ENSSIB
Ecole **N**ationale **S**upérieure des **S**ciences
De l'**I**nformation et des **B**ibliothèques

UCBL
Université
Claude Bernard - Lyon 1

DESS en Informatique Documentaire

Mémoire de Stage

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

Conception du FrontOffice

Jean-Michel MERMET

Sous la direction de Monique Joly - Doc'INSA - INSA de Lyon

1998

Résumés et mots clés

Résumé français

Il est désormais souhaitable de diffuser nos thèses par Internet. D'autres le font déjà. Les aspects de BackOffice comprennent la réception, le traitement et l'archivage des documents. Le FrontOffice inclut l'analyse de la navigation, des accès à l'information et la documentation du projet. L'aspect financier est étudié et des bilans complets d'une telle expérience sont tirés.

Mots clés

Documentation, Diffusion information, Thèse, Internet, Serveur Web

English abstract

We have now the opportunity to publish thesis on the Internet. Other teams have been working on the same subject. BackOffice aspects include reception, treatment and archival of documents. FrontOffice aspects include navigation analysis, access to thesis and to the project information. This document ends with financial aspects and complete assessments of the project

Keywords

Information systems, information dissemination, Thesis, Internet, Web site

Table des matières

1. Introduction générale

L'émergence de nouvelles technologies de publication d'information scientifique, et en particulier l'utilisation d'Internet, permet de concrétiser ce qui semblait être, il y a peu encore, une utopie : la diffusion instantanée et à faible coût des travaux de la recherche. Qui plus est, cette diffusion s'accompagne de services de recherche d'information inédits à la forme papier.

Cette démarche s'inscrit dans l'ensemble des projets qui visent à construire ce qu'on appelle des "bibliothèques virtuelles".

Apparemment simple, la mise en place d'une telle structure soulève des questions passionnantes, tant d'un point de vue technique que de points de vue juridique, organisationnel, etc. Nous les aborderons en y apportant les réponses que nous leur avons apportées à la date d'aujourd'hui et en plaçant en perspectives celles qui restent en suspend.

Ce mémoire présente le projet CITHER (Consultation en texte Intégral des THèses En Réseau) dans tous ses aspects, en renvoyant, le cas échéant vers des documents complémentaires présents sur le site web du projet.

Dans un premier temps, une analyse des besoins dégagera les raisons pour lesquelles un centre de documentation d'école d'ingénieurs[1] comme Doc'INSA s'engage dans un tel processus de diffusion des thèses. Des exigences fondamentales seront posées pour assurer la viabilité d'un tel projet.

Dans un second temps, un bref tour d'horizon permettra de se rendre compte de l'état d'avancement d'autres projets poursuivant des buts similaires. Des conclusions en seront tirées concernant l'apport possible du projet de l'INSA.

En conclusion de cette étude préliminaire, les principaux objectifs et choix techniques sont arrêtés.

La description du BackOffice intègre les documents relatifs au travail de conception informatique précédemment mené, et détaille l'étape d'archivage des données.

L'analyse du FrontOffice comprendra une réflexion sur la conception graphique et sur la navigation d'un site web efficace. Les différents modes d'accès à l'information seront

---

[1] Dans tout le mémoire, ce qui s'applique aux centres de documentation s'applique également aux bibliothèques universitaires.

détaillés, et la "politique" concernant la documentation et la communication autour du projet sera explicitée.

Les aspects financiers d'une telle démarche seront ensuite analysés, avec un souci de transparence.

Un bilan du projet sera finalement proposé.

2. Analyse des besoins

2.1. Favoriser la diffusion des résultats de la recherche

C'est l'une des missions d'un centre de documentation moderne que de participer à la diffusion des résultats de la recherche de sa communauté scientifique.

2.1.1. Diffusion internationale

Pays développés

Pour tenir son rang dans le contexte international, la France se doit d'avoir une politique ambitieuse pour sa recherche et sa capacité d'innovation [COMI-97].

Or, une récente étude gouvernementale, "Forces et faiblesses de la recherche française" [LARE-97], mettait en exergue le fait que, parmi les grandes nations industrialisées, la France se trouve désavantagée par rapport à ses concurrents, parce qu'elle valorise mal sur le marché les découvertes de sa recherche. Elle ajoutait que malgré les multiples relations qui existent entre les entreprises et les laboratoires publics de recherche (organismes et universités), les entreprises recourent encore assez peu aux services des laboratoires publics. Les contrats de recherche représentent un peu plus de 10 % des ressources totales des laboratoires publics, le tiers seulement de ces ressources sur contrats provenant des entreprises.

Il semble que l'industrie française ait du mal à transformer les acquis scientifiques ou à les intégrer dans sa stratégie de garanties en matière de propriété industrielle. En conséquence, si, sur le plan scientifique, la situation de la France est comparable à celle de l'Europe, sur le plan industriel, elle est nettement moins bonne [LARE2-97].

D'autre part, même si ce rapport "Forces et faiblesses de la recherche française" faisait état d'une recherche publique "relativement visible" sur le plan international, les chiffres sont moroses : si la France a réalisé en 1993 près de 22% des dépenses en Recherche

et développement de l'Union Européenne, elle n'est à l'origine que de 16% des publications scientifiques européennes [LARE2-97].

Pour participer à l'amélioration de cette situation, il est possible d'accroître la visibilité de l'information scientifique produite par les organismes publics. En effet, la production scientifique française de qualité existe en véritables "gisements" sur les rayons des centres de documentations et des bibliothèques françaises. Il peut s'agir de rapports scientifiques, de comptes-rendus de travaux de laboratoire, et de thèses. En effectuer une publication large et ouverte peut véritablement aider la recherche française à être mieux identifiée, plus souvent citée et plus souvent utilisée.

L'enjeu est de taille, il participe tout simplement à la survie de la recherche française dans un environnement mondial.

Pays en voie de développement ayant le français en partage

Les pays en voie de développement ont beaucoup de mal à accéder à de l'information scientifique de qualité. On l'oublie trop souvent dans les plaintes adressées aux éditeurs scientifiques pour l'augmentation éhontée du coût des abonnements de revues : les grandes bibliothèques américaines se voient obligées de réduire de manière drastique le nombre d'abonnements proposés à leurs lecteurs. A titre d'illustration, entre 1988 et 1992, cinq universités nord-américaines ont arrêté l'abonnement de 13 021 titres (soit 5.7 % du total). Malgré ces mesures, pendant la même période, ces universités ont vu leur budget consacré aux périodiques augmenter de 30.5 %. Les prix des abonnements ont augmenté de 10 % par an entre 1991 et 1995. Ils ont doublé entre 1985 et 1995 [SENS-95].

Si l'on s'intéresse maintenant aux bibliothèques des pays en voie de développement, il est évident qu'aucune ne peut répondre à cette course effrénée au rendement menée par les éditeurs … Elles se voient donc privées de ce qui fait le moteur de l'avancement de la science : l'information.

Pouvoir leur fournir, à des coûts réduits et sans utiliser les filières terrestres peu sûres, de l'information abondante et de qualité, revêt donc une importance toute particulière. Cette information est d'autant plus précieuse qu'elle est accessible en français, langue encore répandue sur le continent africain pour ne parler que de celui-ci.

Développer un projet de diffusion d'information scientifique par Internet, c'est donc aussi aider directement les pays en voie de développement tout en participant au rayonnement culturel et économique de la France.

2.1.2. Diffusion sur de nouveaux média

L'explosion de la quantité d'information scientifique produite par la communauté internationale et l'état de l'économie du document scientifique ont fait apparaître les limites des procédés et des vecteurs de diffusion actuels. Les revues scientifiques sont reçues par un ensemble de plus en plus réduit de privilégiés, les comptes rendus de congrès sont difficiles à localiser, et, dans le même temps, des technologies mûrissent pour diffuser autrement.

Concevoir une diffusion sur de nouveaux média, et en particulier sur Internet, peut permettre de s'affranchir de ces circuits de diffusion traditionnels, mais les fonctions de tous les acteurs de la diffusion de l'information sont à reconsidérer, depuis l'auteur jusqu'au lecteur final.

En particulier, il est envisageable que certaines fonctions assurées par l'éditeur (diffusion de la recherche, organisation de la validation par un comité scientifique, archivage de l'information, etc.) puissent être assumées, dans le futur, par les centres de documentation.

2.1.3. Servir un public exigeant

L'auditoire auquel s'adresse un site de diffusion de documents scientifiques présente plusieurs caractéristiques :

C'est un public exigeant sur la qualité de l'information qu'il recherche. En particulier, il est très attentif à la validation de l'information qu'il consulte.

C'est un public qui a l'habitude d'utiliser Internet d'une manière utilitaire, et qui va concrètement utiliser l'information qu'il repère.

C'est un public rompu à l'utilisation des outils de recherche du web, et qui exigera de retrouver sur le site de diffusion tout le confort offert par les serveurs d'information commerciaux.

2.1.4. L'importance d'une conception soignée

Diffuser de l'information de qualité sur Internet pour un tel public nécessite de concevoir tout un environnement, schématiquement appelé "FrontOffice", qui permettra au lecteur d'accéder confortablement à l'information qui l'intéresse.

L'accès à l'information se doit d'être multiple et utilisant des outils connus et maîtrisés par le public. Il doit comprendre un accès par listes alphabétiques, un accès par notices catalographiques associées et une recherche sur le texte intégral des thèses.

La navigation dans le site et jusqu'à l'information doit être simple et "ergonomique". Plutôt que de concevoir une navigation de type original qui dérouterait le visiteur, il vaut mieux choisir une structure de navigation connue, que le visiteur saura utiliser.

La conception graphique participe grandement au confort de visite. Elle doit être soignée, discrète, cohérente et guider le visiteur vers l'information importante.

2.1.5. Promouvoir le site de diffusion

Créer un site de diffusion d'information scientifique ne suffit pas. Il faut aussi qu'il soit consulté !

Pour le faire connaître, il est nécessaire d'utiliser tous les moyens classiques de promotion des sites web. Ceci comprend la déclaration auprès des répertoires de sites, des moteurs d'indexation, des annonces dans les forums spécialisés et la coopération des webmasters[2] de l'INSA.

Il est également indispensable de mettre en place des actions ciblées vers la communauté scientifique francophone et internationale (communications dans des congrès, articles dans des revues spécialisées, …). En effet, ce nouvel outil ne sera utile que lorsque les chercheurs eux-mêmes se l'approprieront. Il y a deux raisons pour lesquelles ils ont intérêt à le faire : outre la source d'information qu'il représente, c'est un excellent moyen de mieux faire connaître leur équipe de recherche.

[2] Un webmaster est le responsable éditorial d'un ou de plusieurs sites web

2.1.6. Promouvoir les équipes de recherche

Actuellement en effet, les équipes de recherche ont beaucoup de mal à se faire connaître des partenaires institutionnels et privés. Les documents scientifiques produits en leur sein sont de bonnes vitrines des travaux effectués et du niveau d'excellence du laboratoire, mais ils sont diffusés de manière inadéquate. Un serveur de documents scientifiques peut donc également avoir un impact en terme d'image pour la recherche.

