Proposition d'une nouvelle structure de document pour améliorer la recherche d'information

Rocío Abascal — Béatrice Rumpler — Suela Berisha-Bohé

INSA de Lyon - LIRIS

7 Avenue J. Capelle Bâtiment Blaise Pascal
F69621 Villeurbanne cedex, France
{Rocio.Abascal, Beatrice.Rumpler, Suela.Bohe}@insa-lyon.fr

RÉSUMÉ. *Actuellement l'information contenue dans les bibliothèques numériques n'est pas totalement décrite et son exploitation est insuffisante. La description de l'information en utilisant des métadonnées nous semble une bonne solution pour envisager une recherche d'information plus pertinente. Notre proposition est fondée sur la création et l'introduction au sein du document de « tags sémantiques » capables de décrire, dans notre cas, des thèses doctorales. Nous présentons l'analyse de quatre outils de Traitement Automatique des Langues (TAL) capables d'extraire automatiquement des concepts. Ensuite, nous proposons une base de concepts fondée sur l'analyse des structures logique et sémantique des thèses. Nous présentons ensuite un nouveau modèle du document, en utilisant le XML Schéma, qui contient les nouveaux « tags sémantiques » sur lesquels nous nous appuierons lors de sessions de recherche pour fournir à l'utilisateur l'information pertinente.*

ABSTRACT. *Currently the information contained in the digital libraries is not completely described and its use is still very limited. However, the description of information by using metadata can be of primary importance for the improvement of relevant search. Our analysis is based on the creation and the insertion of « semantic tags » able to describe doctoral theses. In this paper, we present the analysis of four tools for the Automatic Processing of Languages able to extract automatically concepts used then to define the « semantic tags ». We propose a base of concepts, based on the analysis of the logic and semantic structure of the theses. This base will help the user to characterize the thesis during the writing phase. We present a new model of the document by using the XML Schema which contains the new « semantic tags » able to provide to the user relevant information during a search session.*

MOTS-CLÉS : *bibliothèque numérique, document structuré, métadonnées, TAL, XML Schéma, recherche d'information, ontologie.*

KEYWORDS: *digital library, structure of document, metadata, NLP, XML Schema, information retrieval, ontology*

1. Introduction

Notre travail s'inscrit dans le cadre du projet CITHER, projet de mise en ligne d'une bibliothèque numérique de thèses. Le projet CITHER (Consultation en texte Intégral des Thèses en Réseau) de l'INSA de Lyon et sa bibliothèque scientifique, Doc'INSA, permet la diffusion par Internet et l'accès aux thèses scientifiques depuis 1997 en utilisant le format PDF (Portable Document Format). Une des restrictions imposée par ce format est que lors d'une session de recherche il est possible d'accéder au contenu d'une *seule* thèse à la fois, par le biais de chaque chapitre. Il est impossible de sélectionner exclusivement des extraits pertinents. 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 information. Actuellement, ces métadonnées proviennent essentiellement du format Dublin Core (DC, *« auteur »*, *« titre de la thèse »*, *« laboratoire »*, *« résumé »*, etc.). Le problème réside dans l'ambiguïté des métadonnées utilisées qui ne permettent pas de décrire précisément le contenu d'une thèse. Or ces métadonnées servent de base pour la recherche d'information. Dès lors, il est évident que beaucoup de réponses ne correspondront pas aux besoins réels exprimés par les utilisateurs. Le problème ne réside pas uniquement au niveau de la session de recherche d'information, mais plutôt 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 le contenu de la thèse. Ces mots clés ne sont généralement pas des éléments constitutifs d'une thèse, mais bien des éléments externes ajoutés à la thèse.

Notre approche vise à permettre la recherche d'information pertinente en proposant un nouveau modèle de document 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 Traitement Automatique des Langues (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, pour mise à disposition. Donc, la recherche d'information pertinente s'appuiera sur des nouvelles métadonnées rajoutées au sein de la thèse comme des *« tags sémantiques »*. La recherche par métadonnées nous a conduit à : (1) définir un nouveau modèle de thèses et (2) créer une ontologie du domaine permettant de mieux cibler les concepts à utiliser pour faire une recherche.

