

Mise en ligne en texte intégral des thèses de l'université Paul-Sabatier

Gestion de projet

Soraya Salah

Sous la direction de Christian Ollès
Tuteur pédagogique, ENSSIB

Conseiller
professionnel
Ducasse Jean-Paul
Directeur de la cellule édition électronique
Division ERAD Lyon 2

Correspondant
d'établissement
Chourreau Pierre
directeur du SCD
SCD Paul-Sabatier, Toulouse 3

Résumé :

Ce rapport se propose de décrire la mise en place d'un traitement électronique des thèses à l'université Paul-Sabatier de Toulouse.

Descripteurs : Publications électroniques, Thèses et écrits académiques, XML (langage de balisage), PDF (format), LaTeX (logiciel).

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

Remerciements

Un grand nombre de personnes m'ont aidé dans la préparation de ce projet. Je remercie particulièrement Jean-François Vincent de l'Université de Marne-la-Vallée et Jean-Paul Ducasse de l'université Lyon 2 qui ont accepté de m'accueillir en stage et m'ont guidé dans les méandres de XML et du traitement électronique des thèses. Je remercie aussi Pierre Chourreau pour sa relecture attentive. Enfin, un merci tout particulier à Laurent Demay, graphiste, pour ses explications sur les différents formats d'image, son aide dans l'installation de la plate-forme Cyberdocs et sa patience.

Sommaire

INTRODUCTION.....	7
PARTIE 1. L'ÉTABLISSEMENT, CONTEXTE ET ENJEUX.....	8
1. PRÉSENTATION DE L'UNIVERSITÉ PAUL-SABATIER.....	8
1.1. <i>Historique et statuts.....</i>	8
1.2. <i>Une université scientifique et médicale.....</i>	8
1.3. <i>Les différents sites de l'université :.....</i>	9
1.4. <i>La recherche à l'université Paul-Sabatier.....</i>	9
2. LE SERVICE COMMUN DE LA DOCUMENTATION.....	10
2.1. <i>Les bibliothèques.....</i>	10
2.2. <i>Le budget.....</i>	12
2.3. <i>Le fonds documentaire.....</i>	12
2.4. <i>Les projets.....</i>	13
2.4.1 <i>Extension de la bibliothèque de science.....</i>	13
2.4.2 <i>Rénovation de la bibliothèque de santé à Ranguel.....</i>	13
3. LE PROJET DE MISE EN LIGNE DES THÈSES À PAUL-SABATIER.....	14
3.1. <i>Pourquoi mettre en ligne les thèses ?.....</i>	14
3.2. <i>Contexte international : universités américaines (Virginia Tech) et québécoises (Laval).....</i>	16
3.3. <i>En France : recommandations du ministère suite au rapport Jolly</i>	16
3.4. <i>A Paul-Sabatier : un projet inscrit dans le contrat quadriennal de l'université.....</i>	17
3.5. <i>Le choix du format.....</i>	17
3.5.1 <i>Le format PDF.....</i>	18
3.5.2 <i>Le langage XML.....</i>	19
PARTIE 2. LE CIRCUIT DES THÈSES ÉLECTRONIQUES : DU DÉPÔT À LA DIFFUSION.....	21
1. PAUL-SABATIER : ANALYSE DE L'EXISTANT.....	21
1.1. <i>Une production annuelle importante.....</i>	21

1.2. <i>Un circuit papier</i>	24
2. PRÉSENTATION DES PROJETS EXISTANTS DANS D’AUTRES UNIVERSITÉS FRANÇAISES	25
2.1. <i>Lyon2-Cyberdocs/ABES-Sparte</i>	25
2.1.1 Cyberdocs	25
2.1.2 SPARTE.....	27
2.2. <i>Les autres projets</i>	29
2.2.1 Doc’INSA : CITHER	29
2.2.2 Marne-la-Vallée : PELLEAS	29
2.2.3 CCSD : thèses-en-lignes.....	30
2.2.4 Pastel	30
2.2.5 Bilan	30
3. LES DIFFÉRENTS ASPECTS DU PROJET : ANALYSE COMPARÉE.....	31
3.1. <i>Information et formation</i>	31
3.1.1 Information aux écoles doctorales	31
3.1.2 Information à destination des doctorants	31
3.1.3 Formation des doctorants à l’utilisation de la feuille de style.....	32
3.2. <i>Le dépôt numérique</i>	33
3.2.1 Un dépôt mixte obligatoire	33
3.2.2 Droits d’auteur et aspects juridiques.....	34
3.2.3 Vérification de la conformité des thèses papier et numérique	35
3.3. <i>Traitement préalable des documents</i>	35
3.3.1 Un outil de suivi du circuit thèses.....	35
3.3.2 Post-structuration des thèses : le stylage des documents Word et OpenOffice	36
3.3.3 Cas particulier des documents LaTeX.....	37
3.4. <i>La conversion en XML</i>	39
3.5. <i>La diffusion</i>	42
3.5.1 Indexation et signalement.....	42
3.5.2 Consultation de la thèse.....	42
3.6. <i>Les spécificités et contraintes de l’université Paul-Sabatier</i>	43
PARTIE 3. RÉALISATION DU PROJET	45

1.	LES RESSOURCES HUMAINES, MATÉRIELLES ET FINANCIÈRES	45
1.1.	<i>Moyens humains</i>	45
1.1.1	Evaluation des besoins	45
1.1.2	Fiches de postes	45
1.1.3	Réaffectation des ressources internes.....	48
1.1.4	Recrutement de personnel.....	48
1.2.	<i>Aspect matériel</i>	48
1.2.1	Machines : serveurs et postes de travail.....	48
1.2.2	Logiciels :	50
1.2.3	Locaux : salle de formation	51
2.	ORGANISATION DU TRAVAIL AU SEIN DU SCD.....	51
2.1.	<i>Installation de la plate-forme</i>	51
2.2.	<i>Description du nouveau circuit pour les thèses</i>	52
2.3.	<i>Les différents acteurs</i>	53
2.4.	<i>La formation du personnel</i>	53
3.	ECHÉANCIER.....	54
3.1.	<i>Projet pilote 2004-2005</i>	54
3.2.	<i>Généralisation en 2006</i>	54
4.	EVALUATION ET VALIDATION DU PROJET	57
4.1.	<i>Détermination des critères</i>	57
4.2.	<i>Etapas de validation</i>	57
5.	PERSPECTIVES	58
5.1.	<i>Sparte ou cyberdocs ou... ?</i>	58
5.2.	<i>OAI pour le signalement des thèses</i>	58
5.3.	<i>Les thèses de médecine de l'université Paul-Sabatier</i>	59
5.4.	<i>Traitement rétrospectif</i>	59
	CONCLUSION	60
	BIBLIOGRAPHIE	61
	GLOSSAIRE	67
	TABLE DES ANNEXES	68

Introduction

Vitrine au niveau international de la recherche française, les thèses sont aussi au niveau local, l'expression du rayonnement d'une université. Diffuser ces documents sous forme électronique permet de valoriser la production scientifique des établissements et des chercheurs. Une circulaire relative au dépôt électronique des thèses a été émise par le Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche en 2000. Plusieurs projets sont aujourd'hui en phase de production et certains d'entre eux proposent des outils de traitement et de publication en ligne des thèses.

Dans ce contexte, Pierre Chourreau, directeur du service commun de la documentation de l'université Paul-Sabatier m'a proposé comme sujet de projet professionnel personnel, de mettre en œuvre le traitement électronique des thèses dans cette université.

En s'appuyant sur les projets déjà existants, on s'attachera à définir le circuit complet d'une thèse électronique de son dépôt à sa diffusion et l'on étudiera l'adaptation des systèmes de traitement des thèses au sein de cette université, en tenant compte des spécificités de cet établissement.

Partie 1. L'établissement, contexte et enjeux

1. Présentation de l'université Paul-Sabatier

1.1. Historique et statuts

Fondée en 1229, l'université de Toulouse est l'une des plus anciennes d'Europe avec Bologne (Italie), Oxford (Grande Bretagne), La Sorbonne et Salamanque (Espagne). Les 100 000 étudiants qui la composent en font la 2^{ème} ville universitaire de France.

Instituée par le décret 69.1260 du 18 Décembre 1969, l'université de Toulouse III regroupe les anciennes Faculté de Médecine, Pharmacie et des Sciences, les Observatoires de Toulouse et du Pic du Midi et le récent Institut Universitaire de Technologie. Elle prend le nom de Paul Sabatier, un de ses anciens professeurs, prix Nobel de Chimie en 1912 pour ses travaux sur l'hydrogénation catalytique. Dans le paysage universitaire toulousain, l'université Paul-Sabatier prend en charge l'enseignement et la recherche dans les domaines des sciences, de la médecine, de la technologie et du sport, aux côtés de l'université Toulouse 1 Sciences sociales pour le droit, l'économie, la politique et l'administration et de l'université de Toulouse Le Mirail (« Toulouse 2 ») pour les lettres, langues, sciences humaines et sociales, arts, audiovisuel, communication, documentation. Le ministère de tutelle est le Ministère de l'Éducation nationale, de la recherche et de la technologie. Le professeur Jean-François Sautereau est président de l'université Paul-Sabatier depuis le 3 décembre 2002.

1.2. Une université scientifique et médicale

L'ensemble des disciplines scientifiques, de la santé, des technologies et du sport sont enseignées à l'université Paul-Sabatier. L'établissement, structuré en plusieurs Unités de Formation et de Recherche (UFR), Facultés et Instituts, compte près de

26000 étudiants, 2000 enseignants-chercheurs, 600 chercheurs relevant d'organismes nationaux de recherche et 1300 ingénieurs, techniciens, administratifs, ouvriers, personnels de services (ITAOS) et personnels de bibliothèque.

Les effectifs étudiants se répartissent de la manière suivante :

- 10000 étudiants dans le corps des sciences
- 8000 étudiants dans le corps de Santé (dont 2000 étudiants en STAPS)
- 5500 étudiants dans le pôle technologique

Paul-Sabatier s'inscrit aussi dans le contexte plus large du Pôle universitaire européen de Toulouse Midi-Pyrénées, groupement d'intérêt public créé en 1995 et sous tutelle du Ministère chargé de l'enseignement supérieur et de la recherche. Fédération d'établissements d'enseignement supérieur et de recherche, il est composé des trois universités toulousaines, d'établissements d'enseignement supérieur (INSA, Ecole des mines...) et de la mairie de Toulouse, ainsi que la région. Les objectifs du pôle universitaire sont d'animer la vie du site toulousain, de fédérer un réseau de prospective scientifique régionale et de créer les conditions du meilleur développement sur le plan international.

1.3. Les différents sites de l'université :

L'université Paul-Sabatier est installée sur le campus scientifique de Rangueil, à la périphérie sud de la ville de Toulouse (voir annexe 1-1). Outre son campus principal de Rangueil, Purpan et les allées Jules Guesde à Toulouse, l'UPS est implantée à Albi (DEUG Sciences), Auch et Castres (IUT), Tarbes (IUT de plein exercice et DEUG T1), Rodez (DEUG STAPS) sans oublier l'Observatoire du Pic du Midi, Bagnères et Lannemezan. L'ensemble de ces sites couvre une surface de 384 000 m² dont 147 000 m² dédié à la recherche.

1.4. La recherche à l'université Paul-Sabatier

La région Midi-Pyrénées est au 4^{ème} rang des Régions pour le potentiel de recherche après Ile-de-France, Rhône-Alpes et Provence-Alpes-Côte d'Azur. Plus

de 15 500 acteurs de la recherche et développement sont recensés dans cette région. Ils travaillent dans de très nombreux centres répartis entre les laboratoires publics, les centres de formation et les entreprises industrielles.

La recherche et développement profite des nombreuses passerelles qui ont été construites entre les différents domaines de recherche présents en Midi-Pyrénées, par exemple les transports terrestres, le spatial, les matériaux, l'informatique, le calcul. Toulouse est la 2^{ème} ville de France pour la concentration de chercheurs, et l'université Paul Sabatier y occupe une place majeure avec 98 laboratoires dont la plupart sont reconnus au niveau national (CNRS ou Ministère). L'UPS compte 1500 doctorants et 11000 thèses y ont été soutenues depuis 1972 (signalées dans le système universitaire de documentation, SUDOC, le catalogue des bibliothèques universitaires françaises).

2. Le Service commun de la documentation

2.1. Les bibliothèques

Les statuts du Service commun de la documentation (SCD) ont été adoptés le 12 décembre 1994 par le Conseil d'administration de l'université.

Le SCD comprend deux bibliothèques intégrées, la bibliothèque de santé et la bibliothèque de sciences et 35 bibliothèques associées. La bibliothèque de sciences, construite au milieu des années 60, est située sur le campus scientifique de Rangueil et occupe une surface de 7000 m². La bibliothèque de santé est partagée sur 2 sites : l'ancienne bibliothèque qui est située en centre ville (allées Jules Guesde) et l'actuelle bibliothèque centrale d'une superficie de 4000 m² construite dans les années 70, à Rangueil, tout près de l'hôpital.

L'ensemble du SCD est constitué de 62 personnels titulaires (pour un équivalent temps plein de 59 postes), répartis de la manière suivante : 31 pour la bibliothèque de sciences, 21 pour la bibliothèque de santé dont 3 personnes qui travaillent aux allées Jules Guesde et 10 personnes dépendant directement des affaires générales. A ceci s'ajoute des personnels non titulaires : 4 CES et 1 emploi jeune.

Les horaires d'ouvertures correspondent à la moyenne nationale :

- Bibliothèque de santé : ouverte du lundi au vendredi de 9h à 19h30, fermée le samedi (soit 55h par semaine).
- Bibliothèque de sciences : ouverte du lundi au vendredi de 8h30 à 19h, et le samedi de 9h à 13h (soit 56h30 par semaine).

Le SCD est dirigé par Pierre Chourreu depuis 1999.

L'organigramme (voir annexe 1-1) est le même pour les deux bibliothèques intégrées. Elles sont organisées autour de deux pôles principaux : ressources documentaires d'une part et services aux lecteurs de l'autre. On distingue un service central transversal, les affaires générales, dépendant de la direction du SCD et commun aux deux bibliothèques. Il s'agit, outre d'un service administratif, du service de catalogage centralisé, de la formation des usagers et du service informatique.

Au-delà de l'organigramme du SCD, on notera la présence d'un SICD, Service inter-universitaire de coopération documentaire, commun aux trois universités toulousaines. Le SICD gère le système intégré de gestion de bibliothèque (SIGB) actuel, Horizon, de la société Dynix ainsi que le traitement des fonds anciens.

Le service informatique du SCD

Le service informatique où je suis nommée dépend donc directement de la direction du SCD et a vocation à servir les deux bibliothèques intégrées. Cependant géographiquement, ce service ainsi que l'ensemble des affaires générales est localisé dans les locaux de la bibliothèque de sciences. Le service informatique comprend trois personnes : la responsable de l'équipe, ingénieur d'études, un technicien informatique (tout deux ITARF¹) et moi-même bibliothécaire. Une demande a été faite pour obtenir un poste d'ingénieur réseaux. Le technicien divise son temps de travail entre les deux bibliothèques, la responsable du service assure des plages de service public en section sciences, et je

¹ Personnels Techniques et Administratifs de Recherche et de Formation

ferai du service public en section santé, ce qui devrait avoir un effet bénéfique sur la communication et la répartition des ressources informatique au sein du SCD.

2.2. Le budget

Le budget global du SCD pour l'année 2003 était de 2,07 millions euros.

Financement du SCD de l'université Paul-Sabatier. Budget 2003

Part locale : droits de bibliothèque des inscrits à l'université	495 000 euros
Subventions nationales	1,44 millions d'euros
Produits divers (ex. PEB)	115 000 euros
Subvention locale : BQR	25000 euros dédiés à la documentation électronique

Source : rapport d'activité 2003

Budget d'acquisition par section

	Acquisitions Santé	Acquisitions Sciences	Acquisitions SCD Documentation électronique
Ouvrages	96494	193 527	
Périodiques	321986	447 216	
Bases de données			200 000
Couperin			200 000
Total (en euros)	418480	640 743	400 000

Source : rapport d'activité, 2003.

