Documentation et bibliothèques

La publication électronique des thèses : un exemple franco-québécois de coopération à destination de la francophonie

The Electronic Publication of Theses: An Example of Franco-québécois Collaboration Aimed at the Francophone World

La publicación electrónica de las tesis: un ejemplo francoquebequense de cooperación destinado a los francoparlantes

Valérie Boulétreau, Jean-François Gauvin and Jean-Paul Ducasse

Volume 45, Number 4, October–December 1999

Édition électronique

URI: https://id.erudit.org/iderudit/1032721ar
DOI: https://doi.org/10.7202/1032721ar

See table of contents

Publisher(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (print)
2291-8949 (digital)

Explore this journal

Cite this article

Boulétreau, V., Gauvin, J.-F. & Ducasse, J.-P. (1999). La publication électronique des thèses : un exemple franco-québécois de coopération à destination de la francophonie. *Documentation et bibliothèques*, *45*(4), 183–189. https://doi.org/10.7202/1032721ar

Article abstract

CyberThèses is a publishing and electronic dissemination of theses program developed by the Université Lyon 2 and the Presses de l'Université de Montréal. The collaboration aims to produce theses and facilitate their dissemination on the Internet. Using specific technical procedures, it allows any institution of higher learning to be autonomous in the electronic dissemination of knowledge and to contribute to a network of university research projects. This article describes the structures (based on SGML/XML) and referencing (based on the model proposed by the Corporation for National Research Initiatives) of theses. The authors insist on the necessary training of doctoral candidates in order that they acquire the logic used to produce structured electronic documents.

La publication électronique des thèses : un exemple franco-québécois de coopération à destination de la francophonie

Valérie Boulétreau
Chef de projet Édition électronique
Service des nouvelles technologies de l'information et de la réalisation des serveurs (SENTIERS)
Université Lumière Lyon 2
viviane.bouletreau@univ-lyon2.fr
Jean-François Gauvin
Chargé de projet Thèses électroniques
Presses de l'Université de Montréal
jean-francois.gauvin@umontreal.ca
Jean-Paul Ducasse
Maître de conférences
Institut d'études politiques (IEP) de Lyon
Université Lumière Lyon 2
Responsable scientifique du programme
La publication électronique des thèses : pour une diffusion de l'édition savante francophone sur les inforoutes
Fonds francophone des inforoutes
jean-paul.ducasse@univ-lyon2.fr

CyberThèses est un programme d'édition et de diffusion électroniques des thèses conçu par l'Université Lyon 2 et les Presses de l'Université de Montréal. Cette coopération vise à créer une chaîne de production des thèses et de diffusion sur Internet. À l'aide de procédures techniques précises, elle devrait permette à tout établissement d'enseignement supérieur d'acquérir son autonomie sur le plan de la diffusion électronique du savoir tout en participant à un réseau qui mutualise les travaux universitaires de recherche. Cet article présente les modes adoptés de structuration (reposant sur SGML/XML) et de référencement (fondé sur le modèle proposé par la Corporation for National Research Initiatives) des thèses ; et insiste sur la nécessaire formation à mettre en place à l'intention des doctorants pour les initier à la logique de production de documents électroniques structurés.

The Electronic Publication of Theses : An Example of Franco-québécois Collaboration Aimed at the Francophone World
CyberThèses is a publishing and electronic dissemination of theses program developed by the Université Lyon 2 and the Presses de l'Université de Montréal. The collaboration aims to produce theses and facilitate their dissemination on the Internet. Using specific technical procedures, it allows any institution of higher learning to be autonomous in the electronic dissemination of knowledge and to contribute to a network of university research projects. This article describes the structures (based on SGML/XML) and referencing (based on the model proposed by the Corporation for National Research Initiatives) of theses. The authors insist on the necessary training of doctoral candidates in order that they acquire the logic used to produce structured electronic documents.

La publicación electrónica de las tesis : un ejemplo francoquebequense de cooperación destinado a los francoparlantes
CyberThèses es un programa de edición y difusión electrónica de tesis, concebido por la Universidad de Lyon 2 y la Editorial de la Universidad de Montreal. Esta cooperación está destinada a crear una cadena de producción de tesis y de difundirla en Internet. Con la ayuda de procedimientos técnicos precisos, debería permitir que todo establecimiento de enseñanza superior adquiriera su autonomía en el plano de la difusión electrónica del conocimiento, participando al mismo tiempo en una red que facilita el intercambio de los trabajos universitarios de investigación. Este artículo presenta los modos adoptados de estructuración (basándose en SGML/XML) y de referencia (fundado en el modelo propuesto por la Corporation for National Research Initiatives) de las tesis ; e insiste en la formación necesaria de ejecución para que los doctorandos se inicien en la lógica de producción de documentos electrónicos estructurados.