Dans cet article, nous présentons notre proposition d'un nouveau type de document pour les thèses permettant un accès pertinent à l'information (Section 2). Cette proposition est fondée sur l'utilisation d'un outil de TAL (Section 3) et sur l'étude de la structure logique et sémantique des thèses afin de construire une base de concepts du domaine (Section 4). A la fin de l'article nous présentons un nouveau modèle pour la représentation sémantique des thèses et du système de recherche d'information en cours de validation (Section 5).

2. Proposition pour la création d'un nouveau modèle de document pour les thèses

Pour l'amélioration de la diffusion des thèses nous proposons de permettre l'accès à leur contenu de façon précise grâce à l'utilisation de *« tags sémantiques »* rajoutés au sein de la thèse. Nous proposons d'ajouter ces tags lors de la création du document selon trois modalités (Figure 1) :

- manuellement : à partir des choix propres à l'utilisateur,
- à partir d'un outil de TAL permettant l'extraction automatique de concepts du document ou de parties du document, avec une possibilité de sélection par l'auteur parmi les concepts proposés,
- à partir d'une base de concepts du domaine proposée à l'utilisateur.

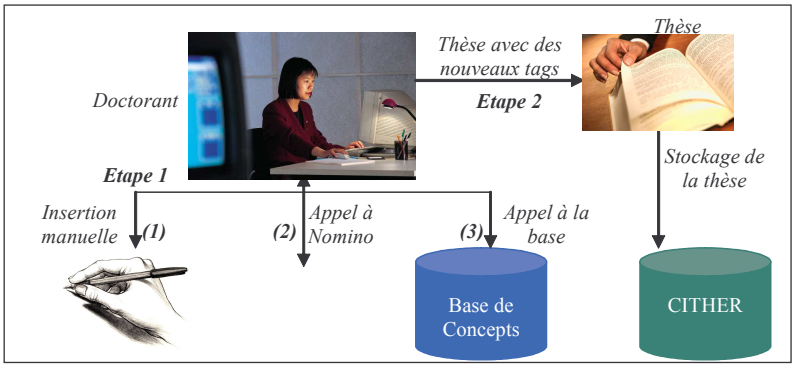

Figure 1. *Processus d'intégration de métadonnées dans les thèses*

En sachant que c'est l'auteur qui connaît le mieux sa thèse, l'insertion manuelle permet de caractériser la thèse en faisant confiance à l'auteur. Dans ce cas, la plupart des métadonnées ajoutées correspondront aux concepts proposés directement par l'auteur de la thèse. 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 mots-clés. Une telle recherche d'informations, que l'on appelle *« sémantique »* a déjà été étudiée dans de nombreux projets et pour différents langages, comme Shoe (Heflin et al., 1998) ou Ontobroker (Fensel, 1998). Ces projets ont démontré l'utilité et l'apport de l'approche de la recherche d'information *« sémantique »*. Une problématique qui se pose actuellement est celle d'une recherche d'information intelligente. Un des intérêts de la construction du Web Sémantique (Berners-Lee et al., 2001) est d'apporter suffisamment de

renseignements sur les ressources, en ajoutant des annotations sous la forme de métadonnées (Soualmia et al., 2003).

Nous présentons dans le paragraphe suivant les étapes de notre approche avant de décrire les fonctionnalités définies pour les modalités (2) et (3).

3. Extraction de concepts à partir d'un outil TAL

Une thèse regroupe un ensemble de concepts. 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 est représentée non plus sous la forme de son articulation physique, mais à partir d'un ensemble de concepts. Ces concepts sont représentés par des métadonnées qui peuvent être caractérisées selon le contexte d'utilisation du document. 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 des nouvelles métadonnées fondées sur les concepts décrivant chacune des thèses. L'extraction manuelle des concepts qui caractérisent un document est une tâche longue et complexe puisqu'il est nécessaire d'avoir connaissance du domaine de spécialité. Si le corpus de documents à évaluer est de petite taille l'extraction manuelle est envisageable, toutefois, dès que l'on manipule des corpus importants il faut que le processus soit automatique ou semi-automatique. Notre approche s'appuie sur l'utilisation d'un outil de TAL. Pour choisir l'outil le plus adapté à nos besoins nous avons évalué des outils existants. Nous présentons donc notre démarche d'évaluation des outils les mieux adaptés et les résultats obtenus.

