Accès au contenu des thèses numériques par leur structure sémantique

Rocío Abascal-Mena* — Béatrice Rumpler**

** Universidad Autónoma Metropolitana – Cuajimalpa*
Avenida Constituyentes 1054, Colonia Lomas Altas
Delegación Miguel Hidalgo, 11950 México D. F., México
mabascal@correo.cua.uam.mx

*** INSA Lyon – LIRIS*
7 avenue Jean Capelle - Bâtiment Blaise Pascal
F-69621 Villeurbanne cedex
Beatrice.Rumpler@insa-lyon.fr

RÉSUMÉ. *Les projets de bibliothèques numériques actuels offrent à l'utilisateur l'accès aux thèses à partir d'une recherche qui ne permet pas d'extraire les parties pertinentes de la thèse et ne renvoie que la thèse intégrale. Ainsi, l'utilisateur doit lire des chapitres entiers pour connaître les parties qui correspondent à son besoin. Le projet CITHER (Consultation en texte Intégral des THèses En Réseau) de l'INSA de Lyon dans lequel s'inscrit cette étude, porte sur la mise en ligne des thèses. Nous proposons de permettre un accès pertinent au contenu des thèses grâce à l'utilisation de « tags sémantiques » rajoutés, par le doctorant, au sein de sa thèse lors de la rédaction. L'exploitation de ces tags permet de cibler la recherche et ainsi mieux satisfaire l'utilisateur. Notre travail porte d'une part sur la constitution d'une base de concepts utilisés pour le « tagage » de la thèse et, d'autre part, sur la définition d'un nouveau modèle de documents à partir des différentes structures de la thèse.*

ABSTRACT. *The current projects of digital libraries offer the user an access to the scientific theses that does not make extraction of relevant parts of thesis possible and that returns only the integral thesis. Thus, the user has to read the whole chapters to know which parts of the thesis correspond to his needs. The project named CITHER, of the INSA of Lyon, in which this study is registered, relates to the setting of the theses online. In CITHER the same problem is to be solved. To improve the diffusion of the theses, we propose to give access to its contents thanks to the use of « semantic markups » added into the thesis, by the PhD student, during the writing step. The exploitation of these markups allows a better accuracy of the research contents, in order to satisfy the user better. Our work focuses on defining a new model of documents based on the different structures of the thesis.*

MOTS-CLÉS : *bibliothèque numérique, recherche d'information, métadonnées, modélisation sémantique, XM, thèses scientifiques, traitement automatique des langues (TAL).*

KEYWORDS: *digital library, information retrieval, metadata, semantic modelling, scientific theses, XML, natural language processing.*

DOI:10.3166/DN.10.2.9-35 © 2007 Lavoisier, Paris

DN – 10/2007. Entreposage de documents, pages 9 à 35

1. Introduction

Actuellement on ressent de plus en plus la nécessité de diffuser les connaissances contenues dans les thèses à partir d'Internet. Les projets de bibliothèques numériques actuels offrent à l'utilisateur l'accès aux thèses à partir d'une recherche en utilisant le titre de la thèse, le nom de l'auteur, le nom du directeur de recherche et la date de soutenance. Ce type de recherche ne permet pas d'accéder directement aux parties pertinentes de la thèse mais seulement à la thèse entière ou au mieux à des chapitres entiers. L'utilisateur doit donc lire des chapitres entiers pour savoir si la thèse correspond à son besoin, tâche fastidieuse et souvent décourageante.

Le projet CITHER[1] (Consultation en texte Intégral des Thèses En Réseau) de l'INSA de Lyon dans lequel s'inscrit cette étude, porte sur la mise en ligne des thèses. CITHER a mis en place une diffusion des documents sous forme de texte intégral, en format PDF (Portable Document Format). De ce fait, lors d'une recherche il est impossible de récupérer en une seule requête toutes les parties pertinentes des thèses. Avec CITHER on accède au contenu d'une seule thèse à la fois, par le biais de chaque chapitre. Ces thèses, lors de leur diffusion, contiennent certaines métadonnées. Une métadonnée est un indicateur porteur de sens qui est rajouté au sein d'un document pour souligner une idée, une information. Ces métadonnées proviennent du format Dublin Core [DC, http://dublincore.org] *(« auteur », « titre de la thèse », « laboratoire », « résumé »*, etc.). Le problème réside dans le manque de précision des métadonnées utilisées qui ne reflètent pas suffisamment le contenu d'une thèse. Or ces éléments servent de base pour la recherche d'information. Dès lors, il est évident que pour des recherches simples (c'est-à-dire des recherches se faisant sur un mot-clé), il y aura énormément de réponses, souvent non pertinentes. Le problème ne réside pas uniquement au niveau de la recherche de l'information, mais aussi dans la manière dont ont été indexés les documents. En effet le moteur de recherche n'est capable de travailler qu'avec les éléments qu'on lui fournit, à savoir les mots-clés qui sont censés refléter la thèse. Ces mots-clés ne sont pas des éléments constitutifs d'une thèse, mais bien des éléments externes ajoutés à cette thèse. On a donc d'un coté la thèse et de l'autre les éléments qui la reflètent.

L'objet de notre travail vise à permettre l'accès à l'ensemble des thèses par leur contenu sémantique. Nous proposons un accès au contenu de façon précise grâce à l'utilisation de *« tags sémantiques »* rajoutés au sein de la thèse. Ceci consiste donc, dans une première phase, à définir les *« métadonnées »* (concepts) qui permettraient une description plus fine du contenu des thèses (Abascal *et al.,* 2004a). Pour définir notre nouvelle structure nous allons analyser finement chaque partie de la thèse afin de connaître son organisation liée à la *« structure sémantique »*. Cette structure permettra d'extraire les éléments les plus porteurs de sens (*« modèle »*, *« méthode »*, *« outil »*, etc.). Grâce à ces éléments nous allons pouvoir définir de nouvelles *« métadonnées »*, puis nous pourrons insérer dans la thèse des *« tags sémantiques »*

1. http://csidoc.insa-lyon.fr/these/

correspondant à ces *« métadonnées »*. Ces indicateurs seront eux-mêmes traduits par des balises XML (eXtensible Markup Language) au sein de la thèse.

Pour ce faire, nous proposons de définir un nouveau modèle de documents en nous appuyant sur l'étude des différentes structures de la thèse (structure logique et structure sémantique). Notre approche est fondée sur la modélisation sémantique des thèses scientifiques à partir des concepts issus des thèses. Nous avons utilisé un outil de Traitement Automatique de Langues (TAL), Nomino (Plante *et al.,* 1997), capable d'extraire automatiquement des concepts pertinents d'un corpus. Cette proposition a été implantée dans notre prototype, pour aider le doctorant à baliser sa thèse pendant la phase de rédaction. Ces *« tags sémantiques »* seront exploités pour effectuer une recherche plus performante et plus ciblée.

Après une présentation synthétique de la Recherche d'Information (RI) dans les bibliothèques numériques de thèses, actuelles, nous proposons dans la section 3, notre méthodologie pour l'identification de la structure de la thèse scientifique. Cette méthodologie s'appuie sur l'utilisation d'un outil de TAL dédié à l'extraction de concepts. Puis nous décrivons la base de concepts du domaine créée qui permettra à l'auteur de la thèse de structurer sémantiquement sa thèse. Dans la section 5, nous proposons une modélisation sémantique de la thèse. Nous utilisons des balises provenant de la structure logique mais aussi de nouvelles balises provenant de la structure sémantique. La structure sémantique intègre de nouvelles balises que le doctorant a ajoutées lors de sa rédaction. Ces balises ont été insérées sous la forme de *« tags sémantiques »*. Nous terminons par la présentation de notre prototype et par notre conclusion.

2. La structuration des documents : un état de l'art

La plupart des systèmes de recherche et de consultation de thèses sont assez limités. En effet, ils n'offrent que des recherches à partir de mots de la thèse, ou à partir de valeurs de quelques mots-clés relatifs à une thèse (comme le nom de l'auteur, la date de soutenance, le titre de la thèse). Un autre inconvénient de ces systèmes est de n'être capables d'afficher que des chapitres ou le contenu entier d'une thèse, alors que la recherche porte sur une expression précise. L'utilisateur est ainsi obligé de rechercher lui même, en lisant la thèse, le ou les passages qui l'intéressent. Ces systèmes ont donc deux limites majeures. La première concerne la recherche d'information : ils ne savent rechercher qu'à partir des mots-clés (titre, auteur, date de soutenance, nom du directeur) décrivant une thèse. La seconde limite est liée à la restitution de l'information : il n'y a pas de possibilité pour restituer les fragments ou les parties de thèses les plus pertinents pour l'utilisateur.

