Département Informatique
Année 1997/1998

Rapport de Synthèse

Serveur de thèses en texte intégral

Marc-Etienne HUNEAU,

Entreprise d'accueil :
Doc'INSA
20, Avenue Albert EINSTEIN
69621 VILLEURBANNE Cedex

Enseignant responsable :
Jean-Marie PINON

Résumé

Doc'INSA, bibliothèque scientifique et technique de l'INSA de Lyon, a souhaité mettre en place un serveur Internet de thèses en texte intégral.
Le format Acrobat d'Adobe ayant été retenu après une étude de l'état de l'art en matière d'édition électronique, les développements se sont organisés autour de la suite Acrobat 3 d'Adobe, dans un environnement Windows. Trois outils ont été développés.
Le premier outil, une chaîne d'édition numérique, permet le traitement par lots des fichiers source fournis par les auteurs dans un format de traitement de texte. Elle génère automatiquement un ensemble de fichiers publiables au format Acrobat (PDF). Ces fichiers comportent un ensemble de liens hypertexte générés automatiquement.
Le second outil est un guide de conversion. Liste de contrôle permettant à l'opérateur de préparer plus facilement un document à une publication en-ligne, elle énumère les opérations et facilite l'accès aux outils de l'édition numérique.
Le troisième et dernier module est une base de connaissances. Celle-ci permet de recenser les incidents de conversion, et leur éventuelle solution. L'opérateur peut chercher une réponse à un problème rencontré dans le texte intégral de cette base de connaissances, et l'enrichir le cas échéant.
La chaîne d'édition inclut également une procédure d'archivage sur cédérom des documents (source et produits).
Le service est d'ores et déjà en phase de test : il s'ouvrira plus largement au public dès que les autorisations de publication auront été obtenues auprès des auteurs.

Mots clefs

INFORMATIQUE, SERVEUR, THESE, BIBLIOTHEQUE VIRTUELLE, DOCUMENT ELECTRONIQUE, TEXTE INTEGRAL, INTERNET, LOGICIEL WORD, LOGICIEL ACROBAT

Abstract

Doc'INSA, science and technology library of the INSA of Lyon, choose to publish thesis on the Internet in full-text.
Adobe's Acrobat file format has been chosen after a study of the state-of-the-art concerning on-line publications. Three software tools have been designed and developed around Adobe's Acrobat existing software, under the Windows system.
The first of these three tools – an electronic publishing line – batch processes 'source' word processor files. The result is a set of ready-to-publish Acrobat (PDF) files. These feature a set of hyperlinks created automatically.
The second tool is a converting guide. Checklist for preparing a document for its conversion, it helps the operator by listing the operations and providing shortcuts to the editing tools.
The third module is a knowledge base. Every conversion tip or trick is added to this base by the operator. One can look for the solution for a problem in this knowledge base, and add new information if necessary.
The publishing line also includes a backup procedure of the 'source' and 'result' files.
This service is already up and running : it will be opened to the public as soon as the publishing authorizations from the authors will be received.

Key words

COMPUTING, SERVER, THESIS, VIRTUAL LIBRARY, ELECTRONIC DOCUMENT, FULL TEXT, INTERNET, MICROSOFT WORD, ADOBE ACROBAT

I. Introduction

L'INSA de Lyon, et sa bibliothèque scientifique et technique Doc'INSA ont souhaité proposer un accès à des thèses numérisées, ou *thèses électroniques*[1].

Doc'INSA, dépositaire officiel des thèses produites dans les laboratoires de l'INSA, reçoit environ 130 documents de ce type chaque année. En supposant que la quasi totalité de ces documents puisse être publiée sur Internet (accord de l'auteur, non confidentialité du mémoire), le volume de données à traiter est conséquent (une thèse à traiter tous les deux jours – en moyenne). Il convenait donc de mettre en place des procédures et de développer des outils efficaces et d'un usage pratique.

