Cahiers GUTenberg

LA PRODUCTION DE DOCUMENTS ÉLECTRONIQUES STRUCTURÉS À GRANDE ÉCHELLE

Viviane Boulétreau, Jean-Paul Ducasse

Cahiers GUTenberg, nº 35-36 (2000), p. 25-35.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_2000___35-36_25_0>

La production de documents électroniques structurés à grande échelle : la diffusion électronique des thèses universitaires.

Viviane Boulétreau[1] et Jean Paul Ducasse[2]

[1] Chef de projet "Edition Electronique".
SeNTIERS, Université Lumière Lyon 2.
Viviane.Bouletreau@univ-lyon2.fr
[2] Maître de Conférences de l'Institut d'Etudes Politiques de Lyon.
Responsable scientifique du programme "La publication électronique des thèses : pour une diffusion de l'édition savante francophone sur les inforoutes" du FFI.
ducasse@univ-lyon2.fr

1. Introduction

Depuis un an, l'Université Lumière Lyon 2 a mis en place un programme de diffusion électronique de ses thèses. Les enjeux d'un tel programme sont multiples, on citera le rayonnement des jeunes chercheurs, une valorisation de la recherche effectuée au sein de l'Université, et une possibilité d'archivage largement distribué, garantie d'une plus grande sécurité.

Les choix technologiques afférents à la diffusion électronique à grande échelle sont primordiaux. Il est en effet indispensable de garantir trois propriétés à l'information électronique : **sa pérennité, l'efficacité de sa diffusion et sa facilité de consultation**. La première de ces contraintes nous a conduit à privilégier l'utilisation de normes et de standards internationaux, de logiciels libres et à limiter autant que possible notre dépendance vis-à-vis de produits commerciaux. La seconde nous a amenés à étudier les différents modes de signalement existant et à développer un schéma de métadonnées propre à la représentation des thèses. Enfin la dernière de ces contraintes implique la possibilité de générer de façon rapide des documents aux formats de diffusion les plus largement utilisés (HTML aujourd'hui, mais aussi XML).

Le programme mis en place à Lyon 2 repose sur une chaîne de production de documents électroniques structurés développée par la cellule édition électronique

de l'Université en collaboration avec les Presses de l'Université de Montréal. Cette chaîne nous permet de produire une version SGML de chaque thèse (DTD TEI Lite[1] [4]). Le format SGML, norme ISO 8879, constitue une garantie de pérennité et, grâce à la notion de document structuré, un apport qualitatif non négligeable aux thèses [1]. A partir de ce format peuvent en outre être dérivés, de façon totalement automatique, d'autres formats plus adaptés à la diffusion tels que HTML ou XML.

Les choix politiques déjà esquissés par la mention de notre orientation vers les *"logiciels libres"* vont au-delà d'une simple utilisation. Notre volonté est de développer des outils aussi génériques que possible ayant vocation à être mis à la disposition de l'ensemble de la communauté scientifique, à être adaptés, développés et améliorés par tous. Notre ambition est de promouvoir une édition électronique structurée s'appuyant sur des normes, de développer de nouveaux modèles de diffusion de l'information scientifique visant à mutualiser les résultats de la recherche, de mettre en pratique le concept d'intelligence répartie, enfin d'ébaucher le cadre d'une future bibliothèque universitaire virtuelle.

Dans cet article nous présentons dans un premier temps l'ensemble de nos réalisations : chaîne de production de documents structurés, schéma de métadonnées, organisation des serveurs... Nous dresserons ensuite un rapide bilan des développements que nous envisageons à court terme. Enfin nous évoquerons les perspectives d'évolution de nos travaux, à partir d'une coopération répartie à l'intérieur d'un réseau.

2. Réalisations

Dans le cadre de la mise en ligne d'un fonds documentaire important, l'homogénéité des documents diffusés est primordiale. D'un point de vue "production", elle permet une économie en termes de développement et de gestion, d'un point de vue "diffusion", elle est le signe extérieur visible de la politique éditoriale de l'institution et de sa cohérence. Le choix des formats d'archivage puis de diffusion est donc primordial. On constate cependant que les documents que nous recevons aujourd'hui sont loin d'être homogènes de par la diversité des éditeurs de texte utilisés lors de la rédaction : Microsoft Word (versions 2 à 2000), StarOffice, WordPerfect, ClarisWorks, QuarkXPress, LotusWorks, et même, de temps en temps... Latex..., mais aussi de par la variabilité de la maîtrise qu'ont les étudiants eux-même de leur traitement de texte.

