Mise en ligne des mémoires de thèse : le cas de CodeX

par Stéphane Chaudiron,
François Role et Majid Ihadjadene*

Les thèses en ligne sur Internet

Cet article présente les premiers résultats du projet CodeX (Consultation et organisation de documents électroniques à Paris-X), qui vise à concevoir et réaliser une plate-forme de visualisation de documents électroniques (mémoires et thèses) adaptée aux différents contextes d'usage. La numérisation des mémoires et des thèses permettra de résoudre les problèmes liés à la conservation des documents, coûteuse en temps et en argent. Elle permet aussi de favoriser la diffusion des travaux de doctorat ou de maîtrise qui sont jusqu'ici accessibles sous les seules formes papier et microfiche..

De plus, la diffusion des thèses sur Internet permettra de valoriser la production des universités, puisque leur consultation ne sera plus confinée à l'université mais accessible de n'importe quel endroit dans le monde. Dans le passé, plusieurs enquêtes ont montré que la thèse est une ressource documentaire qui est sous-exploitée.

Dans le monde, de nombreuses initiatives pour diffuser les thèses sur Internet ont vu le jour. La majorité d'entre elles se limitent encore à des études de faisabilité. En France, on peut citer les initiatives de l'ANRT (Atelier national de reproduction des thèses), les projets Callimaque[1] et CITHER[2], qui proposent l'accès à des thèses en ligne sous forme d'images que ce soit en mode TIFF ou en mode de diffusion PDF. On peut aussi signaler le serveur national de thèses, WebThèses[3], fruit d'un partenariat entre la sous-direction des bibliothèques, l'ABES, l'ANRT et le CINES. Environ une centaine de thèses numérisées en format PDF sont consultables sur ce serveur.

Le projet Cyberthèses[4], fruit d'une coopération entre les Presses de l'université de Montréal et l'université Lumière-Lyon-2, portait en premier lieu sur la réalisation d'une chaîne de production de thèses en ligne s'appuyant sur la norme SGML.

Il existe plusieurs projets en cours de développement aux États-Unis, en Allemagne ou au Canada. Citons par exemple les travaux de NDLTD et de l'UMI.

Le projet NDLTD (Networked Digital Library of Theses and Dissertations), coordonné par Fox (Fox, 1997 ; Phanouriou, 1999), vise à mettre à la disposition des partenaires du réseau NDLTD[5] des outils permettant à chacun de participer à la construction d'une bibliothèque universitaire numérique fonctionnant selon un mode distribué qui met en application le concept d'intelligence répartie. Plus de 2 000 titres sont disponibles en format PDF.

On peut citer les efforts de UMI (Savage, 1999), qui convertit depuis peu les thèses vers le format PDF. Actuellement, plus de 100 000 titres sont disponibles sous ce format.

Tous ces prototypes offrent la possibilité d'effectuer des recherches booléennes par champ. Quelques serveurs permettent un feuilletage alphabétique sur les listes d'auteurs, de mots-clés, de directeurs de thèse. Il est possible de faire une recherche sur le texte intégral des thèses sur les serveurs de WebThèses et de CITHER. Un moteur de recherche, Altavista Search Intranet 97, indexe tous les fichiers de ces site en texte intégral.

Quelques chercheurs ont par ailleurs évalué l'usage que font les utilisateurs de ces documents numériques. De ces expériences, il ressort que :

- Les usagers préfèrent effectuer des recherches analytiques booléennes que de naviguer.
- Les évaluations effectuées dans le cadre des projects TULIP[6] et NDLTD ont montré que très peu d'articles et de thèses numérisés sont lus entièrement. La consultation se fait plutôt sur des parties du document, comme le résumé. Les usagers consultent moins le texte intégral. Ces études montrent aussi le faible usage du feuilletage (linéaire et hiérarchique) comme processus de consultation.

* *Centre de recherche en information spécialisée (CRIS), Université de Paris-X*
{prénom.nom}@u-paris10.fr

– Au niveau du processus de recherche, ils ont identifié plusieurs problèmes. Si la recherche en texte intégral est appréciée par les usagers, ceux-ci ont des difficultés dans l'usage des mots-clés, des recherches booléennes, des opérateurs de proximité, et dans l'élaboration de stratégies de recherche pour réduire le nombre élevé de réponses. Tant que le nombre de thèses disponibles en ligne était réduit, le problème de la recherche documentaire ne se posait pas vraiment.