Les documents électroniques devaient également être accessibles facilement : l'interface WWW devait permettre la recherche de documents par sujet ou par auteur, par année, proposer des listes…

II. Les documents 'source'

Plutôt que de numériser les versions papier des thèses reçues, il est plus judicieux de demander le dépôt d'une version électronique du document par son auteur.

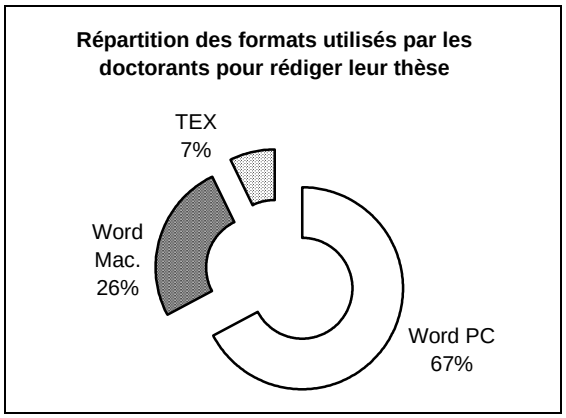

Figure 1 : Résultat d'une enquête menée à l'INSA de novembre 96 à novembre 97, auprès de 124 doctorants

Une enquête réalisée auprès des doctorants (auteurs des thèses) a montré qu'une grande majorité d'entre-eux utilise le traitement de texte Word pour la rédaction de leur document. Quelques-uns emploient T_EX - langage de description de page, plus adapté à la rédaction de documents comprenant des équations ou symboles mathématiques qu'un traitement de texte 'classique' - et peuvent fournir un (ou des) fichier(s) PostScript[2] résultat de leur travail.

Les thèses comportent fréquemment des images, photographies ou graphiques. Ne disposant pas forcément des outils ou des compétences pour intégrer ces éléments au document électronique, certains auteurs incluent (aux ciseaux et à la colle) des documents non numérisés à leur thèse. Le poste de travail destiné à ce projet a par conséquent été doté d'un scanner couleur à plat permettant de numériser ces « extras ».

III. Le dispositif d'édition

La mise en œuvre d'une chaîne d'édition numérique suppose plusieurs phases : le choix d'un média de diffusion (Ici : Internet), d'un format de document (au sens large : format physique et forme des documents) ; le choix d'outils existants et/ou le développement d'outils spécifiques, et la mise en place d'une procédure régissant le cheminement des documents (de la réception des sources à la publication du résultat de la conversion).

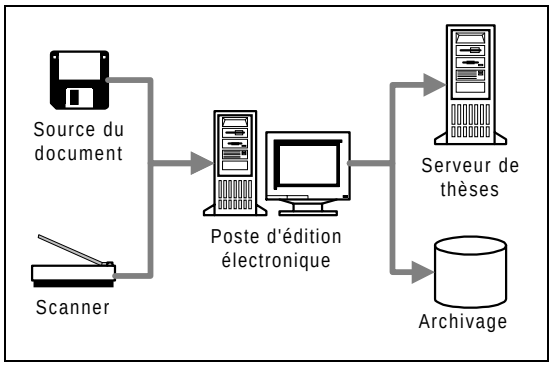

Figure 2 : Vue générale du dispositif

Le choix du format revêt une importance considérable : il décide en partie de l'audience du service (suivant la popularité du format adopté), il conditionne l'aspect des documents, enfin il doit être suffisamment pérenne pour que l'ensemble des documents produits dans ce format reste exploitable.

Le format étant adopté, il reste à le mettre en œuvre. Un service consultable en ligne sous-entend le fonctionnement parallèle de deux systèmes : le système d'édition

[1] *Le terme de thèse électronique (ou de document électronique) – emprunté à des projets similaires – sera employé dans la suite de ce document.*

[2] Langage de description de page développé par Adobe

alimentant le serveur en documents, et le système de recherche/consultation desdits documents.

