Acquisition de connaissances à partir de textes, vers l'élaboration d'une ontologie du droit

Guiraude LAME (*)

(*) Centre de Recherche en Informatique, Ecole Nationale Supérieure de Mines de Paris, 35 rue Saint-Honoré 77 305 Fontainebleau, France, http://cri.ensmp.fr, lame@cri.ensmp.fr

Document EMP-CRI : A/315/CRI

Résumé

Ce travail présente une méthode choisie pour l'élaboration d'une ontologie du droit, méthode fondée sur l'acquisition de connaissances à partir de textes. Nous proposons deux phases d'acquisition de connaissances à partir de textes : une première consistant à relever les termes du domaine au travers d'une base de connaissances terminologique et une seconde consistant à modéliser les connaissances implicites contenues dans les documents.

Cette ontologie a pour finalité de constituer l'ossature conceptuelle d'un système d'information basé sur le Journal Officiel et permettre une navigation par concepts entre les documents.

Mots clés

Acquisition de Connaissances, Ingénierie des Bases de Connaissances, Langage Naturel, Ontologies, Représentation des Connaissances, Terminologies, Droit.

1 Introduction

1.1 Le contexte du travail

Ce travail s'inscrit dans le cadre d'une thèse commencée en octobre 1999 au Centre de Recherche en Informatique de l'Ecole Nationale Supérieure des Mines de Paris sous la direction de Robert Mahl. L'objectif de cette thèse est d'élaborer, le plus automatiquement possible, une ontologie du droit et ce en partant des textes publiés au Journal Officiel (J.O.) de la République Française *Lois et décrets*. L'Ecole héberge en effet depuis 1998 un site Internet (http://droit.org) qui diffuse le J.O.. Les documents sont indexés selon les méthodes classiques des moteurs de recherche et les résultats ne sont pas à ce stade tout à fait satisfaisants du point de vue de la recherche documentaire juridique. L'attente principale des utilisateurs, attente guidée par la matière même, est en effet la possibilité d'une navigation par concepts entre les documents.

1.2 Le choix de l'ontologie

Une telle navigation implique, au-delà de la prise en compte des termes même contenus dans ces documents, d'être envisagée au niveau des concepts du droit. Or envisager un domaine du monde du point de vue de ses concepts se fait par une analyse des connaissances même du domaine, de leurs significations, et ce à l'aide d'une représentation des connaissances, au travers de ce qu'on appelle une ontologie [8]. Une ontologie appréhende en effet le monde au niveau des signifiants [7]. Nous envisageons une ontologie comme constituant à la fois une normalisation du langage de la matière [16] au travers d'une base de connaissances terminologique et un modèle conceptuel du domaine. Dès lors une ontologie peut servir d'ossature dans la structuration d'une base de connaissances et également permettre de proposer une aide à l'interrogation de la base, aide à la formulation des questions.

Ainsi, l'adjonction d'une base de connaissances terminologique – thesaurus - au moteur de recherche permet de proposer une interrogation conjointe du ou des terme(s) saisi(s) à l'interrogation avec celle du concept de niveau supérieur dans l'arborescence (*interrogation généralisée*) ou avec celle des concepts de niveau inférieur (*interrogation spécifiée*). Le système pourra alors proposer à l'utilisateur des documents connexes à sa question (généralisation, spécification) lui permettant d'envisager de façon plus effective le domaine juridique approprié. On se rapproche par là des possibles finalités éducatives des ontologies telles que décrites dans [3].

Une approche de la matière au niveau des concepts et une implémentation de cette approche dans notre base documentaire permettrait nous semble-t-il de proposer ce que nous appelons une *navigation par concepts entre documents*. Une ontologie de la matière est l'outil approprié dans un contexte de base documentaire, c'est une idée largement avancée en ingénierie des connaissances et plus spécifiquement posée par E. Valente et J. Breuker [17] dans les justifications de leur *ontologie fonctionnelle du droit*. Ces outils que constituent les ontologies ont été par ailleurs envisagés ces derniers temps dans le cadre d'Internet et plus généralement dans le domaine de la recherche d'information ([11] et [10]).

Notre ontologie est qualifiable d'*ontologie de domaine* [16] puisqu'elle n'envisage qu'un sous-domaine du droit, le sous-domaine constitué des normes législatives et réglementaires et qu'elle contient une terminologie standardisée du sous-domaine.

Notre ontologie est donc plus spécifiquement une ontologie à objet documentaire au travers des documents normatifs du droit français.

1.3 Une méthode d'élaboration de l'ontologie

La méthode ici proposée est fondée sur l'analyse de corpus [2], notre corpus étant constitué par le Journal Officiel *Lois et Décrets* de la République Française qui publie quotidiennement des textes normatifs tels que les lois, décrets, arrêtés, circulaires ministérielles et avis d'organismes divers. Dans cette perspective d'envisager les documents comme

source de connaissances, nous rejoignons en partie les travaux et objectifs du groupe TIA[1], à la convergence entre terminologie, ingéniérie des connaissances et intelligence artificielle.