2.2. Un des vecteurs importants de la recherche universitaire

Le choix des thèses n'est pas fait au hasard. C'est l'un des types de documents importants produit au sein de l'université. Gérard Losfeld, Président du Conseil d'Administration de l'ADBS, dans un récent article, définissait la thèse comme une : "*Institution universitaire au statut ambigu : épreuve d'initiation (et d'intronisation souvent) et document savant, attestation d'une aptitude à la recherche et en même temps contribution au progrès scientifique par un travail original sur un point déterminé*." [LOSF-98]

2.2.1. Une collaboration de toute l'équipe de recherche

De fait, si la thèse joue les rôles indiqués ci-dessus, elle est également le fruit du travail de toute une équipe autour du doctorant. Celui-ci va largement s'inspirer des travaux antérieurs réalisés au laboratoire. De plus, les collègues chercheurs vont former le doctorant, lui suggérer la forme de son travail et la démarche cognitive à mettre en œuvre. Le contributaire essentiel étant bien entendu le Directeur de Thèse, qui pourrait, dans bien des cas, revendiquer une "parenté partagée" avec le doctorant.

Tout ceci n'enlève aucun mérite au doctorant. C'est la simple réalité d'un travail en équipe dans lequel chacun participe, pour sa part. L'équipe de recherche toute entière est donc responsable des **résultats de la recherche**.

2.2.2. Prise en compte de l'aspect juridique

Malgré tout ce qui est indiqué ci-dessus, d'un point de vue juridique, la législation ne reconnaît qu'un seul auteur : le docteur. L'idée quelque peu naïve, mais très répandue, selon laquelle "les thèses seraient des documents publics et donc libres de droits" est sans fondement.

Comme pour tout document, l'auteur d'un document, quel qu'il soit, bénéficie de toutes les dispositions prévues par la législation du droit d'auteur. Il faut donc prévoir de faire

signer un véritable contrat entre l'auteur et le centre de documentation et de proposer, entre autres, des options de protection des œuvres publiées : protection contre la modification des fichiers publiés ; protection contre le copier/coller et protection contre l'impression.

Protection contre la modification des fichiers publiés

Dans le contexte de l'information numérique, la falsification est facile et indécelable. Il existe des moyens de contrôle de l'intégrité des documents, faisant appel à des technologies de cryptage, mais ces technologies ne sont pas forcément mises en œuvre par les lecteurs. Il faut donc, dès la source, pouvoir interdire la modification des données.

Les œuvres publiées ne doivent donc pas pouvoir être modifiées, ni par un lecteur indélicat, ni d'ailleurs par l'auteur lui-même (diffusion de la version canonique telle qu'elle a été acceptée par le jury).

Protections contre le copier/coller

L'auteur doit avoir le choix d'interdire le copier/coller depuis son document, ce qui pourrait constituer une menace de pillage du travail intellectuel mis à disposition.

Protections contre l'impression

L'auteur peut n'autoriser que la simple consultation sur écran de son document et donc vouloir interdire l'impression de son document (sa représentation sur papier) par un tiers.

2.3. Nécessité de proposer des outils respectant le travail

Concernant le travail à diffuser, nous choisissons de bien distinguer le fond, qui est la substance même de l'information à diffuser, de la forme, qui est la manière de présenter cette information. En pratique, la distinction est parfois difficile à faire tant la présentation d'une information contribue à son sens.

2.3.1. Sur le fond

La diffusion électronique de documents doit permettre de le respecter quant au fond. L'information transmise par la forme papier et par la forme électronique doit être identique en substance. Le centre de documentation doit s'y engager fermement.

2.3.2. Sur la forme

La diffusion électronique de documents doit aussi permettre de le respecter quant à la forme. Or, le changement de mode de diffusion (passage du papier à l'écran) induit forcément une modification de cette forme. Pensons simplement au fait que l'auteur a choisi le type de papier sur lequel son document a été imprimé. Comment reproduire un velin à l'écran ?

L'objectif poursuivi sera donc de réaliser, à l'aide des technologies actuelles, le meilleur fac-similé possible du document papier.

2.4. Tout concevoir pour le long terme

2.4.1. Adresses Internet fixes

Un des problèmes rencontrés dans la consultation des ressources électroniques est la non-garantie de la permanence de leur URL[3]. Le projet doit inclure une procédure très stricte pour l'attribution d'une adresse, et la volonté de ne pas la modifier, même à long terme.

2.4.2. Archivages

Une grande quantité d'information numérique a été perdue depuis le début de l'ère informatique, qui ne date pourtant que des années 1940 ! La diffusion des thèses ne court a priori pas le même risque, puisque la diffusion papier n'est pour l'instant pas remise en cause. Cependant, il convient d'étudier et de mettre en place des procédures qui permettront de conserver, à l'aide des dispositifs les plus sûrs, l'information numérique diffusée par ailleurs. L'étude devra porter sur ce qu'il convient de garder, sur quel médium, et dans quelles conditions.

2.4.3. Mutation entre environnements technologiques successifs

Concevoir une diffusion et une conservation au long terme d'une information numérique, c'est se préparer à faire face aux futures avancées techniques qui imposeront de véritables "mutations technologiques" aux documents électroniques, de l'environnement

---

[3] L'URL (Uniform Ressource Location) est une convention universelle de désignation de ressources sur Internet. Pour en savoir plus, consultez <URL:http://www.w3.org/Addressing/Addressing.html>

informatique *n* à l'environnement *n+1*. Diffuser dans le format du futur se prépare dès maintenant.

3. Etude de l'existant

D'autres projets poursuivent la même finalité. Il est très instructif d'étudier comment les différentes équipes, confrontées aux mêmes problématiques que les nôtres, ont décidé d'agir. Cela ne permettra enfin de nous positionner et de mettre en avant les spécificités de notre démarche.

3.1. Le traitement et la diffusion des thèses papier

Il convient de bien distinguer les thèses d'exercice (notamment celles de médecine, de pharmacie et de vétérinaire) qui ne donnent pas lieu à de telles procédures, des thèses de doctorat.

Actuellement, les thèses de doctorat soutenues sont déposées sous forme papier en trois exemplaires, dans la bibliothèque de l'université de soutenance. Deux exemplaires sont conservés à la bibliothèque et constituent le dépôt réglementaire. C'est à chaque organisme de définir les conditions de consultation (consultation sur place ou prêt ou PEB).

Avec l'accord écrit du docteur, le dernier exemplaire est envoyé dans l'un des deux ANRT[4] (Ateliers Nationaux de Reproduction des Thèses), à Grenoble ou à Lille, et il est reproduit sur microfiches diazoïques pour une diffusion :

Gratuite et systématique dans toutes les bibliothèques universitaires de France, les lieux d'enseignement français à l'étranger et l'INIST (INstitut de l'Information Scientifique et Technique)

Commerciale dans tous les autres cas.

A Grenoble par exemple, le tirage, qui varie selon les disciplines, est en moyenne de 80 exemplaires pour la diffusion gratuite. Un tirage supplémentaire peut être effectué, à tarif réduit pour les étudiants, enseignants et établissements publics [GENO-98].

Les formulaires remplis au moment du dépôt des thèses permettent d'alimenter les banques de données Téléthèses et Pascal.

---

[4] A ne pas confondre avec l'Agence Nationale de la Recherche Technologique, qui finance nombre de thèses françaises.

3.2. Diffusion électronique : les projets français

3.2.1. ANRT Lille

L'ANRT de Lille propose, depuis peu, un service de *Thèse à la carte*, qui offre la possibilité d'acquérir une thèse sous la forme d'un livre. Dans le cadre de ce service, chaque thèse est intégralement numérisée sous forme image (format TIFF). Chaque image est ensuite "encapsulée" dans un fichier PDF[5]. Entre décembre 1997 et juin 1998, 3 500 thèses ont été traitées ainsi.

Les avantages de cette solution sont nombreux :

Simplicité de procédure : la numérisation est effectuée en même temps que le microfilmage des pages de thèses.

Canonicité du résultat : la forme électronique produite est, par définition, de même version que la forme papier.

Identité de la forme : la forme image reproduit exactement la mise en page de l'auteur.

Elle présente néanmoins des inconvénients :

La taille des fichiers : même avec de forts taux de compression, la forme image "pèse" beaucoup plus lourd que la forme "texte". Il en résulte des temps de téléchargement et de consultation beaucoup plus élevés.

L'absence d'information en mode texte interdit la recherche en texte intégral. La structure logique des documents (le plan) ne peut pas être exploitée facilement.

3.2.2. Le projet Callimaque

Fruit d'une collaboration entre l'IMAG, l'INRIA, le CICG et le Centre de recherche de Rank Xerox de Grenoble (RXRC), ce projet propose l'accès à plus de 3 000 documents (dont des thèses) traitant de l'évolution des mathématiques. Callimaque est un projet de gestion électronique de documents qui intègre le traitement, la production de documents et la recherche d'information. Il est basé sur un produit de Xerox appelé XDOD (Xerox Document On Demand) qui permet la numérisation, le stockage et l'indexation des documents. Les documents sont présentés en mode image [BIEN-96].

---

[5] PDF : Portable Document Format, format de diffusion électronique de la société Adobe

3.2.3. Webdoc

Le but du futur service WebDoc est de fournir à des bibliothèques et à des utilisateurs finaux un accès en ligne, via le Web, à des documents numérisés en texte intégral. Le dispositif comportera d'une part un catalogue centralisé, décrivant l'ensemble des documents, et d'autre part des serveurs, sur lesquels seront répartis les documents en question. Ceux-ci seront liés aux notices du catalogue par des pointeurs.

Les thèses diffusées par le projet CITHER de l'INSA de Lyon, aux cotés des thèses fournies par l'ANRT de Lille, figureront parmi les premiers documents de ce système. L'expérimentation devrait avoir lieu à partir de l'automne 1998 [VAISS-97].

3.3. Diffusion électronique : les projets étrangers

Citons, parmi les principaux projets en développement à l'étranger [BEAU-98] :

3.3.1. Le Projet Electronic Theses and Dissertations (ETD)

Depuis janvier 1997, toutes les thèses (maîtrise et doctorat) déposées à Virginia Tech (USA) doivent l'être sous format électronique. **Il incombe aux auteurs eux-mêmes de fournir leur thèse, sous forme de fichiers PDF ou en SGML** en utilisant la DTD ETD Electronic Theses and Dissertations, développée spécifiquement pour ce projet [ELEC-98].

3.3.2. Le Networked Digital Library of Theses and Dissertations (NDLTD)

Le projet Networked Digital Library of Theses and Dissertations (NDLTD) est une extension de celui du Virginia Tech. Ce projet a pour but d'établir un réseau de thèses accessibles sous forme électronique. Il a débuté en septembre 1996 [NETW-98]. Il rassemble 39 membres officiels, essentiellement des universités.

3.3.3. University Microfilms International (UMI)

UMI (USA) s'est engagée à convertir, à partir du 1er janvier 1997, toutes les thèses qui lui sont soumises (que ce soit en format papier ou en format électronique) vers le format PDF. A la fin de 1997, plus de 45 000 titres étaient disponibles en format PDF [PROQ-98].

3.3.4. UW Electronic Theses and Dissertations Server

Projet pilote à l'Université de Waterloo (Canada), ce serveur présente une vingtaine de thèses en format PDF [UWEL-97].

3.3.5. The Joint Electronic Thesis and Dissertation Project

Projet des bibliothèques de l'University of Toronto et de la York University. La phase actuelle est une étude de faisabilité. Quelques rapports d'étude sont disponibles et 8 thèses sont accessibles en ligne [ETDP-98].

3.3.6. Projet de l'Université Laval au Québec

En cours d'étude, ce projet a pour but de traiter la production, la soumission et la diffusion des mémoires de thèses par voie électronique à partir de l'an 2000.