2.3. Le fonds documentaire

La bibliothèque de Santé propose 20000 ouvrages et 900 titres de périodiques. Une particularité du SCD de Toulouse 3 en section Santé mais aussi en section Sciences, est de posséder certains magasins « ouverts » au public. Dans le magasin

« libre-accès » de la section Santé, les usagers peuvent consulter les périodiques de 1985 à 2001.

Le fonds d'ouvrages de la bibliothèque de sciences atteint 90600 volumes. Le fonds en libre-accès regroupe les ouvrages les plus récents et les plus utilisés. Il représente environ 46 % du fonds total d'ouvrages. 317 abonnements en cours à des périodiques complètent cette offre. Les périodiques y sont consultables en magasin « libre-accès » à partir de 1950. On trouve 33720 titres de thèses sur support papier. La section sciences possède de plus un fonds de culture générale de l'ordre de 6500 à 8000 ouvrages et revues généralistes.

Ressources communes au SCD

Les périodiques électroniques sont négociés dans le cadre du consortium Couperin depuis 1999-2000 et comptent près de 3000 titres. Le SCD propose une vingtaine de bases de données. L'université participe au financement de la documentation électronique dans le cadre du Bonus Qualité Recherche (BQR). Pour l'année 2004, cette participation s'élevait à 50 000 euros. Elle devrait se stabiliser dans les années à venir.

2.4. Les projets

2.4.1 Extension de la bibliothèque de science

Dans le cadre d'un plan Etat-Région quadriennal, une extension de 4000 m² est prévue pour la bibliothèque de sciences. Les travaux commenceront à l'automne 2004 pour se terminer en 2006-2007. Dans les nouveaux locaux, une salle informatique accueillera plus de 70 postes informatiques disponibles en libre-accès aux étudiants.

2.4.2 Rénovation de la bibliothèque de santé à Rangueil

Dans l'attente d'une possible extension de la bibliothèque de Santé, des travaux de rénovation sont prévus. Il s'agit d'insonoriser le hall d'accueil et d'agrandir les espaces en offrant notamment une nouvelle salle de recherche bibliographique et

de consultation des ressources électroniques. Par ailleurs, la bibliothèque a beaucoup souffert de l'explosion de l'usine AZF et les travaux de réparation viennent (seulement) de commencer.

3. Le projet de mise en ligne des thèses à Paul-Sabatier

3.1. Pourquoi mettre en ligne les thèses ?

« Signes du dynamisme scientifique des établissements et de la qualité de leur chercheurs, les thèses constituent un ensemble de travaux dont la mise en valeur revêt une importance particulière pour les universités » soulignait Christine Okret-Manville dans un dossier sur la diffusion des thèses daté de 2002². Dès 1999, la conscience que le système de diffusion des thèses en France n'est pas satisfaisant est présente. « A l'heure actuelle, la diffusion et l'utilisation scientifique des thèses académiques, en tant que corpus d'information scientifique est très faible. La circulation du document imprimé est faible et la diffusion sous forme de microfiche reste confidentielle ». (Jean-Paul Ducasse. *Comprendre les usages d'internet*. Colloque des 3 et 4 décembre 1999 à l'Ecole normale supérieure).

Fin 1999, le Ministère de l'Education nationale, de la recherche et de la technologie, demande à la Direction de l'enseignement supérieur de constituer un groupe de travail sur la numérisation des thèses et leur diffusion par voie électronique. Les conclusions de ce groupe de travail donneront lieu au rapport Jolly sur la diffusion électronique des thèses (2000) dont voici un extrait :

« Tout le monde s'accorde aujourd'hui à reconnaître que le dispositif mis en place en 1985 est désormais obsolète. [...] il est clair que les usagers ne sont pas satisfaits de la situation actuelle : cette conclusion est ressortie sans ambiguïté de l'Enquête sur les pratiques des utilisateurs du signalement des thèses et des utilisateurs de thèses, effectuée en novembre 1997 par la société SCP Communication, à la demande du ministère. Ceux-ci souhaitent manifestement avoir accès rapidement et facilement à l'information secondaire (bibliographique)

² « Diffusion des thèses : la voie électronique ». *Arabesques*, 2002

correspondante, pouvoir apprécier dans les mêmes conditions si une thèse repérée est de nature à leur être utile dans leur travail, pouvoir consulter le texte intégral et, le cas échéant, pouvoir passer commande d'un substitut (photocopies, microformes, éditions à la carte). »

La diffusion des thèses sur Internet doit donc se concevoir comme un moyen de rendre visible à l'échelon national et international les résultats de la recherche française, mais elle s'inscrit aussi pour les universités dans le cadre de l'Initiative de Budapest pour l'Accès Ouvert. On trouvera le texte de l'appel de Budapest (14 février 2002) à l'adresse suivante :

<http://www.soros.org/openaccess/fr/index.shtml> [dernière visite le 14/09/04].

La mise en ligne des thèses s'appuie sur un aspect pratique : les thèses sont rédigées sur format électronique à partir duquel est imprimée la version papier des doctorants. Les avantages sont les suivants :

- Rapidité de traitement : une thèse électronique est rapidement disponible après son dépôt et après la soutenance
- Amélioration de l'accès : une thèse électronique est accessible de n'importe quel point du globe, et à n'importe quel moment.
- Enrichissement du contenu : extraits de films, dialogues, passages musicaux, images haute résolution, bases de données, liens vers d'autres documents etc., peuvent être directement intégrés au document.
- Aide à la publicité et à la communication des activités de recherche doctorale : mise à disposition d'une page web institutionnelle, au sein d'un dispositif éditorial valorisant (crédibilité accrue) et sous une adresse électronique pérenne (citabilité de l'URL garantie)

« En assurant un signalement et une diffusion quasi instantanés, une mise à jour immédiate des informations, en facilitant la consultation gratuite du contenu des documents, la mise en ligne des thèses contribuera efficacement au rayonnement

de l'étudiant et de l'université », (Viviane Boulétreau, SeNTIERS Cellule Edition électronique, 1999)

3.2. Contexte international : universités américaines (Virginia Tech) et québécoises (Laval)

Les premiers projets de thèses en ligne sur Internet sont nés dans certaines universités américaines, et notamment à l'université de Virginia Tech où la réflexion sur le sujet a commencé dès 1987. Depuis 1997, les thèses y sont déposées uniquement sous format électronique. Le projet de Virginia Tech se prolonge dans le Networked Digital Library of Theses and Dissertations (NDLTD) (<http://www.ndltd.org/> [dernière visite le 14/09/04]) . Ce projet de recherche, d'une durée de trois ans, a débuté en septembre 1996 et a été subventionné par le Département de l'éducation des États-Unis. Il avait pour but d'établir un réseau de thèses accessibles sous forme électronique. Il fédère aujourd'hui près de 200 projets de diffusion des thèses en texte intégral par Internet.

L'Université de Laval a lancé en 1998 un projet de mise en ligne des thèses et mémoires de ses étudiants. Elle est la première à avoir développé l'aspect formation aux outils bureautiques et à l'utilisation d'une feuille de style. Des formations ont lieu en salle, et une formation en ligne est accessible à l'étudiant (en maîtrise ou en doctorat) pendant toute la durée de son inscription. Les doctorants peuvent choisir entre un dépôt papier ou un dépôt numérique.

3.3. En France : recommandations du ministère suite au rapport Jolly

« En France, pour donner un ordre de grandeur, 11 000 thèses sont soutenues par an, et très peu sont éditées à des fins de valorisation plus large que l'université de soutenance. » (*Bulletin des bibliothèques de France*, 2001 – Paris, t. 46, n° 2, compte rendu de Ghislaine Chartron). Le rapport Jolly sur la diffusion électronique des thèses a été suivi par la circulaire ministérielle du 28 septembre 2000, envoyée aux présidents d'universités et aux directeurs d'établissements d'enseignement supérieur. Plusieurs initiatives ont été lancées par des universités : citons Lyon 2,

l'INSA de Lyon et l'Université de Marne-la-Vallée. Nous détaillerons ces projets dans une deuxième partie. En janvier 2004, un rapport a été commandé par le Ministère à la société Six et Dix. L'objectif de ce rapport était de dresser un état des lieux des projets existants et d'apporter des suggestions d'améliorations au circuit comme aux outils utilisés. Ce rapport a été rendu en juin 2004. Il est consultable sur le site de la sous-direction des bibliothèques³ mais n'a pas encore été validé par le ministère.

3.4. A Paul-Sabatier : un projet inscrit dans le contrat quadriennal de l'université

De par sa taille et son nombre de chercheurs, Paul-Sabatier occupe une place centrale dans la recherche scientifique française. La diffusion des travaux de recherche et plus particulièrement des thèses est un enjeu majeur.

La mise en ligne des thèses est inscrite dans le dernier contrat quadriennal 2003-2006⁴ « L'université souhaite passer à un niveau supérieur de diffusion de sa propre production. Elle mettra pour cela sur l'outil électronique, dans une démarche cohérente de création d'un système d'information, auquel les ressources du SCD seront intégrées. Dans un premier temps, elle propose de mettre en œuvre la mise en ligne des thèses soutenues dès 2003, selon les modalités recommandées par le Ministère. »

La question a aussi été évoquée lors des conseils scientifiques, et semble bien acceptée voire attendue par la communauté universitaire.

3.5. Le choix du format

Sous quel format traiter, archiver et diffuser les thèses ? Les deux formats utilisés par les établissements sont XML et/ou PDF. Le format d'archivage n'est pas forcément le même que le format de diffusion car les enjeux diffèrent.

³ <http://www.sup.adc.education.fr/bib/Acti/These/rapportV5.doc> [dernière visite le 26/09/04]

⁴ au chapitre 1, paragraphe 2.4 intitulé « Diffusion de l'information scientifique »

3.5.1 Le format PDF

Le format PDF ou Portable Document Format est un format propriétaire, créé par la société Adobe⁵ et dérivé du format PostScript servant à l'impression. Il est largement utilisé dans la diffusion de documents en ligne. Son avantage est d'être simple à obtenir tout en respectant la mise en forme du document source. Cependant, dans la majorité des projets, ce format n'a pas été retenu en tant que format de conversion et surtout d'archivage, car il est propriétaire et difficilement modifiable. L'archivage de documents en PDF n'est donc pas pérenne.

Avantages

- Forme fac-similé de la thèse : respect du choix de présentation de l'auteur, aucune modification de forme, respect de la version canonique papier validée par le jury, pas de difficulté pour faire apparaître les graphiques.
- Très largement diffusé, avec un logiciel de lecture gratuitement téléchargeable, Acrobat Reader, qui fonctionne sur la quasi-totalité des plates-formes informatiques.
- Protection des fichiers possible (contre la modification, le copier/coller, l'impression).
- Mode d'affichage ergonomique et fonctionnalités d'aide à la recherche incluses dans Acrobat Reader (fonction recherche, visualisation possible de la structure des documents pour une consultation plus souple ...).
- Bonne qualité d'impression.

Contraintes :

- Format propriétaire
- Poids du fichier qui peut être conséquent

⁵ <http://www.adobe.fr/> [dernière visite le 14/09/04]

3.5.2 Le langage XML

XML⁶, abréviation de Extensible Markup Language, est un langage de balisage qui permet de décrire la structure logique d'un document. Il a été développé sous l'égide du W3C (World Wide Web Consortium). Dérivé du SGML, (Structured Generalized Markup Language), il conserve le vocabulaire standard de ce dernier ainsi que sa méthode de description de structure, tout en étant plus simple. XML a été conçu pour permettre d'échanger et de stocker des données indépendamment des programmes ou des processus qui les produisent ou qui les utilisent. Il permet de séparer le contenu de la présentation.

Un document XML a deux parties : le document proprement dit, obligatoirement présent, et une description formelle optionnelle, appelée Document Type Definition (DTD). Cette description contient les règles syntaxiques que doit respecter le document. Le langage a une grammaire, qui décrit les structures syntaxiques permises et les mots réservés. XML définit le fond indépendamment de la forme. Différentes DTD existent : EAD, (Encoded Archive Document), pour les documents d'archive, Docbook pour les documents électroniques et la TEI, (Text Encoded Initiative), pour les données textuelles.

XML est donc un standard d'archivage et d'échange de documents qui permet des recherches de données puissantes. Il est, de plus, une solution d'interopérabilité intéressante. Enfin il s'agit d'un format non-propriétaire déjà utilisé par l'industrie, le domaine tertiaire (banques) et la documentation en ligne dans le service public et privé (exemple : les Journaux Officiels).

Avantages :

- Structuration des documents réalisée de manière indépendante de la représentation de cette structure (ex. archivage en XML, diffusion en PDF ou HTML possible) format pivot.
- Format normalisé.
- Codage de l'information en UNICODE : portabilité et relecture aisées.

⁶ <http://www.w3.org/TR/REC-xml> [dernière visite le 14/09/04]

Contrainte :

- langage nécessitant de la rigueur. Pour être converti vers XML, un document doit être déjà structuré de façon à correspondre à la grammaire de la DTD.

Pour le traitement des thèses à l'université Paul-Sabatier, nous choisirons XML comme format d'encodage et d'archivage des documents, pour la structuration qu'il apporte et son utilité en tant que format « pivot » vers différents formats de diffusion. Pour autant, nous n'excluons pas l'utilisation du format PDF dans des cas particuliers traités dans la deuxième partie de ce rapport.

Partie 2. Le circuit des thèses électroniques : du dépôt à la diffusion

Nous dresserons un rapide bilan de l'existant à l'université Paul-Sabatier avant de décrire le circuit des thèses électroniques en nous appuyant sur les exemples extérieurs.

1. Paul-Sabatier : analyse de l'existant

1.1. Une production annuelle importante

D'emblée, l'élément quantité apparaît comme le plus marquant et probablement celui qui posera le plus de contraintes. Toulouse III participe à 13 écoles doctorales auxquelles sont inscrits près de 1450 étudiants. Chaque année, environ 300 thèses de doctorats sont soutenues à l'université dans le domaine des sciences, soit 50% des thèses du site toulousain.

Liste des écoles doctorales⁷ :

Neuf Ecoles doctorales dont l'UPS est établissement principal :

- Mathématiques et applications (cohabilitation avec INSA, Toulouse 1 et ENSAE)
- Physique et nano physique (cohabilitation avec INSA)
- Sciences de l'univers, de l'environnement et de l'espace (cohabilitation avec ENM, ENSAE ET INP)
- Chimie (cohabilitation avec INP)
- Biologie, santé, biotechnologie (cohabilitation avec l'INP)
- Matériaux, structure, mécanique (cohabilitation avec INSA, INP et ENSAE)

⁷ Disponible sur <<http://www.ups-tlse.fr/ACCUEIL/Rubriques/RECHERCHE/ecolesdoct.html>> [dernière visite le 14/09/04]

- Informatique et télécommunications (cohabilitation avec INSA, INP, ENSAE et Toulouse)
- Génie électrique, électronique, télécommunications (cohabilitation avec INSA, INP et ENSAE)
- sciences écologiques, vétérinaires, agronomiques et bioingénieries (cohabilitation avec INSA, ENVT, INP E. Mines Albi)

Quatre Ecoles Doctorales pilotées par d'autres établissements :

- Comportement, langages, éducation, socialisation, cognition (cohabilitation avec Toulouse 2)
- Sciences des procédés (cohabilitation avec INP et INSA)
- Energétique, dynamique des fluides (cohabilitation avec INSA, INP et ENSAE)
- Systèmes (cohabilitation avec INSA, INP, ENSAE, Toulouse 1 et Toulouse 2)

A cela s'ajoute les thèses d'exercice soutenues en pharmacie et en médecine : plus de 400 pour l'année 2003. Celles-ci sont des documents courts, d'un contenu de recherche moindre.

Ci-dessous on trouvera la répartition des thèses de l'université Paul-Sabatier extraite du catalogue local (Horizon). Il ne nous a pas été possible d'obtenir la répartition précise pour les thèses scientifiques (information non précisée dans les notices bibliographiques).