1. La version française de la TEI-Lite est consultable à l'URL http ://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/sgml/teintro.html.

Trois critères essentiels ont déjà été mentionnés : homogénéité, pérennité, exportabilité... Trois termes que l'on associe presque immédiatement à l'usage de normes, et dans le monde de l'édition, à la norme SGML/XML [6]. Nous ne pensons pas qu'il soit ici utile de le justifier plus longuement et nous attacherons donc à présenter la mise en oeuvre de ces choix, c'est-à-dire l'ensemble des traitements que nous appliquons au document original fourni par le doctorant pour la production du document SGML.

2.1. Un format intermédiaire : RTF

Nous l'avons dit, les formats d'entrée de notre processus sont multiples. Bien entendu, nous n'avons pas développé une chaîne de traitement complète pour chacun mais avons voulu réduire autant que possible le nombre de formats sur lesquels repose notre travail. Le format *Rich Text Format* (RTF) constitue dans cette optique une solution satisfaisante. Il est sans doute le format d'export proposé par le plus grand nombre d'applications et conserve toutes les informations de mise en page associées à chaque élément du document. Enfin, et surtout, sa syntaxe étant connue, il est possible de développer des automates de conversion vers d'autres formats.

Bien entendu, RTF est un format propriétaire et son utilisation, même comme simple format intermédiaire, nous éloigne de notre objectif : développer une solution logicielle qui repose sur le logiciel libre. Il représente pour nous un compromis temporairement acceptable entre volonté et faisabilité. Il nous semble cependant souhaitable de nous orienter vers une solution qui nous lie moins à la politique commerciale d'entreprises privées.

2.2. Un élément structurant : la feuille de style

Convertir un document en SGML signifie, outre une conversion simple d'un espace de codage vers un autre, un enrichissement du document par le codage explicite de l'ensemble de sa structure. Cette structure n'est souvent pas clairement apparente dans le document original, et sa reconstruction repose sur l'analyse des seuls éléments dont nous disposions : les attributs typographique sou liés de façon plus générale à la mise en page. L'utilisation d'une feuille de style prédéfinie, adaptée au type de document à convertir simplifie énormément cette analyse. La connaissance *a priori* des noms des attributs de style va nous permettre de développer des automates de conversions adaptés à nos documents et dont les résultats seront beaucoup plus fiables qu'une analyse typographique, même fine.

La feuille de style "Thèses" est constituée d'une cinquantaine de styles correspondant chacun à des éléments structurels remarquables : éléments de la

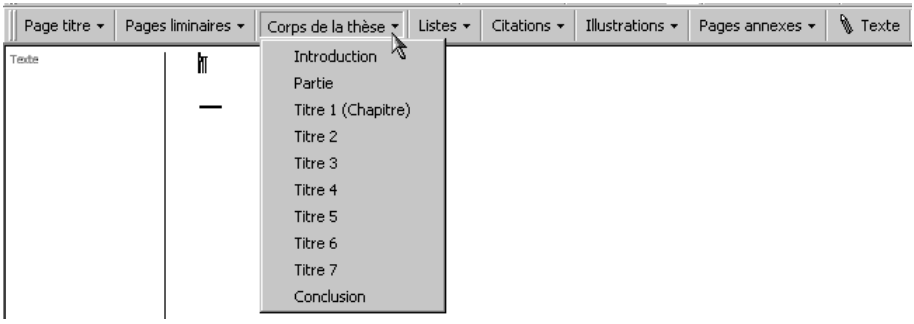

FIG. 1 – La feuille de style "Thèses" : une barre d'outils conviviale.

page de garde, niveaux de titres listes, citations, illustrations, éléments bibliographiques, etc. Pour une utilisation plus conviviale, ces différents styles sont organisés en menus regroupant les éléments de même nature voir figure ci-dessous.

Au-delà de notre problématique de conversion de document, l'usage de la feuille de style s'avère rapidement être une aide précieuse pour l'auteur. Elle lui permet de créer un document de présentation homogène, de générer automatiquement ses tables des matières, listes de figures, de tableaux etc... Enfin, nous sommes convaincus qu'à travers une aide à la structuration physique du document, elle constitue un guide pour sa structuration logique, et permettra donc de clarifier, et d'améliorer de manière sensible, la construction intellectuelle du raisonnement de l'auteur.