À l'heure actuelle, la diffusion et l'utilisation scientifique des thèses académiques, en tant que corpus d'information scientifique, sont très faibles. La circulation des documents imprimés est relativement modeste et la diffusion de ces travaux sous forme de microfiches reste confidentielle. Ce constat a conduit l'Université de Lyon 2 et les Presses de l'Université de Montréal (PUM) à engager en coopération un programme d'édition et de diffusion électroniques de thèses sur info-

routes, en s'appuyant sur la norme SGML : ainsi naquit le programme CyberThèses[1] soutenu par le Fonds francophone des inforoutes[2] dans le cadre de son premier appel à proposition (septembre 1998).

Cette coopération a pour but de mettre en service une chaîne de production et de diffusion électroniques des thèses par la conception et l'élaboration d'un certain nombre de procédures logicielles qui prennent en compte les spécificités de ce genre de travaux de recherche. Ces outils de production et de diffusion ont vocation à être utilisés par tous les membres de la communauté universitaire. Notre objectif est de permettre à tout établissement d'enseignement supérieur d'acquérir son autonomie sur le plan de la diffusion électronique du savoir, tout en participant à un réseau qui mutualise les ressources universitaires en matière de recherche.

La thèse ne doit plus être exclusivement considérée comme l'aboutissement d'une recherche. Il ne s'agit pas non plus, dans notre esprit, de diffuser la version électronique d'un document préalablement imprimé, mais de concevoir un mode d'intégration de nouvelles fonctionnalités qui fasse de la thèse, dans sa version électronique, un véritable instrument de travail satisfaisant la demande des usagers à partir de leur propre logique de recherche et d'investigation. De plus, il faut envisager des solutions d'archivage électronique qui puissent garantir l'accès futur au document lui-même, quel que soit l'environnement logiciel qui existera dans l'avenir. La pérennité du document archivé est ainsi, avec la structuration de ce document, un des éléments clés de notre projet.

Nous avons donc considéré qu'il était possible d'envisager de nouvelles manières de diffuser et de valoriser la production scientifique issue du secteur académique en profitant de l'environnement logiciel de production et de diffusion de ces travaux via les réseaux électroniques. De plus, les outils logiciels de structuration de documents électroniques reposant sur le format « pivot » SGML permettent de mettre en place des instruments très efficaces de recherche et de repérage de l'information.

Mais la démarche adoptée par les équipes québécoise et française ne se limite pas à un projet technologique. Les implications du programme CyberThèses sont multiples. Nous en relèverons deux qui nous sont apparues essentielles à la réussite de ce projet : la nécessité d'une formation des doctorants et la clarification des relations entre le docteur et l'institution universitaire où il a mené sa recherche.

Ce programme impliquait, dès l'origine, un important effort de formation aux technologies de l'information : il fut en effet prévu d'introduire, dans le cursus des doctorants, des programmes de sensibilisation et de formation à la production et à la diffusion de documents électroniques structurés (organisation et rédaction, mise en forme de leur travail, et conversion de leurs documents dans des formats de diffusion intermédiaires). Un plan de formation a été élaboré qui démarrera en janvier 2000. Compte tenu des différences institutionnelles entre le Canada et la France, les modalités d'application n'en sont pas tout à fait identiques à Montréal et à Lyon, mais la logique reste la même. Cette formation se présente sous forme de modules et elle est destinée à l'ensemble des doctorants et étudiants chercheurs de chaque université.

D'autre part, l'étudiant chercheur, parce qu'il maîtrise les mécanismes et les outils de création du document électronique qu'est sa thèse, va se trouver dans la situation d'un producteur d'information scientifique. Il gagne ainsi en autonomie, tout en acquérant un statut personnel de producteur et de diffuseur d'information. Les rapports qui existent entre le doctorant et l'institution dans laquelle il effectue son travail de recherche s'en trouvent modifiés et clarifiés.

Au moment de la soutenance – c'est du moins le cas à l'Université Lyon 2 où le dépôt électronique est maintenant inscrit dans la charte des thèses – le candidat doit fournir, en même temps que les habituels exemplaires imprimés, une version électronique de sa thèse. Après la soutenance, et en fonction de la décision du jury, les différents cas de diffusion prévus par les textes réglementaires sont appliqués : diffusion sans correction, diffusion après correction, diffusion restreinte. L'auteur signe alors un document qui autorise, de façon non exclusive, la diffusion de sa thèse dans sa version canonique, c'est-à-dire la version validée par le jury. Cette autorisation de diffusion est révocable et adaptable selon les cas, en raison de contraintes liées, par exemple, au mode de financement de la thèse. Par cette démarche, l'auteur se réapproprie son travail et définit lui-même les relations qui le lient aux différents diffuseurs publics ou privés.

Cette clarification des relations juridiques entre le docteur et son université qui respecte les droits de l'auteur sur son œuvre permet une diffusion très rapide de la thèse, et donc sa mise à disposition de la communauté scientifique quasi immédiate. C'est un des objectifs du programme : permettre à n'importe quel chercheur dans le monde de repérer une thèse et d'accéder au document proprement dit le plus efficacement possible. Rappelons qu'il s'agit là d'une demande très forte des associations de doctorants en France.

