Les thèses électroniques à l'Université Lumière : le respect des normes

L'Université Lumière, depuis 1999, a mis en place un système d'archivage et de diffusion électroniques des thèses en s'appuyant sur une plateforme logicielle qui convertit les documents que les doctorants ont produit avec un traitement de texte banal, en document structuré au format XML[1]. Cette plateforme de conversion s'appelle Cyberthèses/Cyberdocs ; elle a été conçue initialement en coopération avec les Presses de l'université de Montréal et produisait des documents au format SGML[2]. Elle a ensuite été transformée, toujours avec le soutien de l'Agence intergouvernementale de la *Francophonie* et en collaboration avec la société AJLSM, en une plateforme logicielle entièrement libre sous licence GNU/GPL[3]. Actuellement cette plateforme s'appuie sur une DTD[4] : la TEILite[5] et **produit** des documents au format XML. Elle est couplée à la plateforme SDX[6] qui permet de **diffuser** des documents structurés de manière dynamique en différents formats (Fig.1 La chaîne de conversion et de diffusion des thèses). L'idée principale qui guide les choix techniques c'est : l'accès libre et ouvert aux thèses et pour obtenir ce résultat, il faut respecter les normes.

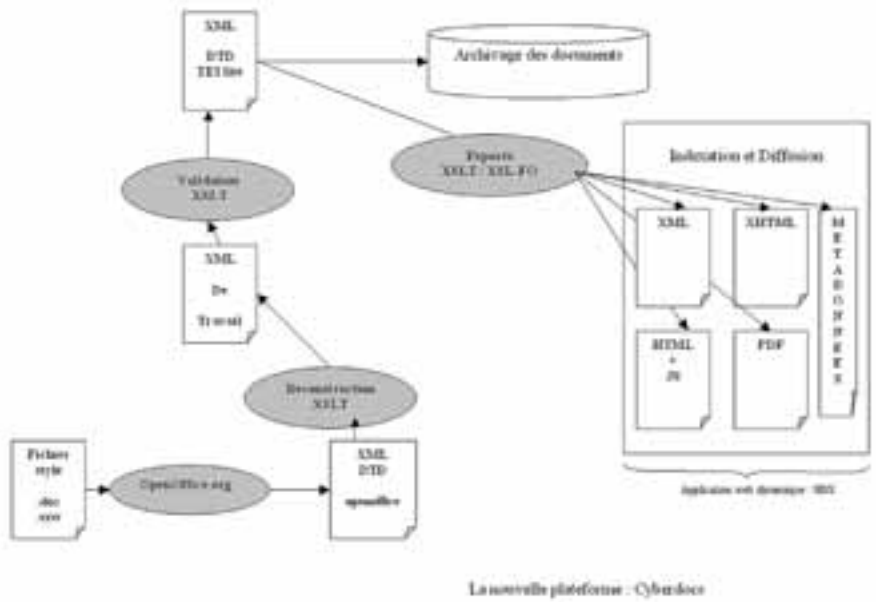
Pourquoi XML ?

La définition de l'accès libre[7] décrit précisément et publiquement les besoins et les attitudes des « utilisateurs » d'informations, en même temps que les principes de mise en forme de l'information scientifique.

« Par "accès libre" à cette littérature, nous entendons sa mise à disposition gratuite sur l'internet public, permettant à tout un chacun de lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces articles, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale, sans barrière financière, légale ou technique autre que celles indissociables de l'accès et l'utilisation d'Internet. La seule contrainte sur la reproduction et la distribution, et le seul rôle du copyright dans ce domaine devraient être de garantir aux auteurs un contrôle sur l'intégrité de leurs travaux et le droit à être correctement reconnus et cités. »

Le projet, tel qu'il est énoncé dans la déclaration de Budapest pour l'accès ouvert est politiquement cohérent ; il énonce une philosophie de la diffusion des résultats de la recherche qui décrit très précisément ce que les chercheurs font dans leur pratique quotidienne :

- lire
- télécharger
- copier
- transmettre
- imprimer
- chercher ou faire un lien vers le texte intégral de ces documents
- les disséquer pour les indexer
- s'en servir de données pour des logiciels.



La nouvelle plateforme : Cyberdocs

Fig. 1 **La chaîne de conversion et de diffusion des thèses**

Que sous-entend cette déclaration en langage documentaire ?