– Les utilisateurs souhaitent une vitesse de téléchargement élevée et une impression de haute qualité. De plus, la lecture de documents à l'écran est souvent difficile. En raison de quelques contraintes de temps (téléchargement), de dispositif (écran de l'ordinateur), de lieu, de coûts, l'usager ne lit pas le document numérique en entier mais essaie de formuler une représentation globale du contenu textuel, en en lisant quelques parties afin de juger de la pertinence des informations qu'il contient.

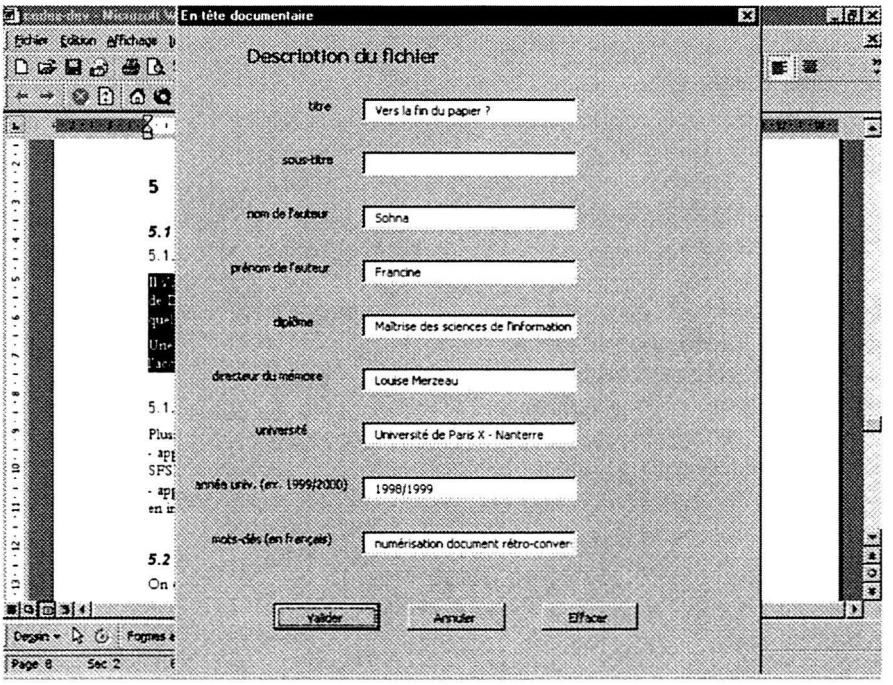
Le projet CodeX

L'objectif central du projet CodeX est de rendre accessible un corpus de documents électroniques selon des modes de consultation adaptés aux différents contextes d'usage. Ce projet s'inscrit dans le contexte pédagogique universitaire, et plus particulièrement du deuxième cycle en sciences de l'information.

Le point de départ du projet a été la volonté de l'équipe d'enseignants de mettre à la disposition des étudiants de maîtrise, sous forme électronique et *via* un serveur, les mémoires soutenus les années précédentes. Ce projet a également une dimension pédagogique. Il s'agit de permettre aux étudiants de maîtriser les outils de publication électronique, et d'utiliser des modèles de documents qui permettent la conversion vers XML.

La base des documents mis en consultation est constituée des mémoires de maîtrise en sciences de l'information et de la documentation des années 1994 à 1999. Pour répondre à la demande de validation des documents mis en ligne, seuls les mémoires jugés d'une qualité suffisante pour être consultés ont été mis en ligne. Au total, le corpus initial est constitué d'environ 50 documents.

Chaque document a été indexé librement par son auteur, sans utilisation d'une liste de mots-clés de référence ni d'un thésaurus. Un résumé d'auteur accompagne également chaque mémoire. Dans un premier temps, un formulaire très simple intégré à l'outil bureautique[7] permet une saisie rapide et systématique des métadonnées dont on souhaite avoir une représentation XML (figure ci-dessous). Une fois ce formulaire validé, une macro Word en récupère les données et génère automatiquement un fichier XML contenant un en-tête documentaire. Dans un second temps, le corps du document est soumis à un convertisseur qui en produit une représentation XML.



L'approche adoptée au sein du projet est donc de convertir le corpus en XML, puis d'utiliser le mécanisme des feuilles de style XSL pour générer des vues adaptées à chaque usage. Complétant la recommandation XML, il existe en effet un langage dit XSL (Extensible Stylesheet Language) qui permet de définir des feuilles de style applicables aux documents XML[8]. Le terme « feuille de style » est un peu réducteur, dans la mesure où une spécification XSL consiste schématiquement en un ensemble de règles de traduction permettant non seulement d'associer des attributs de présentation physique (par exemple une police, une taille, une couleur, etc.) aux éléments d'un fichier XML, mais également de filtrer et de réordonner ces derniers.