Les participants à ce projet remarquent que : "des craintes concernant les droits d'auteurs des étudiants, des objections d'éditeurs, des réticences de professeurs ont été manifestés lors d'expériences similaires." [COMI-97]

3.4. Bilan de l'existant

De cette courte étude il ressort quelques traits marquants :

Tous les projets utilisent le format PDF en mode texte, à l'exception notable de deux projets français, l'un utilisant du format image encapsulé dans du PDF (ANRT Lille), l'autre un mode de diffusion image piloté par une application Xerox.

Les thèses ne sont pas toutes, de loin, disponibles en libre accès. Beaucoup sont disponibles en intranet seulement. L'UMI fait payer la consultation des documents.

Aucun projet ne tente d'offrir des services supplémentaires à la "simple" diffusion des documents. En particulier, aucun n'assiste le lecteur dans sa consultation de documents pourtant volumineux.

Les problèmes de droits d'auteurs sont peu abordés, même si l'on devine derrière le faible nombre de documents diffusés, qu'ils sont très présents à l'esprit des concepteurs.

De ces quelques traits marquants, il nous semble qu'il y a place pour une nouvelle expérimentation de diffusion de thèses électroniques par Internet. Tout en s'appuyant

sur les technologies que la très grande majorité des projets recommande, elle s'en distingue par un certain nombre de points qui touchent à la navigation dans les documents diffusés, au FrontOffice et à la gestion du droit d'auteur.

4. Principaux objectifs et choix techniques

4.1. Un élément essentiel : le choix du nom CITHER !

Appelé à être largement diffusé, le nom du projet revêt une importance particulière. Après d'intenses séances de remue-méninges, nous avons opté pour le nom **CITHER**, pour **C**onsultation en texte **I**ntégral des **TH**èses **E**n **R**éseau.

Une étude réalisée sur la banque de données ICIMARQUES en juin 1998 a permis de vérifier que ce terme n'était pas utilisé dans la catégorie de produits et services qui nous correspond. Il est prévu d'effectuer prochainement un dépôt de ce nom.

4.2. Les objectifs

Nous nous sommes donc fixé pour objectif d'offrir gratuitement, par Internet, l'accès à un maximum de thèses soutenues à l'INSA de Lyon depuis janvier 1997. Les documents devront être optimisés pour la diffusion par le réseau.

Nous voulons en outre proposer un environnement complet de consultation et de recherche d'information, qui prenne en compte les données de référence des thèses.

Des efforts importants seront entrepris afin de conserver l'information dans les meilleures conditions possibles pour le futur.

Toute la documentation du projet, les outils logiciels et conceptuels doivent être accessibles pour de futures expérimentations.

Nous cherchons, en résumé, à mettre en place une véritable bibliothèque virtuelle de thèses.

Nous nous fixons finalement un objectif de promotion de la recherche et des laboratoires de notre Institut.

4.3. Les grandes solutions techniques adoptées

4.3.1. Le format de diffusion

La problématique du choix de format(s)

Tous ceux qui ont un jour récupéré des fichiers créés à partir d'un autre logiciel, ou du même logiciel sur un ordinateur utilisant d'autres polices, d'autres modèles ou feuilles de

styles, etc. savent que l'opération se conclut souvent par une perte d'information de structure et de mise en page. La compatibilité ascendante des documents produits par un même logiciel est loin d'être parfaite, … même des formats très répandus comme Postcript ou HTML existent en plusieurs variantes.

Une diffusion réseau implique aussi que soit prise en compte la taille des fichiers : le format Postcript est d'ores et déjà handicapé par le simple fait qu'il est très volumineux, et la compression n'est pas une solution viable car aucune technique n'est valable pour toutes les plates-formes.

Tous ces exemples militent pour qu'une analyse très minutieuse du choix du format soit effectuée.

Le format actuel

L'analyse du choix effectué par les projets concurrents et l'analyse technique des formats possibles nous ont amené tout naturellement à choisir PDF d'Adobe comme format de diffusion.

PDF peut se concevoir comme étant l'équivalent électronique d'une impression. Il présente les caractéristiques suivantes[6] :

Largement diffusé aujourd'hui (Adobe annonce 250 000 sites diffusant des documents PDF [ABOU-98]), par une société reconnue, il semble qu'on puisse raisonnablement se fier à sa pérennité.

Des logiciels lecteurs de PDF sont disponibles gratuitement pour la quasi totalité des plates-formes informatiques, ce qui en fait un format lisible presque partout sans problème ; c'est une caractéristique qui n'est partagée que par HTML (et encore …) et par le format texte ASCII 7 bits !

C'est un format compact. A titre d'exemple, voici un tableau comparatif de quelques tailles de fichiers de thèse, en format Word ou postcript et leur "équivalent" PDF :

[6] Pour plus d'informations sur le format PDF, consultez le site web d'Adobe à http://www.adobe.com

Ce format permet de mettre en oeuvre une protection des fichiers par clé logicielle :

- Contre la modification : le fichier ne peut être ouvert dans l'application Acrobat Exchange qui permettrait de le modifier.
- Contre l'impression : le fichier ne peut être imprimé. Seule la copie écran peut l'être, avec une forte dégradation de la résolution.
- Contre le copier/coller : le lecteur ne peut sélectionner puis copier les éléments texte du fichier PDF

Ce format permet également de retenir les éléments de mise en page et de police, éléments indispensables au respect de la forme voulue par l'auteur.

Les logiciels lecteurs de PDF incluent des fonctionnalités d'aide à la lecture (outil "loupe", …), à la recherche d'information (outil "jumelles") et à la navigation (déplacements entre les pages,…). Dans bien des cas, grâce à la fonctionnalité "plug-in", l'affichage des fichiers PDF est réalisé directement dans la fenêtre des butineurs[7] classiques, pour un plus grand confort de visite.

Afin d'aider le lecteur dans sa consultation , il est possible "d'enrichir" les fichiers PDF par des éléments de navigation visuels (création de "vignettes" des pages du fichier) et hiérarchique (reprise de la structure hiérarchique du document, c'est-à-dire de son plan, par des liens hypertextes intra et inter-documents). Ces fonctionnalités sont largement mises en œuvre dans le projet CITHER.

Ces enrichissements sont cependant limités et nombre de données de structuration de l'information sont perdues en convertissant un fichier Word[8] en PDF : le plan, les renvois, les signets, etc.

---

[7] C'est ainsi qu'on nomme Netscape et Internet Explorer quand on ne veut privilégier aucun des deux !

[8] Fichier Word bien construit, utilisant des styles de titre hiérarchisés. Les quelques thèses

Les formats futurs

En observant la tendance actuelle, il est possible de repérer une technologie qui permettra, bientôt, d'offrir les meilleures garanties de portabilité entre environnements informatiques, et sans doute un archivage plus sûr.

Il s'agit de XML (eXtended Markup Language)[9]. Dérivé de SGML, le format XML est un métalangage, donc un langage permettant de créer des langages, à l'exemple de HTML. Plus précisément, XML permet de définir des balises de structuration de documents. Les avantages d'XML pour la diffusion de thèses sont nombreux :

Le codage de l'information est effectué en ASCII 7 bits. Ceci garantit une portabilité idéale et une relecture aisée, même dans un futur à moyen terme.

L'information de structure est totalement distincte de l'information de représentation de cette structure. Cela signifie qu'un même document codé en XML peut être représenté d'une certaine manière sur le papier, d'une autre manière sur un écran, et d'une troisième manière à l'aide d'un synthétiseur vocal.

XML est un format "universel" non propriétaire, mais d'ores et déjà adopté par l'industrie. Des logiciels très répandus comme Word de Microsoft permettront, dès 1999, d'enregistrer directement en XML.

Par contre, XML n'intègre pas (encore ?) d'options de protection (interdiction de modifier, d'imprimer et d'effectuer des copier/coller) comme celles de PDF. Cet inconvénient, s'il n'est pas corrigé, restreindra l'utilisation d'XML à l'archivage ; la diffusion continuant à se faire en PDF. XML est un format jeune, il est donc un peu tôt pour se prononcer à ce sujet.

---

que nous avons eu à traiter jusqu'à présent nous laissent à penser que les besoins de formation à l'utilisation de Word sont nombreux. Voir le paragraphe 8.6.1 pour les actions que nous comptons mener en ce sens.

[9] Pour plus d'informations concernant XML, merci de consulter le site de la W3C à <URL:http://www.w3.org/XML/>.

4.3.2. La conversion des documents

Les documents parviennent à Doc'INSA sous forme électronique. Une application[10], pilote toutes les étapes de la conversion en utilisant essentiellement Word 97 et Acrobat Exchange d'Adobe pour créer les fichiers PDF[11]. L'opérateur dispose d'un véritable guide de conversion qui le conduit pas à pas.

4.3.3. Le mode de diffusion

La diffusion s'effectuera à partir d'un simple serveur web. Les documents diffusés ayant la particularité d'être à la fois peu nombreux (au maximum une douzaine de fichiers par thèse, et en moyenne 120 nouvelles thèses par an) et de taille importante, nous pouvons éviter l'emploi d'une solution de GED propriétaire qu'il aurait fallu acquérir et nous contenter d'un simple serveur de fichiers. Nous gérons nous-mêmes l'adressage des fichiers.

4.3.4. Les différents modes de recherche de l'information

La recherche par listes alphabétiques

Pour chaque année de soutenance, nous gérons manuellement une liste alphabétique de références bibliographiques des thèses, triée par noms d'auteurs.

La recherche par catalogue

Doc'INSA utilise DORIS-LORIS de la société EVER pour la gestion des ouvrages, du bulletinage et du prêt. Le projet CITHER a imposé l'achat du module DORIS Web, afin de disposer d'une interface web pour l'OPAC[12]. Grâce à une personnalisation de cette

---

[10] Application développée entre novembre 1997 et juin 1998 par M. Marc-Etienne Huneau, dans le cadre d'un Projet de Fin d'Etudes (PFE) en Génie Informatique à l'INSA de Lyon. Ce PFE a été supervisé par Jean-Michel Mermet.

[11] Voir <URL:http://csidoc.insa-lyon.fr/these/doc/index.html> pour accéder aux documents qui détaillent ces opérations.

[12] Online Public Access Catalog : Catalogue informatisé en ligne grâce auquel nos lecteurs accèdent aux cotes des ouvrages qu'ils désirent consulter et/ou emprunter.

interface et à l'ajout d'un champ MARC[13], les notices de documents disponibles sur Internet disposent d'un champ hypertexte qui pointe vers l'URL du point d'entrée vers le document. La recherche se fait par auteur, mots du titre, sujet et année de soutenance. La recherche par numéro d'ordre est à l'étude.

La recherche dans le texte intégral des thèses

Plusieurs établissements du Campus de la Doua travaillent ensemble au sein du projet VISUEL (**V**éritable **I**ndexation des **S**ites **U**niversitaires et **E**coles de **L**yon), afin de mettre en place une indexation des documents des différents sites web du campus. Depuis juin 1998, Altavista Search Intranet 97 est installé[14] sur un serveur du CISM[15] et indexe intégralement les thèses PDF du projet CITHER. Une interface personnalisée à ce moteur est proposée au sein des pages web de CITHER[16]. Elle intègre un développement en Javascript qui permet de "cacher" aux visiteurs le choix de la sous-base "CITHER" de l'index d'Altavista. La recherche s'effectue par tout mot des documents. Il est possible de restreindre la recherche à une année de soutenance, aux simples ponts d'embarquement[17],… .