Le lecteur-utilisateur veut :

- rechercher des documents dans des corpus, à condition que l'on sache où sont ces corpus ;
- rechercher des informations dans un ou plusieurs documents à condition d'avoir accès au texte ;
- afficher des documents ou des parties de documents ;
- imprimer tout ou partie du document consulté ;
- pouvoir afficher les différentes parties selon des modalités de présentation fonctionnelle correspondant aux besoins du lecteur.

Ce sont les :

- chapitres
- paragraphes
- notes de bas de pages
- notes bibliographiques
- figures
- images, animées ou fixes
- table des matières,
- bases de données externes, etc.

Dans cette énumération, on voit apparaître toutes les exigences d'une communauté d'utilisateurs qui attend des solutions logicielles qui puissent correspondre aux objectifs fixés.

Le schéma suivant représente ces exigences en même temps qu'il fournit une réponse normalisée et standardisée. On note la présence de deux termes qui sont fondamentaux pour fournir une offre corres-

pondant aux exigences des utilisateurs : XML et OAI – Fig. 2 Les fonctionnalités de recherche et de consultation d'un document avec Cyberdocs.

XML : une pierre angulaire

XML est la pierre angulaire de ce nouveau type de publication numérique et convient particulièrement à nos pratiques documentaires.

Rappelons quelques **principes de XML**

- Différencier la forme et la structure logique du contenu
- Standardiser un langage de codage et un langage de description, ce qui a pour conséquence la généricité des échanges ainsi que la généricité des outils de traitement
- Contraindre par des modèles
  - le document est conforme à son modèle
  - les données transportent leur modèle de données
- Faciliter l'automaticité des traitements
  - les traitements sont liés à des modèles
  - les traitements s'appuient sur la structure

L'information devient indépendante des plateformes ou des logiciels qui l'ont créée et indépendante de son utilisation : elle est codée selon un format neutre et structurant ce qui garantit sa pérennité. Les outils qui sont issus de la « famille » XML, tels que XSLT, permettent de transformer la présentation de l'information contenue dans le document en fonction des besoins de l'utilisateur du document à un moment donné.

La formation au document structuré et à la feuille de style

La principale contrainte à la production de documents structurés provient de la forme dans laquelle l'auteur remet son document : celui-ci doit en effet être correctement structuré pour faciliter le passage au format pivot XML. Il est donc nécessaire de sensibiliser et former les doctorants à la pratique du traitement de texte et essentiellement à l'utilisation d'un modèle de document ou feuille de

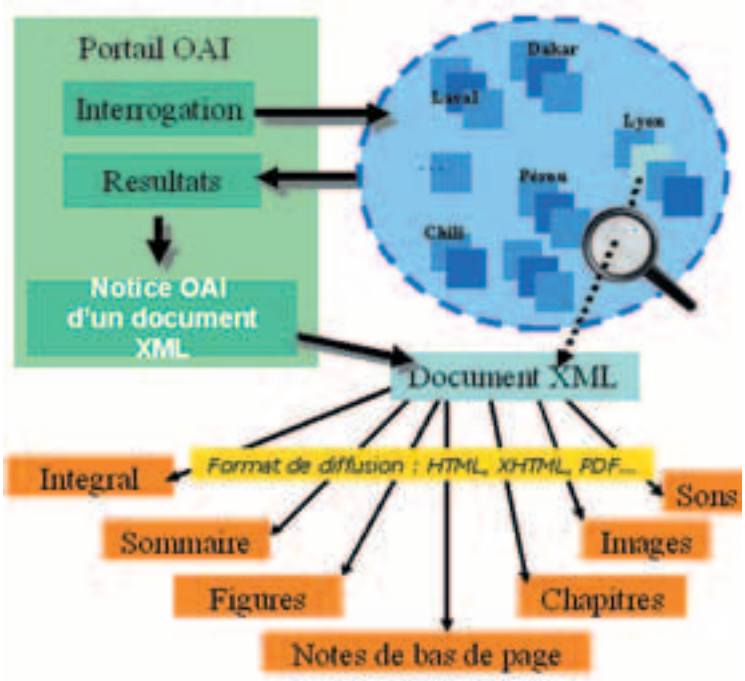


Fig. 2 Les fonctionnalités de recherche et de consultation d'un document avec Cyberdocs