2.3. Des automates...

Nous avons développé deux types d'automates correspondant chacun à une nature de conversion : la conversion d'un document plat (issu d'un traitement de texte classique) vers un document SGML que nous appelons *conversion enrichie*, et la *conversion appauvrie* qui permet la réutilisation d'information encodées en SGML (puisqu'il s'agit d'un des principes de base de la philosophie SGML) et leur exportation vers d'autres formats peu ou pas structurés (HTML par exemple) [2]. Avant d'exposer les caractéristiques de ces automates, nous présenterons l'outil que nous avons choisi pour leur développement : le langage Omnimark.

Omnimark développé par la société Omnimark Technologies est un langage de programmation propre à traiter du SGML et du non-SGML. Il s'agit d'un langage de programmation événementiel basé non pas sur des événement d'origine "utilisateur" ou système, mais sur des "événements de données" (data events).

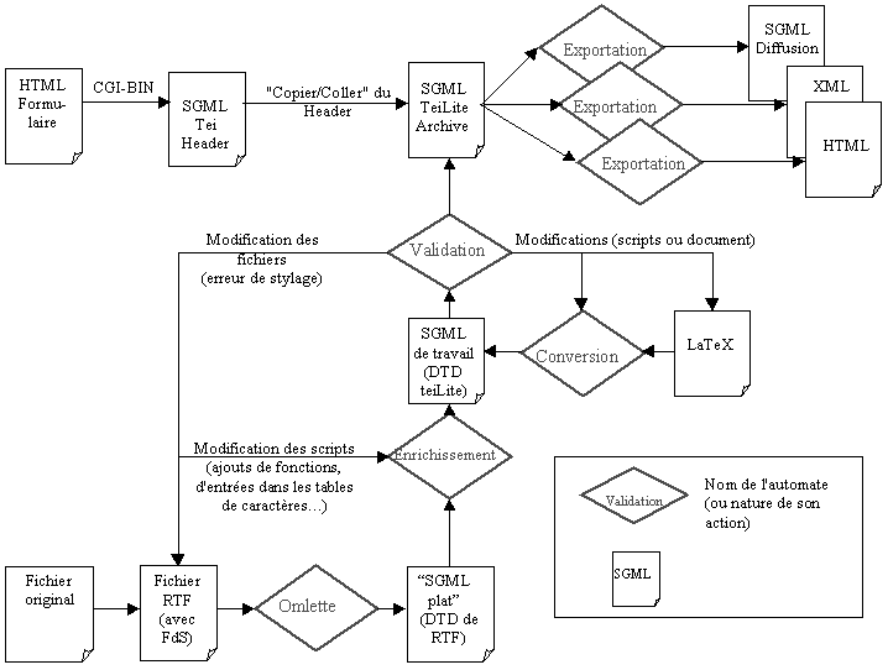

FIG. 2 – Organigramme du processus de production de SGML.

En alliant un système de gestion de flux à un puissant langage d'expressions régulières, il assure la recherche de motifs particuliers et pour chacun d'eux, en fonction du contexte, l'application de règles de conversion ou de production.

Un programme Omnimark se présente donc sous la forme de la donnée d'alphabets de départ et d'arrivée, ce dernier étant l'Universal Character Set (UCS) norme ISO-10646, et d'une suite de règles de production. Soit une grammaire contextuelle de type 1 dans la classification de Chomsky (1957).

2.3.1. ... pour la production de documents SGML

Comme le montre le schéma ci-dessous, l'ensemble de notre processus de production s'appuie en entrée sur un document au format RTF "stylé". La conversion enrichie s'effectue en deux étapes correspondant chacune à un automate distinct.

La première conversion consiste à créer un premier document SGML dont la DTD contient tous les éléments du format RTF. Ce premier automate ("Om-

lette") développé par Rick Geimer[2] n'effectue pas une conversion enrichie, mais il permet d'interpréter les codes RTF et de produire un document SGML valide dont la structure est "plate" puisqu'il est constitué d'une suite linéaire de simples paragraphes. On y trouve en particulier, sous une forme SGML, tous les attributs de mise en page, de polices et de jeux de caractères, styles, tableaux, notes, ... Il s'agit donc d'une DTD attachée à la forme physique du document.

