

Rapport de synthèse

Élaboration d'un cahier des charges pour la définition d'un modèle de document pour les thèses numériques de Doc'INSA

M. IOANNITIS Sébastien

Entreprise d'accueil :

Doc'INSA
60, boulevard Niels Bohr
69621 VILLEURBANNE CEDEX
France

Responsable entreprise : M^{me} JOLY Monique

Enseignant responsable : M^{me} RUMPLER Béatrice

Résumé

De nos jours, les concepts de la documentation structurée et ses implémentations techniques permettent d'apporter une sémantique à l'information contenue dans un document, ceci au moyen d'un balisage (ou marquage) spécifique adapté. Un tel balisage permet de séparer l'information réelle de la présentation (i.e. la mise en forme), ce qui facilite les traitements. La chaîne de traitements des thèses numériques de Doc'INSA doit évoluer dans ce sens. C'est dans cette optique que ce projet a été initié et que nos efforts ont été portés sur l'élaboration du cahier des charges de la solution logicielle à mettre en place.

Actuellement, cette chaîne prend en compte les thèses au format MS Word et L^AT_EX et les convertit en des documents au format PDF dans un but d'archivage et de diffusion. Ceci limite le spectre des applications possibles à partir d'un tel format d'archivage attendu le faible sémantisme que l'information au format PDF revêt. Les développements réalisés lors de ce stage ont permis de mettre en place une chaîne de traitements orientée autour du format XML, ce qui garantit que les informations de structure d'une thèse originale seront préservées. Ceci permet également un archivage pérenne des thèses auxquelles sont associées des métadonnées. En outre, l'automatisation de cette transformation et la diffusion des thèses (structurées) sur des supports comme le papier ou la sortie écran en sont davantage facilitées.

Cependant, une telle entreprise ne serait pas sans soulever des problèmes, notamment des problèmes de réticences au changement de la part des doctorants à qui des efforts de structuration de leur thèse seront demandés. Par ailleurs, ces technologies de représentation de l'information posent d'autres problèmes liés à leur caractère récent et à une maturité non encore atteinte en termes de développements.

Mots-clefs

XML, XSLT, XSL-FO, MathML, MS Word, DocBook, PDF, chaîne de traitements, archivage, métadonnées, conversion

Abstract

Nowadays, the concepts inherent to structured documentation and its technical implementations allow the information in a document to convey a semantics which could be "understood" (or processed) by computers. This semantics is denoted by the use of specific markups. Hence, it is possible to separate the contents of a document from its formatting, enabling an easier information processing. Therefore, the present processing line used by Doc'INSA for numerical theses has to evolve in this direction. It is in this perspective that this project takes part and that our effort focussed on the investigation of the requirements for the software solution to put into place.

At the present time, the afore-mentioned processing line takes into account MS Word and L^AT_EX thesis document formats and converts them into PDF for archival and diffusion purposes. Considering the low semantics contained in a PDF document, the number of applications using such a format is therefore limited. The developments carried out during this project have allowed the setup of an XML-based processing line which preserves the original thesis structure. Therefore persistent archival storage for a thesis and its metadata could be achieved more easily. Not only does the processing line allow this transformation to be carried out automatically, but it also allows the diffusion of structured documents through different types of media such as paper or a computer screen output.

Nonetheless, such an undertaking would not be without raising problems, especially pertaining to the possible reluctance from Ph.D. students to change their habits in writing electronic documents. In addition, these technologies of information representation raise other problems related to their novelty, and therefore their lack of maturity in terms of developments.

Keywords

XML, XSLT, XSL-FO, MathML, MS Word, DocBook, PDF, processing line, archival storage, metadata, conversion