La particularité de notre méthode est de se fonder essentiellement sur le corpus. Ce choix est nous semble-t-il justifié par le fait que notre objet est purement documentaire dans le corpus envisagé. Cette méthode fondée sur le corpus combine deux façons d'acquérir des connaissances à partir de textes : une extraction des termes du domaine contenus dans les documents et l'élaboration de modèles des connaissances explicites portées par nos documents, leurs connaissances endogènes.

Nous présenterons donc dans un premier temps (section 2) la phase d'analyse terminologique du corpus qui conduit à l'élaboration d'une base de connaissances terminologique. Nous proposons dans la section 3 la deuxième phase d'acquisition à partir de textes conduisant aux trois modèles des documents pour enfin envisager (section 4) l'élaboration proprement dite de l'ontologie.

2 Elaboration d'une base de connaissances terminologique

La première étape vers l'élaboration de notre ontologie du droit est la constitution d'une base de connaissances terminologique constituée des termes du domaine et de leurs relations lexicales. Nous rejoignons par là la méthode proposée par [1] et [2].

Puisque l'objectif est la navigation au travers des documents du J.O., nous décidons tout naturellement d'utiliser ces mêmes documents comme support à la construction de notre base de connaissances terminologique. Il n'est en effet pas nécessaire d'aller chercher des informations par ailleurs, que ce soit dans des avis d'experts de la matière que pourraient constituer des cours de droit ou dans d'autres documents. Nous estimons en effet que le J.O. *Lois et Décrets* constitue en lui-même une partie relativement autonome de la matière juridique, sa partie normative qui s'attache, schématiquement, à édicter des règles et à les faire vivre. En effet les lois et leurs décrets d'application ainsi que ce que les juristes nomment les *décrets autonomes* ou les différents arrêtés ministériels posent des concepts normatifs qu'ils s'appliquent à définir et à appliquer. La question de la constitution du corpus, de son éventuelle sélection ne se pose donc pas. Ou du moins nous y répondons ainsi : *« le J.O., tout le J.O., rien que le J.O.*». Plus pratiquement notre corpus est à ce jour constitué des publications quotidiennes du J.O. depuis le 1ier janvier 1998, ce qui correspond environ à 45 000 documents d'une taille moyenne de 5 Ko (dans une fourchette allant de 2,3 Ko pour un simple arrêté de nomination à 30 Ko pour une loi par exemple).

Nous partons donc des documents du J.O. comme support de connaissances stabilisées, connaissances d'autant plus stabilisées qu'elles le sont par la force même du droit. Nos documents sont en effet normatifs, ils définissent des normes et par là, ils posent des concepts, *concepts normatifs*, qui correspondent à des notions juridiques.

La constitution d'une base de connaissance terminologique passe par une analyse linguistique du corpus qui permet d'extraire les termes du domaine et une identification de leurs relations lexicales.

2.1 Extraction des termes du domaine

L'extraction des termes du domaine est une première étape nécessaire à la constitution de la base de connaissances terminologique. Nous utilisons pour ce faire des outils de traitement automatique de la langue (TAL) comme Lexter ou Pertimm® qui sont des extracteurs terminologiques. Ces outils mènent une étude linguistique des documents qui leur sont soumis et permettent l'acquisition des connaissances terminologiques du corpus. Leurs résultats sont à valider par une expertise humaine et nous nous attachons à minimiser le plus possible celle-ci.

Au vu de l'importance de notre corpus et du grand nombre de données fournies par des extracteurs de candidats-termes comme Lexter, le choix a été fait de faire des expérimen-

[1] Groupe « Terminologie et Intelligence artificielle » au sein du PRC-I3 et de l'AFIA

tations sur un échantillon du corpus. A l'issue de ces tests des méthodes de sélection automatique des candidats-termes sont identifiées et définies qui seront ensuite appliquées à la totalité de notre corpus pour proposer une première liste *épurée* de termes du domaine à l'expert.

Deux semaines de publication du J.O., du 4 au 15 avril soit 696 documents (un document par type de texte : loi, décret, arrêté, décision d'organisme divers, circulaire …) ont ainsi été analysées par Lexter. Pour ces documents, l'outil propose exactement 44 318 syntagmes nominaux (SN) candidats-termes du domaine.

Une analyse est à ce niveau nécessaire qui, combinée à une expertise du domaine, va permettre de sélectionner les termes à conserver. Il est tout d'abord nécessaire d'éliminer les candidats-termes que nous estimons hors contexte : tout ce qui est entièrement numérique (pourcentages quelconques, dates, numéros de textes…) et quelques approximations rendues par l'outil. Nous avons à ce stade réduit notre liste de candidats-termes à 37 727, ce qui correspond environ à une réduction de 15% de la liste initiale. Cette étape préalable à l'analyse des résultats étant effectuée, nous devons proposer des regroupements entre termes pour faciliter, et minimiser, l'intervention de l'expert.