Thèses cataloguées à Paul-Sabatier en 2003 (nombre de titres)

	Sciences	Santé				Total santé	Total
		Médecine	Dentaire	Pharmacie	Vétérinaire		
Janvier	25	13	0	10	0	23	48
Février	0	23	17	11	0	51	51
Mars	74	0	0	0	0	0	74
Avril	0	8	0	0	56	64	64
Mai	0	49	15	20	0	84	84
Juin	54	0	0	0	0	0	54
Juillet	4	4	0	9	55	68	72
Août	0	0	0	9	0	9	9
Septembre	12	30	1	22	0	53	65
Octobre	40	0	0	0	0	0	40
Novembre	37	0	23	12	0	35	72
Décembre	0	0	0	0	40	40	40
Total	246	127	56	93	151	427	673

1.2. Un circuit papier

Aujourd'hui, le circuit des thèses de Paul-Sabatier est uniquement papier. Trois exemplaires sont déposés auprès de l'école doctorale.

Selon l'arrêté du 25 sept. 1985 (<http://www.sup.adc.education.fr/bib/Acti/These/textregl.htm#25sept> [dernière visite le 14/09/04]), le dépôt de la thèse s'effectue en deux étapes :

- dépôt d'un formulaire donnant le résumé et les mots-clés déposés par le candidat avant la soutenance auprès du service de doctorat de son établissement.
- 3 exemplaires papiers sont enregistrés après soutenance par la bibliothèque de l'université : le 1^{er} pour l'archivage, le second pour la consultation sur place et le 3^{ème} sur autorisation du jury est transmis aux Ateliers nationaux de reproduction des thèses afin d'y être microfiché. L'ANRT se charge de diffuser les microfiches auprès des universités du domaine scientifique correspondant à la thèse et commercialise à la demande des exemplaires de la thèse.

Au niveau du SCD :

En section sciences, 3 exemplaires sont fournis à la bibliothèque par la scolarité de l'université qui centralise les thèses déposées auprès des écoles doctorales.

- Le 1^{er} exemplaire est envoyé à l'ANRT.
- Les deux autres exemplaires sont équipés (étiquettes et code barre pour l'exemplaire de consultation). La cotation correspond à l'année de soutenance, au code de l'université de Toulouse et à un numéro séquentiel d'enregistrement de la thèse. Cela donne : 2004-TOU3-0001 pour la 1^{ère} thèse cotée de l'année 2004.
- L'exemplaire destiné au prêt est utilisé pour la création d'une notice minimale dans le module d'acquisition d'Horizon, puis il part directement en magasin.

- L'exemplaire destiné à l'archivage (consultation sur place) sera utilisé pour créer une notice bibliographique (création dans le SUDOC et récupération dans le catalogue local Horizon). Le délai est d'environ 1 mois.
- Les thèses sont cataloguées par une même personne pour l'ensemble du SCD (Marie-Joséphine Lamure)

On notera que bien que le dépôt des thèses soit aujourd'hui uniquement papier, il est demandé aux doctorants de fournir avec les exemplaires papier, une disquette contenant les résumés afin de les incorporer à la notice de catalogage. Il est aussi à souligner que si les thèses de santé sont correctement déposées, les thèses de sciences ne sont pas systématiquement envoyées à la scolarité par les écoles doctorales. On estime qu'une cinquantaine de thèses en science ne seraient pas déposées au SCD chaque année.

2. Présentation des projets existants dans d'autres universités françaises

2.1. Lyon2-Cyberdocs/ABES-Sparte

2.1.1 Cyberdocs

Le projet Cyberthèses (1999) résulte d'une collaboration entre les universités de Lyon 2 et Montréal, soutenu par le fonds francophone des inforoutes pour la publication électronique des thèses. Ce projet a donné lieu à la conception et réalisation d'une chaîne de production de documents électronique s'appuyant sur la norme SGML et la DTD TEILite (forme allégée de la TEI). Il avait aussi pour objectif de mettre en place un serveur commun aux différents établissements partenaires permettant l'indexation de leurs thèses. Les thèses sont consultables sur le site de Cyberthèses : <http://www.cybertheses.org/cybertheses/cybertheses.html>.

Plus récemment, en 2002, une évolution de cette plate-forme baptisée Cyberdocs a vu le jour. Elle est construite sur des outils XML, entièrement en Open Source sous licence GPL. Les sources sont accessibles et peuvent être téléchargées. Cela

signifie qu'un établissement peut récupérer ces sources, installer la plate-forme sur ses propres ordinateurs et les adapter à ses besoins. Cyberdocs a été conçue pour traiter différents types de documents. La réussite de ce projet est basée sur la mutualisation des développements informatiques réalisés par les uns et les autres. Plusieurs établissements participent au projet Cyberthèses/Cyberdocs. Des formations ont eu lieu en 2001 et 2002.

On trouvera la documentation concernant le fonctionnement de Cyberdocs, le téléchargement de ses sources et les informations concernant l'installation de la plate-forme sur le site de Cyberthèses⁸.

Plusieurs listes de diffusion relatives au projet sont hébergées sur le CRU, Comité Réseau des Universités. La liste Cyberthèses-users⁹ a vocation à mettre en relation les différents acteurs et résoudre les problèmes d'installation et de fonctionnement.

Fonctionnement de Cyberdocs :

La plate-forme Cyberdocs se divise en trois modules : conversion, publication et gestion.

Le module de conversion consiste à produire, à partir d'un document issu d'un traitement de texte, un document structuré en format XML s'appuyant sur la DTD TEILite, et qui devient ainsi le document de référence qui servira à l'archivage et à la publication. Pour obtenir ce résultat, le document source doit être structuré au préalable à l'aide d'une feuille de style. Les éléments de structure présents dans un document traitement de texte peuvent être :

- De l'enrichissement typographique : caractères gras ou italiques, couleurs, polices particulières, etc.
- Des structures déjà gérées par les traitements de texte : tableaux, illustrations, liens hypertextes, notes de bas de page, etc.
- Des styles (de paragraphes ou de caractères) ayant une sémantique commune à tous les documents : Titre 1, Titre 2, etc.
- Des styles ayant une sémantique particulière pour la plate-forme Cyberdocs : auteur, titre du document, année de publication, etc.

⁸ <http://sourcesup.cru.fr/cybertheses/> [dernière visite le 14/09/04]

⁹ <http://listes.cru.fr/www/info/cybertheses-users> [dernière visite le 14/09/04]

Un fichier de métadonnées¹⁰ XML, essentiellement du Dublin Core (standard composé de quinze éléments qualifiés) doit être joint au fichier source.

Les technologies utilisées s'appuient sur l'environnement JAVA, XSLT et XSL-FO et OpenOffice.

La conversion est automatisée et s'effectue en plusieurs étapes : conversion d'un document en format Doc ou OpenOffice vers du XML à plat (en utilisant OpenOffice), identification de ce qui va être traité ou exclu, transformation de chaque élément en élément TEILite, en fonction de son nom ou du style qui lui est associé, hiérarchisation des éléments en fonction des niveaux de titres, production du document XML final.

Le module de gestion permet, à travers une interface web, de piloter la conversion et la publication des documents.

On obtient un document sous plusieurs formats : HTML, PDF, XML et XHTML.

Le module de publication permet de diffuser, dans une architecture Web dynamique, des documents, avec des fonctions de recherche et de consultation. Elle est bâtie sur la plate-forme SDX¹¹.

Remarque : Cyberdocs ne permet pas à l'heure actuelle de convertir en XML les documents LaTeX. Ces documents doivent être convertis au préalable en format PDF pour entrer dans la chaîne de traitement (gestion et publication).

2.1.2 SPARTE

SPARTE, Système de Production et d'Archivage de Thèses Electroniques, est la chaîne de traitement des thèses électroniques gérée par l'ABES, Agence

¹⁰ Les métadonnées sont « des données sur des données », souvent des éléments de catalogage (titre, auteur, résumé...) auxquels s'ajoutent des données de gestion.

¹¹ Plate-forme documentaire libre pour documents XML. Pour plus d'informations, voir <http://adx.org/sdx/> [dernière visite le 14/09/04]

bibliographique de l'enseignement supérieur et le CINES, Centre Informatique National de l'Enseignement Supérieur, sous l'égide de la Sous-direction des bibliothèques et de la documentation. Elle est mise à la disposition des établissements de soutenance pour la diffusion et l'archivage des thèses déposées sous forme électronique. Elle s'appuie également sur la TEILite. Elle est basée sur la première version de Cyberthèses qui a ensuite été modifiée par la société STILO, éditrice d'OmniMark.

Les premières formations à la chaîne SPARTE implantée à Montpellier ont eu lieu à Montpellier en septembre 2003. 19 établissements sont identifiés dans SPARTE, les plus actifs étant Lille 3, Strasbourg 1, UTC. Un certain nombre d'établissements semblent avoir arrêté les tests. 15 thèses environ ont été traitées jusqu'à la version de diffusion¹² L'ABES fournit une assistance technique via un guichet web commun avec le SUDOC¹³ ou par téléphone. Il existe aussi une liste de diffusion pour les utilisateurs de SPARTE¹⁴ qui est hébergée par le CINES. Cette liste n'est apparemment pas utilisée par les établissements concernés (aucun message d'utilisateur y compris dans les archives au 16/09/04)

Fonctionnement de SPARTE

Accès client-serveur : on envoie la thèse à traiter au format RTF¹⁵ au serveur de l'ABES et on obtient un document au format HTML.

SPARTE doit permettre :

la conversion des thèses du format RTF au format XML,

la gestion des métadonnées administratives et descriptives (fichier XML)

le stockage direct des fichiers XML au CINES,

la production d'un fichier HTML de diffusion pour les établissements

la production d'un fichier HTML de la thèse et des métadonnées afin de faciliter le signalement de la thèse dans le SUDOC.

¹² source : rapport de synthèse de la société Six-et-Dix. Annexe ABES

¹³ <http://susvp.abes.fr> [dernière visite le 14/09/04]

¹⁴ <http://listes.cines.fr/www/info/sparte> [dernière visite le 14/09/04]

¹⁵ format de document texte universel créé par Microsoft pour échanger des fichiers entre différents traitements de texte

La conversion s'effectue aussi en plusieurs étapes : conversion du RTF (provenant du Word stylé) en XML plat (étape 1), puis en XML TEI Lite (étape 2), puis en version de diffusion (en HTML, pas de PDF). Cette version en HTML pouvant être ensuite utilisée dans le client de catalogage WinIBW pour alimenter automatiquement le catalogue SUDOC.

Remarque : SPARTE ne gère pas à l'heure actuelle le format PDF et les documents LaTeX. Il ne semble pas être encore en réelle production.

2.2. Les autres projets

2.2.1 Doc'INSA : CITHER

Doc'INSA, centre de documentation de l'Institut national des sciences appliquées de Lyon, a été l'un des précurseurs en France de la mise en ligne des thèses. Depuis 1999, dans le cadre du projet CITHER, les thèses sont traitées et publiées sur le site web de Doc'INSA. A partir de documents structurés, les thèses sont converties et diffusées sous la forme d'un PDF enrichi (table des matières active, renvois aux notes de bas de page et bibliographie, liens hypertextes).

Remarque : les documents ne sont pas convertis en XML. Le module de conversion réalisé en interne s'appuie sur des logiciels propriétaires (payants).

2.2.2 Marne-la-Vallée : PELLEAS

Pelléas : projet de gestion de l'édition électronique en XML, lancé en 1997 par l'Université de Marne-la-Vallée, sous l'impulsion du SCD, alors dirigé par Christian Lupovici. La réalisation du projet a été confiée à la société Jouve en collaboration avec Cadic. Le projet est en phase de production depuis 2002 mais il ne permet pas la conversion en XML. Pelléas est aujourd'hui utilisé par le SCD de Marne-la-Vallée pour gérer et diffuser des documents (thèses et mémoires) en ligne.

Remarque : la partie conversion en XML s'appuie sur la chaîne Cyberthèses de Lyon 2. Le module de publication est propriétaire (payant).

2.2.3 CCSD : thèses-en-lignes

Le CCSD, Centre pour la Communication Scientifique Directe (CNRS), gère le projet « thèses-en-ligne », TEL. Les doctorants peuvent effectuer un dépôt volontaire en format PDF, ou éventuellement ps (postscript).

<http://tel.ccsd.cnrs.fr>

Le projet TEL diffuse actuellement 2700 thèses (fin avril 2004)

Remarque : les documents ne sont pas convertis en XML. Le dépôt est uniquement volontaire. Il n'y a pas de vérification de conformité entre l'exemplaire papier et électronique.

2.2.4 Pastel

PASTEL (ParisTech service de thèse en lignes) diffuse au format PDF les thèses de ParisTech, grandes écoles d'ingénieurs de Paris (environ 300 thèses en ligne au 15/09/04). Le dépôt est effectué par le lauréat lui-même après soutenance. Ces thèses peuvent être retrouvées dans thèses-en-ligne.

<http://pastel.paristech.org/>

Remarque : les thèses sont déposées et diffusées au format PDF.

2.2.5 Bilan

Les deux projets les plus avancés et utilisables par une autre université sont Cyberdocs et SPARTE. Un certain nombre d'universités utilisent l'une ou l'autre de ces chaînes. Dans l'ensemble, le traitement des thèses ne varie pas énormément d'une université à une autre, mais sur des points pratiques, il sera intéressant de comparer ce qui a été fait, afin de décider ce qui est le plus adapté à notre cas.

Si dans l'approche, Lyon 2 et l'ABES sont partis sur des pistes différentes (décentralisation et mutualisation des ressources d'un côté contre centralisation des données de l'autre), ces deux chaînes de traitement restent assez proches. Que l'on choisisse l'une ou l'autre aura peu d'influence sur le traitement des documents avant conversion. Les divergences apparaissent au moment de la conversion et concernent aussi l'archivage et la diffusion.

3. Les différents aspects du projet : analyse comparée

3.1. Information et formation

3.1.1 Information aux écoles doctorales

La mise en place d'un dépôt électronique aura des répercussions sur les services de l'université, notamment la scolarité, les écoles doctorales, les enseignants et équipes pédagogiques. Les projets ont en général inclus en amont une information en direction de la communauté universitaire, voire une présentation au Conseil scientifique de l'université. La mise en oeuvre de ce projet implique par ailleurs la validation du Conseil d'administration et une modification de la charte des thèses de l'université. Une fois le projet adopté par l'université, des réunions d'informations avec l'ensemble des directeurs des écoles doctorales sont nécessaires afin de mettre en place le nouveau circuit des thèses.

3.1.2 Information à destination des doctorants

La réussite d'un tel projet dépend en grande partie de l'accueil favorable des utilisateurs.

Une double communication est généralement utilisée : les informations relatives au dépôt sont communiquées aux doctorants lors de l'inscription universitaire. Cette information est aussi présente sur le site web des universités, souvent sous la forme d'une rubrique visible intitulée « ressources pour les doctorants ». On y trouve les modalités d'inscription et de soutenance, des informations sur le dépôt électronique (où ?, quand ? quels documents présenter), les feuilles de style à télécharger et des guides pour les utiliser. Lyon 2 propose par ailleurs, un didacticiel (« Produire et diffuser des documents structurés ») en ligne. SPARTE n'est pas encore en production, cependant on trouve aussi ce type d'informations sur le site de l'ABES.

En résumé, on retiendra les points suivants :

- Fournir une information aux doctorants dès leur inscription en conjuguant forme papier et information sur le site du SCD et/ou de l'université.

- Informer de la procédure administrative de dépôt et des aspects techniques.
- Pointer vers les ressources existantes du type didacticiels.
- Proposer des modèles existants.

3.1.3 Formation des doctorants à l'utilisation de la feuille de style

Les établissements qui ont mis en place un archivage numérique et un traitement des thèses en XML ont pour la plupart associé au dépôt numérique une formation à la feuille de style. Le rapport de la société Six-et-Dix de juin 2004 propose d'élaborer une formation ou un modèle de formation au niveau national. D'ici là, nous examinerons les formations déjà existantes et leur incidence sur le temps de traitement des documents. La durée moyenne de ces formations varie d'une demi-journée à une journée complète.

La cellule ERAD de Lyon 2 propose une formation aux doctorants qui présente le cadre théorique et les aspects pratiques.

- Pourquoi mettre des thèses en ligne
- Rassurer sur les risques de plagiat
- Mise en évidence des gains qu'apporte l'utilisation d'une feuille de style (génération automatique du sommaire et de l'index, possibilité de modifier les styles en une seule fois...).