Le document électronique structuré : enjeux et outils de réalisation

Lorsqu'on a affaire à des documents rédigés en vue d'une diffusion électronique, il est possible d'imposer au producteur un certain nombre de contraintes relatives aux outils ou aux formats. En revanche, lorsqu'il s'agit de la diffusion électronique de documents qui, comme des thèses de doctorat, ont été rédigés pour être d'abord imprimés, il n'est pas envisageable de contraindre les étudiants chercheurs à utiliser un outil logiciel particulier pour saisir leur mémoire. Aussi devons-nous traiter des documents produits à l'aide d'outils très divers allant de *Word*, *WordPerfect*, *ClarisWorks* ou *Lotus* à *LaTeX*.

Cependant, dans le cadre de la mise en ligne d'un corpus de documents, quel qu'il soit, il n'est pas pensable de demander à tout lecteur qui pourrait être amené à les consulter de disposer de tous les logiciels associés à chacun des formats de production utilisés. La première étape de notre réflexion a donc porté sur le choix de formats électroniques propres à l'archivage et à la diffusion de documents.

Le choix du format d'archivage, ou « format pivot », est crucial puisqu'il est le garant de la réussite – à long terme – de notre projet :

1. Projet disponible à <URL : http://www.univ-lyon2.fr/sentiers/edition/theses/projettheses.html
2. Pour plus d'information sur l'Organisation internationale de la francophonie et sur le Fonds francophone des inforoutes, consulter le site <URL : http://www.francophonie.org/fonds/>

- il doit, autant que faire se peut, relever du domaine public, afin de ne pas être à l'origine de surcoûts à la production ou à la consultation des documents mis en ligne, et ainsi garantir aux promoteurs du projet une certaine indépendance ;
- il doit assurer la pérennité des documents archivés ;
- il doit permettre la génération de nouveaux fichiers sous de nouveaux formats susceptibles de devenir, un jour, des standards de diffusion (quel format sera l'équivalent de HTML dans dix ans ?) ;
- il doit être suffisamment souple pour s'adapter à la représentation des différentes structures de documents rencontrées selon les disciplines (physique, linguistique, musicologie ou archéologie, par exemple).

Notre choix s'est porté, pour CyberThèses, sur l'utilisation de documents de type structuré, en raison de leur richesse et de la souplesse d'exploitations qu'ils autorisent, et le format que nous avons choisi pour stocker ces documents obéit à la norme SGML/XML. Nous verrons que, correctement mise en œuvre, cette combinaison présente toutes les propriétés que nous venons d'énumérer.

La norme SGML/XML et le document structuré

Il est important tout d'abord de préciser ce que nous entendons par « document structuré ». L'on distingue généralement deux types de structures :

- la structure physique, ou mise en page d'un document : il s'agit des propriétés typographiques (police, taille, graisse, couleur) associées à chaque élément du texte, ainsi que de leur agencement les unes par rapport aux autres ;
- la structure logique décrit, elle, la nature ou le rôle de chacun de ces éléments, et les relations hiérarchiques et/ou logiques qui les lient. C'est ce type de structuration que nous utilisons.

Le *Standard Generalized Markup Language* (SGML) est l'outil adapté au codage de documents structurés[3]. Il s'agit d'une norme internationale (ISO 8879), datant de 1986, conçue pour la définition de méthodes de codage des documents électroniques. Plus précisément, c'est un métalangage permettant la description formelle de langages de marquage descriptif[4].

Par « balisage » ou « marquage descriptif », on entend tout type d'annotation ajoutée au texte qui permettra de savoir comment interpréter, afficher ou imprimer un passage particulier du document. À un niveau d'utilisation « faible », ou du moins courant, on peut considérer que tout texte imprimé est « marqué » par l'utilisation de ponctuation, de lettres capitales, de graisse, de soulignement, etc. ; la fonction de ces « marques » est d'aider à la lecture et à la compréhension du texte. Le métalangage se résume alors à décrire un ensemble de règles typographiques. C'est dans cette perspective que SGML a été utilisé pour définir le langage HTML.

À un autre niveau d'utilisation, on peut envisager que les balises ne servent plus à indiquer l'aspect ou les propriétés de mise en page des différents éléments du texte, mais plutôt leur rôle dans le texte et leurs relations avec les autres éléments. Le métalangage fournit alors un mécanisme d'identification structurelle des éléments textuels ainsi qu'un ensemble de règles permettant de définir précisément comment sont combinés ces éléments.

L'application aux thèses du concept de document structuré permet d'accroître de façon significative la qualité du contenu du document ; elle est aussi le garant de son homogénéité.