Une des premières relations linguistiques entre termes à envisager est la synonymie. Or il s'avère, et cela a été montré notamment par [4], que le corpus traité est relativement pauvre en synonymies. Cette caractéristique est expliquées par la fonction même du corpus et la nature du domaine juridique. Le législateur, ou l'autorité chargée de l'exécution des normes, est en effet moins enclin à la rhétorique que le serait un juge par exemple. En effet le rôle de la norme dans l'environnement social est de définir la règle, et cela passe par la définition de ce que nous nommons des *concepts normatifs*. Le texte portant création de cette règle se doit donc d'être le moins ambigu possible, et s'il l'est, une nouvelle norme inférieure du point de vue de la hiérarchie des normes sera adoptée pour préciser les concepts. Dans tous les cas la dénomination exacte du concept normatif visé sera utilisée, gage de l'applicabilité de la norme. Cela dit, cette remarque, qui nous semble-t-il couvre la très grande majorité de notre corpus, comporte des exceptions. Le Parlement étant en France l'organe législateur, la dimension politique de sa tâche ne peut être ignorée et les cas existent de lois se bornant à poser des déclarations d'intentions et pour ce faire de s'exprimer dans un langage ambigu [4].

La rareté des synonymies dans notre corpus n'empêche pas d'autres types de relations entre termes et notamment les relations d'analogies encore appelées relations d'association étudiée par [4] et [6]. Ces relations permettent de préciser sémantiquement un terme avec des termes hyperonymes et hyponymes.

A partir des candidats-termes résultats de l'extraction terminologique, une première analyse a été menée, celle des regroupements morphologiques des syntagmes nominaux. Une première étape a consisté à ignorer la casse et par là à considérer comme identiques des syntagmes tels que « commission nationale » et « Commission Nationale ».

Une deuxième étape a consisté à travailler sur les *têtes* et les *expansions* des syntagmes nominaux pour regrouper dans un premier temps les candidats-termes présentant des têtes identiques (nous obtenons 9 051 groupes de syntagmes) et ensuite rassembler ceux dont les expansions sont proche morphologiquement (cette proximité morphologique est moins présente : détection de 20 946 expansion de syntagmes différentes). Ces regroupements de syntagmes nominaux permettent de définir des relations hiérarchiques entre termes, ces relations sont certes un peu triviales mais n'en demeurent pas moins potentiellement pertinentes. Les résultats sur notre échantillon sont de ce style :

contrat	**assurance**
contrat de achat	prime de assurance
contrat de transaction	courtier de assurance
contrat de location	contrat de assurance
contrat d'emprunt	attestation de assurance
contrat de crédit-bail	code de le assurance
contrat de association	valeur de assurance
contrat de vente	société de assurance
contrat particulier	**travail**
contrat de assurance faire référence à le décision	durée de le travail
	relation de le travail
contrat de assurance-vie	accord de travail
contrat en vigueur	service travail
contrat régional de plan	coût de le travail
contrat de prestation de service	exécution de travail
contrat de concession	service départemental de le travail
contrat de assurance	ministère de le travail
contrat de travail	entrepreneur de travail
contrat de formation	administration central de le travail
contrat de crédit-bail en double exemplaire	bâtiment travail
	durée maximale de le travail
contrat de cession	confédération français démocratique de le travail
contrat en double exemplaire	
contrat de droit privé	durée de travail
contrat de sous-traitance	charge de travail
contrat de le service à caractère industriel	réception définitive de le travail
contrat constitutif	utilité public de le travail
contrat de assurance	contrat de travail
contrat de assurance exiger	calendrier de travail
contrat de travail	relation collectif de le travail
contrat de travail de le salarié concerner	accident de le travail
contrat de travail temporaire	durée hebdomadaire de le travail
contrat de travail à durée déterminé	service médicaux de le travail
contrat de travail à durée indéterminé	inspecteur de le travail
contrat de travail de le salarié victime	temps de travail
	durée légal de travail

La colonne de gauche présentant des extraits des regroupements de syntagmes nominaux par tête et la colonne de droite par expansion identique.

On peut ainsi lier le SN « contrat d'assurance » à l'ensemble des SN dont l'expansion commence par « assurance » et supposer alors un hypothétique lien à établir entre « contrat d'assurance » et « prime d'assurance », « courtier d'assurance », « attestation d'assurance », « code de l'assurance », « valeur d'assurance » et « société d'assurance ».

Les regroupements que nous proposons automatiquement sont un moyen pour l'expert de faire une sélection plus rapide parmi les candidats-termes extraits du corpus.

A la suite de ces différentes analyses et de la sélection par l'expert des syntagmes considérés comme représentatifs du domaine, nous obtenons une liste des concepts du domaine. Il nous faut maintenant identifier les relations respectives qui les unissent pour élaborer un réseau de concepts, une base de connaissances terminologique.

2.2 Etiquetage des liens lexicaux entre les termes

Afin d'identifier des relations pouvant exister entre des termes du domaine, un premier moyen peut être l'analyse du contexte d'un concept. Notre méthode étant en effet entièrement fondée sur le corpus, celui-là même peut nous permettre d'identifier les termes proches dans le même contexte. La phrase est ainsi définie comme un bon niveau de granularité pour une interprétation référentielle des concepts entre eux qui permettra la définition d'indice de corrélation. Si des termes s'avèrent être souvent associés, dans le sens où leurs occurrences sont corrélées dans un nombre significatif de contextes, on pourra conclure à une relation à établir dans notre réseau conceptuel. Ces algorithme de corrélation entre mots sont à ce jour en cours de développement. Un point de vue contextuel sur les concepts combiné à ceux-ci permet, dans le cadre d'un système d'information qui est le notre, d'améliorer sensiblement les performances de pertinence de la recherche. C'est ce que montrent [14] dans leur étude des techniques appropriées à la recherche documentaire juridique.