Sigles

CALS	<i>Computer-aided Acquisition and Logistics Support</i>
CSS	<i>Cascading Style Sheets</i>
CITHER	Consultation en texte Intégral des Thèses en Réseau
DCMI	<i>Dublin Core Metadata Initiative</i>
DEA	Diplôme d'Études Avancées
DTD	<i>Document Type Definition</i>
Emacs	<i>Eight Megabytes And Constantly Swapping</i>
FAQ	Foire Aux Questions
FOP	<i>Formatting Objects Processor</i>
HTML	<i>HyperText Markup Language</i>
IHM	Interface Homme-Machine
INSA	Institut National des Sciences Appliquées
ISO	<i>International Standards Organization</i>
L ^A T _E X	L ^A mpo ^r 's TeX (TeX pour <i>technology</i>)
MathML	<i>Mathematical Markup Language</i>
OASIS	<i>Organization for the Advancement of Structured Information Standards</i>
PDF	<i>Portable Document Format</i>
PFE	Projet de Fin d'Études
RDF	<i>Resource Description Framework</i>
RTF	<i>Rich Text Format</i>
SDK	<i>Software Development Kit</i>
SGML	<i>Standard Generalized Markup Language</i>
TEI	<i>Text Encoding Initiative</i>
UML	<i>Unified Modeling Language</i>
USDP	<i>Unified Software Development Process</i>
VoiceML	<i>Voice Markup Language</i>
W3C	<i>World Wide Web Consortium</i>
XML	<i>Extensible Markup Language</i>
XPath	<i>XML Path Language</i>
XSL	<i>Extensible Stylesheet Language</i>
XSL-FO	<i>XSL Formatting Objects</i>
XSLT	<i>XSL Transformations</i>

Introduction

Situation

Le projet de fin d'études est effectué en vue de l'obtention du diplôme d'ingénieur de l'INSA de Lyon, parachevant ainsi le cycle entrepris. Il s'est déroulé dans les locaux de Doc'INSA qui est le centre de documentation scientifique et technique de l'INSA de Lyon. Établissement technologique de dimension européenne, l'INSA de Lyon offre une formation de cinq ans après le baccalauréat. Il compte en moyenne 4 000 élèves ingénieurs par année parmi lesquels 800 environ se voient sanctionnés du diplôme d'ingénieur dans l'une des dix spécialités proposées. En sus de la formation d'ingénieur, il est possible de suivre une formation de 3^e cycle (DEA et thèses de doctorat). Par ailleurs, l'INSA est fort de ses 31 laboratoires de recherche et des 24 formations doctorales qu'il propose.

Doc'INSA offre un ensemble de prestations destinées en particulier aux étudiants, aux enseignants et aux chercheurs de l'INSA de Lyon. Ces prestations sont les suivantes :

1. Mise à disposition du fonds documentaire bcal ;
2. Service d'analyse de documents scientifiques ;
3. Service de recherche bibliographique informatisé ;
4. Formation des élèves ingénieurs à la recherche de l'information scientifique et technique ;

5. Service de prêt entre bibliothèques.

Notre¹ stage s'est déroulé dans le cadre de la première prestation sus-dénommée. En particulier, elle concerne la mise à disposition et la consultation en texte intégral des thèses en réseau (projet CITHER).

Intérêt

Un rapport du ministère de l'Éducation nationale, publié en 2000, relève l'urgence de développer en France des dispositifs de production et de diffusion électroniques des thèses. Ainsi ce PFE s'inscrit-il dans cette optique.

Chaque année, environ 120 thèses sont soutenues à l'INSA de Lyon et déposées sous une forme électronique à Doc'INSA qui les met en ligne via son site internet. Pour 95% d'entre elles, les thèses sont soumises dans le format propriétaire Word de Microsoft, le reste étant la part du format L^AT_EX. Les solutions actuelles en termes d'archivage pérenne s'orientent autour des technologies liées au format XML.

Présentation

Ce PFE s'inscrit dans le cadre du projet CITHER et a pour objectif de définir un cahier des charges prenant en compte les trois phases que la figure ci-dessous dénote.

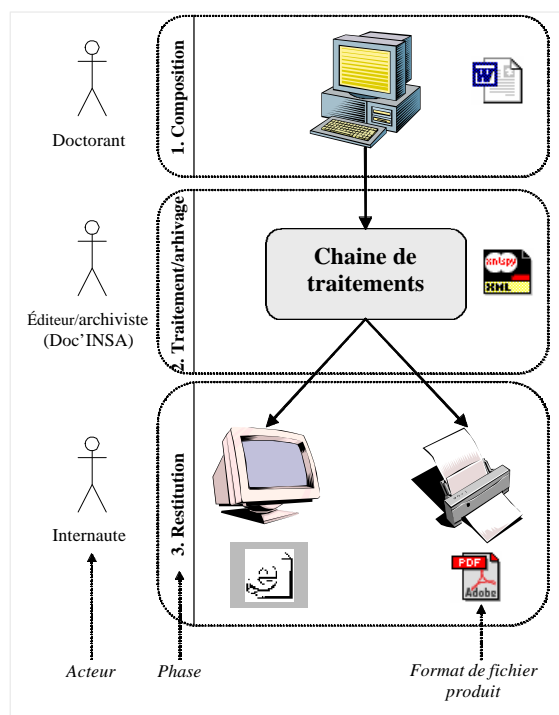


FIGURE 1. - Vue d'ensemble du système.