La structuration : un outil pour la recherche dans le document

Un des atouts du document – logiquement – structuré que nous n'avons pas encore mentionné est son utilité en matière de recherche d'information ; le rôle que joue, dans le document, un élément de texte donné peut en effet être une source d'information très pertinente. Dans un document non structuré, par une simple recherche, on ne peut que savoir si un mot est présent ou non ; dans un document structuré, on peut connaître avec une précision relativement fine le degré d'importance de chaque occurrence d'un mot dans le texte.

Imaginons, par exemple, le cas d'une recherche sur M. Dupont. La présence de son nom dans un document peut avoir un sens et une valeur très différents selon son rôle et sa place dans le texte. En est-il l'auteur ? Le directeur de recherche ou un membre du jury ? S'agit-il d'une citation de M. Dupont ? Son nom apparaît-il dans un titre, auquel cas, il est probable qu'une partie au moins du document porte sur son travail ? etc. Ce type d'information est perçu naturellement lors de la lecture mais, dans le cadre d'une recherche effectuée dans un service électronique d'information, la structuration d'un document est indispensable à une localisation et à une identification rapides et efficaces des différentes occurrences d'un mot dans les différents champs de ce document et de leur signification.

La chaîne de conversion des thèses

La chaîne de traitement que nous avons retenue pour le projet CyberThèses est une adaptation aux thèses de la chaîne développée pour les revues aux Presses universitaires de Montréal. Elle permet d'atteindre l'ensemble de nos objectifs, car elle constitue un système complet pour la production et l'archivage de thèses dans un format électronique structuré (SGML) et pour leur diffusion sous des formats plus courants (HTML, XML et PDF).

L'ensemble du processus de conversion des thèses repose sur l'analyse des attributs de style portés par chaque élément du texte. Nous avons été amenés à définir une feuille de style[5] propre aux thèses, que les doctorants doivent directement appliquer au document lors de sa rédaction. Notre analyse ne portant que sur les noms des styles et non sur les éléments de mise en page (police, alignement, graisse, etc.) auxquels ils correspondent, ces derniers peuvent être modifiés à volonté par les thésards. Ainsi, l'utilisation de la feuille de style constitue pour nous un premier pas vers la structuration logique du document, et pour l'étudiant un outil de rédaction lui permettant de produire des documents de meilleure qualité tout en lui garantissant une grande souplesse dans son choix de mise en page.

3. Éric Van Herwijnen. *SGML pratique*. Paris : International Thomson Publishing (1995) (édition française 1999), 330 p.
4. André Jacques et Vincent Quint. Structures et modèles de documents. *In Le document électronique, cours INRIA, Châtelaillon, 11-15 juin 1990*. (Roquencourt : Institut national de recherche en Pour informatique et en automatique). p. 3-60.
5. Disponible à <URL : http://www.univ-lyon2.fr/sentiers/edition/theses/ressources/These_Lyon2.dot>

La formation des doctorants

La formation dispensée aux doctorants est un élément indispensable à la réussite de notre projet CyberThèses. Les principaux objectifs que nous poursuivons sont les suivants.

Nous voulons sensibiliser les doctorants à l'approche d'une norme de structuration logique du document qui dépasse la présentation simplement physique du corpus. Cette structuration facilitera la recherche d'information à l'intérieur du document, ainsi qu'une recherche du document lui-même facilitée par la présence de métadonnées.

Nous souhaitons initier les étudiants chercheurs aux outils qui permettent la production de documents structurés, notamment à l'emploi des feuilles de style définies par le service de l'université chargé de la diffusion électronique: notions de base, définition du contenu sémantique et structuration syntaxique. Grâce à cette formation, ils apprendront à utiliser un traitement de texte non plus comme une machine à écrire perfectionnée, mais comme un outil qui, à travers des logiques de structuration du texte, permet une réelle intégration des éléments multimédias (les portées musicales d'une thèse en musicologie associées aux éléments sonores, par exemple). En les amenant à mieux structurer leur écriture, ces outils permettront aussi aux étudiants de parvenir à une plus grande maîtrise intellectuelle et scientifique dans leur démarche de recherche.

Enfin, nous abordons avec les doctorants les aspects juridiques liés à la diffusion de documents sur Internet, le risque de plagiat, et plus généralement la question des droits d'auteur.

Cette formation sera proposée à différents publics dès janvier 2000:

- aux étudiants nouvellement inscrits en doctorat afin qu'ils s'approprient ces nouveaux outils dès le début de leur carrière de chercheur;
- aux étudiants chercheurs qui sont sur le point de rédiger ou en phase de rédaction;
- aux directeurs de recherche chargés de l'encadrement des chercheurs, afin qu'ils participent, à leur niveau, à la logique de production de documents électroniques structurés.