Autre possibilité offerte pour étiqueter les liens entre les termes du domaine : l'analyse de ce que [12] dénomme les *schémas lexico-syntaxiques*. Les schémas lexico-syntaxiques sont construits autours de SN simples ou complexes qui comprennent les candidats-termes du domaine, ils représentent des segments de phrase sémantiquement pertinents. Ces schémas sont alors à envisager comme des indices linguistiques agissant comme « marqueurs discursifs pour déclencher du sens ».

Ces schémas se présentent sous cette forme :

SN1	*Vpivot*	SN2

Deux syntagmes nominaux articulés autours de ce qu'on appelle un *verbe pivot*.

Ainsi dans les titres de nos documents, on rencontrera des schémas du type :

- Norme1 *portant modification de* Norme2
- Norme1 *relatif à* SN1 ….

L'analyse de ce type de schémas permet, au niveau des titres de nos documents,

d'identifier le rôle de la norme que nous développerons plus loin.

Dans cette perspective de l'analyse des *schémas lexico-syntaxiques* exprimés dans les documents, une attention particulière peut être portée à une de leur instance : les relations de définition permettant d'identifier des *contextes définitoires* [13]. Dans le corps des documents de notre corpus un certain nombre de marqueurs lexico-syntaxiques ont été isolés tels que :

- SN1 *est fixé à* SN2
- SN1 *comprend* SN2 *et* SN3
- SN1 *résulte de* SN2 …

Le repérage de ces marqueurs permet d'identifier des relations définitoires entre termes ce qui permet de proposer une relation d'association ou à tout le moins une relation de spécification (hyponymes) ou de généralisation (hyperonymes).

Dans cette première phase de construction d'un réseau conceptuel, nous nous sommes attachés à identifier les entités envisageables de notre modèle conceptuel du domaine et leurs relations linguistiques.

3 Plusieurs modèles du domaine

Avant de passer à l'étape de *conceptualisation* qui permet, à partir d'un réseau conceptuel – thesaurus - de développer un modèle conceptuel du domaine – ontologie -, il nous a semblé nécessaire de présenter les modèles que nous avons élaborés à partir de nos documents. Ces modèles sont posés comme des préalables au choix de la méthode de représentation des connaissances, ils sont envisagés comme des composantes du modèle conceptuel final. Ces modèles sont le résultat d'un autre type d'acquisition de connaissances à partir de nos documents que celui présenté précédemment (section 2). Nous nous sommes attachés ici à modéliser les connaissances *endogènes* de nos documents, les connaissances qui ne sont pas directement exprimées comme le sont les concepts du domaine, mais qui sont tout de même à prendre ne compte.

3.1 Le modèle hiérarchique

La particularité première de notre corpus est que les documents sont à envisager les uns par rapports aux autres dans des rapports hiérarchiques. Nos documents sont en effet les textes normatifs de la République et ils faut les considérer comme tels c'est-à-dire hiérarchiquement ordonnancés selon ce qu'on appelle la pyramide des normes.

Nous schématisons dans la *Figure 1* ci-dessous présentée cette hiérarchie juridique des textes de notre corpus, hiérarchie qui, s'appliquant aux contenants, influence les rapports des concepts exprimés dans les contenus des documents. Cette distinction dans les documents juridiques des contenants et des contenus a été avancé par J.P. Buffelan-Lanoré [3]. Nous avons choisi de "mettre à part" (par le biais de lignes pointillées) de cette hiérarchie les décisions des autorités administratives et les circulaires par rapport aux autres normes. Les circulaires sont en dehors de la pyramides puisqu'elles ne sont pas des normes à proprement parler, elles ont du moins une force exécutoire moindre. Les décisions des autorités administratives sont également mises à part mais placées toutefois au niveau réglementaire. Les flèches pleines signifient que ces textes sont modifiables par d'autres textes de même niveau hiérarchique et surtout de même type ; un décret est modifiable par un autre décret et bien évidemment par toute norme supérieure dans la hiérarchie. Enfin les flèches ascendantes signifient que les textes du dessous doivent se conformer aux textes du dessus. L'unique flèche descendante entre les lois et leurs décrets d'application ne signifie pas que les premières doivent se conformer aux seconds mais qu'il peut arriver que des concepts normatifs définis dans des lois soient explicités et précisés dans leurs décrets d'application, ce qui fait qu'en quelque sorte un décret peut influer sur une loi ce que nous ne pouvions pas nous permettre d'omettre dans notre représentation.