¹ Dans ce rapport, nous employons le *nous de modestie* qui justifie l'accord (syllepse) des adjectifs et des participes au singulier.

Description des phases :

- *composition* : permet au doctorant de disposer des facilités nécessaires pour la rédaction et la soumission de sa thèse, en plus d'une aide qui pourrait lui être apportée. Durant cette phase, le doctorant se doit de renseigner convenablement certaines métadonnées² utiles pour l'enregistrement de sa thèse, lequel enregistrement est nécessaire pour qu'elle puisse être diffusée ;

- *traitement/archivage* : permet à l'éditeur/archiviste (Doc'INSA) de traiter les thèses envoyées par les doctorants et de les archiver dans un format pivot, en l'occurrence le format XML. Ainsi est-il possible de transformer le document XML obtenu en d'autres formats dits de *restitution* (ex. : PDF, HTML), ceci notamment pour la consultation sur la toile (*web* en anglais). Cette phase permet également de traiter les métadonnées renseignées par le doctorant dans un but administratif et statistique. Celles-ci servent aussi à l'indexation des thèses;

- *restitution* : permet à l'internaute de rechercher et de consulter une thèse dans un format de restitution adapté à ses besoins (ex. : PDF, HTML). Cette phase concerne en particulier l'interface à partir de laquelle l'internaute accède à une thèse et les outils pour qu'il effectue une recherche.

I. Recommandations/standards

XML est un format de représentation structurée de l'information défini par le W3C³, nonobstant son appellation « langage » abusive et messéante de *langage*. XML⁴ est souvent qualifié de métalangage (i.e. un langage permettant de définir des langages). Les technologies élaborées autour de ce format permettent notamment de réaliser la dichotomie entre le contenu et la présentation (i.e. la mise en forme).

XSL (XSLT, XPath, XSL-FO) est un langage d'expression de feuilles de style défini par le W3C. Il comprend trois parties : un langage pour transformer des documents XML (i.e. XSLT) ; un langage d'expressions utilisé par XSLT pour accéder ou référencer des parties d'un document XML (i.e. XPath) ; et, un vocabulaire XML spécifiant la sémantique de mise en forme (i.e. XSL-FO). C'est grâce à XSL que peuvent être réalisées différentes présentations d'un document XML, lesquelles peuvent être destinées à

² Données structurées sur des données ; informations sur une ressource (porteuse d'information).

³ Le W3C a été fondé pour mener la toile à son potentiel maximal en développant les protocoles communs qui favorisent son évolution et assurent son interopérabilité.

⁴ Par ellipse, nous parlerons de *XML* pour parler du format *XML*. Ainsi un document *XML* désignera-t-il un document dans le format *XML*. Nous ferons de même pour les autres formats.

l'impression (ex. : PDF), à la consultation sur la toile (ex. : HTML) ou à d'autres usages (ex. : VoiceML pour la restitution vocale).

MathML est une recommandation définie par le W3C. Elle permet de représenter des formules mathématiques en XML. Ainsi les formules mathématiques écrites avec MS Word ou MathType peuvent être représentées en MathML.

RDF est une recommandation définie par le W3C. Elle fournit un vocabulaire standard pour représenter les métadonnées en XML (cf. Dublin Core). RDF permet l'interopérabilité entre les applications qui échangent des informations.

Dublin Core est un vocabulaire définissant des éléments de métadonnées pour la description de ressources. Il est à l'initiative du DCMI. Utilisé conjointement avec RDF, il est possible d'offrir à ces métadonnées une structure facilitant l'extraction de ces données. Ainsi avons-nous utilisé RDF et Dublin Core pour représenter les métadonnées liées à une thèse et à son auteur.

DocBook est un système d'écriture (i.e. une DTD) de documents structurés en SGML ou en XML dont la maintenance et l'évolution ont été déléguées au consortium OASIS. C'est la DTD que nous avons utilisée pour la représentation des thèses. Nous justifierons le choix de cette DTD dans la Partie V, *Déroulement du projet*.

II. Outils utilisés

Fors l'environnement XML Spy, l'ensemble des outils utilisés appartient au logiciel libre (*open source* en anglais), la plupart provenant du projet *XML Apache* de la fondation Apache.

XML Spy est un atelier de génie logiciel XML développé par Altova. Nous l'avons utilisé pour la programmation XML (XSL, DocBook, RDF). Très complet, il dispose d'outils d'aide à la programmation, comme un moteur d'expressions XPath.