---

[13] Format MARC : MAchine-Readable Cataloging. C'est le format qui structure les notices du catalogue des ouvrages de Doc'INSA.

[14] Accessible à l'adresse <URL:http://www.univ-lyon1.fr:9000/>

[15] Centre d'Informatique Scientifique et Médicale de l'Université Claude Bernard. Voir <URL:http://www.univ-lyon1.fr/CISM/> pour plus d'informations.

[16] A <URL:http://csidoc.insa-lyon.fr/these/recherche_ti.html>

[17] Accès normalisés aux thèses. Voir le paragraphe 6.1.4 pour plus d'informations.

5. Le BackOffice

Le BackOffice correspond à toutes les opérations qui sont effectuées en dehors de la vue du public et des auteurs. Le BackOffice tel que nous le concevons consiste en la récupération des documents sous forme électronique, leur traitement pour en réaliser une forme diffusable et l'archivage des données.

5.1. Réception des documents et des autorisations

5.1.1. L'organisation du circuit thèse

Récupération des documents - situation actuelle

Avant la mise en place d'un circuit qui garantira que chaque auteur de thèse aura eu l'occasion d'accepter la diffusion de son travail avec CITHER, nous procédons de la sorte :

A l'heure actuelle, les doctorants viennent à Doc'INSA avant la soutenance de leur travail, pour une vérification obligatoire de leurs références bibliographiques. C'est à cette occasion que nous les informons de l'existence du projet CITHER. La procédure de remise des formes électroniques leur est présentée (voir "Conseils techniques pour le dépôt de thèses électroniques." sur le site de CITHER à <URL:http://csidoc.insa-lyon.fr/these/doc/tdepot_conseils.html>).

Nous leur distribuons également une demande d'autorisation de diffusion qu'ils doivent remplir et nous retourner[18].

Nous leur demandons ensuite de nous fournir la forme canonique (c'est-à-dire identique à la forme papier) et électronique de leur thèse par le médium de leur choix (disquettes 3'1/2, disquettes ZIP[19], cédéroms gravés, bandes DITTO 800 Mo[20] et serveur FTP

---

[18] Il est à noter que la version définitive de cette demande n'est pas encore rédigée, et que tous les docteurs devront être ultérieurement contactés pour signer un nouveau contrat.

[19] Développées par la société Ioméga, les disquettes Zip permettent de stocker 100 Mo sur un support compact.

[20] Bandes informatiques utilisées pour les opérations de sauvegarde

anonymous[21]). Cela se produit entre le moment de la soutenance et quelques mois plus tard, suivant la disponibilité des docteurs et la demande ou non, de la part du jury, de modifications à apporter au texte présenté lors de la soutenance.

Quand nous avons un contact direct avec les auteurs, un membre de l'équipe de conversion vérifie avec eux qu'ils ont bien respecté les consignes indiquées sur le guide "Conseils techniques pour le dépôt de thèses électroniques."[22]

Concernant les thèses déjà soutenues depuis janvier 1997, et pour lesquelles le dispositif mentionné ci-dessus n'était pas en place, nous comptons dès la rentrée contacter les directeurs de laboratoires afin d'obtenir les formes électroniques et joindre un par un les docteurs pour obtenir leur accord de diffusion. Si cela n'a pas encore eu lieu, c'est parce que deux obstacles se présentent à nous :

1. Le texte d'autorisation de diffusion n'est pas juridiquement correct (voir paragraphe 5.1.2) ; nous travaillons actuellement à sa révision avec l'aide d'un avocat.

2. Le projet CITHER n'est pas encore connu des directeurs de laboratoire. Nous comptons lancer une action de communication dès la rentrée à ce sujet (voir paragraphe 6.4) et leur demander ensuite de nous fournir les formes électroniques.

S'organiser pour la montée en charge

Nous souhaitons vivement mettre en place, aussitôt que possible, une procédure formalisée de récupération des formes électroniques et des autorisations de diffusion.

Nous comptons pour cela sur une coopération étroite avec le D.E.D., le Département des Etudes Doctorales de l'INSA de Lyon, qui gère notamment les formalités administratives de soutenance des thèses.

Notre souhait est, qu'au moment de la délivrance de l'attestation de soutenance, le docteur se détermine, soit pour :

Refuser de voir son travail diffusé par le projet CITHER.

---

[21] service d'Internet permettant le transfert de fichiers, ici entre la station du doctorant et le serveur de thèses.

[22] <URL:http://csidoc.insa-lyon.fr/these/doc/tdepot_conseils.html>

Accepter la diffusion de sa thèse par CITHER, en déposant la forme électronique (qu'il garantira canonique) et l'autorisation de diffusion

Cela nous semble être le moment idéal d'intervention dans le "circuit de la thèse". Cette procédure peut être mise en place très rapidement. Nous devrons cependant attendre la rentrée universitaire pour qu'elle devienne effective.

5.1.2. Le droit d'auteur

La rédaction d'une autorisation de diffusion

Au commencement de la réflexion, il nous semblait que nous n'aurions aucune difficulté à rédiger une autorisation de diffusion des thèses, car ces documents étaient (naïvement) perçus comme libres de droits et diffusables sans limite.

Il n'en est rien. Le document thèse est protégé, comme tout document, par le Code de la propriété intellectuelle, littéraire et artistique.

Son auteur unique en est le docteur.

Il en résulte que toute diffusion de son travail doit faire l'objet d'un véritable contrat passé avec l'INSA de Lyon, contrat qui définit très clairement l'utilisation qui est faite de ses documents. La législation du droit d'auteur est ainsi faite que tout ce qui n'est pas expressément autorisé par cecontrat est interdit. On comprend donc l'importance que revêt la conception d'un tel contrat.

Nous avons donc décidé de nous adresser à un avocat spécialiste des questions de propriété intellectuelle, pour qu'il rédige avec nous un texte juridiquement valable. Ce travail est toujours en cours. Il sera diffusé sur le site CITHER dès que possible.

Procédure mise en place pour le respect des clauses

Dans le projet de contrat, Doc'INSA s'engage à mettre en œuvre, le cas échéant :

1. Le retrait du document si l'auteur en exprime la volonté. Cela signifie que si les thèses du projet CITHER alimentent le projet Webdoc (voir paragraphe 8.7), il faut en prévoir le retrait le cas échéant

2. Les protections contre la modification des fichiers : il ne s'agirait en aucune façon de la possibilité de modifier les fichiers du site CITHER, mais plutôt d'une modification qui serait faite sur un fichier téléchargé depuis le site CITHER vers le poste de travail

d'un lecteur. Sans mesure de protection, celui-ci pourrait modifier la forme et le fond des documents avant d'en faire une diffusion propre.

3. La protection contre l'impression : il s'agit ici d'interdire au lecteur l'impression, sur une imprimante locale, des documents visualisés à l'écran. Cela ne permet pas d'interdire "l'impression écran", mais cette dernière est d'une qualité nettement inférieure à l'impression directe du PDF et ne permettra pas une rediffusion de l'information.

4. La protection contre le copier/coller : il s'agit ici d'interdire la possibilité de copier un bloc de texte du fichier PDF vers un traitement de texte ou autre. Bien entendu, on ne peut rien faire contre la copie manuelle (tapée). Cette protection gène considérablement un éventuel pilleur. A noter, avec le développement des scanners et les progrès réalisés par l'OCR (Optical Character Recognition), il convient d'interdire également l'impression en sus de la protection contre le copier/coller.

Ces protections (n°2, 3 et 4) sont réalisées grâce à des clés de cryptage non transmissibles (même à l'auteur) et gardées sous clé.

5.2. La conversion des thèses

5.2.1. Conditions à réunir pour le démarrage d'une conversion

Une conversion de thèse peut s'effectuer dès que l'opérateur dispose des éléments suivants :

L'autorisation de diffusion signée par l'auteur

Un exemplaire papier de la thèse

La forme électronique de la thèse

Le formulaire d'enregistrement de la thèse soutenue

Il doit également s'assurer que la thèse a bien été au moins cataloguée à Doc'INSA, c'est-à-dire qu'une notice lui correspond. Les données d'indexation (résumé, mots clés, code de classification Dewey) ne sont pas indispensables pour démarrer une conversion, mais il est préférable qu'elles soient présentes.

5.2.2. Transformation des documents

La conversion proprement dite peut démarrer. Elle est pilotée par le logiciel CEN[23], développé en interne par Marc-Etienne HUNEAU, en stage de fin d'études à l'INSA de Lyon, département Informatique.

Elle est entièrement documentée, d'une part dans le "Guide de conversion" qui assiste l'opérateur tout au long du processus, et d'autre part par la documentation opérateur :

Les "Règles d'édition des documents électroniques" à <URL:http://csidoc.insa-lyon.fr/these/doc/regles_edition.pdf>

La "Chaîne d'édition numérique - Référence utilisateur" à <URL:http://csidoc.insa-lyon.fr/these/doc/cen_ref_utilisateur.pdf>

Le "Manuel utilisateur" à <URL:http://csidoc.insa-lyon.fr/these/doc/manuel_utilisateur.pdf>

5.3. L'archivage des documents

5.3.1. Principes d'archivage

Pour une discussion détaillée des défis de l'archivage électronique, vous pouvez vous reporter à la note de synthèse bibliographique "Le rôle des bibliothèques dans l'archivage des périodiques électroniques scientifiques" [MERM-97].

Les documents électroniques sont, par nature, beaucoup plus fragiles et difficiles à conserver que les documents sous forme papier ou microformes. Depuis le début de "l'ère informatique", de grandes quantités d'information numérique ont été perdues et la réflexion sur ce que doit être un archivage électronique a beaucoup progressé.

Le principe de l'archivage électronique adopté pour le projet CITHER est de permettre la représentation[24], dans un futur arbitrairement lointain, de l'information conservée. Cela implique, nous le verrons, de ne pas se contenter de la conservation de la forme publiée,

---

[23] **CEN** : **C**haîne d'**E**dition **N**umérique

[24] La représentation est entendue ici comme étant la capacité à reproduire l'information signifiante dans un contexte technologique quelconque.

mais de préserver également d'autres formes, dont les fichiers de traitement de texte initiaux.

5.3.2. La sauvegarde des données

Il s'agit ici de la protection contre l'effacement accidentel des données présentées sur le site web de CITHER. Cet "archivage" est réalisé quotidiennement par l'administrateur de la machine csidoc.insa-lyon.fr, sur bandes DAT. Curieusement, peu d'acteurs sont sensibilisés au formidable défi de l'archivage électronique et c'est souvent à cette seule sauvegarde qu'ils se réfèrent en matière de préservation de l'information électronique. Celle-ci est évidemment très importante, mais notoirement insuffisante, pour répondre au cahier des charges que nous nous sommes fixé.

5.3.3. Intégrité des données

Avant même de savoir comment préserver l'information, il est nécessaire de savoir ce qu'il convient de préserver, et qui constitue l'intégrité de l'information. Les caractéristiques qui la déterminent sont le contenu, la fixité, le référencement, la provenance et le contexte.

Le contenu : il ne s'agit pas seulement de l'information textuelle brute. Les thèses scientifiques sont constituées de graphiques, schémas, tableaux complexes, etc. Le plan et la mise en page eux-mêmes révèlent l'intention de l'auteur, ils sont porteurs de sens. Dans l'état actuel des technologies, seuls SGML[25] et XML[26] sont capables de coder ce contenu indépendamment de formats propriétaires. Malheureusement, SGML est un modèle très lourd à implémenter, qui n'est utilisé que dans de grosses administrations et entreprises (notamment le monde de l'édition). Quand à XML, il est encore trop tôt pour l'utiliser. Ceci dit, au rythme des évolutions du monde de l'information électronique, il s'avérera vite être un candidat extrêmement intéressant pour l'archivage dans le futur (voir le paragraphe 4.3.1 pour plus de détails sur XML). Il faut donc pour l'instant conserver les formats propriétaires (la plupart du temps Word de Microsoft).