style. Depuis les origines du programme, l'Université Lumière a mis en place des formations à la feuille de style pour les doctorants. Cet outil de présentation avait été réalisé conjointement avec nos collègues de l'université de Montréal et propose tous les styles qui sont nécessaires à l'écriture d'un document thèse. Ce modèle de document existe pour Microsoft Word, Mac ou PC, pour Star Office et Open Office et est utilisable sous n'importe quel système d'exploitation. Il permet de représenter tous les éléments essentiels qui composent un objet thèse, y compris les informations signalétiques de la thèse contenus sur la première page et qui vont fournir les métadonnées de base.

Ces formations portent leurs fruits : lorsque la thèse déposée par l'auteur est conforme aux prescriptions, l'université prend à sa charge l'impression des exemplaires réglementaires. En 2005, 77 % des thèses déposées ont été prises en charge pour l'impression. Au total depuis le lancement du programme, 1 250 étudiants environ ont été formés aux outils de production de documents électroniques et sensibilisés à l'importance de la diffusion ouverte et libre des travaux de recherche. L'appropriation de cette compétence de base permet aux doctorants de maîtriser la rédaction de leur thèse ; elle sera dorénavant acquise dans le cursus universitaire dès la licence dans le cadre du C2I. On peut donc raisonnablement penser, et notre expérience locale le montre, que la structuration de la thèse par les professionnels ne représente pas dans les années à venir une charge de travail pénalisante comme on a pu le lire ou l'entendre un peu partout. Actuellement, le temps de traitement (préparation, conversion et diffusion) d'une thèse correctement produite par son auteur est de deux heures en moyenne.

Le dépôt électronique est obligatoire depuis 2000 à l'Université Lumière. Le conseil scientifique et le conseil d'administration en ont décidé lors de la mise en place de la charte des thèses conformément à l'arrêté du 3 septembre 1998. L'archivage électronique est obligatoire : il fait partie des missions de l'université, mais c'est l'auteur et lui seul qui peut autoriser la diffusion sur Internet. Lors du dépôt de la thèse qui s'effectue AVANT la soutenance, l'auteur signe une déclaration de conformité du document avec la version qui sera soutenue devant le jury et signe, le cas échéant, une autorisation de diffusion sur les réseaux. Il garde l'intégralité de ses droits de reproduction sur toute forme de support existant ou à venir. La signature de ce document, toujours révisable, permet de sensibiliser les auteurs aux enjeux liés à la propriété intellectuelle et morale dans le monde de la publication scientifique numérique. Nous allons bientôt proposer aux étudiants de diffuser sous licence **Creative Commons**[8]. Les opérations de dépôt et de gestion des thèses ainsi que les relations que nous avons instituées avec les autres services de l'université sont gérées par un outil de workflow : OGET dont la description et les fonctionnalités sont présentées par Magalie Prudon, dans un article paru dans la revue *Documentaliste*, vol. 42, n° 4-5, p.280 à 282.

.../...

Le signalement et l'indexation des thèses conformément au protocole OAI-PMH

L'objectif initial d'archivage et de diffusion des thèses en XML est de faciliter le partage de textes et d'informations structurées en séparant le contenu (les données) du contenant (le support des données). Pour atteindre cet objectif, il faut que les documents puissent être visibles et accessibles partout et par tous. Le protocole **OAI-PMH**[9] permet de répondre à ce besoin d'échange qui est le propre de toute communauté scientifique. Les métadonnées utilisées dans le programme Cyberthèses/Cyberdocs sont conformes au schéma **Dublin Core**[10] non qualifié sur lequel repose le protocole OAI-PMH, et au modèle ETDMS du NDLTD. Notre serveur de thèses est déclaré fournisseur de données et est « moissonnable » par les principaux fournisseurs de services. Il est accessible à travers les principaux serveurs scientifiques et les moteurs de recherche et devrait pouvoir l'être rapidement par les portails de l'ABES (Fig 3 Réponse obtenue sur OAISTER à la question posée « Pinol Jean »). Actuellement, d'après les statistiques de consultation du serveur de l'université, sur la période qui va de mars 2005 à avril 2006, 72 000 **visiteurs différents** soit une moyenne mensuelle d'environ 6 000 visiteurs a été recensée. 715 thèses sont en ligne à la date du 14 mai 2006, soit la quasi totalité des thèses soutenues à l'Université Lumière depuis l'année 2000. Le délai entre la date de soutenance et la mise en ligne après traitement est actuellement inférieur à six semaines.