3.1. *Choix d'un outil pour l'extraction automatique de termes*

Afin de caractériser les thèses par concepts, il faut extraire les termes candidats du corpus. Les termes sont des représentations linguistiques des concepts d'un domaine en particulier. Les termes extraits seront appelés *« candidat termes »*, ce sont généralement des groupes de mots qui peuvent révéler une certaine connaissance traitée dans un document (L'Homme, 2001).

Les logiciels d'acquisition automatique de termes s'appuient sur différentes méthodes. Généralement, ils combinent approches linguistiques avec des approches statistiques (Meilland et al., 2003). Pour notre travail, nous explorons 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).

Nous avons choisi d'évaluer quatre outils : Copernic Summarizer de NRC, Nomino de Nomino Technologies, TerminologyExtractor de Chamblon Systems Inc., et Xerox Terminology Suite de Xerox (XTS). Notre évaluation est fondée sur un corpus de documents scientifiques provenant du domaine de l'informatique. Notre corpus pour l'évaluation de ces outils est constitué de 25 documents scientifiques (20 thèses et 5 articles) d'une taille de 1 105 565 mots. Les 20 thèses proviennent du corpus actuellement disponible et de libre accès en format Word fournis par Doc'INSA. Notre démarche 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 »*. La *« précision »* est le rapport du nombre de concepts pertinents trouvés au nombre total de concepts sélectionnés. Le *« rappel »* est le rapport du nombre de concepts pertinents trouvés au nombre total de concepts pertinents du corpus.

	XTS	Copernic Summarizer	Terminology-Extractor	Nomino
Précision	2.8%	33.9%	6.8%	83.4%
Rappel	90.5%	51%	64.8%	65.1%

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

La Table 1 montre les résultats généraux pour l'analyse de notre corpus. Ces résultats montrent que c'est Nomino qui offre à la fois les meilleurs taux de *« précision »* et le meilleur taux de *« rappel »* (taux > 60%). Suite à cette étude (Abascal et al., 2003a), nous avons retenu Nomino comme étant l'outil le plus adéquat à nos besoins. Nomino est un logiciel développé par l'Université du Québec à Montréal *« capable d'extraire des concepts et des liens conceptuels à partir d'un document donné »* (Van Campendhoudt, 1998). Nomino 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. Ainsi nous proposons que l'auteur de la thèse puisse faire appel à Nomino pendant la rédaction de sa thèse pour extraire les concepts associés à un paragraphe sélectionné. Malgré la qualité des outils tel que Nomino, il subsiste encore des problèmes de sélection de concepts non pertinents. *« Les méthodes utilisées par les outils sélectionnent de nombreux candidats au statut de terme que l'expertise humaine n'aurait pas retenu »* (Beguin et al., 1997). L'utilisateur doit alors accepter ou rejeter les concepts extraits par Nomino pour les intégrer dans la thèse sous la forme de *« tags sémantiques »*.

4. Analyse de la structure sémantique de la thèse

En observant l'organisation des thèses scientifiques on constate que, généralement, elles suivent un plan dont la structure générale 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 sur 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. Nous nous sommes appuyés sur l'outil Nomino, pour essayer de déterminer la répartition des concepts, d'une part selon la structure logique et d'autre part selon l'organisation sémantique des thèses, telle que nous l'avons proposée ci dessus.

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

Un auteur s'intéresse au contenu du document qu'il rédige, mais aussi à son organisation, c'est-à-dire à son découpage en composants et aux relations entre eux. 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. La *« structure logique »* permet un découpage de la thèse en chapitres, sections, sous-sections, etc. On observe que la plupart des thèses suivent cette structuration en permettant un découpage en un ensemble cohérent de parties. 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. Dans une thèse structurée de façon sémantique on pourra envisager un balisage qui reflète ou ajoute du sens au contenu.