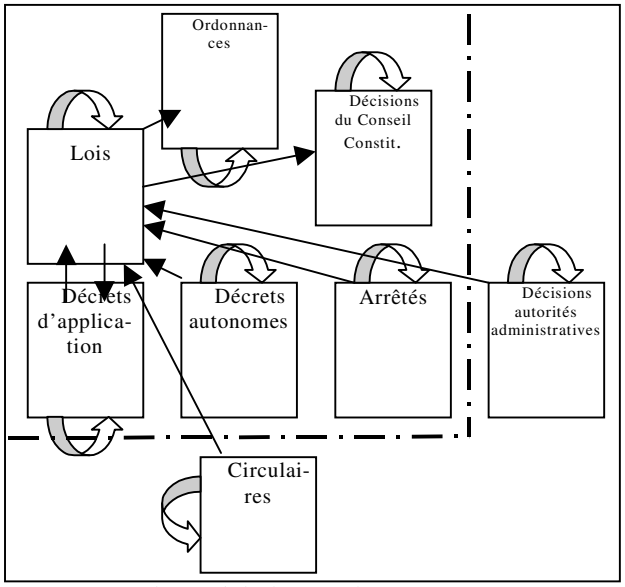

Figure 1 : le modèle hiérarchique

Ce modèle hiérarchique permet d'appréhender notre corpus d'un point de vue *supra-conceptuel*.

3.2 Le modèle structurel

Les documents que nous traitons répondent à un certain formalisme quant à leur structure interne. Nous pouvons donc aisément les segmenter puisque tous répondent à ce formalisme.

Les documents se découpent ainsi tel que schématisé dans la *Figure 2* :

- le titre comprenant le type du document (décret, arrêté …), son numéro (sauf pour les arrêtés), la date et enfin la spécification de l'objet du texte
- le numéro NOR permettant d'identifier le document
- le corps du document qui comprend la spécification de l'organe qui prends la décision, les visas, les articles avec le cas échéant des sous-divisions (I, II, …), pour les décrets et les arrêtés, le dernier article est celui qui précise les organes chargés de l'exécution du texte et enfin le dernier élément du document comporte la signature.

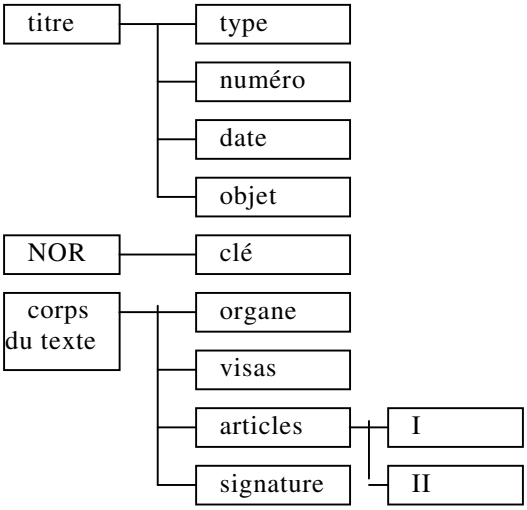

Figure 2 : le modèle structurel

L'identification de tous les segments de nos documents permet donc d'envisager leur structure même. Elle permet également de définir les segments des documents fortement porteurs d'information. C'est ainsi bien sûr dans les articles même de nos documents que se trouvent exprimés les concepts normatifs mais également dans le titre, premier fournisseur d'indices sur le contenu conceptuel des documents. La spécification du type de document et les visas qui renvoient à d'autres documents du corpus nous permettent de nous placer dans le modèle hiérarchique précédemment établi, de situer le document par rapport aux autres et d'établir les liens qui le relient aux autres.

Ce modèle structurel des documents nous permet également d'envisager une amélioration de notre base documentaire par l'utilisation du standard XML qui définit un ensemble de règles pour identifier la syntaxe et l'organisation des documents. Ce standard est appelé à de développer dans la définition des documents – tels que les nôtres – diffusés sur Internet. Un balisage de nos documents peut être envisagé avec des tags correspondants aux segments : <TITRE>, <TYPE>, <NUMERO> …

Ce balisage nous permettra dès lors de considérer nos documents comme organisés dans une base de données comportant les champs titre, NOR ..., l'interrogation en sera d'autant plus facilitée puisque qu'une appro-

che « *sémantique* »[9] des documents est ainsi possible.

3.3 Le modèle fonctionnel

Les documents que nous traitons expriment donc des normes qui sont organisées d'un point de vue hiérarchique et structurel. Nous nous attachons ici à envisager leur caractère fonctionnel.

La prise en considération de la dimension fonctionnelle dans des modèles du droit semble faire l'unanimité parmi les chercheurs. Ainsi A. Valente propose une *ontologie fonctionnelle du droit* [17] et distingue les catégories de connaissances en fonction de leur rôle dans le droit. L'étude de la dimension fonctionnelle des documents juridiques a également été menée par E. Pietrosanti et B. Graziadio [14] qui proposent d'ajouter un contexte fonctionnel aux concepts du domaine.

Notre point de vue, forcement conditionné par l'application visée, est de ne considérer que les connaissances normatives et d'identifier leurs fonctions propres. Nos concepts normatifs du domaine pourront dès lors contenir une dimension fonctionnelle fort utile à leur correcte appréhension par l'utilisateur.

La phase d'identification des *termes-pivot* des titres de nos documents (2-2) nous a permit d'identifier les différentes fonctions des normes traitées.

On peut ainsi distinguer les normes qui agissent sur l'ensemble normatif même des normes agissant sur l'extérieur, principalement en fixant les conditions d'existence d'entités juridiques (*artefacts*) par elles créées.