C'est dans une seconde étape qu'un autre automate assure l'enrichissement du document en reconstruisant la hiérarchie de sa structure. Cet automate va non seulement convertir le document d'une DTD axée sur le formatage du document à une DTD axée sur sa structure, mais aussi ajouter, sans intervention manuelle, des liens entre les appels de références dans le texte lui-même ou entre le texte et la bibliographie.

Le fonctionnement de cet automate repose sur une analyse des attributs de style portés par chaque paragraphe ou élément de texte. A chacun de ces attributs correspond un ensemble de règles de productions qui, en fonction du contexte, va permettre de gérer :

- la structure du document (ajout de nouveaux éléments, fermeture des éléments précédents s'il y a lieu...) ;
- l'identification de chaque élément et ses éventuels liens, ;
- les liens avec les entités externes (figures, sons, documents annexes...) ;
- la conversion des caractères en entités ISO10646.

A l'issue de ces deux traitements, on obtient un document SGML non validé : les règles de production construisent en théorie des documents valides, mais leur efficacité est liée à la qualité du document en entrée et plus particulièrement au soin apporté au stylage. Un dernier automate vérifie donc la validité du document produit, c'est à dire sa conformité à la DTD. Il nous permet de corriger les éventuelles erreurs liées aux styles attribués par l'auteur et de compléter notre jeu de règles lorsqu'une configuration nouvelle est rencontrée. L'ensemble du traitement conversion et validation prend environ 5 minutes pour un document de 600 pages. L'insertion manuelle de l'entête SGML (TEI-Header) généré indépendamment par un formulaire cgi-bin complète le document qui servira d'archive et de pivot pour la réutilisation des données (export vers des formats différents, extraction d'éléments d'information...).

2.3.2. ...pour leur conversion vers d'autres formats

Un fichier SGML n'étant qu'un fichier texte balisé, il est très aisé de le convertir en plus ou moins n'importe quel autre format. Pour les besoins de diffusion

2. Rick Geimer est l'auteur de plusieurs applications Omnimark disponibles gratuitement à l'URL <http ://www.xmeta.com>.

des thèses, nous avons construit deux classes d'automates : la première est un ensemble d'outils permettant la production de documents sous des formats de diffusion courants, la seconde permet d'extraire les méta-données (ou signalement) de la thèse. Outre la diffusion des documents en SGML (lisible avec tous les navigateurs possédant le plug-in adéquat), nous produisons des fichiers au format XML et HTML [5]. Le XML permettant une structuration aussi fine que celle du SGML [3], il s'agit d'une conversion simple ; le HTML par contre résulte d'une conversion appauvrie. Il nous est aussi possible d'exporter nos document sous un format LateX et de générer du PostScript, cependant, en raison du contexte disciplinaire (Sciences Humaines et Sociales) dans lequel nous travaillons, nous ne mettons pas en ligne ce dernier type de document.

2.4. Des outils de signalement

La seconde classe d'automates que nous avons développée nous permet de diffuser le signalement des thèses mises en ligne. Il s'agit d'un élément important pour le succès de la diffusion des thèses puisque c'est par l'interrogation de moteurs de recherche, de bases de données que les chercheurs accèdent aux thèses et les consultent. Nous produisons différents formats de méta-données adaptées chacune à un mode de recherche : Dublin-Core pour l'interrogation par le web, Marc pour l'interrogation par des outils classiques de recherche documentaire, texte formaté pour l'ajout piloté dans des bases de données... Le tableau suivant résume les schémas de conversion que nous utilisons.

Alliées à la mise en place de serveurs "portail" dédiés à la diffusion électroniques des thèses et de l'ensemble de la production universitaire, ces métadonnées assurent une bonne diffusion de l'information et un accès simple et rapide aux documents eux-même (voir figure 3). La création du domaine CyberThèses répond à cet objectif particulier. Doté de deux miroirs desservant l'Amérique du Nord et l'Europe (et prochainement d'un troisième en Amérique du Sud), il centralise les signalements des thèses mises en ligne quelles que soient leur discipline, langue ou origine géographique.

Outre la plus grande visibilité des travaux de recherches sur le réseau, la création de tels portails apporte une aide intéressante aux chercheurs en permettant de comparer dans une même requête les mouvements de pensée ou axes de recherches de chaque établissement dans lequel est abordé son domaine. A titre d'exemple, une thèse indexée sur CyberThèses (qui en recense environ 70 au début de mars 2000) est déjà consultée en moyenne une vingtaine de fois par mois...

