicit anchor at the top left corner. If this is turned off, \tableofcontents will not contain hyperlinks.
Plainpages	boolean	true	Forces page anchors to be named by the arabic form of the page number, rather than the formatted form.
Nesting	boolean	false	Allows links to be nested; no drivers currently support this.

12.3.3. Options étendues

Nom	**Type**	**Défaut**	**Description**
Extension	text		Set the file extension (eg dvi) which will be appended to file links created if you use the xr package.
backref	boolean	false	Adds ‘backlink’ text to the end of each item in the bibliography, as a list of section numbers. This can only work properly if there is a blank line after each \bibitem.
pagebackref	boolean	*false*	Adds ‘backlink’ text to the end of each item in the bibliography, as a list of page numbers.
Hyperindex	boolean	false	Makes the text of index entries into hyperlinks. Easily broken…
Colorlinks	boolean	*false*	Colours the text of links and anchors. The colors chosen depend on the the type of link. At present the only types of link distinguished are citations, page references, URLs, local file references, and other links.
Linkcolor	color	*red*	Color for normal internal links.
anchorcolor	color	*black*	Color for anchor text.
citecolor	color	*green*	Color for bibligraphical citations in text.
filecolor	color	*magenta*	Color for URLs which open local files.
menucolor	color	*red*	Color for Acrobat menu items.
pagecolor	color	*red*	Color for links to other pages.

urlcolor	color	*cyan*	Color for linked URLs.

12.3.4.Options PDF

Nom	**Type**	**Défaut**	**Description**
Bookmarks	boolean	*false*	A set of Acrobat bookmarks are written, in a manner similar to the table of contents, requiring two passes of LaTeX. Some post-processing of the bookmark file (file extension .out) may be needed to translate LaTeX codes, since bookmarks must be written in PDFEncoding. To aid this process, the .out file is not rewritten by LaTeX if it is edited to contain a line \let\WriteBookmarks\relax
bookmarksopen	boolean	*false*	If Acrobat bookmarks are requested, show them with all the subtrees expanded.
bookmarksnumbered	boolean	*false*	If Acrobat bookmarks are requested, include section numbers. pdfhighlight name */I* How link buttons behave when selected; /I is for inverse (the default); the other possibilities are /N (no effect), /O (outline), and /P (inset highlighting).
citebordercolor	RGB color	0 1 0	The color of the box around citations
filebordercolor	RGB color	0 .5 .5	The color of the box around links to files
linkbordercolor	RGB color	1 0 0	The color of the box around normal links
menubordercolor	RGB color	1 0 0	The color of the box around Acrobat menu links
pagebordercolor	RGB color	1 1 0	The color of the box around links to pages
urlbordercolor	RGB color	*0 1 1*	The color of the box around links to URLs
pdfborder		*0 0 1*	The style of box around links; defaults to a box with lines of 1pt thickness, but the colorlinks option resets it to produce no border.

12.3.5.Options intéressantes

Voici la ligne de commande hyperref à utiliser dans le cadre du projet CITHER pour répondre aux exigences définies:

```
\usepackage[colorlinks, linktocpage, pagebackref, a4paper, bookmarks, bookmarksnumbered]{hyperref}
```

L'option **linktocpage** rend actif les liens de la table des matières sur le numéro des pages, et non sur le titre des parties. Autrement, avec dvips, les titres de la table de matières ne sont pas coupés et la mise en page saute.

12.4. Définition du driver pour la conversion

L'extension hyperref utilise une option 'driver' qui permet de spécifier l'usage que l'on va faire du document. Pour une conversion via PostScript c'est le driver dvips qui est utilisé, si on utilise PdfTex, c'est le driver pdftex.

Dans le cadre du projet CITHER, c'est le driver dvips qui est utilisé.

D'autres modules d'extension peuvent nécessiter cette information pour optimiser leur fonctionnement. Il faut donc déclarer le driver comme option globale au niveau du document, par la commande :

```
\documentclass[dvips,…]{…}
```

12.5. La gestion des polices de caractères

La gestion des polices de caractères sous LaTeX est un problème délicat, en effet différents types de police existent, différents standards peuvent être utilisés.