Les normes agissant sur l'ensemble normatif, ayant *fonction organisationnelle*, sont identifiables par la présence dans leur titre même de *verbes-pivot* tels que :

- N1 *abrogeant* N2
- N1 *modifiant* N2
- N1 *relatif à l'extension d'avenants* de N2
- N1 *portant application* de N2 …

Les normes agissant sur l'extérieur ont des *fonctions organiques* dans le sens où elles portent création d'une entité juridique – un *concept normatif* - dont elles déterminent également les conditions d'existence et de fin. Ces artefacts juridiques peuvent être une commission, un fonctionnaire, un statut, un immeuble domanial, une réserve naturelle …

Ces normes seront identifiables dans leur titre par des mentions telles que :

- N1 *portant nomination*
- N1 *portant classement dans le domaine public*
- N1 *portant modification des statuts*
- N1 *portant dissolution …*

La *Figure 3* ci-dessous présentée schématise l'action d'une norme sur le monde réel. L'ensemble normatif comprenant les caractéristiques hiérarchiques et structurelles présentées précédemment.

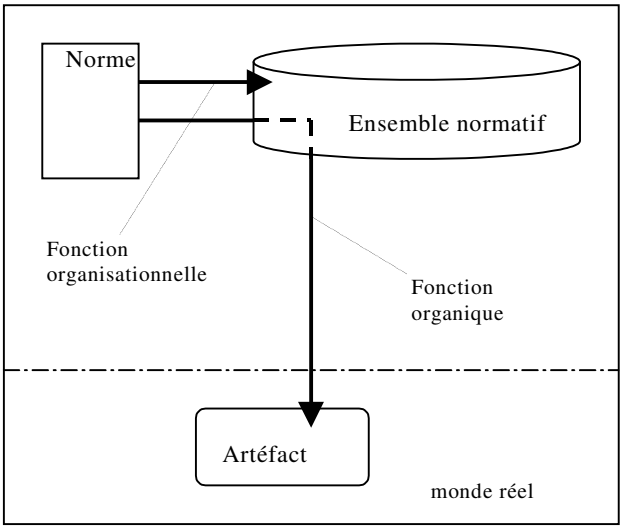

Figure 3 : le modèle fonctionnel

Ces trois modèles formalisent les connaissances implicites de nos documents. Le modèle final – l'ontologie - devra en tenir compte et les combiner.

3 Vers une ontologie du domaine

La phase préalable d'élaboration de la base de connaissances terminologique n'étant pas achevée au moment de la rédaction de ces lignes, nous ne pouvons présenter que les objectifs de ces étapes. Nous pouvons cependant affirmer que les résultats obtenus dans les phases précédentes d'acquisition de connaissances à partir des documents nous guiderons fortement.

4.1 La conceptualisation

Les spécialistes de la matière s'accordent pour avancer que la conceptualisation, ou *normalisation* selon B. Bachimont, est l'étape permettant de passer d'un ensemble de termes du corpus reliés lexicalement à un ensemble de concepts du domaine reliés sémantiquement [1]. Cette étape se divise en deux parties, une première consacrée à l'exploitation des données de la base de connaissances terminologique et une deuxième qui consiste en la conceptualisation proprement dite : structuration des concepts et identification de leurs relations sémantiques.

Dans notre étude, l'analyse contextuelle menée sur les syntagmes nominaux extraits du corpus ainsi que l'étude de leurs cooccurrences nous permet, nous l'avons vu, d'établir une sorte de définition de ceux-ci, de dégager des liens frôlant la sémantique. Nous pouvons dès lors sélectionner les termes en fonction des objectifs du modèle et établir des définitions de ceux-ci en s'aidant éventuellement de dictionnaires du droit (de très bons sont disponibles en ligne) si la méthode uniquement fondée sur le corpus ne propose pas de solutions définitoires satisfaisantes.

L'analyse du caractère fonctionnel des normes, principalement leur *fonction organique*, permet d'identifier des relations sémantiques entre les concepts définis par ces normes.

Ce que nous obtenons à ce stade n'est qu'un modèle semi-formel de la matière qu'il nous faut donc formaliser.

4.2 La formalisation

Cette phase consiste donc à formaliser le modèle intermédiaire obtenu dans l'étape précédente.

Nous nous servirons pour ce faire des trois modèles présentés précédemment. Ces modèles doivent être considérés comme les fondements de notre ontologie, celle-ci devant en respecter les principes. Ainsi le modèle hiérarchique pourra constituer une sorte d'ossature du modèle puisque les concepts exprimés dans les normes sont organisés suivant la hiérarchie même des documents posant ces normes. Le modèle fonctionnel définissant les différentes relations pouvant exister entre des normes, il pose par là les liens à établir entre les concepts exprimés par ces normes. Enfin le modèle structurel servira de modèle à l'organisation des segments de nos documents.