La partie entrant dans le cadre de ce projet de fin d'études était l'adoption d'un format, le dessin des grandes lignes de l'aspect informatique de la procédure d'édition, et le développement d'une chaîne d'édition numérique.

IV. Choix du format de diffusion

La publication de documents scientifiques peut impliquer certaines contraintes : présence d'images, de graphiques, d'équations, document de taille importante, etc. Le format cible devait donc être adapté à ces contraintes, facile à mettre en œuvre, et suffisamment répandu pour que les documents soient consultable sur la plupart des plates-formes.

Les langages permettant de décrire un document ne manquent pas (HTML[3], SGML[4], XML[5]…). Conçus dans des buts différents, ils présentent chacun des avantages et des inconvénients. Ainsi, si SGML semble - au premier abord - idéal pour la publication de documents scientifiques, sa mise en œuvre reste très lourde, en l'absence d'outils intégrés. De plus, il n'existe pas à ce jour de programme – gratuit - permettant de visualiser facilement un document SGML en ligne.

Le langage HTML, qui ne nécessite pas d'autre programme de visualisation qu'un *butineur* Internet classique, est malheureusement peu adapté à la représentation de gros documents scientifiques (pagination difficile, impossibilité de représenter des équations…). Les versions futures de HTML pallieront peut-être à ces manques.

Le format PostScript est rencontré sur de nombreux serveurs Internet pour la diffusion de documents 'complexes' (mémoires, manuels). Cependant, ce format n'est pas adapté à la lecture en ligne : il est encombrant et est conçu pour piloter une imprimante - et non pour être affiché à l'écran.

L'étude de l'état de l'art a rapidement montré l'hégémonie existant autour du format PDF[6] de l'éditeur Adobe. Basé sur le langage PostScript, il est adapté à la consultation en ligne de documents de taille importante, pouvant contenir des images haute résolution et des données multimédia, il permet de définir des hyperliens (au sein d'un fichier ou vers d'autres fichiers), des repères (sorte de table des matières de liens hypertexte), et il peut être 'optimisé' afin d'être consultable page par page. Enfin, le document peut être protégé contre l'impression, le *copier-coller*, la modification.

Le programme d'affichage des fichiers PDF, Acrobat Reader, est gratuit et disponible sur la plupart des plates-formes. Par ailleurs, le format PDF est un format documenté ce qui lui assure une certaine pérennité.

En revanche, le format PDF est un langage de représentation de page, impropre à l'archivage : Ne comprenant pas la notion de structure logique de document (paragraphes, titres, etc.), il ne peut efficacement servir de source à une éventuelle conversion vers un nouveau format. Une solution d'archivage des documents source (fournis par l'auteur et éventuellement retouchés sur le poste d'édition) a donc été retenue, en attendant l'adoption future d'un format tel que SGML ou XML dans le cadre de ce même projet.

Le choix du format PDF effectué, il convenait d'établir un 'modèle' de document électronique. Cette phase a permis de prendre connaissance des possibilités et limitations de PDF et des outils Adobe Acrobat (création et modification de documents PDF).

V. Choix d'un modèle de document

Lors de la conception de la chaîne d'édition électronique, des choix ont été faits quant au format des documents à produire. Ces choix recouvrent aussi bien des aspects purement techniques (tels que les choix de compression et de codage des images contenues dans les fichiers) que des aspects d'interface (choix des hyperliens, etc.).

La diffusion de thèses sous un format numérique comporte des avantages par rapport à la forme papier : possibilité d'inclure des images haute-résolution sur lesquelles le lecteur pourra *zoomer*, possibilité d'ajouter au document des *hyperliens* vers les notes de bas de page ou notes de fin, vers les références bibliographiques, vers d'autres documents (via une URL[7]), affichage d'une table des matières, ou des vues miniatures des pages, facilitant l'accès à une partie ou à un graphique…

[3] **H**yper**T**ext **M**arkup **L**anguage
[4] **S**tandard **G**eneralized **M**arkup **L**anguage
[5] **E**xtensible **M**arkup **L**anguage
[6] **P**ortable **D**ocument **F**ormat

[7] **U**niversal **R**esource **L**ocator, adresse Internet désignant un document.