---

[25] **S**tandard **G**eneralized **M**arkup **L**anguage. Norme ISO 8879

[26] E**X**tensible **M**arkup **L**anguage. Consultez <URL:http://www.w3c.org/XML/> pour plus de détails

La fixité : Identifier et préserver un objet numérique comme un tout intangible est un autre défi. Dans l'environnement numérique, le concept de version 'canonique' a tendance à sérieusement s'estomper dans l'esprit des auteurs et des lecteurs, et il devient possible de présenter plusieurs versions de ce qui sera considéré comme un même document… On pourrait penser que, dans le cas des thèses, la problématique ne se pose pas : la version autorisée à être diffusée par le Président du Jury est la seule version canonique.
Cependant, dans nombre de cas, l'auteur doit effectuer des corrections à son document avant qu'il ne soit accepté. Dans la pratique, s'il est difficile à un centre de documentation comme Doc'INSA de récupérer cette version canonique papier, il est encore plus délicat de récupérer la forme électronique qui correspond effectivement à cette version canonique …
Les documents publiés par CITHER ne peuvent évidemment plus être modifiés postérieurement à la version canonique.

Le référencement : les documents électroniques archivés doivent pouvoir être référencés d'une manière non équivoque, et qui perdure. Les citations de documents électroniques doivent permettre d'identifier sans ambiguïté les travaux mentionnés. L'idéal serait que les œuvres électroniques offrent elles-mêmes les données de références bibliographiques qui permettraient de les identifier. Identifier précisément, sans ambiguïté et sans limite de durée, la localisation d'un document électronique, telles sont les fonctions de l'Uniform Resource Name (URN). A la différence de l'Uniform Resource Locator (URL) d'un document, qui peut évoluer au gré des changements de serveurs, etc., l'URN procure un nom unique et permanent à un document électronique, indépendamment de sa localisation réelle. OCLC a, dans le même ordre d'idée, développé les Persistent URLs (PURLs) pour cataloguer des ressources Internet.
Dans le cas des thèses du projet CITHER, quelques règles simples de choix des URL ont été édictées pour répondre à ces exigences, mais la question de la disponibilité à long terme (plus de cinq ans) des documents n'est pas encore fixée. Nous n'utilisons pas actuellement d'URN ni de PURL.

La provenance : L'intégrité d'un document électronique est également caractérisée par sa traçabilité. Il est fondamental de savoir d'où l'information provient initialement, et par quelles autorités elle a été validée. Dans le cas des thèses, il est relativement aisé de satisfaire à ce critère.

Le contexte : la façon dont le document interagit avec un environnement numérique plus large. Il s'agit de la dépendance envers les logiciels et le matériel qui ont permis son élaboration. L'utilisation d'XML nous affranchira d'une telle dépendance. Pour lors, l'utilisation de formats de documents très répandus (essentiellement le format Word et le format PDF) limitent les risques de perte de données lors des évolutions technologiques.

5.3.4. Formats de données

Nous conservons :

Le format **original** : reçu de l'auteur, avant toute modification. Les résultats d'une enquête menée à l'INSA de novembre 1996 à novembre 1997, auprès de 124 doctorants font état des données suivants [HUNE-98] :

- 67 % en format Word pour PC
- 26 % en format Word pour Mac
- 7 % en format Tex

Le format **source** : c'est le format original enregistré en Word 97 pour PC, modifié pour prendre en compte les paramètres d'impression de PDFMaker[27]. Les fichiers sont revus et souvent légèrement altérés dans leur mise en forme pour homogénéiser la hiérarchie des titres, la place des figures, etc.

Les fichiers numérisés en haute définition par nos soins et fournis comme éléments extérieurs à la thèse par l'auteur : photographies, plans, diapositives, etc.

Les fichiers PDF issus de la chaîne d'édition numérique, **avant leur protection contre la modification, l'impression et le copier/coller** le cas échéant.

Le fichier *.epr* généré lors de la conversion des fichiers par l'application CEN. Consultez "Serveur de thèses en texte intégral - Manuel technique" [HUNE2-98] pour plus d'information sur ce fichier.

---

[27] Macro Word, développée par Adobe, et permettant de créer un document PDF enrichi de la structure hiérarchique des titres présente dans le document Word.

5.3.5. Médium de conservation

Nous avons décidé d'utiliser les cédéroms gravés, pour plusieurs raisons :

Faible coût de stockage au Mo

Simplicité des opérations de stockage

Simplicité des opérations de récupération des informations, à l'aide d'un simple lecteur de cédéroms.

Possibilité de répartir des copies en plusieurs lieux, de manière à assurer une bonne sécurité contre les dommages matériels (voir procédure ci-dessous)

Le stockage n'est pas garanti plus de quinze ans avec cette technologie. Nous prévoyons donc une récupération des données et un nouveau stockage 10 ans après le gravage. La technologie qui sera alors utilisée n'est sans doute pas encore connue.

5.3.6. Procédure d'archivage

Le gravage dans le processus de conversion

Dans le processus de traitement du document pour en produire une version publiable, la procédure d'archivage démarre juste avant la protection par mot de passe des fichiers PDF.

La structure des répertoires archivés

Le logiciel CEN développé par Marc-Etienne Huneau permet de créer, à l'issue de l'optimisation des fichiers, un ensemble de répertoires dont la structure est décrite ci-dessous :

C:\Theses\Archi\auteur_thèse\

Originaux\	*Répertoire contenant les fichiers non modifiés, dans leur format original*
Sources\	*Répertoire contenant les fichiers convertis en word 97 pour PC (exception faite des formats Tex et Postcript) et prêts à être transformés en PDF*
Pdf\	*Répertoire contenant les fichiers prêts à être publiés, non protégés, et le pont d'embarquement*
Scans\	*Répertoire contenant les éventuels fichiers numérisés par nos soins, liés par liens hypertextes aux fichiers pdf*
auteur.epr	*Fichier contenant les données de conversion*[28]

Le choix du nombre de thèses par disque et du nombre de copies à graver

Afin d'assurer un archivage sûr et économique, nous devions prendre en compte les exigences suivantes :

Effectuer plusieurs copies de la même information de manière à diminuer le risque de perte de données par dégradation ou destruction du support.

Répartir les copies en plusieurs lieux pour diminuer le risque de perte de données par endommagement physique (destruction par le feu, …)

---

[28] Il existe un autre fichier contenant des données de conversion, il s'agit du **rapport de conversion.** Celui-ci **n'est pas** rendu public car il contient des données de nature confidentielle, entre autres les clés de désactivation des protections contre la modification, l'impression et le copier/coller. Ce rapport est imprimé à la fin de la conversion et gardé en lieu sûr.

Formaliser la création et la conservation des documents

Contrôler la diffusion des données, soumises à la législation du droit d'auteur.

Minimiser le nombre de cédéroms gravés et le temps de gravage

En fonction de ces contraintes, nous avons décidé d'archiver au maximum 10 dossiers de thèses[29] par cédérom gravé, et de graver chaque cédérom en 3 copies. Chaque copie sera conservée en un lieu différent. Les copies non conservées dans le bureau de conversion seront scellées.

La procédure d'archivage sur cédérom est décrite en annexe 10.2

Cette procédure d'archivage est reprise intégralement dans le guide de conversion.

Cas d'une correction à effectuer a posteriori

Dans certains cas, il sera nécessaire d'effectuer des corrections sur les fichiers PDF publiés[30] postérieurement à un premier archivage. Il convient alors d'effectuer un nouvel archivage du dossier de thèse, en suivant la procédure indiquée en annexe 10.3.

[29] A concurrence de 650 Mo, la capacité de stockage d'un cédérom. Le cas échéant, le cédérom contiendra un nombre plus réduit de dossiers.

[30] Ceci ne s'applique en aucun cas aux modifications effectuées sur les **ponts d'embarquement**

6. Le FrontOffice

Le FrontOffice correspond à toute la partie visible du projet, pour le lecteur et pour l'auteur. Il correspond principalement au site CITHER, situé à <URL:http://csidoc.insa-lyon.fr/these>.

6.1. Les différents aspects de la conception d'un site web

Utilisateurs intensifs d'Internet, nous avons eu l'occasion de visiter un grand nombre de sites web et de constater que les points essentiels à travailler lors du développement d'un site sont la navigation, la structuration de l'information et la conception graphique.

De trop nombreux sites passent, à notre avis, à coté d'un de ces aspects et y perdent en efficacité et en qualité de services rendus. Nous avons donc tenté, de répondre au mieux à ces critères.

6.1.1. La navigation

C'est la fonctionnalité qui permet au visiteur :

De se faire une représentation mentale de la structure du site.

De se déplacer efficacement et rapidement de page en page.

De se repérer à l'intérieur d'un site.

On observe plusieurs grandes catégories de navigation sur le web :

1. La navigation en étoile : une page centrale donne accès à la plupart des autres.
2. La navigation grâce à une table des matières constamment affichée (classiquement, une colonne à gauche).
3. La navigation séquentielle (page à page) : la page *n* donne accès à la page *n+1*, qui donne accès à la page *n+2*, etc.
4. La navigation hiérarchique : un chemin, constamment affiché, indique la place de la page dans l'arborescence.

Compte tenu de la nature des informations que nous proposons, c'est la solution n°4 qui s'est imposée par sa souplesse (l'ajout et le retrait de documents est très simple à

réaliser). Elément non négligeable, c'est celle qui offre l'aspect le plus "scientifique". C'est une solution éprouvée et utilisée dans l'un des sites les plus visités au monde : Yahoo.

Ceci ne signifie pas qu'il faille n'offrir qu'un unique accès aux données, mais simplement que la navigation "statique" (le réseau des hyperliens des pages HTML) présente cette structure.

6.1.2. Conception graphique

Reprenant les exigences formulées au paragraphe 2.1.4, la conception graphique participe grandement au confort de visite. Elle doit être soignée, discrète, cohérente et guider le visiteur vers l'information importante.

Repérage immédiat

Le visiteur doit savoir immédiatement où il se trouve. Des éléments évidents d'identification doivent lui indiquer qu'il est bien à l'INSA de Lyon, sur le site de la bibliothèque, et où se situe la page dans l'arborescence du site.

Unité graphique

Le site doit présenter la même unité graphique de bout en bout. Il s'agit de respecter la charte définie par notre Institut mais de savoir s'en distinguer suffisamment pour que le lecteur repère sa sortie de la zone contrôlée par Doc'INSA.

De nouvelles possibilités sont offertes par l'évolution des standards du web, comme les feuilles de style, qui simplifient la maintenance et améliorent la lisibilité des pages en mode dégradé[31].

6.1.3. Conception structurelle de l'information

L'information présentée par un site web professionnel doit être structurée de manière rigoureuse de façon à pouvoir identifier précisément les mentions de responsabilités, à la fois du travail présenté et de l'organisme diffusant ce travail.

[31] Affichage des pages web par un butineur d'une ancienne génération.

Les pages web doivent en outre inclure des données de repérage, de description et d'indexation formelle des ressources, pour l'indexation dans les moteurs de recherche globaux et locaux.