L'étudiant est invité à venir avec une partie de sa thèse et à la styler. Le retour de ces formations est positif, et aujourd'hui, 60% des thèses déposées à l'ERAD sont stylées correctement, ce qui va permettre de les convertir rapidement.

A Doc'INSA, bien qu'actuellement, les thèses soient converties en PDF, il s'agit d'un PDF structuré et les étudiants utilisent une feuille de style (facultative) à laquelle ils sont formés. D'après mes entretiens avec Dalila Boudia, une grande partie des thèses déposées aujourd'hui sont stylées correctement.

Enfin, l'exemple de Marne-la-Vallée, où il n'y avait pas encore de formation à cette feuille de style pour les doctorants (bien qu'elle existe pour les étudiants en maîtrise) montre que peu de thèses déposées ont été stylées correctement.

La conclusion qui s'impose d'elle-même est que ce type de formation est obligatoire pour ce projet, et ceci dès sa mise en place. Le gain de temps sera déterminant (de 1 à 4h de traitement pour une thèse stylée correctement contre une journée ou bien plus si elle ne l'est pas) pour la réussite du projet.

3.2. Le dépôt numérique

3.2.1 Un dépôt mixte obligatoire

L'ensemble des universités qui ont mis en œuvre le dépôt électronique des thèses ont choisi de pratiquer un dépôt mixte : papier et informatique. Le produit papier étant aujourd'hui le seul reconnu au niveau administratif. Cette tendance est confirmée par le rapport de Six-et-Dix de juin 2004, qui considère que « seule la version papier [est] porteuse du caractère officiel », car il s'agit de l'exemplaire validé par le jury. « Les versions numériques n'étant considérées que comme des versions de substitution de la thèse (pour la diffusion, l'archivage) ». Ce dépôt mixte doit être rendu obligatoire, ce qui demande une intervention du conseil scientifique et/ou du conseil des études et de la vie universitaire et une validation par le conseil d'administration.

Un point reste à déterminer : le dépôt doit-il avoir lieu avant ou après soutenance ?

- Dans le cas d'un dépôt avant soutenance, on est certain d'obtenir le dépôt, la soutenance y étant liée. Cependant, cela suppose de gérer les différentes corrections éventuellement apportées à la thèse par le doctorant (cas fréquent).
- Dans le cas d'un dépôt après soutenance, une seule version de la thèse sera à gérer. Cependant, le doctorant étant alors devenu docteur, le SCD ne dispose d'aucun moyen pour réclamer une thèse non déposée.

Etant donné la difficulté actuelle pour le SCD de l'université Paul-Sabatier d'obtenir la totalité des thèses en sciences, un dépôt avant soutenance me paraît plus approprié.

3.2.2 Droits d'auteur et aspects juridiques

Comme le rappelle le rapport Jolly :

- «en tant qu'œuvre de l'esprit, la thèse relève de la législation sur la propriété intellectuelle (loi de 1957 et textes dérivés) et donc des droits reconnus aux auteurs sur leurs productions ;
- les juridictions administratives assimilent les thèses à un document administratif, auquel s'appliquent les règles d'accès aux dits documents : sauf décision explicite du président ou directeur de l'établissement de soutenance, toute thèse est consultable dans la bibliothèque de l'université de soutenance où elle doit être déposée ;
- selon l'arrêté du 25 septembre 1985, il revient au président ou au directeur de l'établissement de soutenance d'autoriser ou non, sur avis du président du jury, la reproduction d'une thèse.»

Type de diffusion

Les universités proposent aux doctorants de leur donner une autorisation de diffusion sans cession des droits d'auteurs ni exclusivité.

Plusieurs cas sont possibles :

- autorisation de diffusion complète où la thèse sera accessible à partir de tout poste connecté à Internet
- diffusion restreinte en intranet, au sein de l'université
- délais de diffusion que peut imposer un jury ou une entreprise.

Il ne s'agit pas de forcer le doctorant, mais de lui présenter les possibilités existantes, tout en insistant sur la finalité d'un travail de recherche universitaire qui est destiné à une communauté (au sens large, tous les citoyens) et qui a donc pour but d'être diffusé. La principale inquiétude des doctorants à cet égard concerne le risque de plagiat. Ce risque est paradoxalement plus faible si le document est diffusé largement que dans le cas inverse.

Le document concernant l'autorisation de diffusion peut être soumis au service juridique de l'université.

Le retrait

Dans tous les cas, il faut prévoir la possibilité de retirer le document de la diffusion en cas de demande de l'auteur. Une procédure de retrait simple et rapide doit donc être prévue dès le départ, même si ce cas est très rare dans les faits.

3.2.3 Vérification de la conformité des thèses papier et numérique

Outre une déclaration de conformité que doit signer le doctorant, les établissements pratiquent au moment du dépôt mixte une vérification rapide entre le document papier et le document électronique. Il s'agit de s'assurer que tous les éléments sont bien présents : résumés, quatrième de couverture, graphiques, annexes. Dans le cas des documents LaTeX, cette vérification demande de connaître la structure de ce type de documents (souvent un document maître, et un découpage par chapitres, une bibliographie à part, et des fichiers utilisés pour les styles non génériques).

3.3. Traitement préalable des documents

3.3.1 Un outil de suivi du circuit thèses

Entre les différents types de formats, l'étape du traitement (thèse en dépôt, en cours de traitement, en attente de validation, en archivage uniquement, en diffusion intranet et/ou internet), les différents acteurs qui vont intervenir dans ce traitement, mais aussi afin de produire des statistiques, il devient nécessaire d'utiliser un outil de gestion qui permette de suivre la thèse de son dépôt à sa diffusion. C'est ce qu'on appelle un « workflow ». Le SCD de Marne-la-Vallée utilise une base Access créée pour l'occasion. L'université de Lyon 2 a développé un outil de gestion, OGET basé sur le langage PHP. Cet outil sera mis à disposition des établissements qui souhaitent l'utiliser et l'adapter à leur propre besoin, à l'automne 2004 sur le site du CRU. L'ABES intègre un workflow à SPARTE.

Ce workflow doit permettre de gérer les différents états de la thèse, les métadonnées associées et proposer des droits différents selon le niveau de l'utilisateur (administrateur, styleur...).

3.3.2 Post-structuration des thèses : le stylage des documents Word et OpenOffice

Quelle que soit la chaîne de traitement retenue, une conversion en XML impose d'avoir un document structuré a priori, c'est-à-dire avant de débiter la conversion. Les formations jouent un rôle majeur dans la bonne structuration des documents obtenus. Cependant, il est illusoire d'espérer obtenir à court terme 100% des thèses correctement stylées, et même dans ce cas, certains tableaux, graphiques ou équations devront être retravaillés car non pris en charge par les chaînes. Il s'agit par exemple des images copiées dans Word et qui sont ensuite redimensionnées. Converties telles quelles, elles reprennent leur format et taille d'origine. Ce travail de « restylage » est d'une durée très variable. Lyon 2 l'évalue de 1 à 2h dans le cas d'un document parfaitement stylé, à 2 ou 3 jours si ce n'est pas le cas.

Description du traitement du document :

Dans le cadre de la formation initiale des bibliothécaires d'Etat, j'ai effectué deux stages : l'un au SCD de Marne-la-Vallée et l'autre à la cellule d'édition électronique de l'ERAD de l'université Lyon 2. J'ai eu l'occasion de pratiquer le stylage des thèses. Je liste ci-après les points qui m'ont paru les plus importants.

Evaluation de la durée de traitement.

Il n'existe pas à l'heure actuelle d'outil permettant de déterminer à l'avance le temps de traitement d'une thèse, temps de traitement qui peut être très variable. Le rapport de la société Six et dix préconise la création d'un tel outil.

- Vérification rapide du stylage dans Word (niveaux de titres, texte). On utilise pour cela l'affichage en mode normal qui permet d'afficher à gauche tous les styles utilisés.
- Evaluation du nombre de graphiques, équations, etc.
- Choix du format de conversion :
 - o XML si le document est correctement structuré, ou si le travail de re-stylage n'est pas trop important.
 - o PDF si on prévoit une durée de stylage trop longue ou s'il s'agit d'un document LaTeX.

Stylage des documents

- On applique les différents styles
- Les images qui ne sont pas traitées par les chaînes sont exportées (copier/coller ou copie d'écran), puis sauvegardées en format JPEG ou GIF.

Création d'un fichier de métadonnées :

Ces informations peuvent être entrées manuellement dans un workflow (SPARTE, module de gestion de Cyberdocs) ou récupérées à partir de sources extérieures (Lyon 2 récupère les données provenant du logiciel de gestion de la scolarité Apogée). On saisit les informations bibliographiques et administratives.

3.3.3 Cas particulier des documents LaTeX

A l'origine, mon projet professionnel devait être centré sur ce type de documents. Les difficultés posées par une conversion vers XML m'ont conduite à élargir le sujet à tous les formats de documents et à envisager un traitement plus simple pour les thèses LaTeX. Université scientifique, Paul-Sabatier est confrontée aux problèmes de conversions des documents créés en LaTeX.

Qu'est-ce que LaTeX ?

TeX est un langage créé en 1977 par Donald Knuth. Le nom provient du grec et signifie « art, métier ». Il s'agit d'un outil de publication assistée par ordinateur (PAO) plutôt que d'un traitement de texte. Il est utilisé par la communauté scientifique et notamment par les mathématiciens pour mettre en forme et publier des documents. Son grand avantage est sa qualité de rendu d'un point de vue typographique. par exemple resserrer ou élargir les interlignes voire les espaces entre les caractères afin d'éviter qu'un mot ne se retrouve orphelin sur une page suivante, ainsi que le traitement et l'affichage des équations mathématiques. Son grand désavantage, dans notre cas précis, est la difficulté à convertir ce type de documents en XML.

LaTeX et XML

Aucune des deux chaînes de traitement ne sait convertir du LaTeX en XML. Il est a priori possible de le faire, mais cela demanderait des contraintes fortes : un des avantages de LaTeX est de pouvoir créer ses propres macros (ensemble d'actions répétables), et appliquer une feuille de style imposerait d'y renoncer. D'autre part, LaTeX est davantage un logiciel typographique. Il existe une liste de diffusion à ce sujet, Cyberthèses-Latex, et une étude est en cours pour convertir de LaTeX vers RTF, puis de RTF vers XML. Par ailleurs, l'Université de Santiago au Chili est en train de travailler sur une conversion LaTeX/XML.

Etant donné le caractère scientifique de l'université Paul-Sabatier, il semble intéressant de participer à ces travaux. En ce qui concerne la conversion, nous choisirons à l'instar des autres universités, de diffuser les documents LaTeX en format PDF dans l'attente d'avancée significative dans ce domaine. Lors de mon stage de mars 2004 au SCD de l'université de Marne-la-Vallée, j'ai eu l'occasion de travailler sur une conversion LaTeX/PDF « satisfaisante » dans la mesure où l'on obtenait un document PDF qui conservait tous les liens (table des figures, table des matières, index, bibliographie, notes de bas de page) et permettait une navigation aisée. Cette conversion est assez simple à effectuer mais est rarement effectuée de manière complète par les doctorants (une thèse sur 10 dans le cas de Marne-la-Vallée). Cela se comprend aisément étant donné que l'objectif du doctorant est d'obtenir un fichier postscript pour une impression de bonne qualité, et non pas un fichier PDF destiné à une consultation en ligne.

LaTeX vers PDF

Les deux méthodes de conversion vers le PDF sont les suivantes :

- TeX \Rightarrow DVI \Rightarrow PS \Rightarrow PDF

Un fichier TeX n'est pas exploitable immédiatement. Il faut d'abord le transformer en DVI, format indépendant du périphérique de sortie (opération réalisée par le compilateur TeX) puis en PostScript, format d'impression et finalement en PDF (opérations réalisées par des logiciels extérieurs).

- Utilisation du module PDF-LaTeX

Le rendu est de meilleure qualité (notamment en ce qui concerne l’affichage de la table des matières), mais le temps de traitement est un peu plus long et cela peut parfois entrer en conflit avec des macros de LaTeX.

On trouvera en annexe 2-2, le manuel de conversion LaTeX/PDF que j’ai rédigé lors d’un stage professionnel pour le SCD de Marne-la-Vallée. On pourra l’utiliser tel quel dans un premier temps, mais il sera à faire évoluer. On réfléchira à l’adaptation d’une feuille de style LaTeX obligatoire, du type de celle développée par le doctorant Cédric Sibade, à la demande de Jean-François Vincent pour le SCD de Marne-la-Vallée.

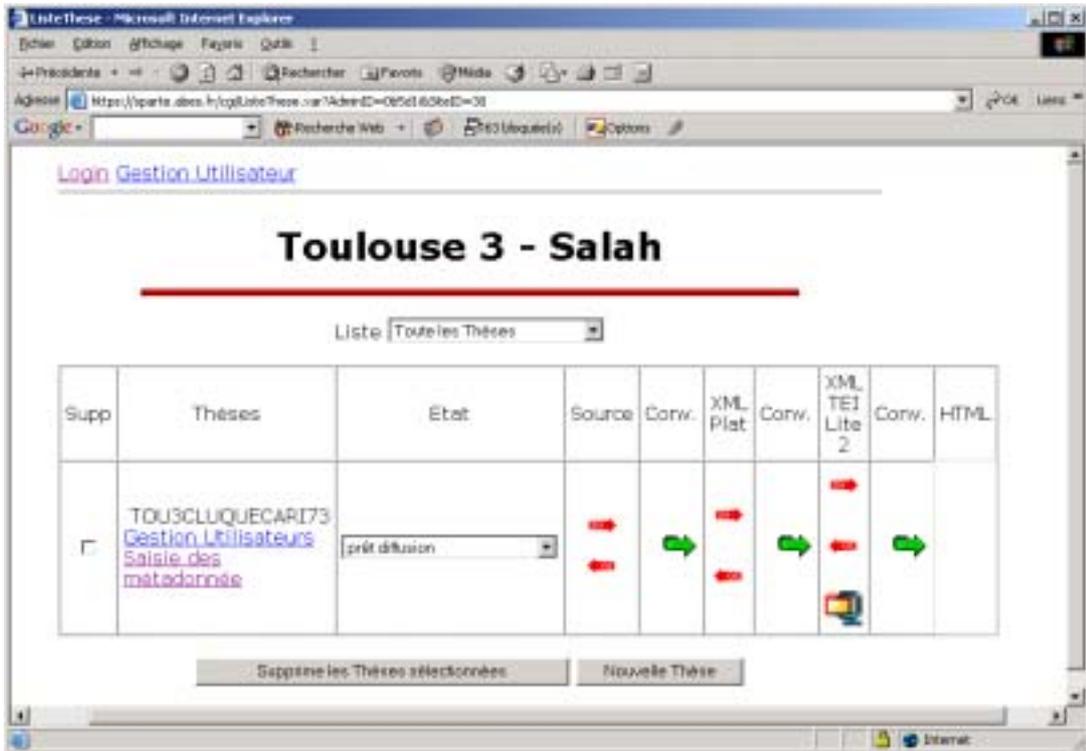
3.4. La conversion en XML

Il existe déjà une littérature abondante sur XML d’une part et sur les chaînes de conversion des thèses et leur fonctionnement en particulier. Je ne ferai donc qu’un bref rappel des opérations effectuées côté utilisateur et renvoie le lecteur aux rapports présents dans ma bibliographie.