Dans le cadre de notre projet, l'affichage à l'écran d'un fichier PDF doit être le meilleur possible. Il est donc nécessaire d'utiliser des polices de "Type1". Ce ne sont pas les polices utilisées par défaut dans LaTeX, il est nécessaire de le lui indiquer.

La méthode la plus simple consiste à utiliser une police comme "Times" ou "Palatino". Ceci se fait de la manière suivante :

Rajouter la ligne

```
\usepackage{Times}
```

Ou

```
\usepackage{Palatino}
```

Ces polices n'étant pas celle utilisées par le doctorant, la mise en page du document peut être légèrement modifiée.

Une autre solution consisterai à utiliser la représentation Type1 des polices usuelles LaTeX (cf. documents en Annexes).

13. Validation des fichiers sources pour le CEN

Voici la marche à suivre pour le traitement d'un fichier LaTeX avec le CEN.

Ouvrir le fichier à l'aide d'un traitement de texte, ou mieux d'un éditeur LaTeX comme WinEdt, puis :

- Vérifier qu'il puisse être compilé avant toute modification. Pour cela déplacez-vous dans le répertoire du fichier et exécuter `Latex fichier.tex.` Ou alors, avec Winedt, cliquez sur l'icône Latex dans la barre d'outil, après avoir défini ce fichier comme fichier principal et son répertoire comme répertoire courant (Menu Project/Main file et Project/Set Current Directory)
  Si le fichier utilise une base bibliographique (présence d'un fichier `*.bib` parmi les fichiers sources ou le code `\bibliography{}` dans le fichier `*.tex`) exécuter `BibTex fichier.tex`

- Les erreurs qui peuvent se produire à ce niveau, doivent venir du changement d'environnement, c'est à dire qu'il faut bien vérifier comment sont inclus les différent fichiers nécessaire au document (fichier image, fichier de la bibliographie, etc.). Changer tout emplacement absolu par l'emplacement relatif.

- Une fois tous problèmes résolus, rajouter les lignes nécessaires pour l'inclusion du module hyperref et la définition des polices.

- Puis tester une nouvelle fois le fichier par une compilation Latex (une seule fois suffit)

Une fois toutes les modifications nécessaires effectuées, il faut vérifier que LaTeX puisse toujours compiler le fichier. Une mauvaise surprise est toujours possible !

- Cette fois, si des erreurs surviennent, elles sont dues à l'ajout du module hyperref qui entraîne certaines incompatibilités avec les autres modules présents. (par exemple, il existe un problème de compatibilité entre hyperref et l'option french de babel qui oblige à changer le caractère ':' dans les références (commandes `\ref{}` ou `\label{}`) par un autre caractère ('_' ou '-')).

14. Modifications apportées au CEN

Le CEN se compose de 7 unités :

- *Projet*
- *Prefs*
- *Main*
- *ListeC*
- *Ole*
- *HTML*
- *About*

Les principales modifications ont été apportées à l'unité Ole, qui s'occupe de la conversion des fichiers. Les nouvelles procédures ou modification ont été commentées afin de faciliter la compréhension par un tiers.

Au niveau de l'interface graphique, dans la fenêtre de conversion, une nouvelle case à cocher permet la conversion de fichier LaTeX. Dans la fenêtre des préférences, un nouvel onglet à été ajouté pour les paramètres des commandes latex et dvips, et de nouveaux champs dans l'onglet 'Programmes', mais, cela n'a pas été utilisé, du fait de l'utilisation d'un script Perl.