6.1.4. La notion de "pont d'embarquement"

Pour le projet CITHER, nous avons conçu un type de document particulier, chargé d'être un point d'accès normalisé à la thèse et contenant un certain nombre d'informations structurées sur celle-ci. Le document inclut en outre des informations sur la diffusion de cette thèse.

Renforçant la métaphore "marine" du site (choix des couleurs, Ile de Cythère, …) nous avons baptisé ce document "pont d'embarquement".

A noter, le pont d'embarquement est généré automatiquement lors de la conversion de la thèse à partir d'un modèle préétabli et des données fournies par l'opérateur ou récupérées dans les formes électroniques des fichiers de thèse. Il est légèrement modifié avant d'être diffusé. La structure du modèle, les informations à inclure et les modifications à effectuer sont détaillées dans les documents d'assistance à l'opérateur (voir paragraphe 5.2.2).

Point d'accès normalisé à la thèse

Listes et catalogue pointent directement vers les ponts d'embarquement. C'est l'accès "officiel" à la thèse. Nommée index.html, le pont d'embarquement "cache" le contenu du répertoire dans lequel il se trouve. On y accède directement en tapant :

`http://csidoc.insa-lyon.fr/these/`**`année`**`/`**`auteur`**

L'indication **`année`** est l'année sur quatre chiffres, pour éviter l'effet "an 2000". L'indication **`auteur`** est le nom d'auteur ramené à une forme en un seul mot sans accents (Le logiciel CEN en propose une version automatiquement. En dernier ressort, la construction de ce mot est confiée à l'opérateur).

Description normalisée de la thèse et de sa diffusion

Description visible

La thèse est décrite par sa référence bibliographique électronique, en respectant le projet de norme ISO/DIS 690-2 -1995. En voici un exemple :

CARBONNEAU, Xavier *Etude des propriétés thermomécaniques de mullite zircone et de zircon* [On-line] Thèse : INSTITUT NATIONAL DES SCIENCES APPLIQUEES DE LYON, 1997 [27.04.1998], 156 p.

Available from internet : <URL:http://csidoc.insa-lyon.fr/these/1997/carbonneau/index.html>

Description invisible[32]

Les ponts d'embarquement incluent également des informations sous forme de métadonnées, invisibles à la simple lecture des pages. La structure des métadonnées utilisées est détaillée à <URL:http://csidoc.insa-lyon.fr/these/doc/meta_dc.html>. Elle s'inspire très étroitement des recommandations du Dublin Core[33]. Un exemple d'applications est donné en annexe 10.5

Possibilités d'extensions

Le pont d'embarquement peut être enrichi de liens vers la "home page" du docteur, du laboratoire, vers son adresse de courrier électronique, vers des articles connexes, etc …

Dans un premier temps, nous comptons ajouter une description succincte du laboratoire d'accueil accompagnée de ses coordonnées.

---

[32] Il serait plus exact de parler d'une description "non directement visible". Tout visiteur peut consulter à sa guise le code HTML des pages de CITHER, et ainsi lire ces données invisibles.

[33] Le Dublin Core Metadata Workshop est une conférence ouverte, initiée en 1995, et qui rassemble des bibliothécaires, des chercheurs dans le domaine des bibliothèques virtuelles, des experts dans le domaine des balises textuelles, des créateurs de contenus … Le but de cette conférence est de promouvoir des normes facilitant la recherche d'information sous forme numérique. Leur proposition vient d'être reprise dans une Request For Comment (RFC) d'Internet (<URL:ftp://ftp.isi.edu/in-notes/rfc2413.txt>).

6.1.5. Résultats

Analyse d'une page de CITHER

1.1.

6.2. Les accès

6.2.1. Prévus par le projet

Comme indiqué plus haut (paragraphe 4.3.4), nous proposons un accès par listes alphabétiques, un accès par catalogue informatisé et un accès par recherche dans le texte intégral des documents diffusés.

6.2.2. Outils externes au projet

Nous comptons prochainement déclarer le site auprès de nombreux outils d'indexation du web, de manière à assurer une visibilité maximale des thèses. Les métadonnées des ponts d'embarquement joueront ici pleinement leur rôle car les versions Internet des grands moteurs d'indexation ne parcourent pas les fichiers PDF, ou bien, quand ils le font, n'en indexent qu'une partie. Les données présentes dans le catalogue informatisé ne sont pas accessibles non plus aux moteurs de recherche. Les ponts d'embarquement deviennent donc l'unique point de repérage des thèses sur l'Internet "global".

6.3. La documentation

Notre politique est de placer et de maintenir toute notre documentation en libre accès sur le site de CITHER, soit en HTML, soit en PDF, de manière à transmettre notre expérience à d'autres équipes[34]. Celle-ci comprend :

La description complète du projet, soit par le présent document, soit, pour la conversion, les documents de Marc-Etienne Huneau.

Le logiciel CEN pour Windows 9x, développé lors de l'étape de conception informatique et toute sa documentation seront disponibles prochainement, dès la fin de notre période de test.

Des documents à destination des doctorants et notamment :

- Les "Consignes de présentation d'une thèse" : remarques générales pour la rédaction des documents

[34] Tout est accessible depuis la page <URL:http://csidoc.insa-lyon.fr/these/doc/index.html>

- Les "Conseils techniques pour le dépôt de thèses électroniques" : tout ce qu'il faut savoir pour transmettre efficacement les documents électroniques à Doc'INSA
- Le document "Transmettre ses fichiers de thèse à Doc'INSA par FTP" : toutes les étapes du transfert sont détaillées.
- Prochainement, le "Modèle d'autorisation de diffusion électronique" (formats RTF et HTML) : document à signer et à remettre à Doc'INSA

6.4. La communication

6.4.1. Les actions

Indispensable aspect de tout projet, la communication autour de CITHER se déclinera, pour notre Institut, sous la forme :

D'entretiens personnalisés avec les doctorants, lors de leur visite à Doc'INSA.

De manifestations à organiser, en direction de la communauté scientifique de l'INSA.

D'un numéro spécial du journal interne de l'INSA consacré aux services offerts par Doc'INSA

De "publicités" sur notre catalogue informatisé

Pour l'extérieur, sous la forme :

De participations à des forums, congrès, …

De rédactions d'articles dans des revues spécialisées

De déclaration du site auprès d'un maximum d'acteurs de l'Internet.

… pour l'extérieur de l'INSA.

Les possibilités sont nombreuses et nous tâcherons de saisir toutes les opportunités de mieux nous faire connaître !

6.4.2. Mesure des résultats

Afin de mieux mesurer les effets de nos actions de communications, nous avons installé un logiciel analyseur de logs[35] sur notre serveur web Apache. Configuré correctement, il nous indiquera quelles sont les parties du site (et donc quelles thèses) sont les plus visitées, par quels types de visiteurs, etc.

Voici un extrait de l'analyse des pages les plus visitées du site, générée le 27 juillet 1998, pour se rendre compte que même sans publicité particulière (autre que l'information faite aux docteurs et à quelques professionnels des sciences de l'information), le site commence déjà à être visité !

Nb : Nombre de requêtes enregistrées par le serveur Web. Il faut noter que de nombreuses demandes sont ignorées par le serveur web, du fait de l'existence de serveurs "cache" locaux.

Dernière date : Dernière demande de ce fichier

Fichier : Arborescence depuis l'adresse http://csidoc.insa-lyon.fr

Nb	Dernière date	Fichier	Commentaires
360	27 Juil 98 10:32	/these/	Page d'accueil de CITHER
…			
149	27 Juil 98 10:32	/these/thaccueil.html	Page d'accueil de CITHER
…			
112	24 Juil 98 10:17	/these/recherche_haut.html	Recherche dans le catalogue
111	22 Juil 98 13:42	/these/1998/cherouali/	Pont d'embarquement
…			
108	27 Juil 98 08:33	/these/pe.html	Liste des ponts d'embarquement
106	26 Juil 98 03:49	/these/recherche.html	Recherche dans le catalogue
…			
98	24 Juil 98 14:04	/these/1997/	Liste des ponts d'embarquement
…			

[35] Les logs sont des fichiers créées par le serveur web. Ils répertorient toutes les demandes de pages.

90	17 Juil 98 09:16	/these/1997/favre/these.pdf	Thèse
…			
86	25 Juil 98 10:26	/these/recherche_ti.html	Recherche en texte intégral
85	27 Juil 98 09:55	/these/doc/	Documentation de CITHER
…			
78	18 Juil 98 12:36	/these/1998/cherouali/chap5.pdf	Un chapitre d'une thèse
…			
71	16 Juil 98 23:10	/these/1997/beretta/12chi_2.pdf	Un chapitre d'une thèse
68	27 Juil 98 08:33	/these/1998/	Liste des ponts d'embarquement
…			
63	13 Juil 98 17:24	/these/1997/girard/these.pdf	Thèse
…			
60	24 Juil 98 18:44	/these/doc/rapport_pfe.pdf	Documentation de CITHER
…			
53	16 Juil 98 13:15	/these/1997/beretta/29titann.pdf	Un chapitre d'une thèse
53	23 Juil 98 11:43	/these/1997/beretta/	Pont d'embarquement
52	17 Juin 98 11:50	/these/1997/debray/These.pdf	Thèse
…			
47	21 Juil 98 12:05	/these/1996/	Liste des ponts d'embarquement

7. Aspects financiers

7.1. Les financements

Ce projet n'aurait pas pu se dérouler sans un financement public important. Nous l'avons obtenu de plusieurs organismes.

Financeurs	Financements (en kF)
Ministère La Sous-direction des bibliothèques et de la documentation du Ministère de l'Education Nationale, de la Recherche et de la Technologie, a doté le projet de 300 kF pour l'achat de matériel.	***300***
Région Rhône-Alpes	***280***
INSA Le service Doc'INSA de l'INSA de Lyon a supporté de nombreux coûts salariaux et frais induits par l'hébergement du projet au sein du service.	***293***
Total des financements	***873***

7.2. Coûts matériels

Matériels	Coûts (en kF)
Csidoc Pour héberger le serveur de thèses, il a fallu acquérir une nouvelle machine Unix	***180***
Poste de traitement Le poste de conversion des thèses, dédié à cette tâche, plus l'achat d'un scanner à plat.	***30***
Total des coûts matériels	***210***

7.3. Coûts logiciels

Logiciels	Coûts (en kF)
Doris-web Le module Doris web permet de réaliser l'interfaçage web de l'application Doris-Loris qui gère notre bibliothèque. Il permet d'accéder aux ponts d'embarquement des thèses à partir d'hyperliens inclus dans les notices Doris.	***92***
Maintenance Oracle Le module Doris-Web nécessitait une mise à jour du SGBD Oracle.	***30***
Acrobat La suite de logiciels Acrobat permet de fabriquer des fichiers PDF. Licence monoposte.	***1***
Word Une partie du traitement de conversion s'effectue à partir de Microsoft Word 97. Licence monoposte.	***1***
Divers Logiciels et outils complémentaires (éditeur HTML, logiciel lecteur de disquettes Mac, …).	***3***
Total des coûts logiciels	***127***

7.4. Coûts salariaux

Même si les coûts salariaux sont quelquefois difficiles à chiffrer, surtout dans la fonction publique, il est intéressant et plein d'enseignements d'en faire une estimation la plus réaliste possible.

On distingue les coûts de mise en place initiale des coûts de fonctionnement prévisibles, au moins pour la première année.

Les coûts de mise en place incluront un an de fonctionnement.