Un document[8] à usage interne a été élaboré, recensant les caractéristiques obligatoires, souhaitées ou optionnelles des documents à produire. La réalisation de la chaîne d'édition s'est ensuite basée - entre autres - sur ce document.

VI. Choix de conception

A l'issue de l'étude préalable, il a été décidé de recentrer ce projet de fin d'études sur la conception et le prototypage du seul BackOffice. Trois outils seraient développés, facilitant la création des fichiers PDF à publier.

Le premier de ces outils est une « chaîne d'édition », permettant le traitement par lots de fichiers source et l'ajout *d'enrichissements* aux fichiers produits (liens inter-documents, titres, sujet…).

Le second est une liste de contrôle interactive aidant l'opérateur à effectuer dans le bon ordre les opérations de conversion – et lui apportant une aide contextuelle (accès aux outils, paramétrages-types de ces outils, etc.).

Le troisième - et dernier - outil est une base de connaissances, alimentée par le ou les opérateurs, interrogeable en texte intégral et facilitant la résolution des problèmes pouvant intervenir au niveau de la chaîne d'édition.

VII. La chaîne d'édition

L'application principale a été baptisée CEN (Chaîne d'édition numérique). Cet outil permet la manipulation de projets d'édition, projets rassemblant un ensemble de fichiers source qui seront traités par lot.

Un projet, au sens de l'application décrite, rassemble plusieurs attributs :

- des informations générales sur le document (Auteur, Titre, Date, Mots-clés),
- un ensemble de fichiers source, chacun d'entre-eux ayant en outre un titre (titre de la partie qu'il représente) et un numéro d'ordre dans le document,
- un ensemble éventuel de documents numérisés *en sus*, à intégrer à la thèse et à archiver avec le reste.

L'application pilote Word, Acrobat Distiller et Exchange[9], et modifie par ailleurs directement une partie des fichiers.

La conversion se déroule en quatre étapes (dont les numéros sont repérés sur la Figure 3).

① Tout d'abord, une macrocommande Word (Adobe PDFMaker[i], légèrement modifiée) crée un fichier PostScript enrichi d'instructions *pdfmark*[10] [ii] à l'intention d'Acrobat Distiller. Cette macrocommande crée (le cas échéant) des liens à partir des champs 'note', 'table', etc. Elle crée également un repère Acrobat pour chaque titre.

② Les fichiers PostScript obtenus sont alors directement modifiés par l'application qui y ajoute des repères (toujours via *pdfmark*) désignant les autres fichiers. Il devient alors possible de parcourir toute une thèse sans se soucier de son découpage éventuel en plusieurs fichiers PDF.

③ Les fichiers PostScript sont ensuite convertis en PDF par Distiller.

④ Enfin, les fichiers PDF sont 'retraités' à l'aide d'Exchange : leurs champs titre, sujet, auteur… sont renseignés ; les miniatures de pages sont créées et les fichiers optimisés pour une lecture en ligne (opération permettant au serveur d'envoyer le document page à page).

Dans le cas où les fichiers fournis par l'auteur sont de format PostScript, la première étape (Word) est ignorée.