Le projet CodeX est découpé en deux phases. Lors de la première phase, il est prévu de rétroconvertir le corpus et de mettre en place les fonctionnalités permettant de disposer du mécanisme de vues[9], ainsi que d'un langage de requêtes exploitant la structure logique des documents. À l'issue de cette phase, il a été décidé de doter CodeX de fonctionnalités supplémentaires. En ce qui concerne tout d'abord les modes de consultation, il est prévu de permettre l'interrogation de la base de documents selon des « vues multiples dynamiques ». Ce mode de consultation de la base permettra aux utilisateurs de lancer des requêtes sur la base de documents correspondant à des traitements linguistiques en temps réel.

Il est ainsi prévu d'offrir les fonctionnalités de filtrage et de résumé automatique d'une section de document ou d'un sous-ensemble de la base. Une étude comparative des différents outils de filtrage et de résumé automatique est en cours, sur les plans technique et des usages. Ces possibilités d'exprimer des requêtes correspondant à des traitements linguistiques dynamiques s'ajouteront aux fonctionnalités actuelles. Les vues multiples du document deviendront ainsi dynamiques. Cette perspective s'inscrit dans le contexte de l'évolution des systèmes de bases de données vers les systèmes de gestion de la connaissance (Chaudiron, 1999).

Un deuxième axe de développement, concernant l'aspect ergonomique, est la réalisation d'une interface permettant aux utilisateurs d'interroger les documents en utilisant le langage naturel. Différents travaux consacrés à la consultation interactive des banques de données, notamment des OPACs (Ihadjadene, 1999), ont montré la nécessité de fournir des aides à l'interrogation (reformulation interactive), surtout quand les documents ont été indexés librement. Enfin, nous comptons mettre en place un processus d'évaluation continu pour étudier l'usage réel de ces documents en vue d'améliorer l'utilisabilité de CodeX et son fonctionnement.

Bibliographie

Stéphane Chaudiron : L'Apport de l'ingénierie linguistique à la gestion des connaissances. in actes de JILA'99 - *Journées internationales de linguistique appliquée*, H. Zinglé, éd., Nice, Université de Nice-Sophia Antipolis, p. 72-76.

Edward A. Fox, John L. Eaton et al. : Networked Digital Library of Theses and Dissertations : an international effort unlocking university resources. *D-Lib Magazine*, septembre 1997. Disponible à : http://www.dlib.org/dlib/september97/theses/09fox.html.

M. Goossens : XML et XSL : un nouveau départ pour le Web. in actes de GUT'99, M. Goossens, éd., Lyon, INPL, p. 3-126.

Majid Ihadjadene : La recherche et la navigation dans un système de recherche d'information grand public. Thèse de doctorat, université de Lyon-1, 1999.

Constantinos Phanouriou, Edward Fox et al. : A digital library for authors : recent progress of the Networked Digital Library of Theses and Dissertations. in *Proc. ACM Digital Libraries '99*, Berkeley, CA, 11-14 août 1999.

François Role et Philippe Verdret : Le Document Object Model, in actes de GUT'99, M. Goossens. éd., Lyon, INPL, p. 155-171.

William Savage : Reflections on a sustainable model for the digital publication of theses and dissertations. Unesco Workshop on the Electronic Dissemination of Theses and Dissertations, Paris, 27-28 septembre 1999. Disponible à : http://www/unesco.org/etd

1. Disponible à : http://callimaque.grenet.fr
2. Disponible à : http://csdoc.insa-lyon.fr/these
3. Disponible à : http://webthese.cnusc.fr:8110
4. Disponible à : http://www.cyberthese.org
5. Disponible à : http://www.ndltd.org
6. Disponible à : http://www.elsevier.nl/homepage/about/resproj/trappdx.htm#AppendixXIII
7. Cette caractéristique nous permet d'utiliser cette méthode pour saisir de façon précise les nouveaux mémoires soutenus à partir de maintenant.
8. Pour une introduction détaillée à XSL et ses différentes composantes, voir Goossens, 1999.
9. Possibilité d'interroger et de consulter des parties du document (introduction, résumé, bibliographie, conclusion, tables des matières, etc.).