SGML TEI	HTML DC	Marc
<Title type="main">	DC.Title	245$a
<Title type="sub">	DC.Title.Alternative	245$b
<Author><Name>	DC.Creator.PersonalName	700$a 700$e
<Author><Date>	-	-
<RespStm><Resp> <Name>	DC.Contributor.PersonalName	700$a 700$e
<Publisher>	DC.Publisher.CorporateName	260$b
<PublicaionStmt><Date>	DC.Date.Accepted	260$c
<PublicationStmt> <Availlability>	DC.Rights	-
<Note type="typedoc">	DC.Type	655$a
<Note type="url">	DC.Identifier	856$u
<SourceDesc>	DC.Source	786$n
<Language>	DC.Language	546$a
<Keywords>	DC.Subject	653$a
<TitlePart type="univ">	DC.CreatorCorporateName	710$a
<Div type="abstract">	DC.Description	520$a

3. Développements et perspectives

L'ensemble de ces outils, feuille de style et automates, nous permet de traiter l'ensemble des thèses soutenues à l'Université Lyon 2 puisqu'il s'agit d'un contexte disciplinaire particulier : les Sciences Humaines et Sociales. Cependant, nous sommes bien conscients de ses faiblesses.

La première réside sans aucun doute dans l'utilisation du format RTF comme format d'entrée du processus de traitement. Même si, à ce jour, nous n'avons pas trouvé d'autre solution, la recherche d'un équivalent reposant sur des logiciels libres est une de nos priorités.

La seconde faiblesse de cette chaîne de traitement est bien entendu l'absence de traitement pertinent pour les document Latex. Parmi les difficultés que nous rencontrons avec ce type de format, on citera :

- le grand nombre de distributions utilisés par les auteurs : toutes produisent du latex, mais chacune y apporte des variantes peu compatibles avec l'objectif de développement d'un traitement unique.
- l'impossibilité de gérer de façon fiable la diversité des macro utilisées par les auteurs.

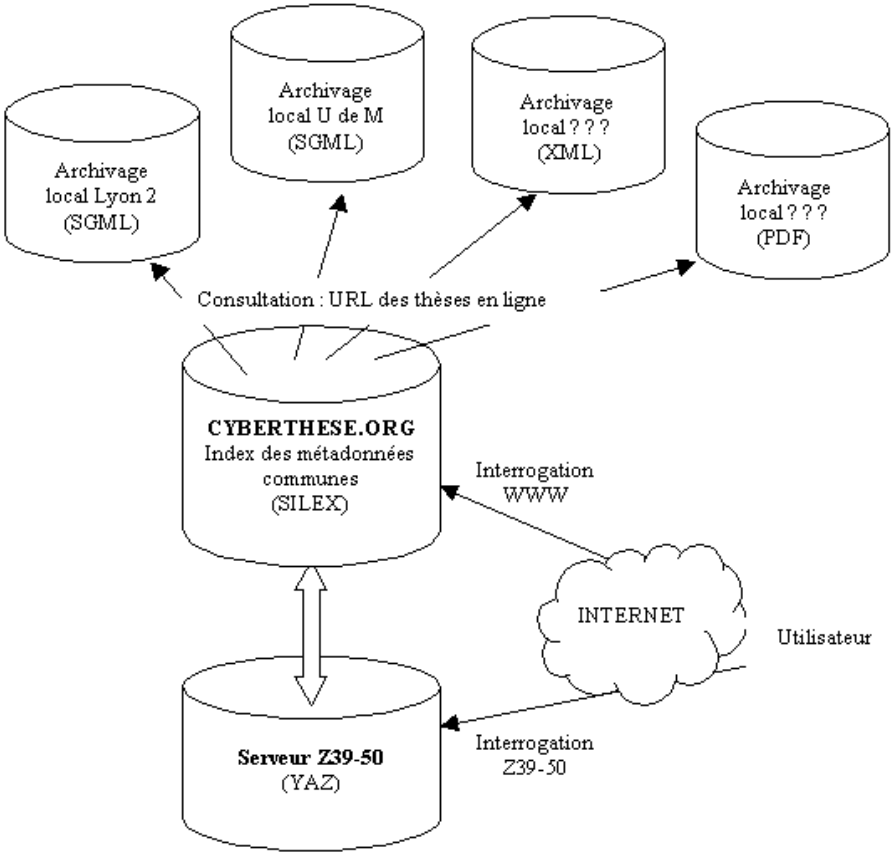

FIG. 3 – Architecture du portail CyberThèses.