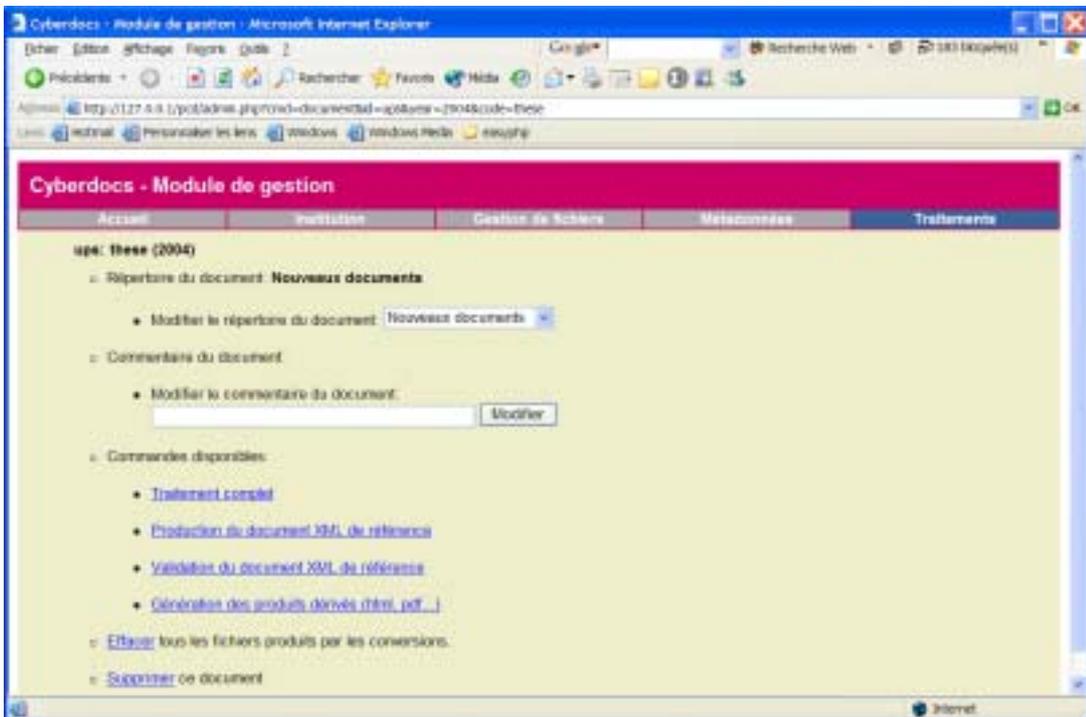
Dans les deux chaînes de traitement, une interface permet de gérer la conversion de la thèse et des métadonnées associées.

- Création d’un espace de travail pour le document
- Importation du fichier de la thèse
 - o Format Word ou OpenOffice ou PDF pour Cyberdocs
 - o Format RTF pour Sparte
- Importation éventuelle de fichiers d’images
- Saisie ou récupération des métadonnées
- Lancement de la conversion

Module de gestion de la conversion sur SPARTE



Module de gestion de la conversion sur Cyberdocs



J'ai testé les deux chaînes sur la thèse d'une doctorante de Paul-Sabatier. Le document avait été réalisé sous Word. Il n'était pas stylé selon les feuilles de style de SPARTE ou Lyon 2 (et pour cause puisque le dépôt est pour l'instant uniquement papier), mais il était tout de même structuré (utilisation de styles pour les titres, génération de la table des matières). Le travail de re-stylage s'est avéré malgré tout assez long : une vingtaine d'heures. J'ai réussi à convertir une partie représentative de la thèse avec les deux chaînes. Dans les deux cas, le traitement a été difficile car lorsque la conversion échoue, les messages d'erreurs sont souvent sibyllins du type :

« An element must not end before its content model is completely satisfied. on line 258.The element with unsatisfied content is "div4".

Ou

“An end tag for an element that is not open is not allowed. on line 345
The element is "p". The open elements are "TEI.2" and "text". The following element can end: "text". The following elements can start: "anchor", "back", "gap", "index", "interp", "interpGrp", "lb", "milestone", or "pb".”

En affichant le fichier XML, on finit par retrouver l'erreur. Les erreurs les plus fréquentes sont liées au style du texte (un style non défini ne sera pas reconnu), aux niveaux des titres (XML fonctionne de manière hiérarchique, on ne peut pas passer d'un titre 1 à un titre 3, le titre 2 intermédiaire est obligatoire) et aux graphiques.

Bilan de la conversion

Ce test comparatif a été effectué sur une seule thèse (difficulté d'obtenir des thèses de la part des écoles doctorales pour les tests en l'absence actuelle de dépôt électronique). Il a cependant permis de suivre une conversion de bout en bout. Des tests plus approfondis sont nécessaires avant une généralisation de ce projet. Avec de la pratique le temps de traitement sera plus rapide, ce qui passera en particulier par la compréhension des messages d'erreurs.

Deux différences notables entre les deux chaînes sont à prendre en compte :

- Le format d'entrée de chaîne de SPARTE est le RTF. Celui-ci est beaucoup plus lourd que le format DOC de Microsoft Word (on passe de 5 MO en .doc à 20 MO en RTF).
- Pour les nombreuses erreurs qui sont apparues lors de la conversion, j'ai pu utiliser les archives de la liste cyberthèses-users, et dans la majorité des cas, cela a suffi à comprendre et résoudre les problèmes. Cela n'a pas été possible pour SPARTE, les utilisateurs de SPARTE n'intervenant pas activement dans la liste de diffusion.

3.5. La diffusion

3.5.1 Indexation et signalement

L'indexation vise à rendre le document accessible sur internet. Elle se base sur les résultats de la conversion et sur le fichier des métadonnées.

Points d'accès au document et types de recherches:

- Notice bibliographique dans le SUDOC et le catalogue local du SCD
- Notice présente sur le site web de l'université
- Recherche par feuillement ou métadonnées
- Recherche en texte intégral

La notice bibliographique présente dans le SUDOC et le catalogue local sera à terme générée par le fichier de métadonnées avec Cyberdocs et SPARTE. Un lien vers le texte intégral est ajouté à cette notice. (zone Unimarc 856)

3.5.2 Consultation de la thèse

Sous quel format diffuser les thèses ? Cyberdocs propose le document sous forme statique en HTML et en PDF. L'HTML a l'avantage de la légèreté, le PDF celui du respect de la mise en forme. Si l'on diffuse du PDF faut-il choisir le PDF converti ou le PDF qui est généralement fourni par le doctorant lui-même ? Lyon 2 propose au lecteur le choix entre les deux.

université scientifique, cela signifie concrètement des documents qui comportent un grand nombre de schémas et d'équations, et le fait qu'une partie non négligeable de ces documents est rédigée en LaTeX.

En deuxième lieu, le nombre très important de thèses pose la question du temps de traitement. En effet, Lyon2, Doc'INSA, Limoges... sont des universités qui gèrent chacune entre 100 et 150 thèses par an.

Au sujet des thèses de médecine, le rapport Jolly indiquait ceci :

« Cette recommandation concerne l'ensemble des thèses de doctorat, à l'exclusion des thèses d'exercice en médecine dont la valorisation ne paraît pas nécessaire pour la recherche. Par ailleurs, les travaux qui accompagnent les habilitations à diriger des recherches (HDR) ne rentrent pas non plus dans le cadre ici dessiné ».

Au vu de cette recommandation et en raison du très grand nombre de thèses en santé soutenues chaque année, il a été décidé, en accord avec le directeur du SCD, Pierre Chourreau, que les thèses d'exercice de santé ne seraient pas traitées dans ce projet.

En ce qui concerne le nombre élevé de thèses scientifiques déposées à l'université, environ 300, il paraît sage de commencer par une école doctorale assez petite pour ensuite élargir progressivement le traitement, en tenant compte de l'évolution des chaînes actuelles qui ne sont pas encore à maturité. L'évolution de Cyberthèses en Cyberdocs peut nous porter à un certain optimisme, car elle a permis de réduire grandement le traitement des documents.

Choix de la plate-forme :

Nous avons choisi d'adapter la plate-forme Cyberdocs, car elle nous paraît aujourd'hui la plus avancée en termes de développements, de documents traités et de nombres d'utilisateurs.

Partie 3. Réalisation du projet

1. Les ressources humaines, matérielles et financières

1.1. Moyens humains

1.1.1 Evaluation des besoins

Dans le meilleur des cas, le traitement complet d'une thèse s'effectue en une demi-journée. Un rapide calcul nous montre que : 300 thèses multipliées par $\frac{1}{2}$ journée = 150 journées de travail supplémentaire pour le SCD en un an. Cela représente donc, un mi-temps supplémentaire dans l'hypothèse la plus favorable. Dans les faits, un plein temps sera très probablement nécessaire.

A ceci, il faut ajouter, la gestion de la chaîne de traitement, l'assistance technique, les formations à donner, le dépôt des thèses. Il faut aussi y ajouter le temps de formation des personnels et les tests. Cela signifiera un travail à plein temps du gestionnaire de la chaîne pendant les tests et la montée en charge, puis un mi-temps, une fois le projet opérationnel. Cela demandera aussi l'investissement de plusieurs personnes au sein du SCD.

1.1.2 Fiches de postes

- Chef de projet
- Technicien informatique
- Personnel dédié au stylage (cat B)

Chef de projet

Cadre A (conservateur ou bibliothécaire)

- Missions principales :
 - o Gestion de la chaîne de traitement
 - o Alimentation, stockage et archivage des données
 - o Mise en ligne d'outils d'aide à la rédaction et au dépôt des thèses (feuilles de style, documents d'aide, procédures) assistance en ligne aux utilisateurs des feuilles de style (chercheurs et étudiants)

- Définition et coordination de la politique de numérisation
 - o Suivi et animation de l'équipe chargée de la numérisation
 - o Personne ressource sur le SCD.
 - o Définition du plan de formation à destination des doctorants en concertation avec le responsable formation du SCD

- Développement des partenariats et de la coopération inter-universitaire
 - o Avec Cyberthèses : utilisation et évolutions de la plate-forme de conversion en XML.

- Compétences techniques particulières :
 - o Maîtrise des potentialités des serveurs : savoir gérer une interface Web, indexation des données.
 - o Maîtrise des DTD et du XML.

Technicien informatique

Agent de catégorie B à mi-temps

Sous la responsabilité du chef de projet, il aura pour principales missions :

- Mise en place et développement de la plate-forme de conversion :
 - o Installation sur les postes dédiés à la conversion à partir de la collecte des sources
- Maintenance de la chaîne :
 - o Maintenance logicielle : sécurité, évolutions du serveur.
 - o Suivi du matériel informatique sur l'ensemble du SCD (serveur, postes clients)
- Assistance technique auprès du personnel chargé de la conversion
- Relations avec le service informatique de l'université

« Styleur »

Agent de catégorie B à plein temps

Missions :

- Assurer des plages horaires de stylage et de conversion de documents stylés en langage XML.

Compétences requises :

- Maîtrise technique de la plate-forme Cyberthèses (V2)
- Maîtrise des outils bureautiques utilisés par les étudiants (Word, PowerPoint, Excel)
- Maîtrise d'1 PC (ouverture fichiers, stockage, gravure)

1.1.3 Réaffectation des ressources internes

On peut envisager une redistribution partielle des moyens existants.

- Mon propre poste à mi-temps pour la gestion de la chaîne au niveau du SCD.
- Le technicien informatique déjà présent pour l'assistance technique et la gestion du serveur.
- Le responsable de la formation pour le SCD.
- Diverses personnes « ressources » pour le dépôt. Il existe déjà un certain nombre de personnes ressources pour l'informatique au SCD.

1.1.4 Recrutement de personnel

En ce qui concerne le stylage des thèses et la conversion dans la chaîne de traitement, il s'agit d'un emploi supplémentaire qui devra être apporté au SCD pour la réalisation du projet. Il peut s'agir d'un recrutement de personnel de bibliothèque (CAT B). Certaines universités ont choisi de faire effectuer le stylage des documents avant conversion par des moniteurs étudiants ou du personnel contractuel. Ceci pourrait être une alternative envisageable. Dans tous les cas, une connaissance approfondie des outils informatiques et notamment du logiciel Word est une condition indispensable.

1.2. Aspect matériel

1.2.1 Machines : serveurs et postes de travail

Le choix s'étant porté sur la chaîne Cyberdocs, des serveurs web sont nécessaires pour la conversion et la publication des thèses. Le serveur de publication sera le même que celui utilisé pour le site web du SCD et qui est abrité par le CICT, (centre interuniversitaire de calcul de Toulouse). Un autre serveur servira en interne pour la conversion et la validation des documents, ainsi que l'archivage qui sera complété par une sauvegarde sur cédérom dans l'attente d'une solution d'archivage nationale qui doit être bientôt proposée par le CINES.

Les ressources nécessaires :

Sur le site de Cyberthèses/Cyberdocs : <http://sourcesup.cru.fr/cybertheses/fr/installation/materiel.html>, on trouve les informations suivantes :

« Pour les serveurs : espace disque requis d'environ 200 MO, incluant OpenOffice.org, PHP, SDX et Java. Les outils Cyberdocs eux-mêmes prennent relativement peu de place, environ 20 MO. La quantité de mémoire vive (RAM) sur le serveur est un facteur critique. Le module de conversion utilise de la mémoire de manière proportionnelle à la taille des documents qu'il doit traiter, de même qu'au nombre d'images comprises dans ces documents. [...] Ainsi, pour un document de 500 pages, il peut être nécessaire d'avoir près de 700 Mo de mémoire pour le traiter, même si très souvent une quantité moins grande sera suffisante. De son côté, la mémoire utilisée par le module de publication dépend largement du nombre d'utilisateurs simultanés que l'on souhaite servir. Pour une application telle Cyberdocs, une mémoire de 200 à 300 Mo sera en général confortable. »

Pour le serveur qui fonctionnera en local, on prévoira un micro-ordinateur actuel, mais avec les caractéristiques suivantes : 1Go de mémoire vive, un graveur et lecteur ZIP. On prévoira aussi l'achat de deux postes informatiques pour la conversion (pour le gestionnaire et le styleur). Les caractéristiques de ces postes seront les mêmes que pour le serveur local, excepté les écrans : on privilégiera des écrans 19 pouces pour le confort visuel lors du travail de stylage.

Matériel informatique	note	Coût TTC (en euros)	quantité	Total TTC
Trois PC avec graveur CD et carte réseau	Capacité 160 Go disque dur. Ram 1 GO	1084	3	3252
Moniteur 17"		200	1	200
Moniteur 19"		300	2	600
Lecteur disquettes 3 1/2 et lecteur ZIP		125,41	3	376,23
Total budget d'investissement TTC				4428

Prix donnés à titre indicatif au 14 septembre 04. Site internet de Dell (avec qui l'université a passé un marché)

1.2.2 Logiciels :

Cyberdocs est composée d'éléments Open Source gratuits. Cependant, l'ensemble des tâches à accomplir nécessite l'achat d'un certain nombre d'applications propriétaires.

- Microsoft Word bien sûr qui ne fait pas partie à proprement parler des outils Cyberdocs mais qui est le format de dépôt le plus fréquent (à plus de 90%) et qui va donc servir au re-stylage,
- La version complète d'Adobe Acrobat,
- Un logiciel d'infographie, type Photoshop pour traiter certains types de graphiques.

Ces logiciels feront l'objet d'une demande d'extension de licence au Centre de ressource informatique de l'université.

Budget logiciels

Logiciel	Prix unitaire (en euros)	Quantité	Prix net à payer
Acrobat Professionnel	52,62	2	105,24
Adobe Photoshop	167,44	2	333,84

Microsoft Office	57,55	2	115,1
Total			554,18

Prix donnés à titre indicatif au 14 septembre 04. Site du CICT (intranet)

D'autre part, les thèses seront mises en ligne sur le site de l'université de façon dynamique en utilisant l'application SDX. Cela signifie concrètement qu'il faudra installer les outils JAVA...sur le serveur du CICT.

On peut utiliser la plate-forme SDX adaptée aux thèses en ligne, proposée par l'AJLSM et disponible sur le site de Cyberdocs, en y insérant le logo de l'université Paul-Sabatier. Cependant un habillage personnalisé pourra par la suite être réalisé soit en interne, soit par l'intermédiaire d'un prestataire extérieur.

1.2.3 Locaux : salle de formation

Les formations des doctorants à l'utilisation des feuilles de style vont mobiliser une/des salles de formation. Jusqu'à cette année, des formations avaient lieu à la bibliothèque de science dans une salle prévue à cet effet. La bibliothèque entrant en phase de travaux à l'automne, il faudra trouver des salles de formation au sein de l'université. Ceci est une contrainte forte à prendre en compte dès la période de tests. En revanche, l'ouverture à venir de l'extension nous est très favorable, puisqu'une salle de formation est prévue, et que des postes informatiques neufs viendront remplacer les postes obsolètes existants.

2. Organisation du travail au sein du SCD

2.1. Installation de la plate-forme

Installation des composantes nécessaires à Cyberdocs.