15. Fichier LisezMoi.txt pour l'installation

```
+-------------------------------------------------------------------------+
                  Chaîne d'édition numérique v2.0
+-------------------------------------------------------------------------+
Fichier Lisezmoi.txt
le 17/09/1999


0.Progammes necessaires a l'application
1.Installation.
2.Presentation des repertoires
3.En cas de problemes



+-------------------------------------------------------------------------+
```

0. Progammes necessaires a l'application
Pour pouvoir utiliser la chaîne d'édition numérique, les logiciels suivants sont indispensable :
Adobe Acrobat 3.0
Ms Word 97
Latex pour Win32
Perl pour Win32
Latexmk (Script perl)

Remarque:
Une distribution Latex pour win32 est fournie avec le Cen ainsi que le script Perl Latexmk (cf. la Presentation des repertoires)
+---+

1. Installation
Lancer le programme setup.exe et suivre les instructions a l'ecran
Le repertoire d'installation propose par defaut est :
c:\DocINSA\Cen\
Le programme d'installation installe aussi les fichiers suivants:

<ProgramFilesDir>\Microsoft Office\Office\demarre\PdfMaker.dot
c'est la macro Word necessaire a l'application.
Si votre version de MS Office n'est pas installer a cette endroit, il vous sera necessaire d'installer la macro vous meme, apres l'installation.
(fichiers fournis)

<ProgramFilesDir>\Microsoft Office\Modèles
une autre macro necessaire a l'application, meme remarque que au-dessus
Ces deux macro rajoutent deux nouvelles barres d'outils dans Word.
La 1ere PDFMaker contenant deux icones (le logo acrobat et Prefs)
La 2eme Theses contenant 1 seul icone (un smiley jaune)
si c'est deux nouvelles barres n'apparaissent pas, reportez-vous a la rubrique "en cas de problemes".

Apres cette procedure d'installation, si votre ordinateur ne possede pas de distribution Latex et Perl deja installer, vous devrez lancer les programmes d'installation des distribution fournies avec le CEN
<INSTALLDIR>\fptex 0.3\setup.exe
<INSTALLDIR>\Perl\APi518e.exe

```
Enfin, ne pas oublier de mettre le script Perl Latexmk dans un repertoire
du PATH. (Par exemple le repertoire bin\win32 de la distribution Latex)

Un editeur shareware de fichier Latex est aussi fourni, il n'est pas
indispensable au CEN.
+-------------------------------------------------------------------------+

2. Presentation des repertoires
Une fois l'installation terminee vous devez etre en presence de
l'arborescence suivante :

Pour l'application
<INSTALLDIR>\Cen
<INSTALLDIR>\Cen\Html
<INSTALLDIR>\Cen\Guide
<INSTALLDIR>\Cen\Guide\images
<INSTALLDIR>\Cen\Projets

Pour les repertoires de travail
<INSTALLDIR>\These
<INSTALLDIR>\These\bruts
<INSTALLDIR>\These\pub
<INSTALLDIR>\These\archi
<INSTALLDIR>\These\travail

Pour les programmes Annexes
<INSTALLDIR>\editeur LaTeX    %editeur Winedt shareware utile pour
                              %l'edition de fichier Latex
<INSTALLDIR>\fptex 0.3        %Distribution Latex
<INSTALLDIR>\Docs             %Fichier de documentation
<INSTALLDIR>\Perl             %distribution Perl
<INSTALLDIR>\Perl\laxtexmk.bat%Le script Perl a installer dans un
repertoire
                              %du PATH
<INSTALLDIR>\PdfMaker         %fichier d'installation de la macro PDFMaker
<INSTALLDIR>\Cacro.reg        %clés de la base de registre pour Acrobat
+-------------------------------------------------------------------------+

3. En cas de problemes
3.1. Configuration de MS Word
2 nouvelles barres d'outils doivent maintenant apparaitre
```

- PDFMaker
- Theses

Vous pouvez verifier le bon fonctionnement de ces macro en imprimant un document sur le "distiller Assistant v3.0". Il est bien evident, que Acrobat
doit deja avoir ete installe!
Remarque importante:
Le repertoire par defaut des modele utilisateur doit etre
...\MS Office\Modèles\

3.2 Probleme de lancement d'Acrobat Exchange a partir du CEN
Si le CEN ne peut pas lancer Exchange, cela signifie qu'il manque les cles interface d'Acrobat dans la base de registre.
Pour y remedier lancer regedit.exe et importer le fichier
<INSTALLDIR>\Cacro.reg (ou double cliquez sur lui)