Au-delà de la chaîne de production

Un des principaux objectifs du programme CyberThèses est de contribuer efficacement au rayonnement des travaux des étudiants et de la recherche scientifique dans nos universités. Il était normal que nous nous interrogions sur la notion de recherche d'information sur Internet. Cela nous a permis de dégager certaines méthodes de description et d'encodage de documents qui permettront un repérage efficace des documents pertinents. Mais, bien plus que cela, notre analyse nous a permis de constater que plusieurs limites des systèmes actuels de repérage d'information sur Internet sont des freins à un processus de recherche d'information à caractère scientifique dont les thèses font partie. Examinons ces limitations et les solutions que nous envisageons pour les dépasser.

Les métadonnées: un outil pour la recherche de documents

Depuis bien longtemps, les professionnels de l'information et de la documentation, et plus particulièrement les indexeurs, savent qu'il est impossible de retrouver un document ou une information s'ils ne sont pas décrits à un moment donné au moyen d'un outil qui permettra ultérieurement de les repérer et de les localiser. Ainsi, de la même façon que les livres des rayons de la bibliothèque sont inscrits au catalogue et que des articles de journaux sont répertoriés au sein d'index, les thèses de doctorats doivent être dotées d'un outil qui leur permette d'être retrouvées par leurs lecteurs potentiels.

Il existe plusieurs façons et méthodes pour décrire un document, mais, de façon générale, les descriptions physique et intellectuelle sont regroupées au sein d'un ensemble d'informations qui regroupe *a minima* l'auteur, le titre et la date de publication. Ces données de base, plus connues sous le nom de «métadonnées», constituent autant de points d'accès grâce auxquels un document peut être facilement identifié. Le concept de métadonnées - information à propos d'information (*data about data*) - n'est pas nouveau puisque ce terme remonte aux années soixante[6]; mais il a pris son envol avec l'arrivée d'Internet et plus particulièrement du Web.

Jusqu'à présent, Internet n'avait pas fait l'objet d'études sérieuses permettant de connaître exactement son ampleur ni surtout le taux d'indexation des millions de pages actuellement disponibles. Dans une récente étude publiée dans la revue *Nature*[7], Lawrence et Giles estiment à environ 800 millions le nombre total de pages sur le Web. Moins de 6 % d'entre elles proviennent des milieux scientifiques et éducationnels. Encore plus alarmant: les auteurs constatent que seulement 34,2 % des pages d'accueil visitées contiennent des métadonnées minimales et que la formalisation de ces métadonnées est des plus anarchiques: plus de 123 balises différentes ont été recensées au sein de leur échantillon.

Cela indique clairement un manque de standardisation en ce qui concerne la description des documents que l'on trouve sur le Web, ce qui a pour conséquence directe l'insuffisance quantitative et qualitative de l'indexation du Web. Si en plus on considère que le moteur de recherche qui a la meilleure couverture ne recense que 16 % de ces quelque 800 millions de pages mises en ligne, il apparaît urgent de se doter de méthodes et d'outils qui facilitent l'accès aux documents du Web, notamment aux thèses.

Pour ce faire, certains groupes de professionnels de l'information se sont déjà constitués au cours des dernières années afin de proposer un schéma commun de description pour les documents accessibles sur la Toile. L'initiative la plus avancée, mais qui est encore peu implantée (elle ne concernerait encore que 0,3 % des documents sur le Web, selon Lawrence et Giles), réunit un ensemble de quinze balises descriptives connues sous le nom de *Dublin Core*. Le tableau 1 présente ce jeu de balises et le tableau 2 les extensions locales que nous proposons afin de mieux couvrir les spécificités liées au repérage des thèses en format électronique.

6. Paul Miller. *Metadata for the masses.* 1996 <URL: http://www.ariadne.ac.uk/issue5/metadata-masses/>
7. Steve Lawrence et C. Lee Giles. 1999. Accessibility of information on the web. *Nature* n° 400: (8 July) 107-109.

Tableau 1 : Les 15 éléments du Dublin Core (DC1)

Élément	Description
TITLE	Titre de la ressource tel que donné par l'auteur (CREATOR)
CREATOR	Nom de la personne physique ou morale responsable du contenu intellectuel de la ressource
SUBJECT	Mots-clés décrivant les grands sujets abordés dans la ressource
DESCRIPTION	Résumé de la ressource ou une description dans le cas d'éléments visuels
PUBLISHER	Nom de l'entité responsable de la diffusion de la ressource dans sa forme présente
CONTRIBUTOR	Personne physique ou morale ayant contribué de façon significative mais secondaire à la ressource
DATE	Date à laquelle la ressource a été rendue disponible dans sa forme actuelle
TYPE	Catégorie de la ressource (exemple : thèse, poésie, essai, etc.)
FORMAT	Format des données (exemple : text/html)
IDENTIFIER	Identificateur unique de la ressource (exemple : adresses url, urn)
SOURCE	Identification de la source de laquelle est dérivée la ressource dans sa forme présente (exemple : ISBN d'un livre dont est issue la version html)
LANGUAGE	Langue(s) utilisée(s) par la ressource
RELATION *	Relation de la ressource avec d'autres ressources
COVERAGE *	Caractéristiques spacio-temporelles couvertes par la ressource
RIGHTS *	Mention de droits d'auteur