La plate-forme sera installée sur le serveur local et les deux postes dédiés à la conversion. Le système d'exploitation sera Windows XP. Les composantes nécessaires sont les suivantes :

- Serveur web Apache
- Environnement PHP

- Machine virtuelle JAVA
- La suite bureautique OpenOffice.org
- Le moteur de servlets Tomcat
- La plate-forme SDX
- Les sources de Cyberdocs

Lors de mes tests, j'ai eu l'occasion d'installer Cyberdocs sur mon poste informatique. J'ai utilisé la documentation présente sur le site de Cyberthèses, ainsi que les rapports de stage de Frédéric Aliotti et Abdrahamane Anne¹⁶. Mon stage extérieur de juillet 2004, à la cellule ERAD de Lyon 2 a permis de régler les problèmes résiduels et de mieux comprendre le fonctionnement de la chaîne, en particulier l'utilisation du module de gestion.

2.2. Description du nouveau circuit pour les thèses

- Dépôt de la thèse

- Dépôt des thèses au SCD avant soutenance.
 - o 2 exemplaires papier
 - o 1 exemplaire électronique sous toutes les formes existantes : accès FTP, méls, diquette, ZIP, clé USB, cédérom.
- Signature par le doctorant d'un contrat de diffusion de la thèse mis au point par le service juridique de l'université.
- Signature par le directeur de thèse d'un avis sur le délai de diffusion.
- Vérification de la conformité papier/électronique
- Délivrance par le SCD d'un quitus, indispensable au doctorant pour soutenir.
- Archivage de la thèse sur le serveur local.
- Après soutenance, si la thèse n'est pas modifiée, le SCD délivre au doctorant une attestation de dépôt de la thèse qui permettra la délivrance du diplôme.

¹⁶ rapports de stage 2003 du DESSID, DESS d'ingénierie documentaire de l'Enssib

- Les exemplaires papiers sont transmis au service de catalogage centralisé du SCD (en pratique à la personne qui catalogue les thèses)

- Traitement de l'exemplaire électronique

- Entrées des données administratives dans l'outil de gestion des thèses. (OGET)
- Ajout des métadonnées.
- Archivage de la thèse sur serveur et cédérom
- Evaluation de la durée de traitement/de la difficulté du document.
- Stylage
- Conversion de la thèse dans la chaîne Cyberdocs par le styleur

- Vérification du résultat par le gestionnaire, validation et mise en ligne

2.3. Les différents acteurs

Formation des doctorants

La cellule ERAD a mis à ma disposition les documents utilisés pour la formation à la feuille de style des étudiants. Ce contenu peut s'adapter facilement à Paul-Sabatier. La préparation de ces formations se fera en concertation avec le responsable de la formation au sein du SCD, Manuel Durand-Barthes. Ces formations auront lieu sur la base d'un rythme trimestriel. Elles seront assurées dans un premier temps par Manuel Durand-Barthes et moi-même. Arrivé à la phase de généralisation, le projet impliquera une participation de plusieurs personnes de la bibliothèque à ces formations, ce qui a déjà été le cas dans le passé pour d'autres domaines (formation des moniteurs étudiants aux visites de la bibliothèque).

Dépôt des thèses : plusieurs personnes doivent être à même de recevoir les doctorants au moment du dépôt.

2.4. La formation du personnel

Il s'agit de former le personnel aux nouvelles tâches :

- Au stylage (il existe des formations données par Lyon 2).

- A la vérification de conformité des thèses.

Cette formation aura lieu en interne. Il faut prévoir une formation particulière pour les documents LaTeX. En effet, il faut non seulement connaître la structure des documents LaTeX afin de vérifier la conformité, mais une personne supplémentaire au moins au SCD doit être capable de manipuler des éditeurs LaTeX et de savoir le convertir vers du PDF dans le cas où le PDF fourni ne remplit pas les exigences de diffusion. Il sera intéressant de se renseigner sur les formations extérieures à LaTeX, voire au sein de l'université. Il existe aussi la possibilité d'utiliser des moniteurs-étudiants de l'école doctorale Mathématiques et applications.

3. Echancier

3.1. Projet pilote 2004-2005

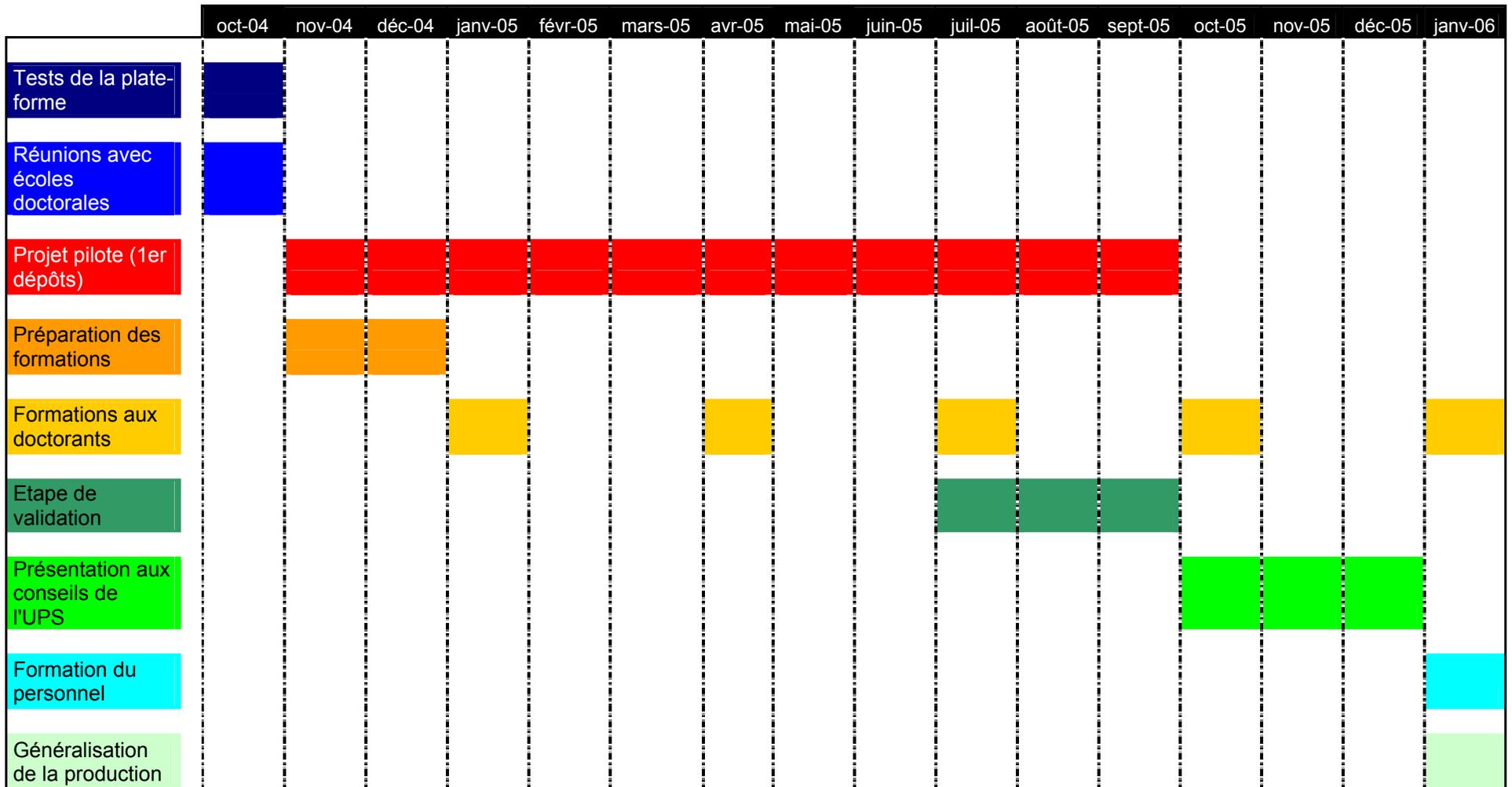
Nous prévoyons une année de tests avec une des Ecoles doctorales acceptant de participer à un projet pilote. Le critère retenu étant une production d'une cinquantaine de thèses par an au maximum. Cela pourrait être l'Ecole doctorale SEVAB, Sciences écologiques, vétérinaires, agronomiques et bioingénierie, la plus réactive en juin dernier. Pendant cette année universitaire 2004-2005, les thèses seront déposées sous forme papier et électronique de façon systématique. Dans l'hypothèse où aucune école doctorale ne souhaiterait participer au test en bloc, nous envisageons un appel au volontariat des doctorants. Ce point sera à préciser lors du Conseil scientifique du mois d'octobre, où le projet doit être validé.

3.2. Généralisation en 2006

A partir de janvier 2006, il est prévu de généraliser le projet à l'ensemble des écoles doctorales scientifiques. Cela suppose son adoption par le conseil scientifique et le Conseil d'administration. La question sera évoquée au conseil scientifique du mois d'octobre 2004. On peut prévoir qu'elle soit aussi évoquée au CEVU. L'obligation de dépôt doit être soumise au Conseil d'administration et votée avant la phase de généralisation, soit au plus tard fin 2005. Cela suppose

aussi que les crédits et moyens nécessaires soient présents (recrutement d'une personne catégorie B).

Planning du projet de mise en ligne des thèses 2004-2006



4. Evaluation et validation du projet

4.1. Détermination des critères

Plusieurs critères vont nous servir pour la validation de ce projet.

- Le nombre de thèses effectivement déposées sous forme électronique, et leur pourcentage par rapport à la population de doctorants de l'échantillon.
- Parmi ces thèses électroniques, le pourcentage de thèses complètes.
- L'adéquation entre l'exemplaire papier et l'exemplaire électronique.
- Le nombre de thèses stylées correctement (ou en grande partie)
- Le temps réel de traitement par document.

Ce critère doit être pondéré. On conviendra que pendant une période de rodage, le temps de traitement soit élevé (pour un document non stylé, de 2 à 3 jours), Le temps de découvrir les cas les plus fréquents et de prévoir un traitement type. (images, tableaux, etc.). Mais ensuite, il faudra établir un temps de traitement maximum pour la conversion en XML. Au-delà de cette durée, on choisira de convertir le document en PDF. La durée proposée dans le rapport de synthèse de la société Six-et-dix est celle proposée par l'université de Laval : 6 heures.

4.2. Etapes de validation

Au bout de six mois de tests, on déterminera :

- si on a assez de thèses pour que les critères retenus soient valides,
- si on est capable d'évaluer le temps de traitement d'un document de façon correcte a priori,
- la durée moyenne de traitement.

Si ces conditions sont remplies, on proposera aux instances universitaires de valider ce test. Si elle ne le sont pas, on prolongera ce test jusqu'à la fin de l'année civile, voire du premier semestre 2005-2006.

Avant de passer en phase de production/généralisation, on doit s'assurer que les critères suivants sont remplis :

- 80% des thèses déposées sont complètes et les documents afférents sont présents.
- La durée de traitement moyenne est inférieure à une journée.

5. Perspectives

5.1. Sparte ou cyberdocs ou... ?

Les chaînes actuelles, ainsi que le souligne le rapport de la société Six-et-Dix, ne sont pas encore arrivées à maturité. Nous avons choisi d'adapter la chaîne Cyberdocs, car elle est la plus avancée à l'heure actuelle (SPARTE ne traite pas encore le PDF). Cela peut être amené à changer, le Ministère peut choisir d'imposer l'une ou l'autre des chaînes existantes, ce n'est pas notre objet ici. En revanche, nous devons prévoir l'évolution vers une autre chaîne. Et surtout, l'objet de mon point suivant, la possibilité de faire pointer vers nos ressources.

5.2. OAI pour le signalement des thèses

Au-delà du signalement des thèses dans le SUDOC et de la présence d'un lien dans la notice du document imprimé vers la version électronique en texte intégral, l'utilisation des métadonnées va permettre de signaler des ressources qui sont stockées dans des lieux différents. L'Open Archive Initiative (OAI) vise à favoriser l'accès aux publications électroniques (e-prints) en favorisant l'interopérabilité des différents dépôts de documents (bibliothèques virtuelles).

L'interopérabilité est rendue possible grâce à la normalisation du format d'échange des métadonnées (ici le Dublin core). Cette normalisation permet à des applications informatiques conformes au protocole de rassembler puis d'exploiter les métadonnées de différents dépôts hétérogènes via un procédé nommé «metadata harvesting» (en français, récolte de métadonnées). En stockant leur adresse URL, il permet de pointer directement sur les ressources. La plate-forme Cyberdocs est compatible OAI.

5.3. Les thèses de médecine de l'université Paul-Sabatier

Au cours d'entretiens informels avec le personnel de la section santé, il est apparu que les thèses d'exercice en santé, que nous avons exclus de notre projet, étaient fréquemment demandées par le biais du prêt entre bibliothèque et qu'il existait une attente, de la part du public de Santé, de pouvoir se procurer ces documents en ligne. A moyen terme, on peut envisager le dépôt et la diffusion de ces thèses, au moins sous format PDF. Etant donné le volume, une enquête menée auprès des étudiants/doctorants permettrait de cerner les besoins (domaines particulièrement demandés).

5.4. Traitement rétrospectif

Il faudra bien à un moment donné se poser la question du traitement rétrospectif des thèses papier déjà archivées. L'université Paul-Sabatier possède un très grand nombre de thèses sous forme papier. On peut envisager le traitement d'une partie de ces documents à moyen terme sur la base d'une sélection. Ce travail de grande envergure ne sera pas réalisé en interne par le SCD. L'exemple de Marseille qui a commandé à Jouve la numérisation des thèses papiers sera peut-être la solution. On peut imaginer qu'à terme, les coûts de numérisation vont continuer à baisser, rendant la chose possible.

Conclusion

La mise en ligne des thèses est un projet qui concerne l'ensemble des acteurs de l'université. Sa réussite dépend de l'implication et de la coordination des différents services. Cette nouvelle tâche se place pour le service commun de la documentation dans la continuité de ses missions et lui permettra de resserrer ses liens avec les différentes composantes de l'université en se plaçant au cœur de la chaîne éditoriale.

Au-delà de l'aspect institutionnel, la mise en œuvre d'un tel projet demande des compétences techniques qui ne semblent pas, a priori, être celles d'un bibliothécaire. Cependant, si les techniques changent, l'enjeu reste le même : conserver, signaler et diffuser. Pour traiter un fonds ancien, il faut connaître les techniques touchant à la fabrication, à la reliure et à l'édition. Pour traiter des documents numériques, il faut connaître les formats et les techniques multimédias. Cela correspond à l'évolution du métier, mais n'est pas sans poser de questions sur les spécialisations éventuelles dont les futurs bibliothécaires pourraient avoir besoin pour être opérationnels que ce soit dans le cadre de la formation initiale ou celui de la formation continue.

Bibliographie

Sur la documentation électronique

"Le document numérique : un objet fédérateur de recherche en sciences de l'information". *Documentaliste*, 2002, vol. 39, n° 6, p. 298-305.

BRULEY, Caroline. "Les sites web des bibliothèques universitaires". *Bulletin des bibliothèques de France*, 2003, n° 4, p. 14-23.

CAVALIER, François. "La préservation des documents numériques". *Bulletin des bibliothèques de France*, 2002, n° 1, p. 97-99.

CHARTRON, Ghislaine. "Le nouveau monde numérique". *Bulletin des bibliothèques de France*, 2003, n° 3.

DESRICHARD, Yves. "Successes and Failures of Digital Libraries". *Bulletin des bibliothèques de France*, 2001, n° 3.

LUPOVICI, Christian. "La chaîne de traitement des documents numériques". *Bulletin des bibliothèques de France*, 2002, n° 1.

OKRET-MANVILLE, C. "Diffusion des thèses : la voie électronique". *Arabesques*, 2002, n° 26, p. 17-18.

Guide pratique : les publications électroniques. Paris : Archimag, 2003.

CHARTRON, Ghislaine. *Le nouveau monde numérique.* Paris : éd. La Découverte, 2002. (Recherches). ISBN 2-7071-3893-2.

JACQUESSON, Alain, RIVIER, Alexis. *Bibliothèques et documents numériques : concepts, composantes, techniques et enjeux.* Paris : éd. Du Cercle de la librairie, 1999. ISBN 2-7654-0716-9

LUPOVICI , Catherine. *La pérennité des documents numériques.* Grenoble : ISDN, 2000.