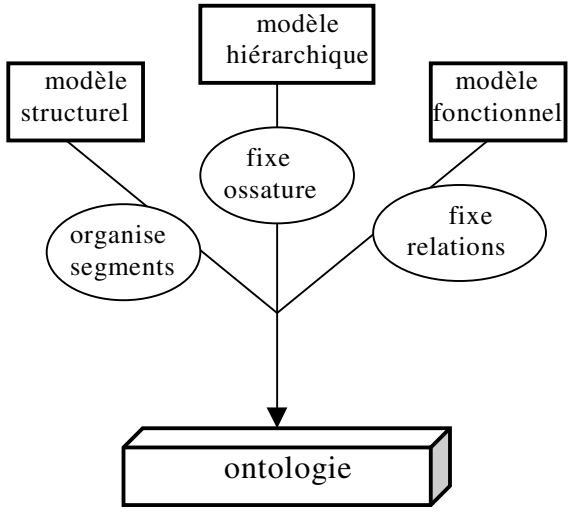

Figure 4 : combinaison des modèles

A l'issu de cette formalisation, nous devrions donc obtenir un modèle, une *ontologie du domaine* telle que définie par J. Breuker [16].

4.3 La réactualisation du modèle

Le problème est récurent dans notre domaine, de nouveaux concepts normatifs apparaissent régulièrement, résultats de l'édiction de nouvelles normes. Ainsi sont apparus dernièrement les concepts de « pacte civil de solidarité » ou de « réduction du temps de travail ». Une réactualisation de notre modèle est donc à prévoir si on veut que celui-ci demeure pertinent au regard du corpus.

La réactualisation devra être envisagée en deux phases : d'une part l'insertion du nouveau concept normatif dans la base de connaissances terminologique – le thesaurus - et ensuite une insertion éventuelle du concept comme nouvelle primitive du modèle.

L'intégration du nouveau concept dans le thesaurus ne devrait pas poser de problème majeur. Si le concept correspond à une nouvelle entité juridique, il suffit de lui trouver une place auprès de ses congénères. Ainsi le

PACS serait à classer parmi les types de contrats. L'apparente facilité de la tâche ne doit pas faire oublier qu'elle repose sur une solide analyse juridique préalable. Il faudra donc qu'un expert décide de l'insertion d'un nouveau concept dans le système, et où. L'expert sera en cela aidé par la norme même. En effet le droit est une discipline plutôt rétive aux grands bouleversements et la création d'une nouvelle entité juridique sera le plus souvent accompagnée d'une qualification de celle-ci dans une *supra-entité* déjà existante. Ainsi en 1998, quand le législateur a décidé de considérer juridiquement en tant quel tel l'objet base de données, il s'est tout d'abord attaché à le catégoriser, ce qui a conduit à la qualification des bases de données en « œuvre de l'esprit » au sens du Code de la propriété intellectuelle. Le concept « base de données » figure donc désormais dans les thesaurus juridiques comme instance du concept « œuvre de l'esprit ».

Suite à l'intégration du concept dans le thesaurus – la base de connaissances terminologique – la deuxième étape consiste en son insertion éventuelle dans le modèle conceptuel. Dans ce cas, soit le concept est mineur du point de vue de son importance dans l'ordonnancement juridique et il peut prendre la place d'un étiqueteur de concept, soit il est majeur et il faudra là envisager une éventuelle modification du modèle. L'hypothèse étant, nous semble-t-il rarement remplie du fait de cette caractéristique de la discipline que nous avons expliquée plus haut.

Quoi qu'il en soit, la réactualisation du modèle et de la base de connaissances terminologique doit être conférée à un expert du domaine qui sera aidé en cela par les outils que nous avons décrits dans le deuxième point, c'est-à-dire les outils qui servent à l'élaboration de la base.

Dans notre cas d'une augmentation journalière du corpus et en tenant compte des particularités du domaine on peut envisager un processus hebdomadaire de réactualisation du modèle.

5 Discussion

5.1 Positionnement du travail

Les premières visions ontologiques du droit, qui sont l'œuvre de théoriciens du droit, ont consisté à envisager la matière comme un ensemble de règles à combiner à un ensemble de faits. Ces points de vue ont été des points de départs pour une grande partie des travaux combinant l'intelligence artificielle et le droit jusqu'à la moitié des années 1990 et la thèse de A. Valente [15]. A. Valente a en effet montré qu'il y a dans le droit différentes catégories de connaissances qui devront être envisagées séparément, H. Kelsen, théoricien autrichien du droit les a partiellement identifiées au début des années 1990. Globalement une distinction peut être faite, en envisageant l'ensemble du domaine juridique, entre les connaissances normatives qui fixent un standard, les normes de commandement, les normes conférant pouvoir et les normes dérogatoires.

Notre travail se situe dans la continuité de celui de A. Valente dans la mesure où nous envisageons les connaissances juridiques en tant que telles et non plus les faits séparés des règles. Notre travail se focalise sur la partie normative des connaissances juridiques puisque notre corpus est tel. Par ailleurs, chaque modèle répond à des questions précises et est fortement conditionné par le corpus de référence. Notre objet est documentaire tandis que celui de A. Valente était d'élaborer un système décisionnel juridique. C'est en cela que nous nous démarquons du modèle qu'il présente et également par le fait que notre méthode s'efforce de se fonder uniquement sur les documents, par le fait que la connaissance utilisée est acquise des documents eux-mêmes.