4.1.1. Analyse des principaux concepts extraits selon la structure logique

L'analyse des concepts extraits selon la *« structure logique »* de chacune des thèses est fondée sur le découpage du document en chapitres ou sections et sur l'organisation de ceux-ci. Nous avons utilisé Nomino pour extraire les concepts correspondant aux différents découpages logiques de la thèse : 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 de toutes les thèses analysées les parties correspondant aux préliminaires (page de titre, liste de professeurs, liste de figures, etc.), les parties correspondant aux post-liminaires (annexes, bibliographie, etc.) et les parties relevant des aspects plutôt administratifs (folio administratif, etc.). Nos thèses contiennent donc uniquement l'information liée au sujet traité par l'auteur.

Une première analyse a consisté à extraire tous les concepts de chacune des thèses complètes de notre corpus. Il est important de mentionner que Nomino utilise le *« calcul de saillance »*, lequel repose sur deux principes : le *« gain à la portée »* et le *« gain à l'expressivité »* (Nomino, 2001). 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. De cette manière, lorsque l'on appliquera Nomino à l'ensemble de la thèse il y aura des concepts très répétitifs qui ne seront pas extraits.

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

Table 2. *Présentation du nombre de concepts extraits pour la thèse complète et du nombre de concepts extraits pour l'ensemble des chapitres*

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 de tous les chapitres nous pouvons remarquer que : dans 70% des thèses, en enlevant l'introduction et la conclusion le nombre de concepts extraits (des chapitres exclusivement) augmente (Table 2).

Nous avons également calculé, la moyenne des pourcentages de concepts qui apparaissent dans chaque chapitre analysé, par rapport à l'ensemble de la thèse. Ceci nous a permis de connaître la partie de la thèse qui contient le plus grand nombre de concepts. Cette indication est intéressante car la recherche de l'information pertinente pourra s'envisager à partir de certaines sections ou chapitres de la thèse. La Table 3 résume les résultats obtenus en analysant des thèses constituées de 5 chapitres (18 des 20 thèses analysées sont composées par au minimum 5 chapitres). Le chapitre 2 (C_2) apparaît comme le chapitre qui contient plus de concepts pertinents. Ce chapitre est la plupart du temps consacré à la présentation des thèmes principaux traités dans la thèse.

Table des matières	Introduction	C_1	C_2	C_3	C_4	C_5	Conclu-sion.
9,5%	12,9%	20,7%	25,9%	22,2%	25,7%	23,6%	13,4%

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

Nous pouvons donc conclure à partir 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. L'analyse de la structure logique vérifie nos premières suppositions sur l'importance d'analyser en priorité les chapitres. La table des matières, l'introduction et la conclusion n'apportent pas beaucoup de concepts pertinents. Le chapitre 2 apparaît comme le plus pertinent.

4.1.2. Analyse des principaux concepts extraits selon la structure sémantique

Un *« segment sémantique »* est 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.

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 nous appuierons sur quelques uns de ces segments repérés par voie expérimentale (Table 4), afin de proposer notre modèle. Il peut exister d'autres segments à partir desquels une thèse peut être découpée sémantiquement.

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.

Table 4. *Présentation de quelques « segments sémantiques » d'une thèse scientifique*

4.1.2.1. Découpage sémantique appliqué à notre corpus

La première partie de notre analyse a consisté à découper manuellement les thèses analysées en *« segments sémantiques »*. Le découpage sémantique varie selon la thèse analysée. Un exemple illustrant le cas de deux thèses est présenté dans la Table 5. L'exemple montre que la première thèse est composée de trois *« segments sémantiques »*. En revanche, la deuxième thèse est composée de 7 segments dont 3 correspondent à différents *« états de l'art »*. Ce type de découpage, une fois validé expérimentalement, nous permettra de créer un modèle (par exemple en XML Schéma) en examinant les différentes possibilités de structuration sémantique de la thèse.

4.1.2.2. 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 définis par observation, pour chacune des thèses du corpus. Les concepts ainsi obtenus et validés pourront alors être utilisés en tant que *« tags sémantiques »*. Nous présentons par la suite les résultats de l'extraction de concepts pour deux découpages sémantiques distincts : *« état de l'art général »* et *« modèle »*. Afin de comparer ces deux découpages, et de

montrer l'importance du découpage correspondant à l'*« état de l'art général »*, nous traitons seulement les thèses contenant ces segments (9 thèses au total), car, comme nous l'avons déjà souligné (Table 5), les thèses n'ont pas toutes la même structure sémantique.