COLLOQUE INTERNATIONAL SUR LE DOCUMENT ELECTRONIQUE (6 ; 2003 ; Caen). *Document électronique dynamique.* Paris : Europa productions, 2003. 325 p. ISBN 2-909285-24-3.

Sur les aspects juridiques

FARCHY, Joëlle (éd.), RALLET Alain (éd.). *Droit d'auteur et numérique.* Paris : Hermès science, 2002. ISBN 2-7462-0429-0

Droit d'auteur et accès à l'information dans l'environnement numérique. Paris : Éd. Unesco, 2000. (Bulletin du droit d'auteur, n° 4, 2000)

Sur l'information et la formation aux doctorants

Guide pour la rédaction et la présentation des thèses à l'usage des doctorants [en ligne]. [Paris] : Ministère de l'éducation nationale , Ministère de la recherche, 2001. [consulté le 25/02/04]. Disponible sur : <<http://www.sup.adc.education.fr/bib/Acti/These/guidoct.rtf>>.

INSA de Lyon, Doc'INSA. *Documentation du projet CITHER.* [en ligne]. [Villeurbanne] : [Centre de documentation de l'INSA de Lyon], [s.d.]. [consulté le 25/02/04]. Disponible sur : <http://csidoc.insa-lyon.fr/these/doc/special_doctorants.html>.

Université de Marne-la-Vallée. Bibliothèque. [*Guide pour les thèses en ligne*] [en ligne]. Champs sur Marne : Bibliothèque de l'université de Marne-la-Vallée, [s.d.]. [consulté le 25/02/04]. Disponible sur : <http://www.univ-mlv.fr/bibliotheque/bib_ligne/theses/theses.htm>.

Université de Lyon 2, cellule édition électronique. *Didacticiel : produire et diffuser des documents structurés.* [en ligne]. Lyon, cellule édition électronique, [s.d.]. [consulté le 25/02/04]. Disponible sur : <<http://sophia.univ-lyon2.fr/didacticiel/>>.

Sur les projets de publication de thèses en ligne

BOULETREAU, Viviane, DUCASSE, Jean-Paul, GILLIERON-GRABER, Marie-Pierre. "Cyberthèses en Europe". *Bulletin des bibliothèques de France*, 2001, n° 6, p. 122-125.

BARRAL, Sabine [et] BRZUSTOWSKI Katie. " L'ABES, coordinatrice et animatrice de réseaux". *Bulletin des bibliothèques de France*, 2003, n° 2, p. 32-42.

CHARTRON, Ghislaine. "Rapport sur la diffusion électronique des thèses". *Bulletin des bibliothèques de France*, 2001, n° 2

Mise en place d'une chaîne de traitement des thèses électroniques Service Commun de la documentation de l'Université de la Méditerranée (Aix-Marseille 2). Villeurbanne : Enssib, 2003.

Diplôme de conservateur des bibliothèques : enssib, 2003. [consulté le 02-09-2004]. Disponible sur : <<http://www.enssib.fr/bibliotheque/documents/dcb/M-2003-PRO-15-Aix-Marseille2-vol1.pdf>>

ALIOTTI, Frédéric. *Etude de 2 chaînes d'édition numérique XML : projet de diffusion électronique de la production scientifique de l'INSA.* [en ligne]. Villeurbanne : Enssib, 2003.

Rapport de stage : Enssib : 2003. [consulté le 25-02-2004]. Disponible sur : <http://csidoc.insa-lyon.fr/these/doc/dess_2003.pdf>.

ANNE, Abdrahamane. *Appropriation d'une plateforme d'édition électronique basée sur XML : Cyberdocs.* Villeurbanne : Enssib, 2003.

Rapport de stage : enssib : Université Claude Bernard Lyon 1 : 2003. [consulté le 24-02-2004]. Disponible sur : <<http://www.enssib.fr/bibliotheque/documents/dessid/rsanne.pdf>>

CCSD. Centre pour la Communication Scientifique Directe [en ligne]. 2003. [consulté le 23-02-04]. Disponible sur: <<http://www.ccsd.cnrs.fr/>>

Centre pour la communication scientifique directe. *Serveur de thèses multidisciplinaire* [en ligne] 2003. [consulté le 23-02-04]. Disponible sur: <<http://tel.ccsd.cnrs.fr/>>.

ETD. *Projet ETD.* [en ligne]. [S.l.] : ETD, 1997. [consulté le 24-02-2003]. Disponible sur: <<http://etd.vt.edu/>>.

JOLLY, Claude (réd.). *Rapport sur la diffusion électronique des thèses.* [en ligne]. [Paris] : Ministère de l'Education Nationale, Direction de l'enseignement supérieur, Sous-direction des bibliothèques et de la documentation, 1999. [consulté le 25/02/04]. Disponible sur : <<http://www.sup.adc.education.fr/bib/Acti/These/jolly/entete.htm>>.

CLERC, C. *Contribution au développement d'un serveur de thèses électroniques.* Villeurbanne : Enssib, 1999.

Rapport de stage : enssib : Université Claude Bernard Lyon 1 : 1999.

MERMET, Jean-Michel. *Coordination et mise en place d'un serveur de thèses en texte intégral à l'INSA de Lyon.* Villeurbanne : Enssib, 1998.

Rapport de stage : enssib : 1998.

ROUSSEAU, Emilie. *Etude d'une chaîne de publication Web de texte intégral au format XML.* [S.l.] : [s.n.], 2001.

VAN DOOREN, Bruno. *La documentation électronique dans les universités françaises.* Paris : Ministère de l'éducation nationale, de la recherche et de la technologie 1998.

Sur le format XML

"le XML expliqué à vos enfants ". *Archimag*, 2002, n° 159, p. 26-28.

MIREMONT, Liliane. « eXtensible Markup Language ». *Bulletin des bibliothèques de France*, 2001, n° 2, p. 109-110.

XML : dossier documentaire. [en ligne]. [Paris] : Ministère de la jeunesse, de l'éducation nationale et de la recherche, Direction de la technologie, 2003. [consulté le 25/02/04]. Disponible sur : <<http://www.educnet.education.fr/dossier/xml/default.htm>>.

MORGAN, Eric Lease. *XML pour les bibliothécaires : un manuel et un atelier*. S.n. : s.l., 2003.

Sur les formats LaTeX/PDF

"Acrobat et PDF : un duo de haut vol ". *Archimag*, 2003, n° 162 , p. 26-27.

REMIZE, Michel. "Opinions : pour le PDF, sans trop de réserves". *Archimag*, 2003 , n° 162.

Adobe Acrobat 6.0 standard. Paris : Peachpit, 2003. ISBN 2-7440-8089-6.

DESGRAUPES, Bernard. *Metafont : guide pratique*. Paris : Vuibert informatique, 1999. ISBN 2-7117-8642-0.

DESGRAUPES, Bernard. *LaTeX : apprentissage, guide et référence*. 2e éd. Paris : Vuibert informatique, 2003. ISBN 2-7117-4809-X.

FABRE, Patrick. *PDF, exploitez toutes les ressources d'Acrobat 5*. [S.l.] : Aliss informatique, 2002. ISBN 2-84782-012-4.

KENT, Gordon. *Publication sur Internet avec Acrobat : un ouvrage de référence pour créer et intégrer des fichiers PDF sur Internet ou sur des intranets*. Paris : Simon & Schuster-Macmillan, 1997. ISBN 2-7440-0224-0.

Documentation pratique

DCMI. *About the Dublin Core Metadata Initiative*. [en ligne]. Dublin Core Metadata Initiative, 2003. [consulté le 24-02-2004]. Disponible sur : <<http://dublincore.org/about/>>.

Extensible markup language (XML) 1.1, W3C recommendation [en ligne]. [S.l.] : W3C, 2004. [consulté le 25-02-2004]. Disponible sur :

<<http://www.w3.org/TR/2004/REC-xml11-20040204/>>

Aperçu général de la plate-forme Cyberdocs. [en ligne]. [S.l.] : Cyberthèses, 2003. [consulté le 25-02-2004]. Disponible sur :

<<http://sourcesup.cru.fr/cybertheses/fr/apercu.html>>.

Installation [de la plate-forme cyberdocs]. [en ligne]. [S.l.] : Cyberthèses, 2003. [consulté le 25-02-2004]. Disponible sur :

<<http://sourcesup.cru.fr/cybertheses/fr/installation/installation.html>>.

Les composantes sous-jacentes à la plate-forme Cyberdocs. [en ligne]. [S.l.] : Cyberthèses, 2003. [consulté le 25-02-2004]. Disponible sur :

<<http://sourcesup.cru.fr/cybertheses/fr/installation/composantes.html>>.

Cyberthèses-users : liste de discussion des utilisateurs de plateformes cyberthèses. [en ligne]. [S.l.] : Comité réseau des universités. 2004. [consulté le 25-02-2004]. Disponible sur : <<http://listes.cru.fr/wws/arc/cybertheses-users>>.

Cyberthèse-LaTeX : groupe de travail, [liste de diffusion sur] LaTeX. [en ligne]. [S.l.] : Comité réseau des universités. 2003. [consulté le 25-02-2004]. Disponible sur : <<http://listes.cru.fr/wws/arc/cybertheses-latex>>.

Glossaire

CSS (Cascading Style Sheet) : feuille de style en cascade.

Dublin Core : ensemble d'éléments de métadonnées, constitué de 15 éléments descriptifs relatifs au contenu et à la propriété intellectuelle.

DTD (Document Type Definition) : élément constitutif d'un document SGML permettant de définir le modèle de structure d'un document et le vocabulaire de balisage adéquat.

HTML (HyperText Transfer Protocol) : langage dérivé de SGML et utilisé pour mettre en forme les pages web à l'aide de balises.

GIF (Graphics interchange format) : format de compression d'image.

JPEG (Joint Picture Expert Group) : format de compression d'image.

LaTeX : application de publication assistée par ordinateur permettant de formaliser des documents.

OAI (Open Archive Initiative) : protocole qui exploite les métadonnées.

PDF (Portable Digital Assistant) : format des fichiers du logiciel Acrobat

RTF (Rich Text Format) : format de document texte universel pour échanger des fichiers entre différents traitements de texte.

SGML (Standart Generalized Markup Language) : norme ISO de structuration de document utilisant une grammaire particulière nommée DTD.

TEI (Text Encoding Initiative) : recommandations en vue de l'échange de données électroniques textuelles (principalement sciences humaines et textes littéraires).

TEI Lite : version simplifiée de la TEI.

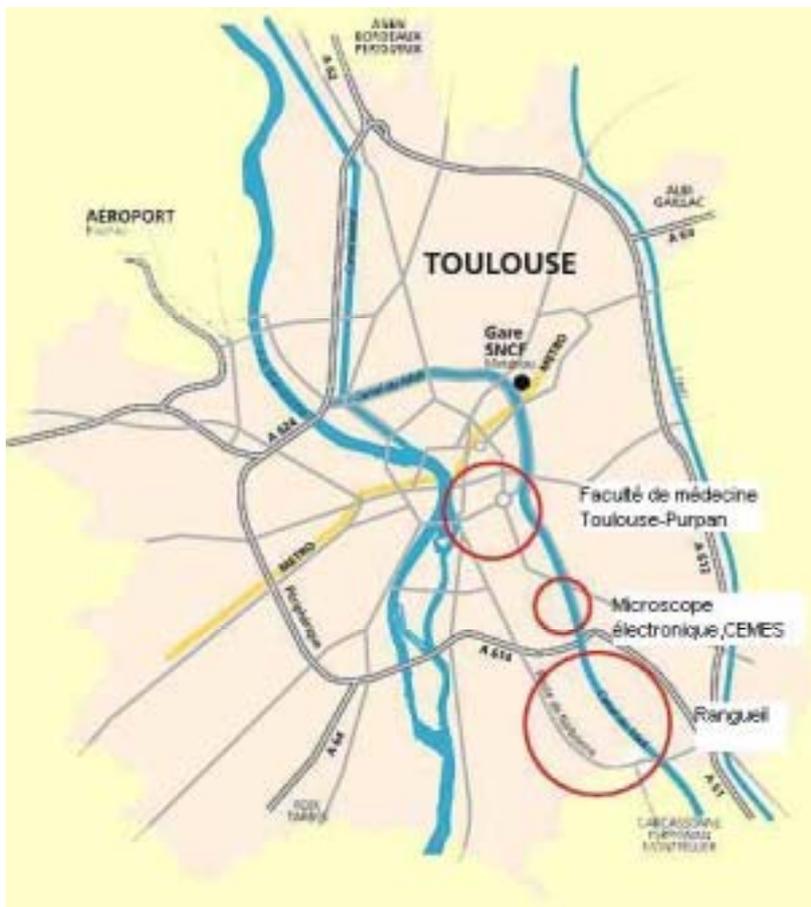
TeX : langage de programmation

XML (eXtensible Markup Language) : métalangage dérivé de SGML. Repose sur l'utilisation de feuilles de styles. Permet de décrire des informations partageables.

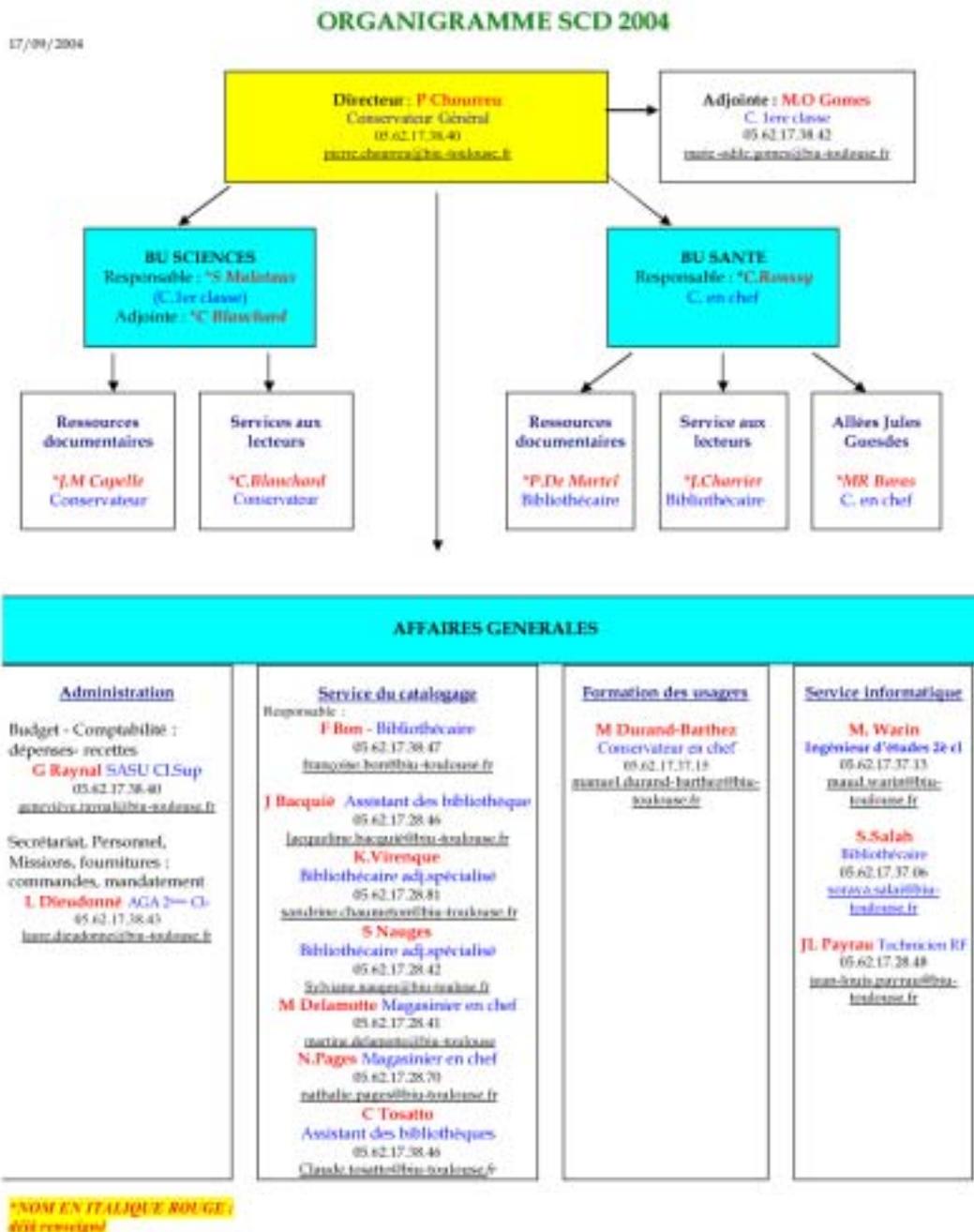
Table des annexes

ANNEXE 1-1 : PLAN DES DIFFÉRENTS SITES DE L'UPS À TOULOUSE.....	I
ANNEXE 1-2 : ORGANIGRAMME DU SCD.....	II
ANNEXE 3-1 : MANUEL DE CONVERSION L ^A T _E X/PDF	III

Annexe 1-1 : plan des différents sites de l'UPS à Toulouse



Annexe 1-2 : organigramme du SCD



Annexe 3-1 : manuel de conversion LaTeX/PDF

Conversion LaTeX->PDF

Préambule :

Dans tous les cas, vérifier si les fichiers sont en lecture seule. Les sélectionner tous (CTRL+A), clic droit propriété et décocher lecture seule.