L'utilisation de standards associés aux documents pour structurer l'information facilite le partage de l'information contenue dans les documents numériques (Tsinaraki *et al.,* 2005). Cependant, cette utilisation est encore réduite et les études proposées actuellement ne permettent pas d'affirmer que les standards développés

sont bien adoptés par tous les utilisateurs (Heath *et al.,* 2005). Le choix des métadonnées constitue une phase importante pour la structuration d'une bibliothèque numérique. Les métadonnées peuvent être définies comme étant des données relatives à d'autres données (Amerouali, 1999). Une métadonnée permet de donner du sens au contenu des ressources afin que leur localisation et l'interrogation soient plus aisées et plus pertinentes. Pour les utilisateurs, le choix des métadonnées est critique pour trouver l'information pertinente, particulièrement lorsque l'information se situe au niveau du web (Heath *et al.,* 2005). Actuellement, la plupart des bibliothèques numériques commencent à planifier l'utilisation des métadonnées. Dans certaines propositions on utilise des standards tels que : Dublin Core, DTD (Définition de Type de Document) Open eBook (http://openebook.org), DTD DocBook (http://www.oasis-open.org/docbook/), entre autres. Ces standards permettent de définir différentes structures d'un document, telles que la structure logique, physique et sémantique. Par exemple, la structure logique d'un document décrit l'organisation d'un document par chapitres, sections, paragraphes, etc. De plus, cette structure peut être manipulable, indépendamment, du document, en permettant l'identification des différents segments de texte (Goecke *et al.,* 2006). Une des possibilités offertes en décrivant et identifiant la structure logique est, par exemple, la possibilité de trouver des *« anaphores »* aux termes précédemment utilisés dans un document (Goecke *et al.,* 2006). Dans ce cas, le cycle de vie d'une expression linguistique peut être suivi à partir de la structure logique. De cette manière, il est possible de trouver l'anaphore d'une expression. Cette approche s'appuie sur la DTD DocBook pour ajouter des balises logiques aux documents scientifiques et le résultat est extrait en documents XML.

Cependant, les standards associés à la définition des métadonnées, ne permettent pas l'insertion des connaissances propres au domaine (Tsinaraki *et al.,* 2005). De ce fait, la recherche documentaire s'appuie encore sur seulement quelques éléments significatifs du document, à savoir le titre de la thèse, le nom de l'auteur et la date d'apparition. Ces éléments ne permettent pas d'accéder de manière précise au contenu sémantique du document. Afin de résoudre ce problème, quelques auteurs proposent d'introduire une ontologie pour guider, de façon standardisée, la définition de métadonnées du domaine (Tsinaraki *et al.,* 2005 ; Singh *et al.,* 2004 ; Tenier *et al.,* 2006).

D'autres travaux portant sur la création automatique de résumés s'appuient sur la structure de documents (Leskovec *et al.,* 2005). Par exemple, les résumés sont produits à partir d'un graphe sémantique issu du document original et à partir, aussi, de l'identification de la sous-structure afin d'extraire les phrases qui vont composer le résumé. Le graphe sémantique est composé de phrases (sous la forme de triplets) construites à partir de termes reliés, en s'appuyant sur WordNet. Ceci permet d'établir des relations de synonymie entre les termes et de trouver des équivalences entre les autres triplets (Leskovec *et al.,* 2005).

Dans le cas des documents historiques (anciens) le processus de structuration des documents commence lorsque le document est numérisé, et les historiens explorent

le passé à partir des annotations trouvées dans les documents numérisés. Le document ancien est composé de la structure physique, sémantique, structurale, fonctionnelle et légale. Ceci est modélisé en utilisant le format XML (Antonacopoulos *et al.,* 2004).

Dans les architectures client-serveur, le format XML est essentiellement utilisé dans la partie serveur. Pour qu'il soit accessible au client, XML est transformé en XHTML. Afin d'éviter ce processus de transformation, la plupart des utilisateurs préfèrent utiliser directement XHTML pour produire leurs documents. Pour faciliter la tâche de création et d'édition de documents bien structurés et sémantiquement riches, dans la partie serveur comme dans la partie client, on peut utiliser le langage Xtiger (Campoy *et al.,* 2006). Ce langage permet la communication des documents tagués, avec XML, entre le serveur et le client.

La plupart des systèmes qui indexent des documents structurés supportent des documents fondés sur la structure des éléments et ne considèrent pas les documents fondés sur la structure des attributs. Dans ce cas, quand les éléments sont utilisés pour décrire la structure du document, cette structure devient statique et peu extensible (Seung-Kyu *et al.,* 2002). Les auteurs proposent une nouvelle méthode pour indexer les documents fondée sur la structure des attributs. Ainsi, l'information contenue dans une structure fondée sur les éléments et sur les attributs est intégrée afin de produire une structure générale du document.

L'exigence d'une nouvelle génération du web, le web sémantique, est d'intégrer des documents bien structurés afin de rendre la recherche d'information plus précise. McCalla (2004) pense que le web sémantique doit aussi ajouter de la pragmatique au contexte afin de donner à l'utilisateur l'information pertinente. Il propose d'ajouter des balises contenant le profil de l'utilisateur (par exemple, la façon dont l'utilisateur apprend et réagit à certaines situations). La création de documents structurés est une tâche qui peut être faite en partie par l'auteur du document. La conceptualisation et catégorisation de documents sont, à notre avis, des caractéristiques à intégrer dans les systèmes de thèses en ligne.

Face aux problèmes classiques de la structuration d'information, les standards actuels essaient de pallier ces problèmes en apportant du sens aux documents. L'apport du sens doit aussi s'appuyer sur des techniques de TAL. Notre approche est fondée, entre autre, sur l'utilisation d'un outil de TAL capable d'extraire automatiquement des concepts pertinents. Cependant, les outils de TAL nécessitent encore l'avis d'un utilisateur expert pour différencier un concept non pertinent d'un concept pertinent. Ces points font l'objet de notre étude dans le paragraphe suivant.

3. Méthodologie pour l'identification de la structure de la thèse scientifique

Une thèse est un ensemble de données semi-structurées. Elle est composée d'éléments (chapitres, sections, paragraphes, etc.) de taille et de format variables.

Une donnée semi-structurée est une donnée dont la structure est incomplète ou irrégulière (Abiteboul, 1997). Plus généralement, un document semi-structuré comporte à la fois des informations sur le contenu du document et sur son organisation. Une thèse est aussi un document scientifique très codifié. Sa représentation se fait habituellement selon deux structures : la structure physique et la structure logique. Nous proposons d'identifier une troisième structure : la « *structure sémantique »*

3.1. *Sélection d'un outil de TAL pour l'extraction automatique de concepts*

Un des problèmes essentiels pour envisager la structuration sémantique d'une thèse réside dans l'acquisition des métadonnées pertinentes. La plupart des projets de thèses en ligne reposent sur l'existence d'une base de métadonnées associées aux ressources. Le défi d'aujourd'hui est d'essayer d'automatiser le plus possible les processus d'acquisition des métadonnées. Actuellement, l'acquisition des métadonnées est encore une tâche essentiellement manuelle, assez lourde. Il est indispensable que les bases de métadonnées soient complètes et suffisamment renseignées. Il existe, pour des domaines précis, des thesaurus et des ontologies sur lesquels peut s'appuyer le processus d'automatisation de l'acquisition des métadonnées. Cette acquisition est principalement fondée sur l'utilisation des techniques et des outils de TAL.

Parmi les différents types d'outils de TAL dédiés à l'analyse de textes nous trouvons les outils dédiés à l'acquisition automatique de termes. Les termes sont des représentations linguistiques de concepts d'un domaine en particulier. Aujourd'hui les travaux de divers groupes de recherche liés au TAL sont orientés vers la constitution de ce que l'on nomme *« bases de connaissances terminologiques »*. L'objectif de ces travaux est de construire des outils capables d'extraire des connaissances d'un texte et de produire une information terminologique solide (Biebow *et al.,* 1997). Les logiciels d'acquisition automatique de termes utilisent généralement des techniques d'analyse sémantique et des statistiques du corpus en exploitant la fréquence d'apparition de mots. Nous parlerons plus précisément des logiciels fondés sur une analyse morphosyntaxique. Cette analyse considère que *« la construction d'unités terminologiques obéit à des règles de formation syntaxique bien stables »* (Séguéla, 2001).

Les outils d'acquisition automatique de termes reposent sur l'analyse statistique de termes trouvés dans un corpus (Beguin *et al.,* 1997). Ils trouvent un grand nombre de termes et nécessitent une assistance manuelle pour choisir les termes adéquats. Nous trouvons plusieurs exemples d'extracteurs de termes. Afin de choisir un outil pour notre travail, nous avons choisi d'évaluer, selon leurs fonctionnalités requises, quatre outils :

– 1) Copernic Summarizer de NRC (http://www.copernic.com/en/products/summarizer/),

– 2) Nomino de Nomino Technologies (http://www.nominotechnologies.com/),

– 3) TerminologyExtractor de Chamblon Systems Inc. (http://www.chamblon.com/terminologyextractor.htm), et

– 4) Xerox Terminology Suite de Xerox (XTS, http://www.xerox.com/go/xrx/equipment/product_details.jsp?prodID=XTS).