Découpage sémantique	T_1	T_2
	Etat de l'art général *-Chapitre 1 à 3*	**Etat de l'art général** *-Partie 1 Section 1-3*
	Proposition *-Chapitre 4*	**Etat de l'art de méthodes** *-Partie 2 Section 1-2 et 2.1*
	Prototype *-Chapitre 5*	**Expérimentation** *-Partie 2 Section 2.2*
		Etat de l'art d'outils *-Partie 3*
		Modèle *-Partie 4*
		Etat de l'art de techniques *-Partie 4 Section 2*
		Prototype

Table 5. *Exemple de découpages sémantiques différents*

La Table 6 indique le nombre de concepts extraits pour chaque thèse selon certains découpages sémantiques. 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. Cette table illustre la différence qui existe entre le nombre de concepts extraits pour chacun de ces deux découpages. Grâce à ces exemples, nous pouvons affirmer que le découpage sémantique correspondant à l'*« état de l'art général »* est très significatif car il apporte beaucoup d'informations sur le contenu de la thèse.

Thèse	ETAT DE L'ART GENERAL		MODELE	
	Nb concepts	***Dans chapitres***	***Nb concepts***	***Dans chapitres***
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

Table 6. *Concepts extraits pour les segments sémantiques correspondant à l'« état de l'art général » et au « 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. La Table 7 présente les résultats de cette comparaison. Nous pouvons voir que le pourcentage de concepts apparaissant comme pertinents dans l'ensemble de la thèse est très important pour le segment *« état de l'art général »*. Ainsi, au lieu d'étudier l'ensemble de la thèse, il peut être intéressant d'analyser avant tout, la partie concernant l'*« état de l'art général »*.

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
Etat de l'art général	107 36%	6 16%	29 44%	26 37%	17 42%	26 52%	21 46%	18 78%	15 42%
Modèle	45 15%	6 16%	29 44%	14 20%	17 42%	12 24%	19 41%	5 22%	10 28%

Table 7. *Pourcentage de concepts extraits pour les deux segments sélectionnés par rapport à la totalité des concepts de la thèse*

Enfin, à partir d'une comparaison manuelle des concepts extraits, nous avons constaté, que 57 concepts réapparaissent régulièrement dans la plupart des thèses analysées. Parmi les plus pertinents nous avons trouvé : *« annotation textuelle », « base de connaissance », « base de données », « capacité de stockage », « document multimédia », « document vidéo », « échange d'information », « environnement d'apprentissage », « formation à distance », « graphe conceptuel », « intelligence artificielle », « modèle de données », « recherche d'information », « séquence vidéo », « système d'information ».*

L'analyse de la structure sémantique nous a permis de valider l'intérêt du découpage de la thèse en *« segments sémantiques »*. Cette analyse nous a également permis 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.

4.2. Création d'une base de concepts du domaine

Dans notre proposition nous allons donc intégrer l'utilisation d'une base de concepts, que le doctorant pourra utiliser lors de la rédaction de sa thèse. Cette base, construite de manière hiérarchique, sera fondée sur l'analyse des concepts extraits des thèses scientifiques du domaine de l'informatique.

Nous avons comparé 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 2126 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 509 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 701 concepts pertinents du domaine de l'informatique.

Afin de rendre plus aisée l'utilisation de notre base de concepts pour le balisage de la thèse nous avons organisé la base de manière hiérarchique. La hiérarchisation 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).

4.2.1. Limites de notre base de concepts

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 »*[1]. Ce document est le résultat d'une compilation des divers arrêtés issus des travaux de la Commission ministérielle de terminologie informatique.

Notre base de concepts sera régulièrement mise à jour avec l'intégration de nouveaux concepts issus de nouvelles thèses. En partant du corpus actuel, nous avons évalué expérimentalement la progression du nombre de concepts apportés par chaque nouvelle thèse et la progression reste voisine de 2,24%. Ces

[1] http://www-rocq.inria.fr/qui/Philippe.Deschamp/CMTI/glossaire.html

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 et constitue donc une composante essentielle de notre proposition. L'ensemble de ces résultats nous a permis de définir un nouveau modèle de documents intégrant la dimension sémantique.