Le nombre de documents Latex à traiter à Lyon 2 ne dépassant pas une ou deux thèses par an (soit environ 1 % des thèses), nous n'envisageons pas de pallier à cette lacune de notre chaîne de traitement.

Un de nos axes de travail privilégié est orienté vers la généralisation de nos outils de production : nous pensons que leur utilisation ne doit pas, à terme, entrainer de contrainte pour les auteurs et les éditeurs, aussi avons nous engagé une phase de test et d'adaptation pour assurer leur compatibilité avec le plus grand nombre d'environnements possible :

– l'ensemble de la chaîne de traitement fonctionne sous différentes plateformes : windows 9X et NT, HP-Unix, Solaris et Linux ;

– un premier transfert de technologie financé par l'UNESCO au profit de l'Université du Chili à Santiago nous a permis de vérifier la portabilité de nos outils, leur rapidité d'adaptation à une langue et à des structures de documents différentes et leur relative simplicité de prise en main.

A partir de ces outils génériques, des produits dérivés peuvent être facilement développés, permettant à chaque utilisateur d'adapter ses outils à son propre contexte : changement de DTD, de langue ou d'alphabet, spécialisation de la feuille de style et/ou des programmes de conversion par discipline...

4. Conclusions

Le traitement et la production de documents structurés est la première étape du programme des thèses électroniques en ligne. Le signalement de ces données est l'autre volet complémentaire et indispensable à la diffusion des résultats de la recherche universitaire.

La création du serveur Cyberthèses est une réponse à ce problème. Il fonctionne selon un mode distribué puisque chaque établissement partenaire assurera lui-même la mise en ligne de ses thèses sur son site et produira les métadonnées correspondantes qui seront hébergées sur le serveur central et les sites miroirs continentaux. Ces serveurs n'hébergent que les métadonnées et les liens vers les documents mis en ligne localement. La constitution d'un réseau de producteurs et de diffuseurs permettra également mutualiser les développements futurs qui devraient permettre de solutionner les problèmes liés à la diffusion de documents multimédia. Il faut, en effet, en plus de l'intégration des formats en vigueur dans les sciences "dures" (LateX), envisager le traitement des documents sonores (musicologie par exemple), vidéo, et des textes en caractères spéciaux : alphabets grec, chinois,arabes, polices de caractères propores au traitement linguistique.

Notre objectif est de constituer un espace universitaire public, ouvert à tous, qui bénéficiera, par un effet d'intelligence répartie, des efforts de chacun de ses membres au profit de la collectivité toute entière.

La tenue à Paris, en septembre 1999, sous l'égide de l'unesco, d'un groupe de travail sur le thème de la diffusion électronique des thèses confirme que ce mouvement de production et de diffusion de documents scientifiques structurés s'étend à l'ensemble de la planète.Toutes les énergies ne seront pas de trop pour la généralisation et l'ouverture à tous de ce projet ambitieux.

5. Références

[1] BEAUDRY, Guylaine. La Text Encoding Initiative : les moyens pour ajouter de la valeur à un texte numérisé. In : *Cursus*, ol. 1, n° 2, printemps 1996. Consultable à l'URL <http ://www.fas.umontreal.ca/EBSI/cursus/vol1no2/beaudry.html>

[2] BEN LAGGHA, S. *Modélisation et réutilisation de documents structurés*, thèse de doctorat, Ecole Nationale des Sciences de l'Informatique, Tunis, 1998.

[3] BEN LAGJA, S. SADFI, W. & BEN AHMED, M. Une comparaison SGML-XML. In *Cahiers GUTenberg*, n° 33-34, mai 1999, 28 pages.

[4] BURNARD, Lou & SPERBERG-McQUEEN, C.M. La TEI simplifiée : une introduction au codage des textes électroniques en vue de leur échange. Traduction François ROLE. In : *Cahiers GUTenberg*, n° 24, juin 1996, pp. 23-151.

[5] HUDRISIER, Henri. SGML, HTML, XML : l'ère des machines grammatologiques. In : *Passerelles*, Numéro spécial Recherche Paris 8, n° 24, 1999, pp. 42-44.

[6] VAN HERWIJNEN, Eric. *SGML pratique.* Edition française 1999 : Paris, International Thomson Publishing. 330 pages. 1995.