Notre évaluation s'est faite sur un corpus de documents scientifiques provenant du domaine de l'informatique (Abascal *et al.,* 2003b). Le corpus d'évaluation est constitué de 25 documents scientifiques (20 thèses et 5 articles) d'une taille de 1 105 565 mots. Notre démarche s'est fondée sur la comparaison de la liste de concepts produite par chaque outil avec une liste de concepts extraite manuellement par un expert du domaine pour chaque document. Les mesures utilisées pour l'évaluation viennent du domaine de la Recherche d'Information (RI) (Salton *et al.,* 1983 ; Baeza-Yates *et al.,* 1999). Ces mesures sont : la *« précision »* et le *« rappel »*.

	XTS	**Copernic Summarizer**	**TerminologyExtractor**	**Nomino**
Précision	2.8 %	33.9 %	6.8 %	83.4 %
Rappel	90.5 %	51 %	64.8 %	65.1 %

Tableau 1. *Résultats de « précision » et de « rappel » obtenus en appliquant les quatre outils au corpus d'entrée*

Le tableau 1 montre les résultats généraux pour l'analyse de notre corpus. Ces résultats montrent que c'est Nomino qui offre à la fois le meilleur taux de *« précision »* et le meilleur taux de *« rappel »* (taux > 60 %). Suite à cette étude (Abascal *et al.,* 2003b), nous avons retenu Nomino comme étant l'outil le mieux adapté à nos besoins. Nomino est un logiciel développé par l'Université du Québec à Montréal (Van Campendhoudt, 1998), il est fondé sur l'utilisation extensive du pouvoir d'attraction des UCN (Unités Complexes Nominales). Ces unités sont des expressions composées qui permettent de clarifier le sens de certains mots et permettent la structuration du sens. C'est à partir de ces unités que nous définissons les nouvelles métadonnées à ajouter.

3.2. *Analyse des principaux concepts extraits selon les différentes structures de la thèse*

L'analyse des concepts extraits de la thèse nous permet de connaître les parties dans lesquelles l'auteur utilise la plupart des concepts qui caractérisent le mieux la thèse. Notre analyse est fondée sur l'analyse de la structure logique et de la structure sémantique, Nomino sera l'outil d'extraction de termes.

3.2.1. *Analyse des principaux concepts extraits selon la structure logique de la thèse*

Un auteur s'intéresse au contenu du document qu'il rédige, à son organisation, c'est-à-dire à son découpage en composants et aux relations entre ces composants. La structuration est perçue avant tout par l'auteur comme un des critères de bonne présentation visant à apporter au lecteur, un confort pendant la consultation du document. Aussi essaie-t-il d'associer à chaque composant un rôle défini. Une thèse peut alors être découpée en : titre, auteur, résumé, chapitres, sections, annexes. Elle peut aussi comporter des notes, des figures, etc. La notion de *structure logique*, du point de vue de l'auteur, prend également en compte l'ordre des composants. C'est ainsi que le titre d'un chapitre précède toujours son contenu. En effet, non seulement un titre renseigne sur le contenu de la suite mais il contribue aussi à une lecture plus aisée en offrant la possibilité de ne lire que les parties pour lesquelles on a de l'intérêt. Les composants d'un document et leurs relations déterminent l'organisation du document : c'est la *structure logique* du document.

Nous travaillons avec des thèses composées selon les recommandations du ministère de l'Education (Jolly, 2000). Ces recommandations stipulent qu'une thèse peut se décomposer en trois parties : (1) Préliminaires (page de titre, dédicace, remerciements, table des matières, table des illustrations, table des annexes, résumés et mots-clés, éléments d'indexation spécialisée), (2) Corps du texte (introduction, chapitres, sections, paragraphes, conclusion) et (3) Postliminaires (bibliographie, glossaire, index, annexes).

L'analyse des concepts extraits selon la *structure logique* de chacune des thèses est fondée sur le découpage du document en chapitres ou en sections et sur l'organisation de ceux-ci. Nous avons utilisé Nomino pour extraire les concepts correspondant aux différents découpages logiques de la thèse que nous avons retenus: thèse complète, introduction, chapitres, conclusion. Concernant la *« thèse complète »*, nous exploitons, dans notre analyse, uniquement l'introduction, les chapitres et la conclusion. Nous avons supprimé les parties correspondant aux préliminaires (page de titre, liste de professeurs, liste de figures, index, etc.), les parties correspondant aux post-liminaires (annexes, bibliographie, etc.) et les parties relevant des aspects plutôt administratifs (folio administratif, etc.).

Une première analyse a consisté à extraire tous les concepts de chacune des *« thèses complètes »* de notre corpus. Il est important de mentionner qu'en appliquant Nomino avec le *« calcul de saillance »* à l'ensemble du corpus, il y aura des concepts très répétitifs qui ne seront pas extraits. Ce calcul repose sur deux principes: le *« gain à la portée »* et le *« gain à l'expressivité »*. Le principe du gain à la portée stipule qu'une information sera d'autant plus *« payante »* qu'elle est rare. Le gain à l'expressivité, quant à lui, classera les arbres en fonction du caractère spécifique de l'information qui s'y trouve.

Nous avons effectué une deuxième analyse qui a consisté à extraire les concepts de l'ensemble des chapitres (sans l'introduction ni la conclusion). En faisant la comparaison entre le nombre de concepts extraits pour la thèse complète et le

nombre de concepts extraits des chapitres seuls, nous pouvons remarquer que: dans la plupart des thèses, en enlevant l'introduction et la conclusion, le nombre de concepts extraits (des chapitres exclusivement) augmente (tableau 2).

Une troisième analyse a consisté à extraire les concepts qui apparaissent dans des parties très particulières de la thèse (introduction, chapitres et conclusion) afin d'y étudier la répartition des concepts et leur poids. Après observation de la structuration généralement faite des thèses, nous avons considéré que les thèses de 5 chapitres constituaient une organisation souvent retenue. Nous avons calculé la moyenne des pourcentages de concepts qui apparaissent dans chaque partie analysée (table de matières, introduction, chapitres, conclusion), par rapport à l'ensemble de la thèse. Ceci nous a permis de connaître les parties de la thèse qui contiennent le plus grand nombre de concepts. La recherche de l'information pertinente pourra donc s'envisager à partir de certaines sections ou chapitres de la thèse.

Thèse	**Nombre de concepts extraits de la thèse complète**	**Nombre de concepts extraits de l'ensemble des chapitres seulement**	**Thèse**	**Nombre de concepts extraits de la thèse complète**	**Nombre de concepts extraits de l'ensemble des chapitres seulement**
$\mathbf{T_1}$	293	296	$\mathbf{T_{11}}$	50	57
$\mathbf{T_2}$	36	38	$\mathbf{T_{12}}$	36	40
$\mathbf{T_3}$	66	64	$\mathbf{T_{13}}$	46	54
$\mathbf{T_4}$	45	43	$\mathbf{T_{14}}$	47	51
$\mathbf{T_5}$	69	73	$\mathbf{T_{15}}$	81	85
$\mathbf{T_6}$	42	42	$\mathbf{T_{16}}$	23	24
$\mathbf{T_7}$	38	42	$\mathbf{T_{17}}$	36	43
$\mathbf{T_8}$	115	124	$\mathbf{T_{18}}$	17	14
$\mathbf{T_9}$	40	38	$\mathbf{T_{19}}$	29	32
$\mathbf{T_{10}}$	52	54	$\mathbf{T_{20}}$	35	33

Tableau 2. *Nombre de concepts extraits de la thèse complète et de l'ensemble des chapitres seulement*

Le tableau 3 résume les résultats obtenus en analysant des thèses constituées de 5 chapitres (C_1 à C_5). Ce tableau présente la moyenne obtenue dans chacune des analyses effectuées. La table des matières (TM) représente (en moyenne) 9,51 % de concepts pertinents pour toute la thèse. L'introduction (I) et la conclusion (CO) apportent moins de 14 % de concepts chacune. En revanche, les autres chapitres

apportent plus de 20 % de concepts pertinents chacun. Le chapitre 2 est le chapitre qui contient le plus de concepts pertinents. Ce chapitre est la plupart du temps consacré à la présentation des thèmes principaux traités dans la thèse.

TM	I	C_1	C_2	C_3	C_4	C_5	CO
9,51%	12,98%	20,77%	25,93%	22,23%	25,71%	23,60%	13,42%

Tableau 3. *Comparaison de la moyenne de pourcentages obtenus pour chacune des parties qui constituent la structure logique de la thèse*

Nous pouvons conclure de ces expérimentations que les parties correspondant à l'introduction et à la conclusion sont d'un intérêt moindre, puisqu'elles sont seulement un résumé de toute la thèse (Abascal *et al.,* 2005).