Saxon a été écrit par Michael KAY⁵ de Software AG. Cet outil est ce que l'on appelle un *processeur XSLT* permettant d'extraire et de restructurer des informations présentées dans le format XML. De plus, Saxon intègre un *parser XML*. Ce dernier vérifie qu'un document XML est bien structuré (i.e. si le code produit respecte le format XML). Par ailleurs, il est qualifié de *validant* car il vérifie si un document XML est conforme au schéma (ex. : DTD, XML Schéma) auquel il est lié. Saxon a été choisi pour sa bonne conformité aux recommandations du W3C.

FOP est un outil prenant part au projet *Apache XML*. À partir d'un document XSL-FO, il permet de produire un document dans d'autres formats (ex. : PDF, HTML, RTF).

⁵ Michael KAY est l'auteur de nombreux ouvrages techniques, dont un sur XSLT.

MS Word et VBA : MS Word est le logiciel de traitement de texte le plus utilisé au monde. Il est développé par Microsoft. Dans 95% des cas, les thèses de l'INSA de Lyon sont rédigées avec MS Word. Nous l'avons donc utilisé pour la création d'un modèle de document afin d'aider le doctorant dans son travail de composition. Ce modèle a été amélioré au moyen de macros VBA, langage intégré à MS Word. Nous avons aussi utilisé ce langage pour l'extraction des équations mathématiques et leur conversion en MathML (ceci en conjonction avec MathType), et pour la conversion d'un document MS Word en RTF.

MathType est un éditeur d'équations mathématiques développé par Design Science. Par défaut, MS Word en intègre une version allégée. Il existe un kit de développement (SDK) MathType permettant d'utiliser les fonctionnalités de MathType *via* des macros VBA. En l'occurrence, MathType permet de convertir en MathML des équations écrites avec MS Word ou MathType lui-même.

UpCast est un outil développé par Inifinity-Loop. Il nous a permis de convertir une thèse au format RTF en un document XML selon une DTD propre à UpCast⁶.

Java et Borland JBuilder Personal : Java est un langage orienté objet qui est développé par Sun Microsystems. L'exécution d'un programme Java est indépendante de la plate-forme d'exécution. Il a été utilisé afin d'assembler tous les composants de la chaîne de traitements, et d'offrir à l'éditeur/archiviste une interface graphique permettant l'automatisation et le paramétrage de la chaîne de traitements. Nous avons utilisé JBuilder Personal de la société Borland pour la programmation en Java.

Chaîne (de traitements) de Norman WALSH : Norman WALSH de Sun Microsystems fait partie du comité DocBook d'OASIS. Il a créé l'implémentation XML de DocBook et participé à la spécification du langage XSL. En sus, il a développé ce que nous appellerons la *chaîne (de traitements) de Norman WALSH*. Elle comprend un ensemble de feuilles de style XSL paramétrables (environ 650) qui permettent en particulier de convertir un document DocBook en HTML ou en XSL-FO.

Gemini Solo est un logiciel développé par Inceni Technology. Nous avons utilisé cet outil pour convertir un document PDF en un document HTML/CSS pour préserver sa mise en forme.

III. Description de la chaîne de traitements

La FIGURE 1 a défini les grandes lignes du système. Nous allons maintenant en donner le détail.

⁶ Nous appellerons cette DTD *XML-UpCast*.

La chaîne de traitements se décompose en deux sous-chaînes que nous présentons ci-après.