Coûts salariaux de mise en place initiale	Coûts (en kF)
Un ingénieur stagiaire - 4 mois Il s'agit d'un stagiaire en informatique à l'INSA, qui a développé le logiciel de conversion des thèses.	***6***
Un ingénieur documentation - 6 mois Ingénieur d'études en poste à Doc'INSA chargé de piloter le projet, de réaliser le FrontOffice et de suivre la montée en charge du projet.	***128***
Un technicien "opérateur de conversion" - 2 mois Chargé de tester le logiciel de conversion des thèses.	***36***
Deux ingénieurs système pendant 1 mois Maintiennent le serveur web et le fonctionnement de csidoc	***40***
Réunions concernant le projet En complément des personnes dont les coûts salariaux ont été évoqués ci-dessus, quelques intervenants ont participé à des réunions de CITHER : un Professeur d'université, un ingénieur de recherche et un assistant-ingénieur. Nous avons répertorié 15 réunions pour un temps estimé de 30 heures soit environ 4,5 jours. Afin de calculer les dépenses entraînées par ces réunions, nous avons schématiquement appliqué un coût unique, soit 4 kF par jour[36] pour toutes ces personnes. Soit 3 personnes pendant 4,5 jours à 4 kF par jour.	***54***
Experts extérieurs Un avocat spécialiste des questions de propriété intellectuelle, chargé de rédiger le contrat d'autorisation de diffusion des thèses.	***20***

[36] Correspond aux tarifs d'INSAVALOR.

Un an de fonctionnement (voir ci-dessous)	***172***
Total des coûts de mise en place initiale	***456***

Coûts salariaux de fonctionnement, par an	Coûts (en kF)
Un technicien mi-temps Pour tous les aspects de la conversion des thèses, nous évaluons raisonnablement d'employer un technicien à mi-temps.	***89***
Un ingénieur d'étude quart-temps Pour tous les aspects de maintenance et d'amélioration des services offerts actuellement (développements informatiques, documentation, …).	***53***
Ingénieur système 1/8 temps	***30***
Total des coûts salariaux de fonctionnement, par an	***172***

7.5. Coûts divers

Coûts divers de mise en place	Coûts (en kF)
Déplacements	***10***
Frais de formation des élèves de troisième cycle	***45***
Protection de marques	***5***
Frais généraux	***20***
Total des coûts divers de mise en place	***80***

8. Bilans et perspectives

Le bilan que l'on peut tirer d'un tel projet est contrasté. Afin d'en rendre compte, nous avons choisi de proposer une série d'analyses sur des points précis du projet.

8.1. Bilan des choix effectués

L'un des choix de départ était de partir des formes numériques des documents. Cela nous semblait être une démarche logique et moderne. Or, les difficultés rencontrées furent nombreuses :

Hétérogénéité des formats de fichiers reçus, incompatibilités entre différentes versions d'un même logiciel, etc. nous conduisent à de trop nombreuses interventions "manuelles" qui interdisent une automatisation poussée du processus de traitement avant diffusion.

Difficulté de vérifier la canonicité des documents : comment vérifier que les fichiers fournis recèlent bien l'exact contenu de la version papier diffusée ?

Fragilité du système : les logiciels utilisés sont complexes et leur comportement n'est pas toujours aussi prédictible qu'annoncé ! Or nous combinons l'utilisation de plusieurs de ces logiciels …

Par contre, le choix de départ permet ce qu'interdisent les projets de numérisation sans OCR[37] : la recherche en texte intégral.

Un autre choix de départ était l'utilisation du format PDF pour la diffusion. Ce choix n'est pas, pour l'heure à remettre en question. Il satisfait au cahier des charges proposé, même si nous avons encore des difficultés à diffuser les documents dans le mode "byte serving" (page à page).

8.2. Bilan technique

Compte tenu des mentions faites au paragraphe précédent, le bilan technique est globalement positif. Nous avons pu résoudre la plupart des problèmes signalés au

[37] OCR : Optical Character Recognition : reconnaissance optique de caractères, qui permet, à partir de l'image d'une feuille imprimée, d'obtenir la suite des caractères qui composent le texte.

début de l'étude, par des moyens somme toute classiques. Nous avons été grandement aidés par l'arrivée à une certaine maturité de certaines technologies du web. Citons pêle-mêle les métadonnées, les feuilles de style, les moteurs de recherche pour intranet, les interfaces web de SGBD[38], le Javascript[39], etc.

8.3. Bilan juridique

Nous avons finalement été freinés par des aspects juridiques que nous étions mal préparés à affronter. Notre expertise dans ce domaine s'est largement développée, mais nous aurons encore besoin des conseils de spécialistes à l'avenir.

8.4. Transfert de technologies

En développant ce projet, nous avons eu le constant souci de bien en documenter chacun des aspects, de manière à ce que notre expérience dans la conception d'une bibliothèque virtuelle de thèses puisse être utilisée par d'autres. Il est bien sûr trop tôt pour mesurer l'impact et la réutilisation de nos informations par d'autres équipes, mais nous avons concrètement l'espoir qu'elles servent au mieux la communauté.

8.5. Aspects ressources humaines

A l'occasion de ce projet, plusieurs personnes ont été amenées à travailler ensemble et à développer des relations professionnelles à l'extérieur de leur cadre habituel de travail. Certains ont eu l'occasion, pour la première fois, d'encadrer du personnel.

L'expérience a été riche d'enseignements et il en a résulté une plus grande ouverture d'esprit et une meilleure cohésion du personnel impliqué dans ce projet à Doc'INSA.

8.6. Aider les doctorants à fournir de meilleurs documents

8.6.1. Formations à mettre en place

Nous ressentons la nécessité de proposer, le plus tôt possible, des formations à la rédaction de longs documents dans une perspective de diffusion électronique. Elle

---

[38] SGBD : **S**ystème de **G**estion de **B**ases de **D**onnées

[39] Langage de programmation des pages Web

s'adressera aux doctorants en priorité, dès le début de leur travail de recherche. Un plan de formation sera élaboré dès la rentrée universitaire.

8.6.2. Doc'INSA, partenaire du travail de recherche

En développant de tels projets en effet, Doc'INSA se positionne comme un partenaire important du chercheur de l'INSA. Cette position peut véritablement nous aider à atteindre un public parfois peu motivé aujourd'hui pour utiliser nos compétences. C'est aussi l'occasion de nous remettre en question, de mieux comprendre les besoins de ce public et de proposer des services mieux adaptés.

8.7. Intégration des documents dans le projet Webdoc

L'ABES (Agence Bibliographique de l'Enseignement Supérieur), à la demande du Ministère de l'Education Nationale, de la Recherche et de la Technologie met actuellement en œuvre, de manière anticipée, le module Webdoc et le catalogue Webcat de Pica[40] pour la publication de documents électroniques.

Les thèses de CITHER seront prochainement diffusées dans le cadre de ce projet, si les problèmes d'ordre juridique afférents à cette seconde diffusion sont correctement résolus.

8.8. Pour en savoir plus

N'hésitez pas à consulter le site CITHER à <URL:http://csidoc.insa-lyon.fr/these> et particulièrement l'espace documentation à <URL:http://csidoc.insa-lyon.fr/these/doc/index.html> pour plus d'informations sur le projet.

---

[40] Plus d'information sur Pica, Webdoc et Webcat à l'adresse <URL:http://www.pica.nl/frames/index.html>

9. Références bibliographiques

Les références bibliographiques sont rédigées selon le guide proposé par Doc'INSA : **BURLAT, J.M., PRUDHOMME, B.** *REFERENCES BIBLIOGRAPHIQUES - Rédaction et lecture.* [On-line]. Villeurbanne (Fr) : Inst. Nat. Sci. Appl., Doc'INSA, Sep. 1997 [Visité le 8 septembre 1998] Available from internet : <URL:http://www.insa-lyon.fr/Insa/Departements/DocInsa/refbibli.html>

[ABOU-98] *About PDF.* [On-Line]. Mars 1998. [Visité le 21 juillet 1998] Available from Internet : <URL:http://www.adobe.com/prodindex/acrobat/adobepdf.html>

[BEAU-98] **BEAUDRY, G.** *Projet de publication et de diffusion électroniques des thèses - Proposition d'une collaboration entre les universités québécoises.* [On-line]. Montréal (CA) : Les Presses de l'Université de Montréal. Jan. 1998. [Visité le 20 juillet 1998] Available from Internet : <URL:http://www.pum.umontreal.ca/publ_electr/rapports/bea02/projet_theses.html>

[BIEN-96] *Bienvenue sur le serveur CALLIMAQUE*. [On-line]. Oct. 1996. [Visité le 20 juillet 1998] Available from Internet : <URL:http://callimaque.grenet.fr/>

[COMI-97] *Comité de travail sur les thèse électroniques - page d'accueil.* [On-Line]. Nov. 1997. [Visité le 23 juillet 1998] Available from Internet : <URL:http://www.bibliothèque.ulaval.ca/doelec/theses/>

[COMI-97] **Comité interministériel de la recherche scientifique et technique.** *LA RECHERCHE : UNE AMBITION POUR LA FRANCE* [On-line]. Paris : Sans éditeur. Juin 1997. [Visité le 27 juillet 1998] Available from Internet : <URL:http://www.recherche.gouv.fr/gouv/cirst/ambi.htm>

[ELEC-98] *Electronic Thesis and Dissertation Initiative.* [On-line]. Fev. 1998 [Visité le 20 juillet 1998] Available from Internet : <URL:http://etd.vt.edu/>

[ETDP-98] *ETD Project - The Joint Electronic Thesis and Dissertation Project.* [On-Line]. Juil. 1998. [Visité le 20 juillet 1998] Available from Internet : <URL:http://www.fis.utoronto.ca/etd/>

[GENO-98] **GENOVESE, D.** L'Atelier national de reproduction des thèses de Grenoble. *Arabesque*, 1998, n°11, p. 11

[HUNE2-98] **HUNEAU, M-E.** *Serveur de thèses en texte intégral - Manuel technique* [On-Line]. Juin 1998. [Visité le 15 juillet 1998] Available from Internet : <URL:http://csidoc.insa-lyon.fr/these/doc/manuel_technique.pdf>

[HUNE-98] **HUNEAU, M-E.** *Rapport de Synthèse - Serveur de thèses en texte intégral* [On-Line]. Juin 1998. [Visité le 15 juillet 1998] Available from Internet : <URL:http://csidoc.insa-lyon.fr/these/doc/rapport_synthese.pdf>

[LARE-97] *LA RECHERCHE : UNE AMBITION POUR LA FRANCE - Les activités de R. & D. dans le monde* [On-line]. Paris : Sans éditeur. Juin 1997. [Visité le 27 juillet 1998] Available from Internet : <URL:http://www.recherche.gouv.fr/gouv/cirst/forfa.htm>

[LARE2-97] *LA RECHERCHE : UNE AMBITION POUR LA FRANCE - Forces et faiblesses de la recherche française.* [On-line]. Paris : Sans éditeur. Juin 1997. [Visité le 27 juillet 1998] Available from Internet : <URL:http://www.recherche.gouv.fr/gouv/jaune/monde.htm>

[LOSF-98] **LOSFELD, G.** Editorial. *Arabesque*, 1998, n°11, p. 2

[MERM-97] **MERMET, J-M.** *Le rôle des bibliothèques dans l'archivage des périodiques électroniques scientifiques* [On-Line]. Sep. 1997. [Visité le 15 juillet 1998] Available from Internet : <URL:http://www.insa-lyon.fr/Insa/Departements/DocInsa/jmm/rrbfinal.html>