L'analyse de la *structure logique* vérifie nos premières suppositions sur l'importance d'analyser essentiellement les chapitres. La table des matières, l'introduction et la conclusion apportent peu de concepts pertinents. Le chapitre 2 apparaît le plus pertinent.

3.2.2. *Analyse des principaux concepts extraits selon la structure sémantique de la thèse*

La structure sémantique sera, elle, capable de fournir des passerelles vers une interprétation cohérente du document. Dans la structure sémantique, les données sont organisées selon leur sens et leur définition respective. La structure sémantique est étroitement liée à la notion de concept. Une thèse regroupe un ensemble de concepts et ceux-ci sont ordonnés. Par exemple le concept *« équation »* est composé d'une *« hypothèse »*, de *« conditions d'applications »* et d'un *« raisonnement »*. Ce dernier utilise des *« variables »* et des *« théorèmes »* pour mener à bien son *« équation »*. Un *« théorème »* a un *« titre »* et un *« auteur »*. Une thèse peut contenir plusieurs *« équations »*. Ainsi, une thèse peut être représentée non seulement par le biais de son articulation physique ou logique mais aussi sous une forme structurée d'un ensemble de concepts. Ces concepts sont représentés par des métadonnées caractérisées selon le contexte d'utilisation du document. Ainsi, une thèse pourra contenir plusieurs *« équations »* chacune ayant la même structure mais traitant de sujets différents. Une thèse peut alors être modélisée sous la forme d'un arbre de concepts (ou de métadonnées).

Dans le but de mieux caractériser le contenu des thèses, nous avons décidé d'utiliser de nouvelles métadonnées ou *«tags sémantiques »* définis à partir des concepts décrivant chacune des thèses. Afin d'identifier des tags pertinents nous avons commencé par extraire des concepts pertinents issus de la thèse. Par *« concepts pertinents »* nous faisons référence à des concepts significatifs de la thèse, capables d'apporter une information pertinente sur le contenu de la thèse.

La définition de concepts associés à une thèse peut s'effectuer manuellement ou assistée d'outils de TAL, ce sera Nomino dans notre cas.

En observant l'organisation des thèses scientifiques on constate que, généralement, la thèse suit un plan dont la structure est fondée sur le découpage logique sous forme de chapitres et de sections. Les chapitres sont eux mêmes souvent en partie liés à une structure plutôt *« sémantique »*. Ainsi on retrouve généralement en début de thèse une partie consacrée à *« l'état de l'art du domaine »*, puis un ou deux chapitres proposant une nouvelle approche, souvent présentée sous forme de *« modèle »* plus ou moins formel. Ensuite vient une partie où sont décrites les *« implémentations »* et la *« mise en œuvre »* des nouvelles *« techniques »* du domaine. Enfin, la dernière partie est plutôt dédiée à la validation et à *« l'évaluation de la proposition »*. Si ce découpage s'appuie principalement sur la notion de chapitres, il apparaît de façon sous jacente, mais bien perceptible, une structure sémantique du document de type *« thèse »*. Nous avons étudié de manière expérimentale comment s'articulaient ces découpages logiques et sémantiques à partir de l'analyse des concepts extraits dans les différentes parties de la thèse.

3.2.2.1. Découpage de la thèse en segments sémantiques

Un *« segment sémantique »* est issu d'un découpage permettant d'accéder au contenu des thèses par le biais des thèmes ou sujets traités, ce qui diffère de la section précédente où l'on s'appuyait sur le découpage logique.

Segments sémantiques	Présentation du segment
Etat de l'art	On le retrouve dans différents chapitres de la thèse mais la plupart du temps c'est le deuxième chapitre qui est consacré à l'état de l'art général. Ensuite on peut trouver dans certains chapitres, des états de l'art plus ciblés comme par exemple: *« état de l'art de méthodes »*, *« état de l'art d'outils »*, …
Méthodologie	On la retrouve pour la représentation d'une démarche proposée en vue de la résolution d'un problème.
Modèle	Ce segment peut se retrouver dans plusieurs chapitres.
Algorithme	Une des approches trouvées dans la plupart de thèses consiste à modéliser un problème en utilisant des algorithmes.
Architecture	Concerne les principales caractéristiques du prototype créé.
Prototype ou Etude de cas	Partie généralement présentée dans les derniers chapitres décrivant l'expérimentation.

Tableau 4. *Quelques « segments sémantiques » d'une thèse scientifique*

En analysant manuellement le contenu des thèses scientifiques, nous avons détecté des *« segments sémantiques »* particuliers traitant de manière ciblée un aspect particulier de la thèse (*« état de l'art », « méthodologie », « modèle »,*

« algorithme », « architecture », « prototype ou étude de cas »). Nous exploiterons quelques uns de ces segments repérés par voie expérimentale (tableau 4), pour proposer notre modèle. Il peut exister d'autres segments à partir desquels une thèse peut être découpée sémantiquement.

La première partie de notre analyse a consisté à découper les thèses analysées (20 thèses du domaine informatique) en *« segments sémantiques »*. Le découpage sémantique varie selon la thèse analysée. Ce découpage provient de notre observation des thèses. Ce type de découpage, une fois validé expérimentalement, nous permettra de modéliser la thèse selon les différentes possibilités de structuration sémantique.

3.2.2.2. Analyse des segments sémantiques à partir des concepts

La deuxième partie de notre analyse a consisté à extraire tous les concepts de chaque *« segment sémantique »* que nous avons défini par observation pour chacune des thèses du corpus analysé. Les concepts ainsi obtenus et validés pourront alors être utilisés en tant que *« tags sémantiques »* et insérés dans le document. Il est important de préciser que toute l'analyse sémantique menée pour les différents *« segments sémantiques »* a été faite manuellement. Seule l'extraction de concepts a été faite de manière automatique en utilisant l'outil Nomino.

Nous présentons par la suite quelques résultats de l'extraction de concepts pour certains *« découpages sémantiques »* distincts : *« état de l'art général »*, *« état de l'art de méthodes »*, *« modèle »* et *« prototype »*.

Thèse	**nombre de concepts issus de l'état de l'art de méthodes**	**nombre de concepts issus de la thèse complète et similaires à ceux de la partie dédiée à l'état de l'art de méthodes**	**Pourcentage de concepts extraits de l'état de l'art de méthodes par rapport à la thèse complète**
T_2	23	7	19,44%
T_6	34	12	28,57%
T_{10}	17	14	26,92%
T_{19}	26	8	27,58%

Tableau 5. *Nombre de concepts extraits pour le découpage sémantique correspondant à la partie de l'état de l'art de méthodes de chaque thèse*

Le tableau 5 présente 4 thèses qui possèdent le *« segment sémantique »* nommé *« état de l'art de méthodes »*. Dans ce cas, le pourcentage de concepts extraits pour l'état de l'art de méthodes par rapport au nombre de concepts extraits pour la thèse

complète se situe entre 19 % et 29 %. Donc plus de 70 % restant de concepts pertinents pour chaque thèse ne se trouvent pas dans l'état de l'art de méthodes.

L' « *état de l'art général »* est l'état de l'art le plus important. Il se situe presque toujours dans le premier ou le deuxième chapitre de la thèse. Cette partie présente tous les travaux du domaine existants sur lesquels l'auteur s'appuie pour traiter son sujet de recherche. En général, cet état de l'art situe la problématique et les travaux de recherche en cours du domaine.

Thèse	**Nombre de concepts issus de l'état *de l'art général***	**Nombre de concepts issus de la thèse complète et similaires à ceux de la partie dédiée à l'état *de l'art général***	**Pourcentage de concepts extraits de la partie état de l'art général par rapport à la thèse complète**
T_1	241	107	36,51 %
T_2	22	6	16,66 %
T_3	51	29	43,93 %
T_5	59	26	37,28 %
T_6	24	5	11,90 %
T_7	24	23	60,52 %
T_8	100	38	33,04 %
T_9	32	17	42,50 %
T_{10}	47	23	44,23 %
T_{11}	56	26	52 %
T_{12}	14	11	30,55 %
T_{13}	46	21	45,65 %
T_{14}	54	20	42,55 %
T_{15}	65	49	60,49 %
T_{16}	26	18	78,26 %
T_{17}	65	15	41,66 %
T_{18}	7	4	23,52 %
T_{20}	25	10	28,57 %

Tableau 6. *Nombre de concepts extraits pour le découpage sémantique correspondant à la partie état de l'art général pour chaque thèse*

Le tableau 6 présente les résultats de l'analyse du segment « *état de l'art général »*. Nous voyons que seulement deux thèses *« T_4 et T_{19} »* des 20 analysées ne contiennent pas ce segment. De plus, on constate qu'en moyenne 40 % des concepts pertinents se trouvent dans ce segment de la thèse. Pour les thèses *« T_3 »*, *« T_7 »*,

« T_9 », *« T_{10} »*, *« T_{11} »*, *« T_{13} »*, *« T_{14} »*, *« T_{15} »*, *« T_{16} »* et *« T_{17} »* nous avons obtenu un pourcentage supérieur à 40 %. Les thèses *« T_7 »*, *« T_{15} »* et *« T_{16} »* contiennent plus de 60 % des concepts pertinents dans ce segment. C'est-à-dire, qu'il suffirait d'analyser ce segment pour trouver un panorama général des concepts qui caractérisent le mieux chaque thèse. Parmi les concepts extraits de ce segment nous trouvons : *« algorithme de compression »*, *« apprentissage à distance »*, *« base de donnée »*, *« échange d'information »*, *« interaction homme/machine »*, *« interface H/M »*, *« système d'exploitation »*, *« système d'information »*, *« travail coopératif »*, etc.