1. Sous-chaîne *Traitement du contenu*

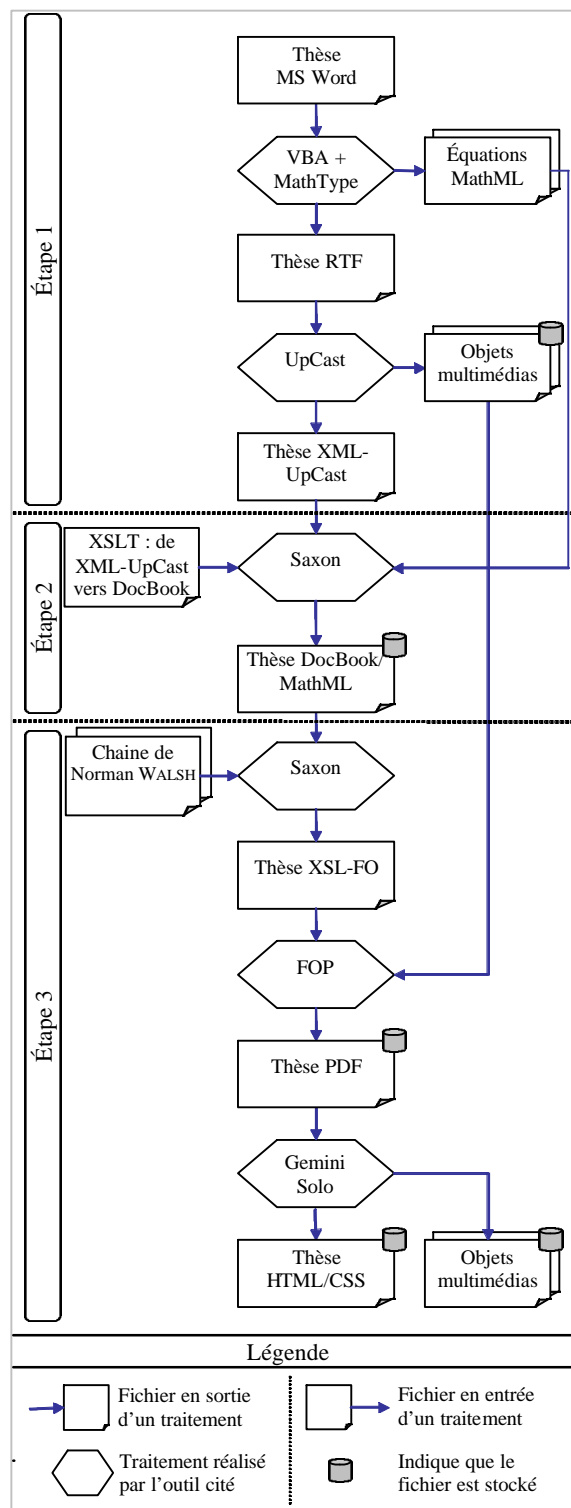


FIGURE 2. - Sous-chaîne *Traitement du contenu*.

Cette sous-chaîne comprend trois étapes :

- *étape 1* : le document de thèse original au format MS Word est traité par un ensemble de macros VBA qui extraient les équations mathématiques écrites avec MS Word ou MathType. Celles-ci sont regroupées dans un fichier, puis transformées en MathML. Afin de pouvoir reconstruire le document final, chaque équation

MathML est pourvue d'un identifiant auquel le document MS Word fait référence. Ce document est ensuite converti au format RTF. UpCast convertit ce document RTF en un document XML-UpCast. Il extrait et sauvegarde aussi les objets multimédias (i.e. images, graphiques, etc.) dans des fichiers distincts. Notons qu'UpCast convertit les tableaux RTF en tableaux CALS, format de tableaux utilisé par DocBook.

- *étape 2* : le document XML-UpCast est converti en DocBook grâce à une feuille de style XSLT que nous avons conçue et qui crée une correspondance – parfois difficile – entre les éléments du fichier XML-UpCast et ceux relatifs à DocBook. Les références vers les objets multimédias sont préservées dans le document DocBook et les équations MathML y sont intégrées ;

- *étape 3* : la chaîne de Norman WALSH associée à un document DocBook permet au processeur XSLT, Saxon, de générer un document XSL-FO. Ce dernier est transmis au processeur XSL-FO, FOP, qui génère un document PDF. La chaîne de Norman WALSH a été modifiée afin de prendre en compte les spécificités de présentation des thèses de l'INSA. Le document PDF intègre les objets multimédias extraits par UpCast. Enfin, Gemini Solo nous a permis de transformer le document PDF en un document HTML/CSS.

2. Sous-chaine Traitement des métadonnées

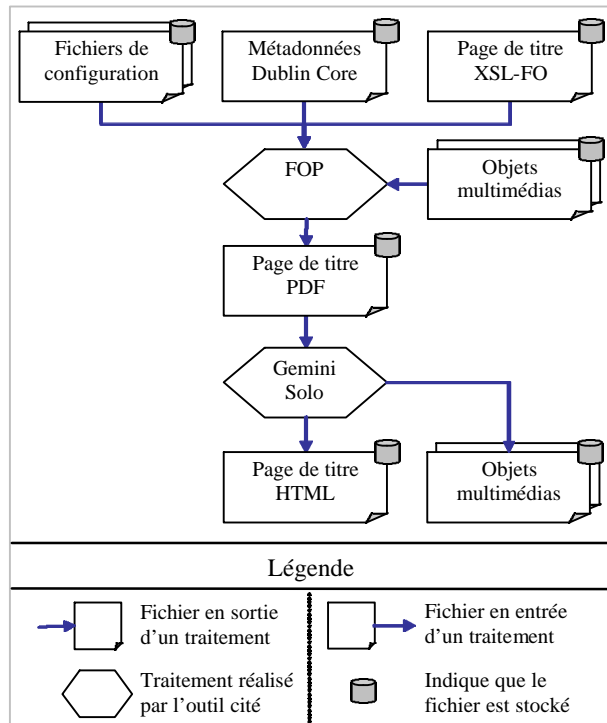


FIGURE 3 - Sous-chaine Traitement des métadonnées.