Fig 3 Réponse obtenue sur OAISTER à la question posée " Pinol Jean "

Un nouveau paradigme ou un retour aux sources ?

Le choix qui a été fait à l'Université Lumière, avec Cyberthèses/Cyberdocs, de privilégier, à travers le document structuré XML et le signalement OAI, le respect des normes et des standards est un gage de pérennité, d'**interopérabilité**[11] et d'accessibilité dont nous commençons à recueillir les fruits. Depuis toujours, ce sont les universités et les établissements habilités à délivrer le grade de docteur qui ont en charge la gestion, la soutenance, l'archivage, la diffusion et le signalement de ces travaux. Les potentialités offertes par l'apparition des documents numériques, ne changent pas la donne de manière fondamentale. La géographie institutionnelle permet de construire des unités de production et d'archivage, réparties sur le territoire selon une logique décentralisée propre aux réseaux, proches des chercheurs qui n'est pas contradictoire avec l'autonomie institutionnelle des universités. L'édification de portails thématiques ou nationaux qui prennent en charge dans le respect des normes existantes, les opérations de signalement permettra une visibilité encore plus grande des thèses produites dans les établissements, sans en supporter la charge de traitement.

Grâce à ces choix, les docteurs et chercheurs de notre université dont les travaux peuvent être consultés en toute ouverture et liberté, sont visibles et accessibles dans la communauté mondiale des chercheurs – Fig. 4 Les résultats de l'interrogation d'un portail de thèses OAI-PMH.

Par delà l'accessibilité du document, on commence à évaluer les effets de l'accès ouvert des documents scientifiques électroniques structurés sur le mode de fonctionnement de la recherche.

Nous sommes encore dans un paradigme de recherche individuelle où toute la chaîne est contrôlée par le chercheur et où la thèse constitue le produit final alors que la documentation (la fourniture de la preuve) reste inaccessible à la communauté. **La diffusion des thèses selon un format qui respecte un format structuré nous permet d'accéder aux *outils et sources utilisés*** par l'auteur ouvrant la voie à une recherche ouverte et collaborative a posteriori quant à l'élaboration du travail et l'administration de la preuve. Le chercheur avec les moyens que nous mettons à sa disposition peut présenter sa problématique, ses outils et soumettre ainsi sa démarche à l'évaluation de ses lecteurs. N'est-ce pas un retour aux origines de la communication scientifique ?