[NETW-98] *Networked Digital Library of Theses and Dissertations.* [On-line]. Juin 1998 [Visité le 20 juillet 1998] Available from Internet : <URL:http://www.ndltd.org/>

[PROQ-98] *Proquest digital dissertations (pilot site)* [On line]. Juil. 1998 [Visité le 20 juillet 1998] Available from Internet : <URL:http://wwwinfo.umi.com/solutions/2.0.html]

[SENS-95] **SENS, J.C.** *Electronic Publishing in Science*. [On-line]. Geneva (Ch) : Sans éditeur. Mar. 1995. [Visité le 15 juillet 1998] Available from Internet : <URL:http://epswww.epfl.ch/ene/ene_apr96_sens_text.html>

[UWEL-97] UW *Electronic Theses and Dissertations Server.* [On-Line]. Oct. 1997. [Visité le 20 juillet 1998] Available from Internet : <URL:http://library.uwaterloo.ca/~uw-etpt/pilot.html>

[VAISS-97] **VAÏSSE, P., ROBERT, F.** *WebDOC : un partenariat documentaire au coeur du Web* [On-Line]. Nov. 1997. [Visité le 20 juillet 1998] Available from Internet : <URL:http://www.enssib.fr/eco-doc/com.vaisse.html>

10. Annexes

Table des matières des annexes

10.1. Glossaire

ADBS	L'association des professionnels de l'information et de la documentation
ANRT	Dans ce document, signifie : Atelier National de Reproduction des Thèses
ASCII	American Standard Code for Information Interchange
BackOffice	"Zone" du projet hors de la vue du lecteur.
CEN	Chaîne d'Edition Numérique

CISM	Centre d'Informations Scientifiques et Médicales.
CITHER	Consultation en texte Intégral des THèses En réseau
DED	Département des Etudes Doctorales
FrontOffice	"Zone" du projet visible par le lecteur.
GED	Gestion Electronique de Documents
HTML	HyperText Markup Language : langage de description des pages web
ICIMARQUES	Banque de données de l'INPI recensant toutes les marques en vigueur en France : les marques françaises communautaires et internationales désignant la France
IMAG	Institut d'Informatique et de Mathématiques Appliquées de Grenoble
INIST	Institut de l'Information Scientifique et Technique
INRIA	Institut National de Recherche en Informatique et en Automatique
INSA	Institut National des Sciences Appliquées
MARC	Machine-Readable Catalogue
OCLC	Online Computer Library Center
OPAC	Online Public Access Catalogue : interface d'accès au catalogue informatisé
PDF	Portable Document Format
PEB	Prêt Entre Bibliothèques
Pont d'embarquement	Fichier HTML. Point d'entrée normalisé vers la thèse
Postscript	Langage d'impression développé par la société Adobe

PURL	Persistent URL
RTF	Rich Text Format : format développé par Microsoft pour faciliter l'échange de documents entre traitements de texte.
SGML	Standard Generalized Markup Language
TIFF	Tagged Image File Format
UMI	University Microfilms International
URL	Uniform Resource Location
URN	Uniform Resource Name
Webmaster	Responsable d'un site web
XML	EXtended Markup Language

10.2. Procédure d'archivage sur cédérom

1. Déplacez les dix répertoires d'archives choisis dans le répertoire c:\Theses\gravage.

2. Ajoutez-y le fichier lisezmoi.txt dont le modèle est donné en annexe 10.

3. Choisissez une lettre de lecteur libre sur votre station de travail (exemple ici : z). En ligne de commande, tapez
   `subst z: c:\Theses\gravage`

4. Démarrez le logiciel de gravage, choisissez le répertoire z: et sélectionnez tout ce qu'il contient.

5. Gravez 3 cédéroms avec le même contenu, en vérifiant à chaque fois la bonne qualité du gravage (bonne lecture des répertoires et ouvertures de quelques fichiers pris au hasard).

6. Imprimez, 3 exemplaires de la jaquette personnalisée (modèle en annexe 10.6) en changeant à chaque fois, le numéro d'exemplaire (1 à 3) et le numéro d'inventaire.

7. Reprenez chacune des notices de Doris correspondant aux thèses archivées, et ajoutez-y, dans le champ GESTION>Etat>Type>Cote_CDROM la cote du cédérom d'archives : CD1.(numéro)

8. Effacez le contenu du répertoire c:\Theses\gravage.

9. Remplissez le carnet d'inventaire local pour répertorier ces trois copies.

10. L'exemplaire 1 sera scellé par bande adhésive et sera envoyé au Département des Etudes Doctorales (DED). Il ne devra pas être descellé. L'exemplaire 2 sera scellé par bande adhésive et conservé dans le magasin de Doc'INSA. Il ne devra pas être descellé. L'exemplaire 3 sera conservé dans le bureau de conversion des thèses.

10.3. Procédure de ré-archivage des données

1. Recopiez, dans c:\Thèses\Archi, le dossier d'archivage complet de la thèse en question, à partir du cédérom gravé antérieurement.

2. Modifiez puis enregistrez les fichiers PDF désirés. Seuls les fichiers PDF peuvent être modifiés. Il est interdit de toucher aux autres.

3. Ne faites plus d'opérations sur ces fichiers, ils seront à nouveau archivés lors de la prochaine opération de gravage.

4. Copiez-les dans un répertoire temporaire (par exemple le bureau). La suite de la procédure s'applique à ces nouveaux fichiers.

5. Protégez chacun d'entre eux suivant les désirs exprimés par l'auteur sur son autorisation de diffusion, et à l'aide du mot de passe défini pour la thèse.

6. Renvoyez les fichiers modifiés sur csidoc par le moyen habituel. Ils remplaceront les anciennes versions.

7. Effacez les fichiers sur lesquels vous venez de travailler. GARDEZ BIEN LE DOSSIER D'ARCHIVAGE ET SON CONTENU.

8. Facultatif : demandez à l'Altavista du campus de venir réindexer les documents modifiés en lui soumettant l'URL du pont d'embarquement de la thèse. Ceci n'est à faire que dans le cas d'une modification du fond.

9. Lors de la prochaine opération de gravage, incluez le dossier d'archivage de la thèse modifiée.

10. Mettez à jour la référence au cédérom d'archives dans la notice de la thèse, sous Doris-Loris.

10.4. Format du fichier lisezmoi.txt

Ce document est un modèle du fichier d'avertissement qui se trouve à la racine de tous les disques d'archives gravés pour le projet CITHER. En pratique, les informations à modifier sont indiquées en gras. Ce modèle de fichier est directement accessible depuis le guide de conversion.

```
                cédérom d'archivage des thèses du projet CITHER

Ce disque archive des formes électroniques de thèses de l'INSA de
Lyon, publiées dans le cadre du projet CITHER.

Important : les informations contenues sur ce disque sont propriété et
à usage interne de Doc'INSA, service de documentation de l'INSA de
Lyon et certaines d'entre elles sont confidentielles. Toute
reproduction non autorisée par écrit des données de ce disque est
rigoureusement interdite. De même, ce disque ne peut être prêté ou
loué. Pour tout renseignement, contactez :

Responsable projet CITHER
Doc'INSA, Bât.220
20, Av Albert Einstein
69621 Villeurbanne

Tél : +4 72 43 85 64
Mél : cither@insa-lyon.fr
Fax : +4 72 43 85 02

================================================================
Cote Doc'INSA : CD1.(remplacer par le numéro du cédérom)
Date : date au format jj-mm-aaaa

================================================================
Références bibliographiques électroniques des thèses présentes sur ce
disque :

----------------------------------------------------------------
Référence bibliographique électronique de la thèse 1 (à reprendre
depuis le pont d'embarquement)
----------------------------------------------------------------
Référence bibliographique électronique de la thèse 2 (à reprendre
depuis le pont d'embarquement)
----------------------------------------------------------------
...
----------------------------------------------------------------
Référence bibliographique électronique de la thèse 10 (à reprendre
depuis le pont d'embarquement)
----------------------------------------------------------------
```

10.5. Métadonnées de la thèse de M. Carbonneau

```
<!-- META information -->
<!-- modèle modifié le mardi 16 juin 1998 08:57:01 par Jean-Michel Mermet -->
<!-- Classique -->
<META NAME="keywords" CONTENT="(LANG=fr) CERAMIQUE, MULLITE ZIRCONE, FLUAGE,
FISSURATION, PROPAGATION FISSURE, PROPRIETE MECANIQUE, PROPRIETE THERMOMECANIQUE,
FLEXION, TORSION, SPECTROMETRIE MECANIQUE, MICROSCOPIE ELECTRONIQUE, PHASE VITREUSE"
>
<META NAME="description" CONTENT="(LANG=fr) Le travail porte sur la caracterisation
du comportement mecanique de mullite zircone et de zircon a haute temperature dans le
but de mieux comprendre le comportement a la rupture et de montrer l'existence d'un
seuil dans le mode de propagation des fissures." >


<!-- Dublin Core -->
        <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
        <meta name="DC.creator.personalname" content="(LANF=fr) CARBONNEAU, Xavier">
        <meta name="DC.creator.email" content="">
        <meta name="DC.creator.origin" content="GEMPPM">
        <meta name="DC.contributor" content="FANTOZZI, G.">
        <meta name="DC.title.main" content="(LANG=fr) Etude des propriétés
thermomécaniques de mullite zircone et de zircon">
        <meta name="DC.title" content="(LANG=fr) ">
        <meta name="DC.title.main" content="(LANG=en) ">
        <meta name="DC.title" content="(LANG=en) ">
        <meta name="DC.type" content="thesis">
        <meta name="DC.subject.keywords" content="(LANG=fr) CERAMIQUE, MULLITE
ZIRCONE, FLUAGE, FISSURATION, PROPAGATION FISSURE, PROPRIETE MECANIQUE, PROPRIETE
THERMOMECANIQUE, FLEXION, TORSION, SPECTROMETRIE MECANIQUE, MICROSCOPIE ELECTRONIQUE,
PHASE VITREUSE">
        <meta name="DC.subject.keywords" content="(LANG=en) ">
        <meta name="DC.description" content="(LANG=fr) Le travail porte sur la
caracterisation du comportement mecanique de mullite zircone et de zircon a haute
temperature dans le but de mieux comprendre le comportement a la rupture et de
montrer l'existence d'un seuil dans le mode de propagation des fissures.">
        <meta name="DC.description" content="(LANG=en) ">
        <meta name="DC.identifier" scheme="ISAL" content="97 ISAL 0105">
        <meta name="DC.format" content="ADOBE ACROBAT PDF">
        <meta name="DC.language" content="fre">
        <meta name="DC.date.creation" content="1998">
        <meta name="DC.identifier" scheme="URL" content="http://csidoc.insa-
lyon.fr/these/1997/carbonneau/">
        <meta name="DC.publisher" content="CITHER - Doc'INSA - INSA de Lyon">
        <meta name="DC.publisher" content="Doc'INSA - INSA de Lyon">
        <meta name="DC.publisher" content="B&acirc;t. 220,
                                           20, Av. Albert Einstein,
                                           69621 Villeurbanne Cedex (France),
                                           t&eacute;l : +33 4 72 43 81 40,
                                           Fax +33 4 72 43 85 02,
                                           M&eacute;l : doc@insa-lyon.fr">
```

10.6. Jaquette des cédéroms

Cette maquette de jaquette (vue réduite) est directement accessible depuis le guide de conversion.