** Éléments dont la spécification est en cours d'étude*

Tableau 2 : Les extensions locales pour les thèses

Élément	Description
AVAILABILITY	Information sur l'accès à la thèse (exemple : free, restricted)
FACULTY	Nom de la faculté
DEPARTMENT	Nom du département
JURY	Noms des membres du jury de la thèse

Référencement permanent des thèses : de l'URL à l'URN

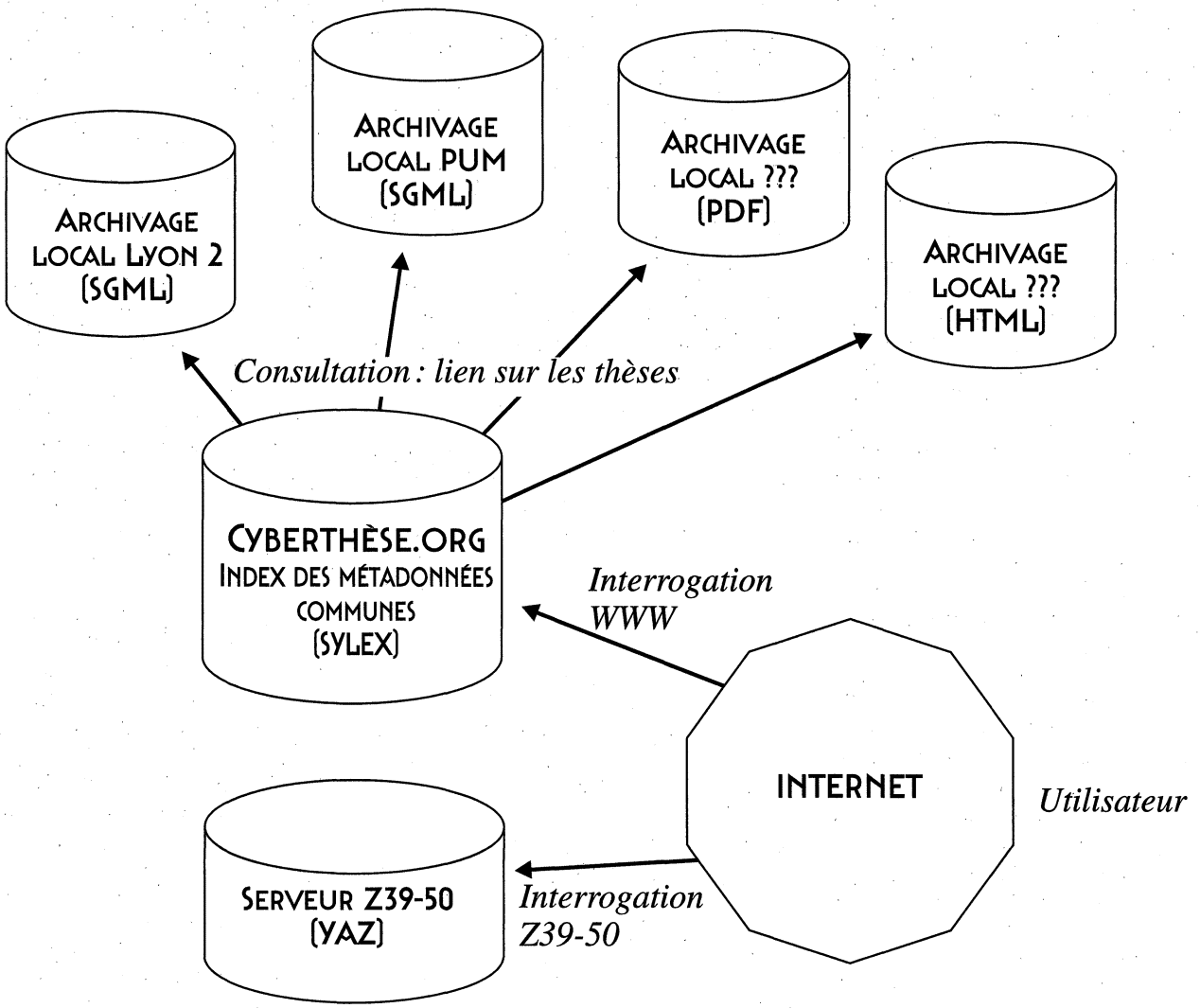
Les mécanismes actuels de référencement par liens hypertextuels sur Internet utilisent une syntaxe qui correspond à la localisation physique d'une ressource donnée. Cette syntaxe est définie par la RFC 1738[8] et est connue sous le nom d'Uniform Resource Locator (URL). Elle présente certains inconvénients que nous sommes souvent à même de constater : qui n'a jamais rencontré, par exemple, la fameuse erreur « HTTP 404 Not Found »[9] qui signifie que le serveur ne trouve pas la localisation de la ressource demandée ? Cela ne signifie en rien que cette ressource n'existe plus sur ce serveur, car elle peut avoir été simplement déplacée vers une autre localisation sur le même serveur ou sur un autre. Comme il n'y a aucun moyen automatique de mettre à jour l'URL d'une ressource qui a été transférée à un autre endroit, il est tout à fait compréhensible que l'on rencontre fréquemment cette fameuse erreur HTTP.