Dans la sous-chaine *Traitement du contenu*, précédemment décrite, les documents générés ne comprennent pas de page de titre. Au moment de l'écriture de ce document, c'est une sous-chaine dédiée au traitement des métadonnées qui a la charge de générer cette page.

Les métadonnées sont décrites au moyen du vocabulaire Dublin Core et représentées dans le format RDF. Une page de titre écrite en XSL-FO récupère les métadonnées à partir du fichier RDF précité. Le fichier de configuration permet à l'éditeur/archiviste de paramétrer la présentation de certains objets sur cette page. Il se peut que des objets multimédias y figurent, comme le logo de l'école doctorale. FOP produit à partir de tous ces fichiers un document PDF. Son équivalent HTML/CSS est ensuite généré par Gemini Solo.

Par la suite, ces deux sous-chaînes fusionneront ; les métadonnées seront directement incluses au format RDF dans le document DocBook. Cette « ségrégation » entre le contenu et les métadonnées a permis une plus grande facilité et rapidité de développement. Ainsi avons-nous pu obvier aux difficultés de configuration de la chaîne de Norman WALSH, d'autant que les modifications à y apporter sont peu ou prou nombreuses et délicates.

IV. Sources d'informations

Attendu les orientations du projet autour des technologies XML, une grande partie du temps a été consacrée à la recherche d'informations, ceci en utilisant principalement deux sources : l'internet et les ouvrages imprimés.

Internet : les technologies liées à XML, et plus généralement à l'internet, sont en plein essor. Aussi seyait-il que nous nous tinssions au courant de leurs évolutions. De plus l'internet est ici le média de diffusion le plus approprié au contraire des ouvrages imprimés qui n'offrent qu'une vue statique d'une recommandation/standard ou d'un outil, la pertinence de ces ouvrages pouvant être entamée par l'obsolescence qu'ils subissent. Concernant les technologies XML, nous avons principalement consulté le site internet du W3C [W3C02]. Nous avons aussi consulté les sites [ZVO02] et [XML02] pour leur contenu didactique.

Enfin, nous nous sommes inscrit aux listes de diffusion de DocBook, nous permettant non seulement de suivre régulièrement ses évolutions et celles des outils supportant cette DTD, mais aussi d'obtenir une aide lors des problèmes rencontrés.

Ouvrages imprimés : riche d'un fonds documentaire d'environ 85 000 ouvrages, Doc'INSA a constitué, après l'internet, notre seconde source d'informations. Nonobstant l'obsolescence susmentionnée que ces ouvrages peuvent essayer face aux évolutions des recommandations/standards et des outils en pleine « adolescence », ils nous ont été profitables pour l'acquisition des notions fondamentales liées aux technologies utilisées, en particulier les recommandations et les langages de programmation (ex. : XML, XSLT, Java, VBA).

V. Déroulement du projet

Le projet s'est déroulé en suivant la première phase de la méthode USDP, à savoir la phase d'*étude préliminaire*. USDP est un processus de développement itératif, incrémental et basé sur les cas d'utilisation. Il repose sur le langage de modélisation UML.

Phase de veille technologique

Cette phase peut être qualifiée de transversale au projet. Elle nous a permis de suivre les dernières évolutions des recommandations/standards et des outils utilisés.

Phase 1 : étude de l'existant

Durant les deux premières semaines de notre PFE, nous avons effectué une étude de l'existant. Une revue a été organisée à la suite de cette étude afin de constater que notre connaissance de l'existant était à jour.

Phase 2 : collecte des besoins

Cette phase nous a amené à élaborer un questionnaire soumis au groupe de travail des thèses de Doc'INSA et discuté en détail lors de la première revue. Nous avons ainsi pu recenser les besoins pour le système à mettre en place.