Pour la partie dédiée au *modèle,* nous avons non seulement analysé les chapitres ou sections qui décrivent le *modèle* à utiliser mais aussi les parties décrivant la manière selon laquelle le modèle est appliqué : *modélisation*. Nous parlons de *proposition du modèle* quand l'auteur de la thèse reprend un *modèle* déjà utilisé dans la littérature pour l'améliorer et l'appliquer à la validation de sa recherche. Le tableau 7 décrit les résultats issus de 13 thèses analysées. En général, le nombre de concepts extraits n'est pas très significatif mais la plupart des concepts extraits apparaissent comme pertinents. Par exemple pour la thèse *« T_3 »*, 31 concepts ont été extraits et de ces concepts, 29 sont pertinents.

Thèse	**Nombre de concepts pour le *Modèle***	**Nombre de concepts issus de la thèse complète et similaires à ceux de la partie dédiée au *Modèle***	**Pourcentage de concepts extraits du *Modèle* par rapport à la thèse complète**
T_1	Proposition du modèle : 54	45	15,35 %
T_2	Modèle : 8	6	16,66 %
T_3	Proposition du modèle : 31	29	43,93 %
T_4	Modélisation : 30	30	66,66 %
T_5	Proposition du modèle : 19	14	20,28 %
T_6	Proposition du modèle : 17	16	38,09 %
T_9	Proposition du modèle : 17	17	42,5 %
T11	Modélisation : 20	12	24 %
T_{13}	Proposition du modèle : 20	19	41,30 %
T_{14}	Modélisation : 19	16	34,04 %
T_{16}	Modèle : 6	5	21,73 %
T_{17}	Proposition du modèle : 10	10	27,77 %
T_{19}	Proposition du modèle : 3	2	6,89 %

Tableau 7. *Nombre de concepts extraits pour le découpage sémantique correspondant aux parties « Modèle » de chaque thèse*

La dernière partie de la plupart des thèses du domaine de l'informatique présente un *prototype de validation*. Dans notre corpus, 19 thèses consacrent un chapitre à la présentation du prototype. Généralement, le chapitre consacré au prototype présente aussi la résolution de la problématique et les résultats de la recherche faite pendant la thèse. Le tableau 8 présente les résultats de l'évaluation de ce segment.

Thèse	**Nombre de concepts issus de la partie *Prototype***	**Nombre de concepts issus de la thèse complète et similaires à ceux de la partie dédiée au *Prototype***	**Pourcentage de concepts issus de la partie *Prototype* par rapport à la thèse complète**
T_1	105	60	20,47 %
T_3	3	13	19,69 %
T_4	16	11	24,44 %
T_5	30	21	30,43 %
T_6	18	7	16,66 %
T_7	6	3	7,89 %
T_8	77	45	39,13 %
T_9	19	8	20 %
T_{10}	12	11	21,15 %
T_{11}	23	20	40 %
T_{12}	23	12	33,33 %
T_{13}	8	8	17,39 %
T_{14}	26	19	40,42 %
T_{15}	57	29	35,80 %
T_{16}	15	14	60,86 %
T_{17}	18	12	33,33 %
T_{18}	4	3	17,64 %
T_{19}	6	5	17,24 %
T_{20}	13	12	34,28 %

Tableau 8. *Nombre de concepts extraits pour le découpage sémantique correspondant à la partie PROTOTYPE de chaque thèse*

L'objectif des analyses présentées précédemment est de détecter quels sont les segments sémantiques contenant le plus de concepts, donc les segments contenant le

plus de sens et qui pourront être utilisés pour extraire l'information précise et pertinente.

Pour bien comprendre les résultats décrits dans la section précédente, nous présentons par la suite la comparaison des résultats de l'extraction de concepts pour deux segments sémantiques distincts : *« état de l'art général »* et *« modèle »*. Afin de comparer ces deux segments sémantiques, et de montrer l'importance du découpage correspondant à l'*« état de l'art général »*, nous avons choisi de traiter seulement les thèses contenant au moins ces segments car, comme nous l'avons déjà souligné : les thèses n'ont pas toutes la même structure sémantique. Le tableau 9 indique le nombre de concepts extraits pour chaque thèse selon ce découpage sémantique. Par exemple, pour la thèse *« T_1 »* le segment *« état de l'art général »* se trouve réparti dans les chapitres 1, 2 et 3. Pour ce segment, nous avons obtenu 241 concepts. En revanche pour le segment du *« modèle »* qui correspond au chapitre 4, nous avons obtenu 54 concepts. La thèse *« T_5 »* présente un autre cas d'étude où les segments *« état de l'art général »* et *« modèle »*, sont imbriqués dans le chapitre 3. Pour l'*« état de l'art »* nous avons obtenu 32 concepts provenant donc du chapitre 3, alors que les 17 concepts du segment *« modèle »* proviennent des sections 3.3 et 3.4 du chapitre 3 ainsi que la section 4.1 du chapitre 4. Le tableau 9 illustre la différence qui existe entre le nombre de concepts extraits pour chacun de ces deux découpages.

Thèse	**concepts extraits pour la partie dédiée à l'état de l'art général**		**concepts extraits de la partie dédiée au modèle**	
	Nb concepts	**Dans chapitres (sections)**	**Nb concepts**	**Dans chapitres (sections)**
T_1	241	1, 2 et 3	54	4
T_2	22	1(1, 2, 3)	8	4
T_3	51	1, 2 et 3	31	4 et 5
T_4	59	1, 2, 3 et 4	19	5, 6, 7 et 8
T_5	32	3	17	3 (3.3, 3.4), 4(4.1)
T_6	56	2	20	3
T_7	46	1	20	2
T_8	26	2	6	3(4)
T_9	65	2	10	3

Tableau 9. *Concepts extraits pour le découpage sémantique correspondant aux segments : « État de l'art général » et « Modèle » de la thèse*

Afin de souligner l'importance de l'analyse de la structure sémantique, nous avons comparé le poids (en pourcentage) des concepts extraits selon notre découpage sémantique par rapport à la totalité des concepts extraits de la thèse incluant l'introduction, les chapitres et la conclusion. Nous avons constaté que le pourcentage de concepts apparaissant comme pertinents dans l'ensemble de la thèse est plus important pour le segment *« état de l'art général »* par rapport aux autres segments sémantiques. Donc, le segment, *« état de l'art général »*, permet d'extraire une grande partie des concepts des thèses. Ainsi, avant d'étudier l'ensemble de la thèse, il peut être intéressant d'analyser la partie concernant à l'*« état de l'art général »*.

L'analyse de la structure sémantique nous a permis de valider l'intérêt du découpage de la thèse en *« segments sémantiques »,* de localiser les parties de la thèse les plus riches en information sur le contenu de la thèse et d'extraire les concepts présents dans la plupart des thèses.

Les analyses présentées précédemment ont servi à déterminer les fragments les plus intéressants de la thèse mais aussi à extraire les concepts pertinents de chaque thèse. Nous avons ensuite construit une base de concepts du domaine que nous allons présenter dans la section suivante.

4. Création d'une base de concepts du domaine

Dans notre proposition, nous allons intégrer cette base de concepts, que le doctorant pourra utiliser lors de la rédaction de sa thèse. Elle sera organisée en catégories à partir des concepts extraits expérimentalement.

Pour la création de la base, nous avons tout d'abord extrait tous les concepts de toutes les thèses à l'aide de l'outil Nomino. Ensuite, nous avons classé les thèses selon le nombre de concepts extraits de manière descendante. Puis, un expert du domaine a analysé les concepts extraits afin d'éliminer ceux qui ne correspondaient pas au domaine. De cette manière, l'expert a éliminé manuellement environ 200 concepts, dont certains ne correspondaient pas au domaine de l'informatique, par exemple : *« communauté urbaine »*, *« compacité CoH »*, *« enquête de satisfaction »*, *« modulation par impulsion »*, etc., l'expert a aussi éliminé certains termes extraits par Nomino qui n'étaient pas des concepts comme par exemple : *« besoin d'information »*, *« caractéristique globale »*, *« situation de travail »*, *« tâche de création »*, entre autres.

A partir des concepts extraits, nous avons classifié les concepts en pertinents et non pertinents afin de créer notre base de concepts. Nous avons comparé les concepts extraits des chapitres (sans l'introduction ni la conclusion, uniquement les chapitres) des thèses analysées avec les concepts extraits de la partie dédiée à l'état de l'art général.