Pour faire un parallèle, une URL correspond à l'adresse postale d'une personne tandis que l'URN (Uniform Resource Name)[10] correspond à son numéro d'assurance sociale (sécurité sociale). Ainsi, ce numéro est attaché à une ressource et non à une adresse physique. En connaissant ce numéro ou identificateur, il est donc possible de remonter à cette ressource même si son adresse physique n'est plus la même qu'au moment où elle a été référencée. On comprend alors facilement pourquoi une URN est indispensable pour le référencement permanent des ressources sur Internet.

Dans le cadre du projet CyberThèses d'édition et de diffusion électroniques des thèses, nous avons mis en place un système de production d'URN fondé sur le modèle proposé par le CNRI (The Corporation for National Research Initiatives)[11]. Ainsi, un serveur global basé au CNRI gère des *Naming authorities* qui font référence à des numéros d'éditeur. Un serveur local est installé chez l'éditeur et héberge, quant à lui, une base de données qui assure la gestion des associations entre les URN et les URL. Tout ceci ressemble beaucoup au système mis en place par InterNic pour la gestion des DNS[12] qui régissent l'adressage IP des ordinateurs reliés à Internet, à ceci près que le référencement se fait ici vers des ressources (documents HTML, SGML, XML, PDF, multimédias, etc.) et non vers des ordinateurs.

8. Pour plus d'explications techniques sur les principes qui régissent les URL, on pourra consulter la RFC 1738 à <URL : http://www.faqs.org/rfcs/rfc1738.html>
9. Pour plus d'explications sur le protocole de transfert HTTP (*HyperText Transfert Protocol*), on pourra consulter la RFC 1945 : <URL : http://www.faqs.org/rfcs/rfc1945.html>
10. Pour plus d'explications techniques sur les principes qui régissent les URN, on pourra consulter la RFC 1737 à <URL : http://www.faqs.org/rfcs/rfc1737.html>
11. Voir le site <URL : http://www.cnri.net>
12. *Domain Name Server* ou serveur de nom de domaine. C'est par ce programme que sont gérés les noms de domaine d'Internet. Pour une explication plus détaillée, se référer à la RFC 882 à <URL : http://www.faqs.org/rfcs/rfc882.html>

Schéma 1 : Architecture du serveur CyberThèses



Le modèle proposé par le CNRI est le système *handle*[13]. Ce système sert aussi de support au système à valeur ajoutée de la DOI (*Digital Object Identifier*) Foundation[14]. La construction du *handle* se découpe en deux parties. Le préfixe de l'URN correspond au numéro d'éditeur (le numéro d'éditeur des Presses de l'Université de Montréal, par exemple, est le *1012*). Ce numéro est unique et ne peut être utilisé par aucun autre éditeur. Des *sub-names* peuvent être ajoutés derrière ce numéro pour subdiviser en unités plus précises (par exemple : « *thèses* »). Cette séquence est suivie du caractère barre oblique (« / ») et d'une séquence alphanumérique au choix de l'éditeur.

Pour constituer l'identificateur URN de type *handle* d'une thèse, nous avons pour notre part choisi l'année de soutenance de la thèse considérée, le nom de l'auteur, l'année de naissance de celui-ci et le format du fichier. Cet URN se lit ainsi : *hdl:1012.Theses/1999-Albert.Mathieu (1959)-[HTML]*

Cette méthode de référencement est encore expérimentale pour les thèses de l'Université de Montréal et sera bientôt implémentée à Lyon 2 dans le courant de l'année prochaine. Elle mérite cependant notre attention puisqu'elle offre une solution pour l'archivage à long terme de documents qui, contrairement à la plupart des pages web classiques, demeureront figés dans le temps.

13. Pour plus de précisions, voir à <URL : http://www.handle.net>
14. Pour plus d'information sur la DOI Foundation, consulter le site <URL : http://www.doi.org>

Le programme CyberThèses

Un de nos objectifs était la création d'un point d'accès à l'ensemble des thèses converties au format SGML à Lyon et à Montréal par le biais d'un site unique indexant nos métadonnées. Dans un souci d'ouverture, nous avons décidé d'élargir la participation à ce serveur à tous les établissements d'enseignement supérieur diffusant leurs thèses en texte intégral sur Internet, sans contrainte ni restriction relatives à la langue utilisée ni au format de diffusion choisi.