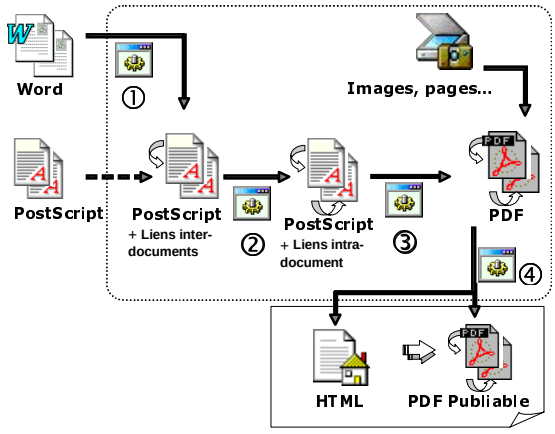

Figure 3 : Opérations de la chaîne d'édition

A ce point, le traitement par lot est terminé, et un rapport de conversion a été généré.

L'application génère en outre un 'pont d'embarquement' vers la thèse, page HTML rassemblant la référence bibliographique du document et des liens vers tous les fichiers PDF. Enfin, elle peut préparer les fichiers à un

[8] Document intitulé 'Règles d'édition des documents électroniques'

[9] Acrobat Distiller et Exchange font partie de la suite Acrobat d'Adobe.

[10] Opérateur du langage PostScript, destiné à Acrobat Distiller

archivage en les rassemblant dans un répertoire.

L'opérateur peut facilement modifier les fichiers (source ou produits) depuis l'application (appel de Word ou d'Acrobat Exchange). Il peut ainsi vérifier la forme des documents Word, ou ajouter des éléments non numérisés ou des séquences multimédia aux fichiers PDF.

Cette application programmée en Delphi 3 dans l'environnement Windows contrôle les autres applications via plusieurs mécanismes : Word et Exchange[iii] sont pilotés via COM/OLE[11], alors que Distiller[iv] est contrôlé par des messages Windows[12].

L'application reproduit le *look & feel* des application Windows 95/NT4, afin de faciliter son utilisation.

VIII. Le guide de conversion

La seconde application développée pour ce projet est un guide de conversion. Il revêt la forme d'une liste de contrôle dont les points peuvent être cochés durant la procédure.

La présence de ce guide de conversion se justifie par le fait que le traitement des documents se fait par lots : Il est important de ne pas omettre d'opération avant de lancer le traitement. Par ailleurs, un certain nombre d'opérations ne peuvent être automatisées, les API[13] des outils Acrobat ne le permettant pas. Ces opérations sont signalées à l 'opérateur par le guide de conversion.

Le guide de conversion n'impose pas l'ordre d'exécution des tâches. En revanche, il est capable de détecter si une modification du projet nécessite d'effectuer à nouveau certaines opérations.

Enfin, le guide de conversion propose un accès facilité aux outils dont peut avoir besoin l'opérateur : programme de numérisation d'images, programmes de lecture d'archives (les fichiers fournis par les auteurs étant souvent compressés), programme de création de cédérom, etc.

Le contenu du guide de conversion doit être facilement modifiable : il est par conséquent implémenté en HTML.

IX. La base de connaissances

Dernier élément du dispositif d'édition électronique, la base de connaissances permet de capitaliser une certaine expertise acquise par le ou les opérateur(s).

Etant donnée une certaine disparité entre les documents fournis par les auteurs (différences de forme, contenus particuliers, types de support), il est inévitable que des incidents surviennent lors de la conversion.

L'opérateur confronté à un tel problème pourra rechercher une solution dans la base de connaissances. Pour ceci, il disposera d'un outil de recherche en texte intégral dans l'ensemble des fiches d'incident référencées.

S'il n'existe pas de solution dans la base de connaissances, l'opérateur rédigera une courte fiche d'incident, qui sera ajoutée à cette base. Lorsqu'il identifiera la solution, il l'ajoutera à cette fiche.

Une certaine forme d'assistance se constituera ainsi lors de l'utilisation de l'application. Dans l'hypothèse ou un nouvel opérateur utiliserait la chaîne d'édition, il pourrait ainsi trouver les réponses à certaines de ses questions.

X. Archivage des documents

L'archivage des thèses électroniques sur cédérom fait partie intégrante de la chaîne d'édition. Les fichiers publiables résultat de la conversion seront ainsi protégés contre la perte de données.