Thèse	Nombre de concepts pertinents uniquement pour les chapitres	Nombre de concepts pertinents de *l'état de l'art général*	Thèse	Nombre de concepts pertinents uniquement pour les chapitres	Nombre de concepts pertinents de *l'état de l'art général*
T_1	296	52	T_{11}	57	56
T_2	38	71	T_{12}	81	80
T_3	64	31	T_{13}	54	46
T_4	43	21	T_{14}	51	43
T_5	73	59	T_{15}	85	67
T_6	42	37	T_{16}	24	26
T_7	40	20	T_{17}	43	65
T_8	124	52	T_{18}	14	7
T_9	38	32	T_{19}	32	34
T_{10}	54	19	T_{20}	33	22

Tableau 10. *Nombre de concepts extraits pour l'ensemble des chapitres et pour le segment sémantique « état de l'art général »*

Thèse	Nombre de concepts extraits de l'ensemble de la thèse	Nombre de concepts pertinents qui appartiennent au domaine de l'informatique	Thèse	Nombre de concepts extraits de l'ensemble de la thèse	Nombre de concepts pertinents qui appartiennent au domaine de l'informatique
T_1	296	89	T_{11}	43	23
T_2	85	53	T_{12}	40	21
T_3	124	46	T_{13}	54	20
T_4	81	35	T_{14}	73	20
T_5	51	30	T_{15}	38	18
T_6	43	29	T_{16}	38	18
T_7	57	29	T_{17}	64	16
T_8	54	28	T_{18}	33	11
T_9	32	24	T_{19}	24	10
T_{10}	42	24	T_{20}	14	3

Tableau 11. *Nombre de concepts extraits et nombre de concepts retenus comme pertinents pour le domaine de l'informatique*

Le tableau 10 dresse l'état des lieux résultant de l'analyse de tous les chapitres (sans l'introduction ni la conclusion) et de l'analyse du segment sémantique *état de l'art général* pour toutes les thèses analysées. Dans 50 % des cas, la quantité de concepts extraits dans *l'état de l'art général* est très proche de l'analyse de l'ensemble de la thèse.

Nous avons comparé manuellement la liste des concepts issus de tous les chapitres et celle issue du segment *état de l'art général* afin de compter le nombre de concepts différents. Nous avons extrait au total 2 126 concepts pour l'ensemble des chapitres. Nous avons épuré la base en éliminant les doublons et les concepts dont le sens est proche (exemple : *« domaine de l'informatique »* et *« domaine informatique »*). Ainsi, nous avons retenu 241 concepts significatifs. En comparant les deux listes (liste de concepts pour l'ensemble de chapitres et liste de concepts pour le segment état de l'art général), l'analyse montre que le segment *état de l'art général* apporte 192 concepts différents de ceux issus de l'ensemble des chapitres. Au total, notre base de connaissances contient donc 433 termes pertinents issus du domaine de l'informatique.

La deuxième colonne du tableau 11 présente le nombre de concepts extraits pour chaque thèse. La troisième colonne présente le nombre de concepts retenus comme pertinents, c'est-à-dire le nombre de concepts correspondant au domaine de l'informatique.

Catégorisation des concepts extraits des thèses

La catégorisation des concepts aura pour but d'aider l'utilisateur à trouver aisément les concepts sémantiques à insérer dans le texte comme des tags sémantiques.

Cette catégorisation est établie manuellement, alors que l'extraction initiale de concepts est effectuée automatiquement. Nous avons défini, à partir d'une analyse de concepts existants dans la plupart des thèses, 17 catégories de base : *« Algorithme »*, *« Apprentissage »*, *« Base de données »*, *« Documents »*, *« Groupware »*, *« IHM »*, *« Intelligence artificielle »*, *« Langages »*, *« Logiciel »*, *« Matériel »*, *« Multimédia »*, *« Programmation »*, *« Recherche d'information »*, *« Réseaux »*, *« Système d'information »*, *« Web sémantique »* et *« Workflow »*. Ces catégories contiennent des sous-catégories regroupant les concepts obtenus par Nomino. Un concept peut être classé dans plusieurs catégories (ou sous-catégories). La classification est fondée sur la syntaxe du concept mais surtout sur la sémantique, le sens du concept. Par exemple, dans la catégorie *« Algorithme »* nous trouvons le concept *« Algorithme génétique »* lequel est également dans la catégorie *« Intelligence artificielle »*.

Faute d'existence de thésaurus du domaine de l'informatique, nous avons complété la base de concepts en utilisant le *« Glossaire informatique des termes de*

la Commission ministérielle de terminologie informatique »[2]. Ce document est le résultat d'une compilation de divers arrêtés issus des travaux de la Commission ministérielle de terminologie informatique ainsi que du projet d'arrêté qui était en cours lorsque le dispositif terminologique a fait l'objet d'une profonde réforme.

Notre base de concepts n'est toutefois pas complète, en effet, durant l'établissement de cette base, nous avons rapidement constaté, qu'il y aurait d'autres concepts du domaine de l'informatique à intégrer. Par exemple des concepts de psychologie cognitive peuvent être très fortement liés à des concepts informatiques. C'est pourquoi, nous avons décidé d'implémenter physiquement la base de connaissances sous forme de différents fichiers correspondant aux concepts de l'informatique proprement dite et aux domaines d'applications. Ainsi, chaque fois qu'une communauté crée son thésaurus, nous pouvons l'intégrer à notre base grâce à son mode de gestion modulaire.

En ce qui concerne, l'évolution de la base de concepts nous avons évalué expérimentalement la progression du nombre de concepts apportés par chaque nouvelle thèse et la progression reste voisine de 2,2 %. Ces expérimentations sont à poursuivre pour confirmer nos premiers résultats mais nous pouvons déjà prétendre que la taille de la base de concepts évoluera très progressivement.

L'ensemble de ces résultats va nous permettre de définir un nouveau modèle de documents intégrant la dimension sémantique.

5. Un modèle pour représenter la structure de la thèse scientifique

La construction du document de type *« thèse »* repose sur deux étapes : (1) la mise en place de la structure logique et (2) l'ajout des éléments sémantiques utilisés comme indiqué dans les paragraphes précédents. Pour la première étape, nous avons suivi, à quelques détails près, les recommandations du ministère de l'Éducation (Jolly, 2000). Pour la seconde étape, nous avons exploité le mécanisme de XML Schéma[3] pour formaliser la structure globale d'une thèse scientifique.

A partir d'une étude de différentes normes utilisées pour la structuration des documents, nous avons retenu XML Schéma. Certaines caractéristiques ont attiré notre attention, ce sont les possibilités de création des types complexes, d'attribution de cardinalité maximum, de manipulation de fichiers etc. Par exemple, nous devons traiter la notion de spécialisation pour gérer les établissements, car certains doctorants préparent leurs thèses en cotutelle avec un autre établissement, et les soutiennent dans les établissements d'origine. Ainsi la structure générale *« Etablissement »* possédant un *« nom »* et des *« coordonnées »* sera définie comme un type complexe qui sera attribué à chaque établissement. Le format DTD

2. http://www-rocq.inria.fr/qui/Philippe.Deschamp/CMTI/glossaire.html
3. http://xmlfr.org/w3c/TR/xmlschema-0/

(Définition de Document Type) ne nous permet pas d'implémenter cette spécificité. Le ministère de l'Éducation recommande, au maximum, neuf subdivisions dans une thèse. Mais dans une DTD, il n'est pas possible d'implémenter cette cardinalité maximale, alors que XML Schéma permet de mettre en place une structure de neuf *« Sous-sections »* dans une *« Section »*.

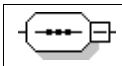
Etant donné que de nombreux auteurs rédigent leurs thèses à partir du logiciel Word, nous aurions tout à fait pu choisir d'utiliser le schéma *« XML document 2003 »* fourni par Microsoft dans la bibliothèque des schémas WordprocessingML pour Office 2003[4]. Mais celui-ci est un format propriétaire et les fichiers générés sont assez lourds à manipuler. Nous avons donc préféré créer notre propre modèle en utilisant XML Schéma (Abascal *et al.,* 2005). Avec ce modèle, nous pouvons générer uniquement des fichiers de données, ce qui rend plus aisé le traitement et la manipulation des documents dans le cadre de la recherche de l'information. Les éléments les plus utiles pour la recherche d'information sont les métadonnées insérées par le rédacteur, qui sera assisté d'une part par Nomino pour extraire s'il le souhaite des concepts des paragraphes sélectionnés, et d'autre part par la base de concepts que nous avons définie.

En se basant sur la description de la structure semi-formelle d'une thèse, nous avons créé une structure formelle exprimée par XML Schéma à l'aide de l'outil de développement XMLSpy (Berisha-Bohé *et al.,* 2005). Ce schéma comprend 9 types complexes globaux qui seront repris par différents éléments ou types du document et 15 éléments globaux qui apparaîtront dans le corps d'une thèse.