Le serveur CyberThèses fonctionne selon un mode distribué. Chaque établissement partenaire assure lui-même la mise en ligne de ses thèses sur son propre site et produit les métadonnées correspondantes. Aucune contrainte n'a été imposée ni sur le format sous lequel sont diffusées les thèses ni sur leur mode de production. Seules les métadonnées doivent obéir à un schéma commun. Le serveur n'héberge que ces métadonnées et le lien vers le document en ligne. Leur gestion et leur indexation sont assurées par un système de gestion de bases de données: SILEX (Serveur d'indexation lexicale, © SGBI Entreprise)[16]. Il autorise l'interrogation soit directement par le Web, soit au moyen d'un serveur client Z39-50. L'architecture globale du serveur CyberThèses est résumée par le schéma à la page 188.

Outre le serveur de métadonnées, CyberThèses se veut un lieu d'échange et de discussion. L'ensemble des ressources techniques et pédagogiques mises en œuvre y sont à la disposition des établissements d'enseignement supérieur désireux de participer à ce programme.

La mise en place du site <www.cybertheses.org> permet de structurer la coopération naissante en s'appuyant sur les modes de fonctionnement et de répartition des compétences propres à Internet. Notre souhait est qu'il puisse rapidement assurer une meilleure diffusion des travaux de recherche effectués au sein des établissements partenaires et constituer ainsi un outil efficace pour l'ensemble de la communauté des chercheurs.

Afin d'améliorer les possibilités offertes à chacun des établissements membres d'accéder aux serveurs de documents et aux documents eux-mêmes, sans qu'ils soient soumis aux aléas du réseau, nous souhaitons généraliser et développer leurs sites miroirs, au moins par grand secteur géographique: Asie, Afrique, Amérique(s), Europe. Cette solution aurait pour conséquence non négligeable, en référence à l'actualité bibliothéconomique lyonnaise[17], de multiplier les sauvegardes de chaque site à travers un réseau mondial de serveurs. La diffusion électronique de la littérature savante doit aussi prendre en compte cet aspect de la réalité.

Au-delà de la francophonie, au-delà des thèses

À la fin de l'année 1999, le programme de coopération franco-québécois respecte le calendrier que ses promoteurs avaient fixé. La chaîne de production est stabilisée, le serveur de diffusion cyberthèses.org est opérationnel, et les deux pôles sont entrés dans une phase de production « industrialisée », en même temps que la formation aux doctorants se met en place. Il reste encore beaucoup à faire, notamment l'intégration en amont des différents formats de traitement de texte utilisés (particulièrement en sciences), et les développements logiciels qui permettront de s'affranchir des formats propriétaires en s'appuyant sur les logiciels dont les sources sont libres. Il s'agit d'un chantier en évolution constante.

Nous avons fait la preuve que la mise en commun de deux pôles de compétences permettait d'avancer de manière conséquente; la qualité des contacts que nous avons noués avec d'autres universités francophones ou appartenant à d'autres ensembles linguistiques nous montre que notre projet recueille l'intérêt et l'adhésion. Il faut multiplier et généraliser cette logique d'intelligence répartie qui permet à chacun de donner ce qu'il a de meilleur et faire rebondir le processus.

La tenue à Paris, fin septembre 1999, sous l'égide de l'Unesco, d'un groupe de travail sur le thème des « thèses électroniques » a confirmé que ce mouvement de diffusion électronique des thèses était en train de s'étendre à toute la planète et qu'une coopération était souhaitée au-delà du monde francophone. Une action de formation et d'implantation de la chaîne de production pilotée par l'Université Lyon 2 a eu lieu à l'Université du Chili (Santiago) à la fin du mois de novembre. L'Université du Chili joue maintenant le rôle de formateur et d'expert pour la diffusion de cette chaîne de production dans le continent sud-américain.

Les choix logiciels, aussi bien en production qu'en diffusion, sont tous tournés vers SGML/XML; l'indexation et le signalement à partir des métadonnées, unanimement proposés, valident *a posteriori* les choix de départ des promoteurs du projet CyberThèses.

L'engagement des universitaires dans un processus autonome de production et de diffusion électroniques des thèses permet d'envisager des développements importants et coordonnés dans le domaine de l'édition électronique savante; ils pourraient esquisser les contours d'une nouvelle économie politique de la connaissance. Si cette coopération franco-québécoise a pu y contribuer, elle aura atteint ses objectifs.

15. Pour un exemple d'application d'URN pour le projet des thèses de PUM, on pourra consulter le catalogue des thèses de l'Université de Montréal à <URL: http://www.pum.umontreal.ca/theses/> et cliquer sur le symbole rond marqué d'un triangle orange pour activer le référencement. On notera que, pour être en mesure d'utiliser le système handle, l'on doit préalablement télédécharger le module externe du CNRI.
16. Pour plus d'information, consulter le site de SGBI Entreprise à <URL: http://www.cei-sgbi.insa-lyon.fr>
17. Rappelons que la Bibliothèque de l'Université de Lyon a été détruite par un incendie en juin 1999: 350 000 ouvrages, dont la totalité du fonds des thèses, ont disparu.