Phase 3 : familiarisation avec XML

La phase de collecte des besoins nous a donné une vue d'ensemble du système à concevoir et des technologies à utiliser. Aussi ceci nous a-t-il permis de nous renseigner davantage sur les technologies *ad hoc* et les outils disponibles. Cette phase de familiarisation a concerné plus particulièrement la nébuleuse XML, notamment XML, XSL, MathML et DocBook.

Phase 4 : états de l'art

Cette phase a compris deux états de l'art.

État de l'art des DTDs : il existe trois DTD généralement utilisées pour la représentation des thèses en XML : TEI, ISO 12083, DocBook. Nous avons dû réaliser une étude comparative afin de déterminer celle correspondant le mieux aux besoins de représentation des thèses de Doc'INSA. La DTD DocBook a été retenue pour les raisons suivantes :

- elle répond aux besoins structureaux des thèses ;
- elle est adaptée aux documents techniques ;
- elle est largement répandue ;
- ses éléments sont clairs et bien définis ;
- le support est bon : sites internet, livre de Norman WALSH, listes de diffusion, etc. ;
- de nombreux outils sont développés (en grande partie par Norman WALSH) ;
- les évolutions sont quotidiennes ;
- elle comprend un sous-ensemble pour la description des présentations multimédias qui peut intéresser Doc'INSA par la suite.

État de l'art des projets existants : force projets visant à l'archivage des thèses existent déjà. Toutefois, très peu utilisent XML comme format pivot de représentation, le fait étant que les technologies XML sont trop récentes. La plupart des projets utilisent le format PDF pour l'archivage des thèses, ce qui est actuellement le cas des thèses déposées à Doc'INSA.

Phase 5 : cahier des charges et analyse

Il nous a paru utile d'adopter une vue systémique de la solution à mettre en place. Nous avons défini trois sous-systèmes : le sous-système de composition, le sous-système de traitement et d'archivage, et le sous-système de restitution. Ils ont été compendieusement décrits dans l'introduction. Le sous-système de traitement et d'archivage comprend la chaîne de traitements à proprement parler, les modules pour le transfert et l'archivage des thèses, ainsi que l'interface graphique. Les besoins de ces trois sous-systèmes ont été retranscrits en détail dans le cahier des charges. En dépit de son intitulé, le cahier des charges comprend non seulement l'étude des besoins fonctionnels mais aussi une grande partie de la spécification fonctionnelle du système (i.e. l'analyse).

De plus, il inclut une étude des métadonnées dites *descriptives* et *structurales*, ceci dans le but de concevoir un modèle de document approprié. Ces métadonnées sont utilisées à plus forte raison dans un but administratif et statistique. Elles servent aussi à l'indexation des thèses.

La charge principale de travail a été consacrée à la définition des besoins et à la spécification, le reste ayant été distribué entre la conception et le prototypage, ceci conformément à la méthode USDP. Nous pouvons considérer que 90% des cas d'utilisation ont été définis.

Plusieurs réunions avec le groupe de travail des thèses de Doc'INSA ont été tenues pour qu'il [le cahier des charges] fût examiné, amendé et validé.

Phase 6 : conception et prototypage

La phase de prototypage – en cours – a été définie en fonction des priorités fixées pour chaque sous-système, le sous-système *Traitement et archivage* étant prioritaire car il requiert le plus d'analyse et d'efforts. Viennent ensuite les sous-systèmes *Composition* et *Restitution*.

Conclusion

Résultats obtenus

La chaîne de traitements produit à l'heure actuelle des documents DocBook/MathML à partir d'un document MS Word, ceci en conformité avec le modèle de document élaboré. Elle produit aussi à partir du document DocBook généré des documents PDF et HTML. Restent toutefois certains problèmes liés aux limitations d'UpCast (ex. : gestion des objets sans cadre, dits *floatants*). D'autres éléments n'ont pas été pris en compte, comme les pages à colonnes et les pages en paysage (ou à l'italienne). Enfin, d'autres problèmes sont liés à MS Word, comme le fait qu'un tableau ou un objet ne soient pas associés par un lien logique avec leur légende.

De plus, les modules de la chaîne de traitements sont en train d'être assemblés afin que les traitements puissent être contrôlés *via* une IHM.

D'autres développements restent encore à réaliser, notamment pour que le doctorant puisse enregistrer électroniquement sa thèse auprès de Doc'INSA, et de fait renseigner les métadonnées utiles pour la création de la page de titre.

Perspectives d'évolution de la chaîne

Au moment de la rédaction de ce rapport la phase de prototypage a atteint les 70% des développements nécessaires, la sous-phase *Traitement et archivage* étant pratiquement aboutie.