5.1. *Description du nouveau modèle pour les thèses scientifiques*

Une *« Introduction »* de thèse est constituée de son *« Titre »* et d'un ou plusieurs paragraphes (figure 1). Les paragraphes peuvent être *« tagués »* ou *« non tagués »*, et c'est pourquoi nous avons introduit le terme générique *« BlockParagraphe »* pour les deux types de paragraphes. Nous les intitulons *« tagué »* quand ils sont entourés par une ou plusieurs métadonnées (concepts). Donc, au début du paragraphe réside l'entête de la métadonnée *« EnteteMetadata »*. Dans l'entête, nous trouvons le *« TagOuvrant »*, un (ou plusieurs) *« Concept »*, et la variable booléenne *« Précédent »*. Le pied de la métadonnée *« PiedMetadata »* réside à la fin du paragraphe. Cet élément est constitué par la variable booléenne *« Suivant »*, le (ou les) *« concept »* déjà apparu(s) dans l'entête et le *« TagFermant »*.

4. http://www.microsoft.com/office/xml/default.mspx



Une séquence d'éléments. Les éléments doivent apparaître exactement selon la séquence dans laquelle ils apparaissent dans le schéma.

Un choix d'éléments. Seulement un élément doit apparaître dans cette position.

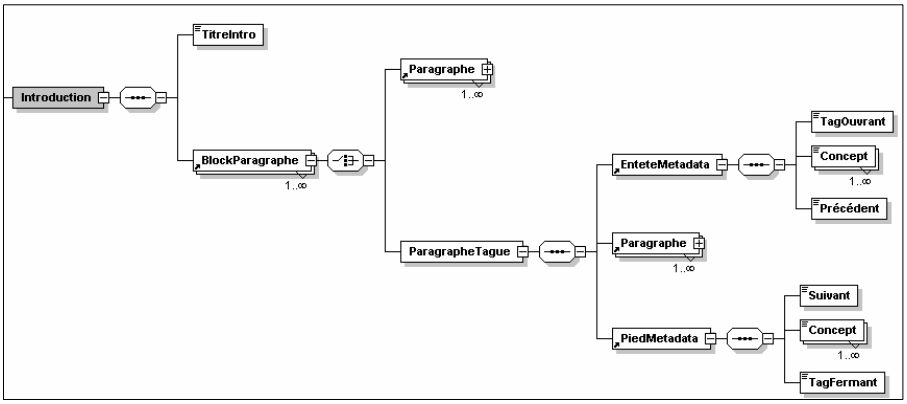


Figure 1. *Exemple de la modélisation de la structure logique et de la structure sémantique d'une thèse en utilisant XML Schéma*

La présence de tous les éléments (entête de métadonnée, paragraphe et pied de métadonnée) est obligatoire dans un paragraphe *« tagué »*. C'est pourquoi nous avons défini les cardinalités minimum de chaque élément à 1. Par contre, un ou plusieurs concepts peuvent entourer un ou plusieurs paragraphes successifs. De même, pour les structures *« EnteteMetadata »* et *« PiedMetadata »*, tous les éléments les constituant sont obligatoires. Comme nous venons de le souligner, plusieurs concepts peuvent apparaître dans un paragraphe ou dans une suite de paragraphes. De cette manière, une introduction contenant plusieurs paragraphes peut contenir plusieurs métadonnées. De la même façon, seront construites les *« Parties »* (un groupement de chapitres traitant la même approche comme par exemple l'état de l'art, ou le développement du prototype), les *« Chapitres »*, les *« Sections »* et les *« Sous-sections »* de la thèse, qui sont des regroupements de paragraphes. Les paragraphes eux-mêmes peuvent également contenir plusieurs métadonnées, au début, mais aussi dans le corps. Cela est possible par le regroupement de plusieurs blocs de texte (qui sont la plus fine partie de la structure du document *« THESE »*) entourés par des métadonnées de la même façon que les blocs de paragraphes.

L'utilisation des variables booléennes *« Précédent »* et *« Suivant »* est nécessaire pour la gestion des blocs de texte. Si, par exemple, un segment sémantique est constitué de la dernière phrase d'un paragraphe courant, et des deux premières

phrases du paragraphe suivant, nous allons être capables de reconstituer le segment au delà de la structure logique *« Paragraphe »* grâce à ces éléments booléens. Ainsi, le rédacteur pourra insérer des métadonnées dans n'importe quelle partie du corps de la thèse en fabriquant un document bien décrit. Grâce à ces métadonnées, l'application pourra localiser l'information pertinente durant un processus de recherche.

Afin de simplifier l'utilisation des métadonnées dans notre modèle, nous avons décidé d'utiliser des attributs. Pour cela, un *« ParagrapheTague »* peut être composé d'un ou de plusieurs *« Paragraphe »* qui peuvent eux mêmes posséder (être composés de) une ou plusieurs *« Metadata »* avec un attribut nommé *« Concept »* (figure 2).

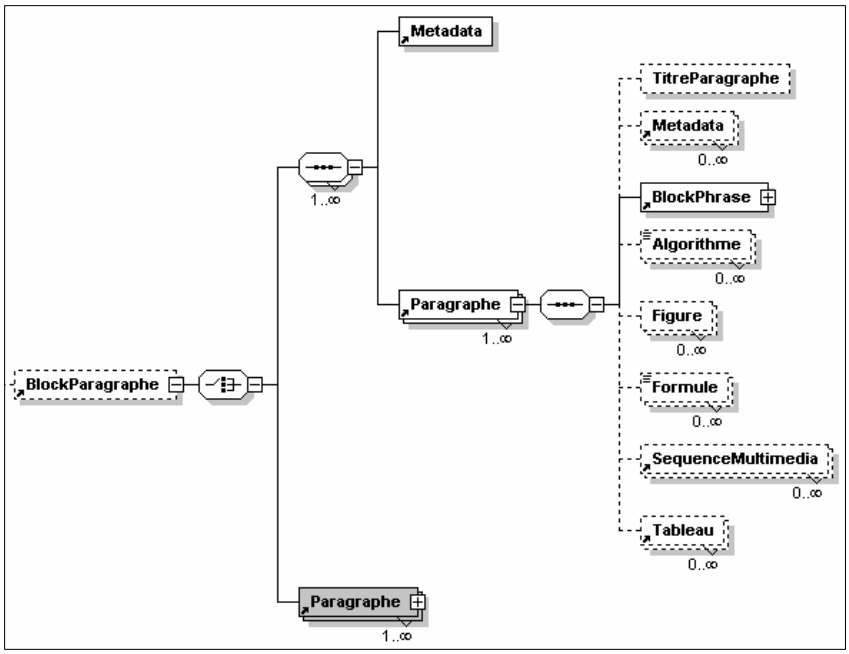


Figure 2. *Exemple de la modélisation de la thèse en utilisant la balise « metadata » avec un attribut « concept »*

6. Proposition d'un système pour la structuration sémantique de la thèse

Pour améliorer l'accès au contenu des thèses, nous proposons d'ajouter les tags sémantiques lors de la création du document selon trois modalités : (1) Manuellement : à partir des choix propres à l'utilisateur, (2) A partir d'un outil de TAL permettant l'extraction automatique de concepts du document ou de parties du document, avec une possibilité de sélection possible des concepts, par l'auteur, (3) A partir d'une base de concepts du domaine proposée à l'utilisateur.

En sachant que c'est l'auteur qui connaît le mieux sa thèse, l'insertion manuelle des tags sémantiques permet de caractériser la thèse en faisant confiance à l'auteur. Au cours de la saisie, les annotations peuvent être utilisées pour ajouter des informations non prévues par le concepteur du document (Bringay *et al.,* 2004). La recherche d'information en exploitant les métadonnées comme des *« annotations »* permettra d'accéder aux ressources selon leur contenu plutôt que par simples mots-clés.

Notre prototype permet d'effectuer balisage logique et sémantique de la thèse. A partir des résultats obtenus, nous avons constaté que la thèse rédigée de cette manière est mieux organisée. L'auteur doit prendre en compte le schéma XML afin de valider sa thèse. Les balises sémantiques permettent de donner plus de sens à la thèse. C'est par le biais de ces balises que l'utilisateur pourra ensuite obtenir l'information pertinente. Le prototype construit pour la recherche d'information concerne les fonctionnalités actuelles du projet CITHER (recherche par titre, par nom de l'auteur, par date de soutenance) et propose de nouvelles fonctionnalités pour extraire des fragments pertinents des thèses. Ainsi, l'utilisateur peut choisir entre la restitution par fragments, par chapitres, par résumé, etc., mais aussi par la thèse entière.

Nous avons effectué quelques tests afin de valider les possibilités obtenues en utilisant notre prototype par rapport à la version initiale de CITHER. Même si les tests restent encore limités, les premiers résultats de nos travaux semblent tout à fait prometteurs. La prochaine étape est de passer à une plus grande échelle.