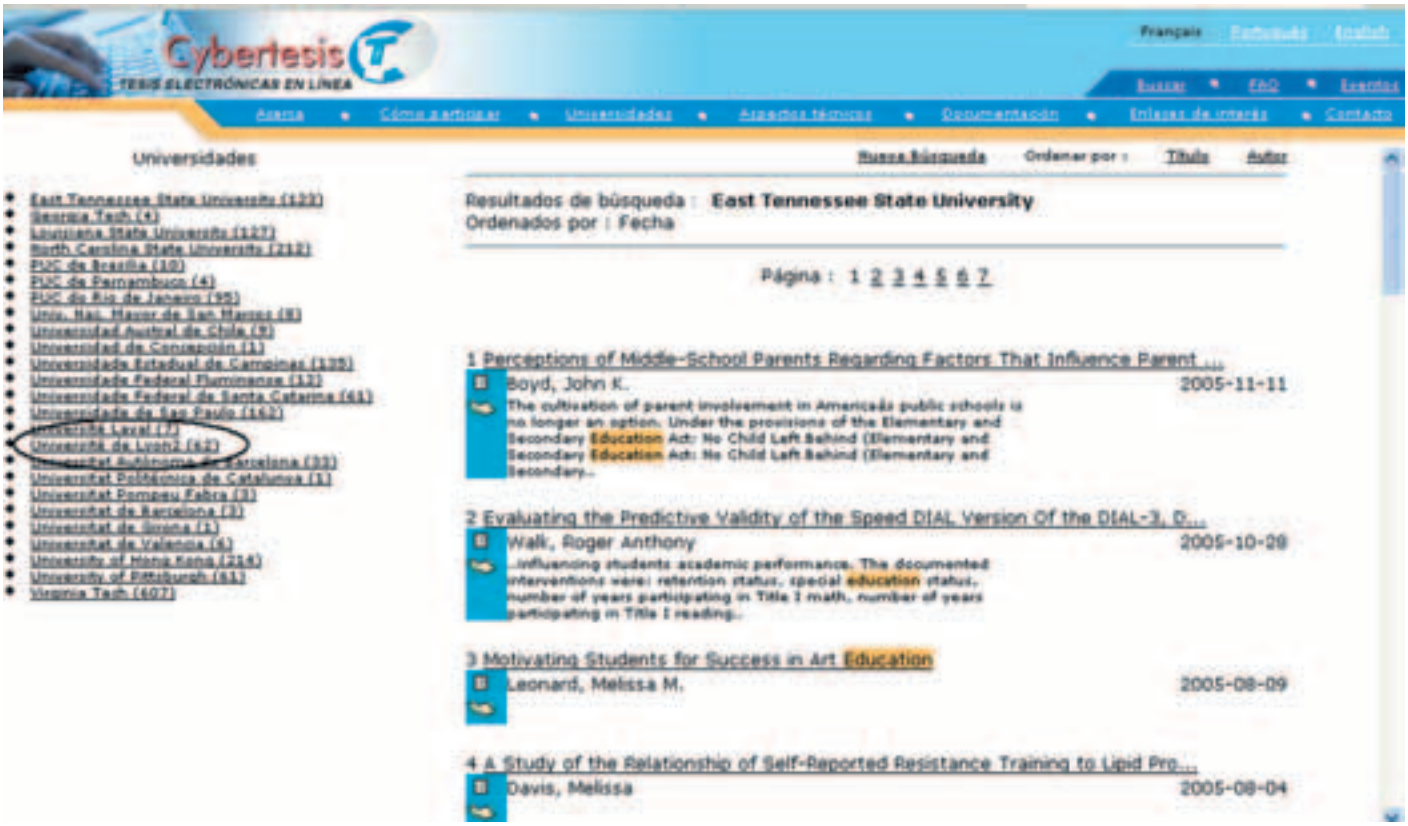


Fig. 4 Les résultats de l'interrogation d'un portail de thèses OAI-PMH

Les seules adresses de référence du programme Cyberthèses/Cyberdocs sont les suivantes :
le site de Lyon-II : http://theses.univ-lyon2.fr
le portail francophone des thèses : http://cybertheses.francophonie.org
le site collaboratif : http://sourcesup.cru.fr/cybertheses/fr/
le site de l'Université du Chili : http://www.cybertesis.cl/
le site du réseau Cyberthèses : http://www.cybertheses.org

Jean-Paul Ducasse
Jean-Paul.Ducasse@univ-lyon2.fr

Jean-Paul Ducasse ✆ 04 78 69 74 25 🖹 76 51
Service général des publications scientifiques
Programme Cyberthèses/Cyberdocs

Claude Journès, président de l'Université Lumière – Lyon-II
✆ 04 78 69 71 52 🖹 56 01 ✉ 86 rue Pasteur 69007 LYON

1 http://fr.wikipedia.org/wiki/XML, http://www.w3.org/XML/1999/XML-in-10-points.fr.html
2 http://fr.wikipedia.org/wiki/SGML
3 http://fr.wikipedia.org/wiki/GNU_GPL, http://www.gnu.org/home.fr.html
4 http://fr.wikipedia.org/wiki/DTD
5 Lou Burnard et CM. Sperberg : La TEI simplifiée : une introduction au codage des textes électroniques en vue de leur échange - http://www.gutenberg.eu.org/publications/autres/TEILITE/
6 http://adnx.org/sdx/
7 http://www.soros.org/openaccess/fr/read.shtml
8 http://fr.creative.commons.org
9 NAWROCKI François : le protocole OAI et ses usages en bibliothèque : http://www.culture.gouv.fr/culture/dll/OAI-PMH.htm
10 http://fr.wikipedia.org/wiki/Dublin_Core
11 http://fr.wikipedia.org/wiki/Interopérabilité