5. Création d'un nouveau modèle de documents

La construction du document de type *« thèse »* repose sur deux étapes : la mise en place de la structure logique et l'ajout des éléments sémantiques tel que présenté dans les paragraphes précédents. Pour la première étape, nous avons suivi, à quelques détails près, les recommandations du Ministère de l'Education (Jolly, 2000).

Pour la seconde étape, nous avons utilisé XML Schéma[2] pour formaliser la structure globale d'une thèse scientifique. En se basant sur une étude de différentes normes existantes pour la structuration des documents, nous avons retenu XML Schéma car il répondait mieux que les autres à nos besoins. 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[3]. Mais celui-ci reste encore un format propriétaire et les fichiers générés sont assez lourds à manipuler. Un extrait du modèle est présenté en Figure 2. Grâce à ce modèle, nous pouvons générer uniquement des fichiers de données, ce qui rend facile 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 vont être les métadonnées insérées par le rédacteur, qui est assisté d'une part par Nomino pour extraire s'il le souhaite des concepts de paragraphes que le rédacteur aura sélectionné, et d'autre part par la base de connaissances qui lui propose des listes de concepts du domaine.

5.1. *Description du modèle*

Une *« Introduction »* de thèse est constituée de son *« Titre »* et d'un ou plusieurs paragraphes. Les paragraphes peuvent être *« tagués »* ou *« non tagués »*, et c'est pourquoi, nous avons introduit le terme générique pour les deux types de paragraphes, *« BlockParagraphe »*. 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

[2] http://xmlfr.org/w3c/TR/xmlschema-0/
[3] http://www.microsoft.com/office/xml/default.mspx

paragraphe. Cet élément est constitué par la variable booléenne *« Suivant »*, le (ou les) *« concept »* déjà apparu(s) dans l'entête et le *« TagFermant »*.

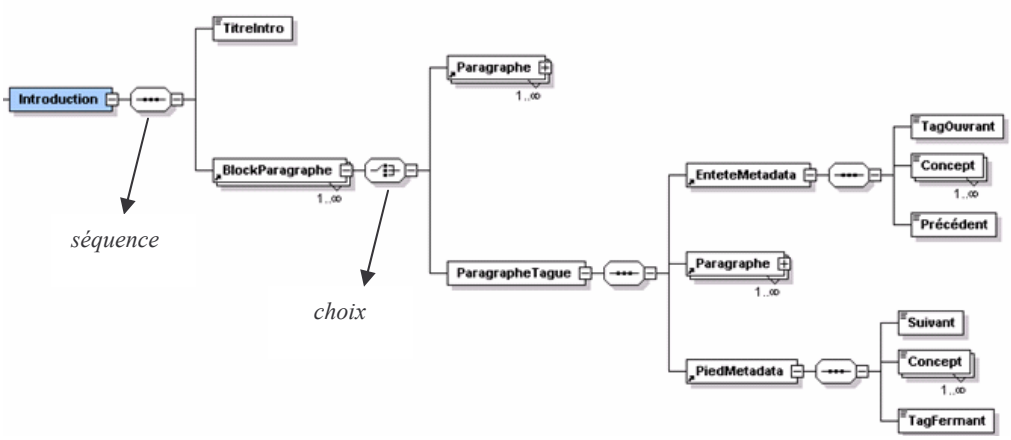

Figure 2. *Exemple de la modélisation de la structure logique et la structure sémantique d'une thèse en utilisant XML Schéma*

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, doivent être obligatoires. Comme on vient 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 blocks de texte (qui sont la plus fine particule de la structure du document *« THESE »*) entourés par des métadonnées de la même façon que les blocks de paragraphes.

L'utilité de l'utilisation des variables booléennes *« Précédent »* et *« Suivant »* est manifeste pour la gestion des blocks de texte. Si, par exemple, un segment sémantique est constitué par la dernière phrase d'un paragraphe courant, et les 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 fabricant un document bien décrit. Grâce à ces métadonnées, l'application pourra localiser l'information pertinente durant un processus de recherche.