À court terme, l'on peut s'attendre à ce que les thèses commencent à être transformées en DocBook/MathML, mais ceci ne pourra se faire sans la participation des doctorants qui sont le premier maillon de la chaîne. Aussi une formation *ad hoc* devra-t-elle leur être dispensée.

Par la suite, il sera possible d'envisager que les doctorants rédigent leur thèse directement en DocBook/MathML avec un éditeur adapté comme Emacs associé au module DocBook écrit par Norman WALSH.

Par ailleurs, la chaîne de traitements devra prendre en compte les thèses au format L^AT_EX. Afin de réduire les développements, il serait envisageable de convertir un document L^AT_EX en RTF.

Enfin, l'intégration d'une base de données, telle qu'Oracle 9i qui gère le XML, serait souhaitable - solution que nous avons étudiée lors de ce PFE.

Bilan personnel

Ce projet nous a beaucoup apporté car il nous a permis de mettre en pratique des méthodes ainsi que d'autres acquis souvent restés au stade de la théorie. En outre, il nous a permis de réaliser l'importance des premières phases d'un projet qui établissent les fondations pour son bon déroulement.

L'intérêt de ce projet a été de travailler avec les technologies de la nébuleuse XML et d'en apprécier leur puissance. Toutefois, cet intérêt n'a pas été sans susciter de nombreuses difficultés notamment dues au grand nombre de technologies et outils utilisés.

Le projet étant terminé, nos perspectives sont maintenant d'aboutir à un prototype opérationnel couvrant principalement les étapes de composition, et de traitement et d'archivage, même s'il n'est pas entièrement fonctionnel. Enfin, il convient également de terminer la documentation pour faciliter la maintenance et l'évolution du système.

Références bibliographiques (principales)

- [AMM02] AMMAN (B.) et RIGAUX (P.), *Comprendre XSLT*, Paris, O'Reilly, 2002, 517 pp.
- [APA02] *Site internet d'Apache : projet Apache XML*, 2002, <http://xml.apache.org>
- [CHA98] CHASE (N.), *XML et Java*, Paris, Campus Presse, 1998, 430 pp.
- [JAV02] *Site internet Java de Sun*, 2002, <http://java.sun.com>
- [RAY01] RAY (Erik T.), *Introduction à XML*, Paris, O'Reilly, 2001, 352 pp.
- [W3C02] *Site internet du W3C : recommandations*, 2002, <http://www.w3c.org>
- [WAL02] *Site internet de Norman WALSH : chaîne de N. WALSH*, 2002, <http://nwalsh.com/docbook/>
- [WAL01] WALSH (N.) et MUELLNER (L.), *DocBook, La référence*, Paris, O'Reilly, 2001, 685 pp.
- [XML02] *Site internet de XML.fr : documentation XML, FAQ et articles*, 2002, <http://xmlfr.org>
- [ZVO02] *Site internet de ZVON : ressources didactiques et outils XML*, 2002, <http://www.zvon.org>
- [ROQ02] ROQUES (P.) et VALLÉE (Franck), *UML en action*, Paris, Eyrolles, 2002, 385 pp.

Note de rédaction

Ce document a été rédigé avec un grand souci du bon usage de la langue française. La grammaire, le vocabulaire, ainsi que la typographie utilisés s'y conforment autant que possible. En sus, ce document respecte *Les rectifications de l'orthographe* [REC90] élaborées par le Conseil supérieur de la langue française et approuvées par l'Académie française, rectifications parues au Journal officiel de la République française le 6 décembre 1990.

Références

- [ACA02] *Dictionnaire de l'Académie française*, 9^e éd., 2002, <http://www.academie-francaise.fr>
- [GRE00] GREVISSE (M.) et GOOSE (A.), *Le bon usage*, 13^e éd., Paris, Duculot, 2000, 1762 pp.
- [INF90] *Lexique des règles typographiques en usage à l'Imprimerie nationale*, Paris, Imprimerie nationale, 1990, 196 pp.
- [LAR99] *Le petit Larousse illustré*, Paris, Larousse, 1999, 1784 pp.
- [REC90] *Les rectifications de l'orthographe*, Journal officiel de la République française, n° 100, Direction des journaux officiels, 1990, 17 pp., http://www.academie-francaise.fr/langue/rectifications_1990.pdf
- [TRE97] *Le Trésor de la langue française informatisé*, 1997, <http://atilf.inalfr.fr/tf3.htm>