7. Conclusion

Notre travail s'inscrit dans le cadre du projet CITHER, projet de mise en ligne d'une bibliothèque numérique de thèses en utilisant le format PDF. Une des restrictions imposée par ce format est que lors d'une session de recherche il est impossible de sélectionner exclusivement des extraits pertinents. Le problème réside dans le manque de précision des métadonnées utilisées.

Notre approche vise à permettre la recherche d'information pertinente en proposant un nouveau modèle de documents pour les thèses, fondé sur l'utilisation de nouvelles métadonnées. Nous proposons donc à l'auteur de la thèse de décrire son document avec des métadonnées caractérisant le contenu de sa thèse. Afin d'aider l'auteur dans sa démarche de description, nous avons envisagé : (1) l'utilisation d'un outil de TAL capable d'extraire automatiquement des concepts d'une thèse et (2) la construction d'une base de concepts, à partir de thèses du domaine, disponible pour proposer de nouvelles métadonnées. La recherche d'information pertinente, telle que nous l'envisageons, s'appuiera sur de nouvelles métadonnées rajoutées au sein de la thèse comme des *« tags sémantiques »*. Dans cet article, nous avons présenté notre proposition d'un nouveau modèle de document pour les thèses permettant un accès pertinent à l'information. Cette proposition est fondée sur l'utilisation d'un outil de TAL et sur l'étude de la structure logique et sémantique des thèses.

Nous avons également conçu un système qui permet au doctorant, pendant la phase de rédaction de sa thèse, d'ajouter des *« tags sémantiques »* à sa thèse selon trois modalités : (1) sur choix propre de l'utilisateur, (2) en s'appuyant sur la base de concepts et (3) en utilisant le logiciel Nomino pour l'extraction des concepts d'un fragment de texte sélectionné. Dans notre système, les traitements sont transparents à l'utilisateur. L'utilisateur n'a pas besoin de connaître XML pour ajouter les balises. De plus, grâce à l'utilisation des balises le doctorant est capable de mieux organiser sa thèse et évitera les répétitions des concepts. Notre système permet ainsi de restituer à l'utilisateur plusieurs fragments de thèse(s) pertinent(s).

Nous travaillons actuellement sur l'aspect recherche d'information. Nous proposons d'utiliser une ontologie qui permettra au système de réaliser une expansion de la requête en utilisant des concepts proches de ceux proposés par l'utilisateur (Gruber *et al.,* 1993 ; Maedche *et al.,* 2002 ; Abascal *et al.,* 2003a).

8. Bibliographie

Abascal R., Rumpler B., "Conceptualización de Tesis Científicas dentro del Contexto de una Biblioteca Digital mediante el uso de metadatos", *Conferencia Ibero-Americana IADIS WWW/Internet 2004 (CIAWI 2004),* IADIS Press, Madrid, Spain, October 2004a, p. 40-48.

Abascal R., Rumpler B., Berisha-Bohé S., « Proposition d'une nouvelle structure de document pour améliorer la recherche d'information », *Proceedings of the CORIA'05 Grenoble,* IMAG, 2005, p. 389-404.

Abascal R., Rumpler B., Pinon J-M., "An Analysis of Tools for an Automatic Extraction of Concept in Documents for a Better Knowledge Management", *IRMA 2003 International Conference,* Philadelphia Pennsylvania, USA, May 2003b.

Abascal R., Rumpler B., Pinon J-M., « Conception d'une Ontologie dans le Contexte d'une Bibliothèque Numérique », *ISKO 2003*, Grenoble, France. July 2003a.

Abascal R., Rumpler B., Pinon J-M., "Information Retrieval in Digital Theses Based on Natural Language Processing Tools", *VICEDO J.L. et al., España for Natural Language Processing (EsTAL'04),* LNAI 3230, Springer Berlin Heidelberg, Alicante, Spain, October 2004b, p. 172-182.

Abiteboul S., "Querying Semi-Structured Data", *AFRATI F., KOLAITIS P., editors, Database Theory – ICDT'97*, vol. 1186 of LNCS, Springer, 1997, p. 1-18.

Amerouali Y., « Métadonnées basées sur des éléments de description de ressources et profils d'utilisateur », *Colloque ISKO*, oct. 1999, Lyon, France. 1999, p. 43-48.

Antonacopoulos A., Karatzas D., Krawczyk H., Wiszniewski B., "The lifecycle of a digital historical document: structure and content", *Proceedings of the 2004 ACM symposium on Document engineering DocEng '04*, ACM Press, October 2004.

Baeza-Yates R., Ribeiro-Neto B., *Modern Information Retrieval*, Addison-Wesley-Longman, Harlow *et al.* The ACM Press/The MIT Press, 1999.

Beguin A., Jouis C., Widad M., « Évaluation d'outils d'aide à la construction de terminologie et de relations sémantiques entre termes à partir de corpus », *Premières Journées Scientifiques et Techniques (JST) du Réseau Francophone de l'Ingénierie de Langue de l'AUPELF-UREF*, Avignon, France, 1997, p. 419-425.

Berisha-Bohe S, Rumpler B., Abascal R., « A semantic structure to improve information retrieval using XML », Dobreva M., Engelen J., Peeters Publishing Leuven, *9th ICCC International Conference– Elpub2005*. Belgium, June, 2005, p. 319-321.

Biebow B., Szulman S., « Avancée sur le concept de base de connaissances terminologique », *Actes des 6e journées nationales du PRC-GDR intelligence artificielle (PRC-GDR IA'97)*. Ed. Hermès, 1997, p. 357-370.

Bringay S., Barry C., Charlet J., « Les documents et les annotations dans le dossier patient hospitalier », Salaün M., Charlet J., (Eds.), *Le document numérique, Cépaduès: Toulouse. Numéro thématique Le document numérique de la revue I3*, vol. 4, n° 1, 2004, p. 191-211.

Goecke D., Witt A., "Exploiting Logical Document Structure for Anaphora Resolution", *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006),* Genoa, Italy. 2006.

Gruber T. R., "A Translation Approach to Portable Ontology Specifications", *Knowledge Acquisition*, 1993, 5, (2), p. 199-220.

Heath B. P., McArthur D. J., McClelland M. K., Vetter R. J., "Metadata lessons from the iLumina digital library", *Communications of the ACM*, vol. 48, n° 7, July 2005.

Jolly C., Rapport sur la diffusion électronique des thèses. [En ligne] Paris: Ministère de l'Éducation nationale–SDBD, 2000. Disponible sur ;
<http://www.sup.adc.education.fr/bib/Acti/These/jolly/entete.htm>

Leskovec J., Milic-Frayling N., Grobelnik M., « Impact of Linguistic Analysis on the Semantic Graph Coverage and Learning of Document Extracts ». *12th National Conference on Artificial Intelligence (AAAI2005).* Pittsburgh, PA. June 2005

Maedche A., Staab S., Studer R. *et al.,* "SEAL — Tying Up Information Integration and Web Site Management by Ontologies", *IEEE Computer Society Data Engineering Bulletin, Special issue on Organizing and Discovering the Semantic Web*, vol. 25, n° 1, March 2002, p. 10-17.

McCalla G., "The Ecological Approach to the Design of E-Learning Environments: Purpose-based Capture and Use of Information About Learners", *Journal of Interactive Media in Education*, (7), Special Issue on Educational Semantic Web. 2004.

Plante P, Dumas L., Plante A., Nomino version 4.2.22., 1997. [En ligne] Disponible sur:
<http://www.nominotechnologies.com/>

Salton G., McGill M. J., *Introduction to Modern Information Retrieval*. New York *et al.*: McGraw-Hill, 1983, 400 p.

Séguéla P., Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques, Thèse Université Toulouse III, March 2001.

Seung-Kyu K., Yoon-Chul C., "A structured documents retrieval method supporting attribute-based structure information", *Proceedings of the 2002 ACM symposium on Applied computing SAC'02,* ACM Press, March 2002.

Singh S., Gaba S. K., Pandita N., "Architecture and Building of Medical Digital Library at NIC [of India]: What Exists and What is Required for MeDLib@NIC?" *Proceedings International Conference on Digital Libraries,* New Delhi, India, 2004.

Tenier S., Toussaint Y., Napoli A., Polanco X., "Instantiation of Relations for Semantic Annotation", *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence WI '06*, IEEE Computer Society, 2006.

Tsinaraki C., Polydoros P., Christodoulakis S., "GraphOnto: A Component and a User Interface for the Definition and Use of Ontologies in Multimedia Information Systems", *Proceedings of AVIVDiLib 2005*, Cortona, Italy, April, 2005.

Van Campenhoudt M., « Les Voies de Recherche Actuelle en Terminologie et en Terminotique », *7e Université d'Automne en Terminologie, En bons termes*, Paris, La Maison du dictionnaire, 1998, p. 109-119.