Nous ajoutons cela s'ajoute le fait que chaque chercheur a sa propre conceptualisation, c'est ce que constatent [18] dans leur étude comparative de différentes ontologies du droit existantes.

5.2 Une éventuelle réactualisation de notre ontologie

Nous venons de présenter une méthode pour élaborer une ontologie, méthode fondée sur l'analyse des documents d'un corpus. Notre ontologie est, nous l'avons dit, une ontologie de domaine puisqu'elle est construite à partir des documents constituant un sous-domaine du droit, le sous-domaine normatif. Dès lors se pose la question de la réutilisabilité de notre modèle : peut-on l'appliquer à d'autres sous-domaines du droit ?

Selon [18], plus une ontologie est dépendante d'un domaine, plus elle est détaillée, moins elle est transposable à d'autres domaines. Or notre ontologie, par la méthode utilisée pour la concevoir, est fortement dépendante du domaine et sa finalité documentaire fait qu'elle est également très fouillée. La réutilisabilité de notre ontologie n'en est pas condamnée pour autant.

Ainsi, les documents de notre corpus expriment des *concepts normatifs* du droit qui ont vocation à s'appliquer juridiquement. L'application passera par le juge (nous envisageons là les documents constituant un sous-domaine du droit qui est la jurisprudence) et cette application sera commentée et expliquée (via les documents constituant la doctrine). On pourrait alors avancer que les *concepts normatifs* exprimés dans les documents de J.O. seront forcement repris dans des documents de jurisprudence et de doctrine. Il s'avère que cette déduction est vraie, mais elle ne l'est que partiellement. Les tribunaux et les chaires de droit sont en effet des lieux privilégiés de l'exégèse du droit, de la création de nouveaux *concepts normatifs*. Ils appliquent et commentent bien évidemment les notions juridiques telles que définies par le législateur mais vont également bien au-delà.

On peut donc envisager une adéquation partielle de notre ontologie à d'autres documents juridiques, à d'autres sous-domaines du droit. Cela dit, on peut tout à fait appliquer la méthode ici développée à ces documents et élaborer ainsi de nouvelles ontologies comprenant des concepts du droit non *normatifs.* Ces ontologies de domaine étant éventuellement agrégeables en une ontologie globale.

5 Références

[1] P. Aussenac-Gilles, B. Biébow et S. Szulman. Modélisation du domaine par une méthode fondée sur l'analyse de corpus. *Actes de la Conférence IC'2000*, pages 91-104, 2000.

[2] D. Bourigault, L. Charlet. Textes et ontologies. *Actes de la Conférence IC'2000*.

[3] J. Breuker, A. Muntjewerff, B. Bredeweg. Ontological modelling for designing educationnal systems. Dept. of Computer Science & Law, University of Amsterdam, June 17, 1999

[4] J.P. Buffelan-Lanoré. Informatique juridique documentaire. Edition de l'Espace Européen, la Garenne-Colombes 1991.

[5] P. Catala. Linguistique et informatique juridique. *Ecrits en hommage à Gérard Cornu*, pages 39-45, Editeur, Paris 1994.

[6] P. Catala. Le droit à l'épreuve du numérique., vol 2, pages 39-45, Editeur, 1998.

[7] N. Guarino. The ontological Level. *Proceedings of the 16th Wittgenstein Symposium*, Kirchberg, Austria, August 1993.

[8] T. Gruber. What is an ontology ? *A translation approach to portable ontologies.* Knowledge acquisition, 5(2) : 199-220, 1993

[9] R. Hellman. A Semantic Approach Adds Meaning to the Web. *Computer*, pages 13-16, IEEE Computer Society, December 1999.

[10] C. H. Hwang. Incompletely and Imprecisely Speaking : Using Dynamic Ontologies for representing and retrieving Information. InfoSleuth™, Microelectronics & Computer Technology Corp. (MCC), Austin, Texas, USA, June 1999.

[11] S. Luke, L. Spector, D. Rager. Ontology-based Knowledge Discovery on the World-Wide Web. *Proceedings of the Workshop on Internet-based Information Systems*, AAAI-96.

[12] E. Morin. Extraction de liens sémantiques entre termes à partir de corpus de textes techniques, thèse IRIN, Nantes, 1999.

[13] J. Rebeyrolle. Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. *Actes de la Conférence IC'2000*, pages 105-114, 2000.

[14] E. Pietrosanti, B. Graziadio. Advanced techniques for legal document processing and retrieval. *Artificial Intelligence and law*, vol 7, pages 341-361, Kluwer Academic Publishers, 1999.

[15] A. Valente. Legal knowledge enginering : a modelling approach, University of Amsterdam, The Netherlands, IOS Press, 1995.

[16] A. Valente, J. Breuker. Toward principled core ontologies. Proceedings KAW'96, Banff (Canada), November, 1996.

[17] A. Valente, J. Breuker. A functional ontology of law. *Artificial Intelligence and law*, vol 7, pages 341-361, Kluwer Academic Publishers, 1999.

[18] P.R.S. Visser, T.J.M. Bech-Capon. A comparison of four ontologies for the design of legal knowledge systems. *Artificial Intelligence and law*, vol 6, pages 27-57, Kluwer Academic Publishers, 1998.