Cependant, afin de conserver le plus possible d'informations sur ces documents, les documents 'source' de l'auteur (éventuellement retouchés) seront également archivés.

Le groupe de travail 'Time & Bits'[v] constatant la disparition quotidienne d'informations numériques (et citant l'exemple de la NASA - qui a perdu les enregistrements de la sonde Voyager), a formulé une liste de conseils pour un archivage sûr.

Appliqué à notre projet, cela revient à : Choisir un format populaire pour l'archivage des documents (ici : Word), enregistrer le plus de métadonnées possible avec les documents (inclusion du rapport de conversion), ajout d'échelles (de couleur et de contraste) aux numérisations de documents.

Ces consignes (format d'enregistrement des documents à archiver, procédures d'archivage) pourront avantageusement être rappelées à l'opérateur par le *guide de conversion*.

Une véritable stratégie d'archivage des documents sera mise en place ultérieurement dans le cadre de ce projet.

---

[11] **C**ommon **O**bject **M**odel / **O**bject **L**ink **E**mbedding : modèle objet de Windows.

[12] Mécanisme de base de communication entre les entités de Windows

[13] **A**pplication **P**rogramming **I**nterface

XI. La question du droit d'auteur

La question de la titularité des droits d'auteur sur un mémoire d'étudiant n'est à priori pas simple : une thèse est le fruit d'un travail dirigé par un enseignant, et est à ce titre une œuvre complexe. Toutefois, la jurisprudence a tranché, considérant que le rédacteur d'une thèse doit en être considéré comme l'auteur unique[vi].

La thèse est un document déposé : ce dépôt est obligatoire. L'auteur (titulaire des droits moraux sur cette thèse) peut alors en autoriser expressément la diffusion. En l'absence de cette autorisation, la thèse ne sera pas diffusée.

Chaque usage de la thèse doit être expressément autorisé par son auteur. L'auteur doit par conséquent autoriser la diffusion de son document sous forme électronique avant que le document puisse être publié sur le serveur de thèses.

Cette autorisation de diffusion sur le serveur de thèses de Doc'INSA sera également visée par le directeur de thèse.

XII. Conclusion

Le projet de serveur de thèses en texte intégral est encore au stade expérimental. Cependant, les procédures se mettent en place, les outils peuvent d'ores et déjà fonctionner, et plusieurs thèses ont été converties et mises à disposition[14] des lecteurs sur le serveur[15].

Il est probable que le choix du format Acrobat soit remis en question d'ici quelques années, et que le service évolue vers une nouvelle forme. Ceci n'est pas contradictoire avec la publication de nombreux documents dans l'intervalle, ce qui fera de ce service un des premiers serveurs de thèses opérationnel en Europe.

[14] L'accès à ces documents est pour l'instant limité aux acteurs du projet, en attendant l'autorisation des auteurs pour une publication en ligne.

[15] http://csidoc.insa-lyon.fr/these/

Références bibliographiques

[i] **Adobe** *Adobe PDFMaker 1.0 for Microsoft Word 97* [On-line]
http://www.adobe.com/supportservice/custsupport/LIBRARY/4d9e.htm

[ii] **Adobe Developer Support**, *pdfmark Reference Manual*
Technical Note #5150

[iii] **Adobe Developer Support**, *Acrobat Viewer Interapplication Communication Support Overview*
Technical Note #5164

[iv] **Adobe Developer Support**, *Acrobat Distiller Control Interface Specification*
Technical Note #5158

[v] **Time & Bits**, *Managing Digital Continuity* [On-Line]
http://www.ahip.getty.edu/timeandbits/

[vi] **Marter, Alain**, *A propos du droit d'auteur sur les mémoires et les thèses* [On-line]
http://www.enssib.fr/Enssib/resdoc/droit/droitmem.html