Regarder la structure des documents.

En général : 1 fichier principal .tex, plusieurs autres fichiers .tex ou .bbl ou autres (annexes, bibliographie, chapitres...). Il faut déterminer lequel est le document complet source (ex. de nom : dissert, thèse...).

I. Ajout de commandes nécessaires à la génération d'un document PDF avec images, signets et liens

A. Copier les lignes suivantes dans le fichier principal tex

A faire quelle que soit la méthode retenue pour la compilation des sources.

- 1^{ère} ligne = `\documentclass{...}`

Si on ne trouve pas *a4paper*, il faut l'ajouter entre les accolades.

Ex. : `\documentclass[12pt,a4paper,oneside]{book}`

- après la ligne `\documentclass...copier` :

```
\ifx\pdfoutput\undefined
\usepackage[colorlinks,dvips]{hyperref}
\hypersetup{bookmarks=true,bookmarksnumbered=true}
\else
\usepackage[pdftex]{hyperref}
\hypersetup{bookmarks=true,bookmarksnumbered=true}
\fi
```

```
\ifx\pdfoutput\undefined
%%%%%%%%%%%%%% LATEX
%%%%%%%%
\usepackage[dvips]{graphicx,rotating}    %%% graphics for dvips
\DeclareGraphicsExtensions{.eps,.ps}    %%% standard extension for included graphics

\else
%%%%%%%%%%%%%%
%%%%%%%% PDFLATEX
%%%%%%%%
```

```
\usepackage[pdftex]{graphicx,rotating}    %%% graphics for pdfLaTeX
```

```
\DeclareGraphicsExtensions{.jpg,.png,.pdf}
```

```
%\DeclareGraphicsExtensions{.pdf}
```

```
\pdfadjustspacing=1
```

```
\fi
```

- *Ligne après*

```
\begin{document}
```

Copier

```
\pdfbookmark[0]{Page de titre} {deb}
```

- *Ligne précédant la bibliographie, on introduit*

```
\pdfbookmark[-1]{Bibliographie} {deb}
```

Remarque : le signet ne pointe pas toujours sur la bonne page dans le PDF obtenu. Si cela arrive, il faudra déplacer le signet dans Acrobat en faisant un clic droit->set destination)

- *Vérifier s'il y a une table des figures ou une liste des tableaux*

Chercher (CTRL+F)

```
\listoffigures
```

Ligne suivante, copier

```
\addcontentsline{toc}{chapter}{Liste des figures}
```

Chercher

```
\listoftables
```

Ligne suivante, copier

```
\addcontentsline{toc}{chapter}{Liste des tables}
```

- *Si on trouve la ligne*

```
\usepackage{graphicx}
```

La mettre en commentaire/rendre inactive (ajout d'un %)

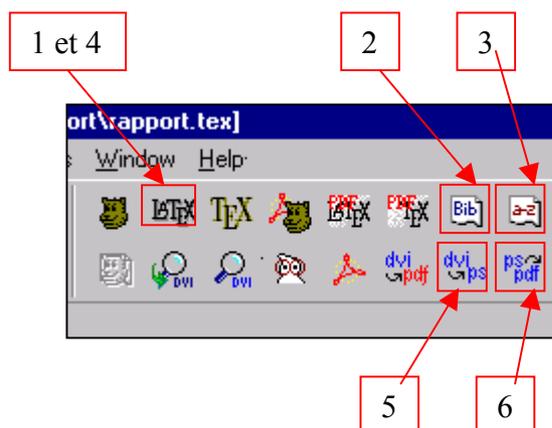
B. Méthode 1 : conversion LaTeX->DVI->PS-PDF

Méthode la plus rapide (mais ne donne pas le meilleur résultat)

A essayer en 1^{er}

Compiler avec LaTeX. Cf. copie d'écran sur Winédit.

Méthode 1 : LaTeX->dvi->ps->pdf
Boutons de raccourcis à utiliser dans
Winedt



- (1) Compiler une 1^{ère} fois avec LaTeX
- (2) Compiler la bibliographie [bib]
- (3) Compiler l'index, si le bouton n'est pas grisé [a-z]
- (4) Recompiler avec LaTeX. Parfois, il faudra répéter cette dernière opération plusieurs fois.

- (5) Lancer la conversion de dvi vers ps
- (6) Lancer la conversion de ps vers PDF

Cette dernière opération va ouvrir Acrobat avec le document PDF obtenu.
Vérifier particulièrement la table des matières.

En cas d'erreur, ou de documents incomplets, voir la F AQ.
Si le document obtenu n'est pas satisfaisant, on peut tenter de convertir avec PDF LaTeX
(mais plus long)

+ *Ajout ou sur Acrobat des signets Index et résumé.*

C. Méthode longue : PDFLaTeX

PDFLaTeX ne sait pas traiter les images .ps et .eps. Si on trouve des fichiers .ps et .eps, et que les fichiers ne sont pas aussi présents sous un format jpg ou pdf.

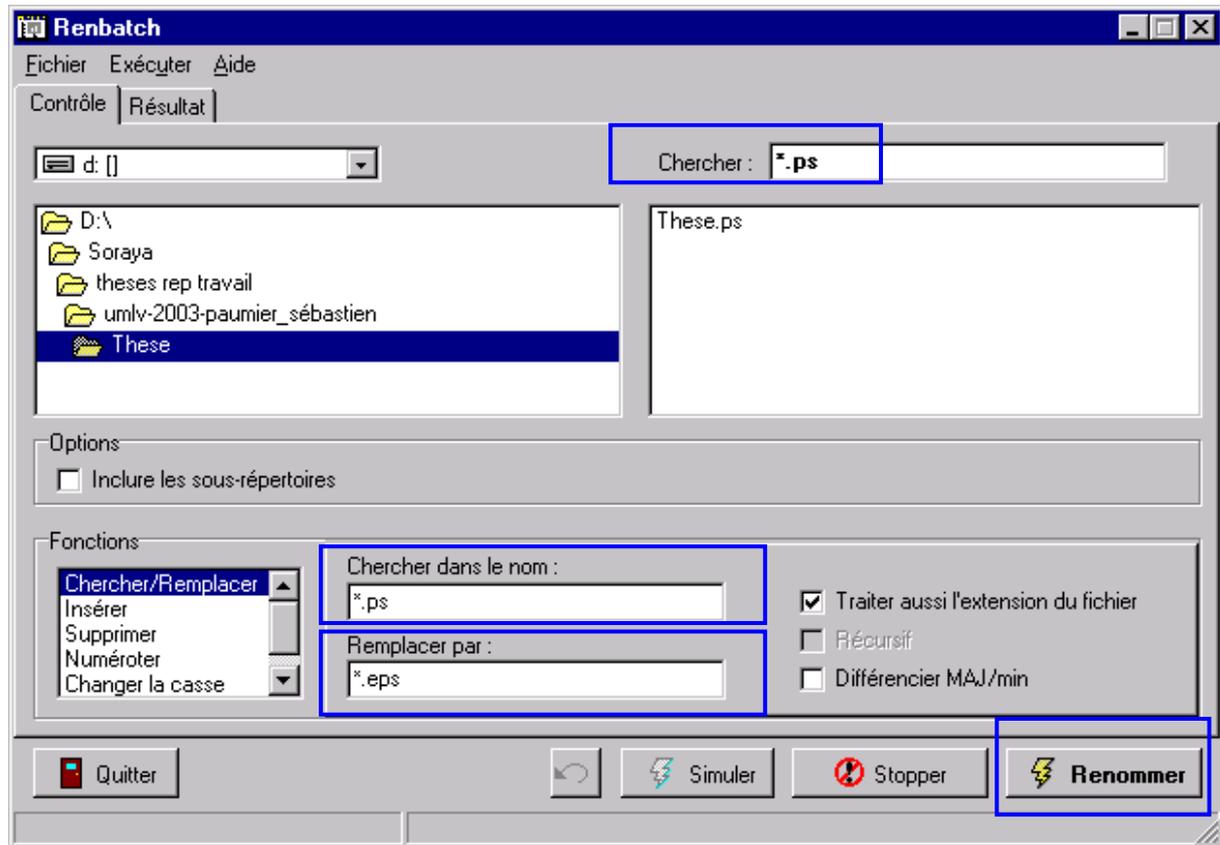
1) Renommer les .ps en .eps

- Chercher les .ps

Pour voir si on trouve des .fichiers .ps, on peut faire un clic droit sur le répertoire de la thèse ->recherche *.ps

- Les renommer :

Les fichiers ps doivent être renommés en eps. S'il y a beaucoup de fichiers, on utilisera l'utilitaire Renbatch pour renommer par lots.



Lancer Renbatch.exe

Aller chercher le dossier que l'on veut traiter et l'ouvrir

Chercher : *.*ps

Chercher dans le nom : *.*ps

Remplacer par : *.*eps

On peut simuler (éventuellement) puis

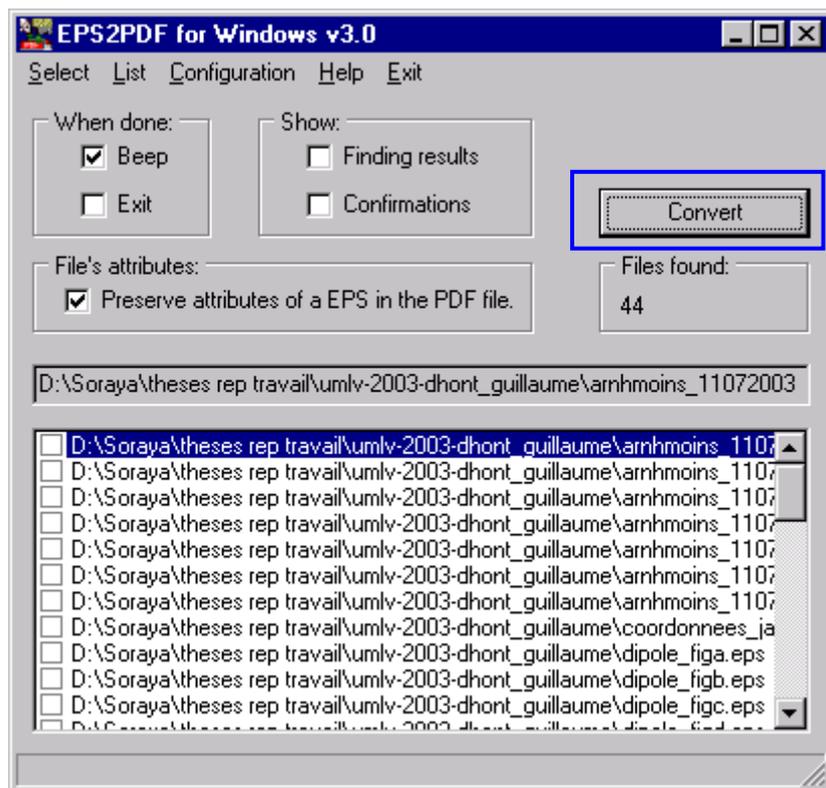
Renommer

Quitter sans sauvegarder les résultats

2) Convertir les .eps en .pdf

Si l'on trouve des fichiers .eps, il faut utiliser eps2pdf

Lancer l'exe



A) Pour la 1^{ère} configuration de eps2pdf :

Lancer eps2pdf.exe

Chercher l'exe de Ghostscript (pour la version 7 : C:\gs\gs7.04\bin\gswin32c.exe)

Menu configuration->save as->eps2pdf.config

Bouton Accept

B) Utilisation de eps2pdf

Menu select->Files : aller chercher les fichiers .eps qu'on veut convertir

Sélectionner tous les fichiers (CTRL+A)

Appuyer sur le bouton convert.

Si pb lors de la conversion (message d'erreur), voir FAQ.

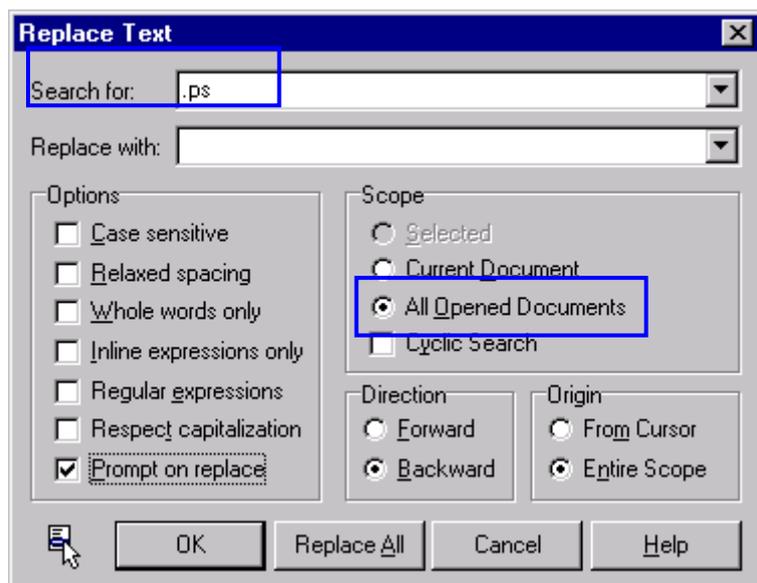
3) Retirer les extensions .ps et .eps dans le/les document(s) source(s)

[Dans Winedt]

Ouvrir les document tex

Chercher si il y a des .ps ou .eps dans TOUS les documents Tex

Pour aller plus vite, on peut ouvrir en une fois tous les documents tex du répertoire (CTRL+A)



Puis CTRL+R ou menu Search->replace

Search for : .ps

Replace with : (ne pas remplir)

Scope : vérifier que la case « all opened documents » est cochée

Options : Vous pouvez décocher « Prompt on replace », mais à vos risques et périls (vous ne verrez pas ce que WinédT va supprimer). Conseillé s'il y a énormément de graphiques.

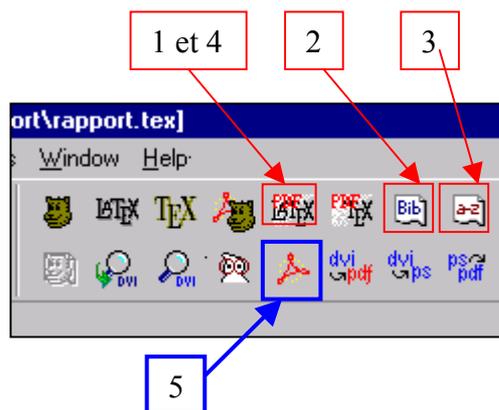
Faire la même opération pour les .eps

Si le doctorant a utilisé des consignes récentes, il n'aura pas indiqué les extensions de fichier d'images dans ses documents tex, cependant, il faut vérifier. (ex pour thèse de Panten : 35 occurrences trouvées dans l'ensemble des documents tex)

Fermer tous les fichiers en les sauvegardant : CTRL+F4 .

4) Compiler avec PDFLaTeX

- (1) Compiler avec PDFLaTeX
- (2) Compiler la bibliographie [bib]
- (3) Compiler l'index, si le bouton n'est pas grisé [a-z]
- (4) Recompiler avec LaTeX (plusieurs fois si nécessaire)
- (5) Lancer Acrobat pour voir le résultat



+ Ajout ou sur Acrobat des signets Index et résumé.