6. Conclusion

La recherche d'information dans une bibliothèque numérique est focalisée sur l'utilisation de mots clés qui sont construits au fur et à mesure que l'utilisateur obtient des résultats. Généralement, l'outil de recherche propose des options pour améliorer la recherche. Dans la bibliothèque numérique CITHER, notre travail consiste à offrir à l'utilisateur la possibilité d'interagir avec la collection de thèses au moyen de *« concepts »* qui ont été insérés comme des *« tags sémantiques »*. L'insertion de ces étiquettes est fondée sur l'utilisation d'un outil de TAL capable d'extraire automatiquement les concepts. Dans cet article nous avons présenté une évaluation de quatre outils afin de choisir le plus adéquat à nos besoins : Nomino. L'objectif principal de notre travail a consisté à créer une base de concepts (à partir de l'analyse de la structure logique et de la structure sémantique des thèses) qui servira d'appui à l'auteur pendant la rédaction de sa thèse. Les concepts choisis sont modélisés comme des étiquettes XML dans chaque thèse. Ainsi, nous proposons un modèle de thèse qui est fondé sur l'analyse des concepts communs et pertinents d'un corpus de thèses scientifiques.

Une autre étape de nos travaux a été la création d'une ontologie, non présentée dans cet article, laquelle s'appuie sur l'utilisation des *« tags sémantiques »* qui ont été ajoutés aux thèses doctorales (Abascal et al., 2003b). Nous travaillons actuellement sur la conception d'un système *« évolué »* qui utilisera cette ontologie pour effectuer la recherche d'information pertinente. Une évaluation de l'intérêt d'utiliser une ontologie pour améliorer la recherche d'information sera réalisée. Les résultats de cette évaluation nous permettront d'envisager on non l'introduction de synonymes ou de mots complémentaires aux concepts qui auront été préalablement introduits par l'auteur de la thèse.

Une des nos perspectives de travail à futur porte sur l'utilisation des *« tags sémantiques »* comme des *« annotations »* pouvant être utilisées pour comprendre les intérêts du lecteur et lui suggérer des documents similaires aux passages annotés.

7. Bibliographie

Abascal R., Rumpler B., Pinon J-M., An analysis of tools for an automatic extraction of concept in documents for a better knowledge management. *Proceedings of 2003 IRMA International Conference*, Philadelphia Pennsylvania, Ed. Mehdi Khosrow-Pour, IDEA Group Publishing, May, p. 201-204, 2003a.

Abascal R., Rumpler B., Pinon J-M., Conception d'une ontologie dans le contexte d'une bibliothèque numérique. *4ème conférence ISKO (International society for knowledge organization)*, Grenoble, 2003b.

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

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.

Berners-Lee T., Hendler J., Lassila.O, The Semantic Web, *Scientific American*, May 2001.

Bringay S., Barry C., Charlet J, Les documents et les annotations dans le dossier patient hospitalier, Vol. 4, Num. 1, *I3 Information, Interaction, Intelligence*, 2004.

The Dublin Core Home Page, [online] URL: <http://purl.oclc.org/metadata/dublin_core> [06/02/2005].

Fensel D., Decker S., Erdmann M., Studer R., Ontobroker : Or how to enable intelligent access to the www. *Proceedings of KAW 98*, Banff, Canada, 1998.

Heflin J., Hendler J., et al., Reading between the lines: Using shoe to discover implicit knowledge from the web. In *Proceedings of the AAAI Workshop on Artificial Intelligence and Information Integration*, p. 51–57, 1998.

Jolly C., Rapport sur la diffusion électronique des thèses, Ministère de l'éducation nationale – SDBD, 2000.

L'Homme M. C., *Nouvelles technologies et recherche terminologique, Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe*. L'impact des nouvelles technologies sur la gestion terminologique, University York, Toronto, August 2001.

Meilland J-C., Pellot B., *Extraction automatique de terminologie à partir de libellés textuels courts*. Linguistique de corpus. Presses Universitaires de Rennes, 2003.

Nomino 4.2.22, updated in July 25, 2001. [online] URL: <http://www.ling.uqam.ca/nomino/> [06/02/2005].

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 de doctorat, Université Toulouse III, 2001.

Soualmia L. F., Darmoni S., Projection de requêtes pour une recherche d'information intelligente sur le Web. *Recherche Jeunes Chercheurs en IA, (RJCIA)*, Laval, p. 59-72, 2003.

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.