ÉCOLE DES MINES DE PARIS

École des mines de Paris
Collège doctoral

N° attribué par la bibliothèque
/__/__/__/__/__/__/__/__/__/__/

Thèse

pour obtenir le grade de

Docteur de l'École des Mines de Paris
Spécialité : Informatique Temps Réel, Robotique et Automatique

présentée et soutenue publiquement
par
Guiraude LAME

le 6 décembre 2002

CONSTRUCTION D'ONTOLOGIE A PARTIR DE TEXTES
Une ontologie du droit dédiée à la recherche d'information sur le Web

Jury

Mme.	D. Bourcier	Rapporteur
M.	D. Bourigault	Examinateur
M.	J. Breuker	Examinateur
M.	S. Cottin	Examinateur
M.	C. Fluhr	Rapporteur
M.	R. Mahl	Directeur de thèse

Remerciements

Je tiens tout d'abord à remercier Robert Mahl, Directeur du Centre de Recherches en Informatique de l'Ecole des mines de Paris, Directeur de ma thèse, sans qui ce travail n'aurait pu exister. Je lui suis fort reconnaissante d'avoir bien voulu croire en ma candidature dans cette discipline et sur ce sujet de thèse. Robert Mahl a mis à ma disposition un ensemble de moyens qui m'ont permis de mener à bien ces travaux, qu'il trouve dans ces quelques mots l'expression de ma gratitude.

Je remercie également les membres du Centre de Recherches en Informatique de l'Ecole des mines de Paris de m'avoir accueilli en leur sein.

Mes remerciements s'adressent également à Didier Bourigault, chercheur au CNRS, qui m'a accompagné durant les principales avancées de ces travaux. Son outil Syntex m'a permis d'obtenir des résultats structurants et de qualité.

Je remercie Danielle Bourcier, Directrice de recherche au CNRS, d'avoir rapporté mes travaux et présidé le jury de soutenance. Je remercie également Christian Fluhr, Professeur au CEA, d'avoir accepté de rapporter mes travaux.

Je remercie également Joost Breuker, Professeur à l'Université d'Amsterdam, et Stéphane Cottin, Greffe du Conseil Constitutionnel, pour m'avoir fait le plaisir de faire partie du jury. J'exprime toute ma reconnaissance à chacun des membres du jury pour les remarques constructives qu'ils ont apporté à mes travaux.

Enfin, merci à mes proches, parents et amis, à Alexandre, d'avoir été de mes plus fervents supporters tout au long de cette course de fond qu'est une thèse.

Introduction

Nul n'est censé ignorer la loi.

Nous serions tentés d'ajouter à la célèbre maxime : *encore faut-il être en mesure d'y accéder !*

L'informatique tient, depuis quelques temps déjà, une place importante dans la problématique de l'accès au droit. Les premières bases de données juridiques électroniques[1] se sont étoffées avec le temps et l'augmentation des capacités de stockage informatique. Ces initiatives ont ainsi posé les principes d'une discipline encore jeune[2] : l'informatique juridique dont l'objet est le traitement informatique de données juridiques (documents, systèmes de règles etc).

Internet, de par ses capacités de diffusion et ses facilités d'accès, devient conjointement un acteur majeur de cette problématique. La création, en France, d'un service public d'accès au droit par l'instauration du *service public de la diffusion du droit par l'internet* (Décret n° 2002-1064 du 7 août 2002[3]) est, de ce point de vue, significatif.

Dès lors qu'un grand nombre de documents relatifs au droit français sont gratuitement consultables sur Internet (l'essentiel des textes législatifs et réglementaires ainsi que de nombreuses jurisprudences), la question devient celle-ci : comment accéder à ces documents ? Si nul ne peut se soustraire à la loi sous prétexte qu'il l'ignore, toute personne devrait être en mesure d'y accéder. Au-delà de la question de l'intelligibilité de la loi et du rôle majeur que peuvent tenir les professionnels du droit, cette injonction induit que chacun devrait être en mesure de trouver les textes juridiques répondant à ses interrogations.

Le principal moyen d'accéder aux documents sur Internet, et *a fortiori* aux documents juridiques, réside dans les moteurs de recherche. Schématiquement, les moteurs de recherche sur Internet indexent les documents ; ils stockent dans des fichiers les informations relatives aux mots présents dans ces documents et relient ces mots à l'adresse des documents[4]. Ainsi, quand un utilisateur soumet une requête composée d'un ou plusieurs mots à un moteur de recherche, celui-ci lui renvoie en réponse la liste des adresses des documents contenant ces mots.

Nos travaux se situent à cette étape cruciale de la recherche d'informations, celle où un utilisateur formule une requête à un moteur de recherche. Cet instant peut en effet être le siège d'une contradiction majeure, celle prévalant entre le langage de l'utilisateur et le langage des documents indexés

[1]Jurisdata par exemple.
[2]La trentaine !
[3]http ://www.legifrance.gouv.fr/citoyen/jorf_nor.ow ?numjo=PRMX0205836D
[4]Cette adresse est l'adresse *URL*, *Unified Ressource Locator*, du type http ://www.site.net/document.html

par le moteur[5]. Cette distinction peut engendrer de nombreuses frustrations pour l'utilisateur constatant assez vite qu'il n'est pas en mesure de trouver les documents susceptibles de répondre à sa question. L'inadéquation de la requête de l'utilisateur est en effet un problème récurrent de la recherche d'informations, et ce pour plusieurs raisons. D'une part, les termes[6] choisis par l'utilisateur pour formuler sa requête peuvent ne pas correspondre à ceux des documents dans lesquels sa recherche s'effectue. Les résultats de la requête sont alors un nombre insuffisant de documents, ou des documents inadéquats, ne traitant pas le thème de la requête de l'utilisateur. D'autre part, le nombre de mots composant les requêtes est bien souvent trop faible, il varie en moyenne entre 2.3 et 2.5 mots. Il en résulte des requêtes souvent peu précises auxquelles correspondent un nombre trop important de documents, en ce sens qu'ils ne peuvent tous être lus ou au moins consultés.

Pour pallier ce type de problème dans le domaine particulier du droit français, nous avons mis au point un système d'aide à la reformulation de requête. Ce système s'insère entre la requête de l'utilisateur et la réponse que peut en faire le moteur de recherche. Il consiste à analyser cette requête, et, sur la base de connaissances du domaine, suggérer à l'utilisateur des termes plus ou moins liés à ceux de sa requête. Celui-ci a alors toute latitude pour choisir parmi eux ceux lui permettant de formuler une nouvelle requête dont le thème peut être plus précis ou plus étendu, en tous les cas en meilleure adéquation avec le langage utilisé dans les documents recherchés. L'objectif étant pour l'utilisateur d'être en mesure d'accéder aux documents juridiques répondant à ses questions.

L'objet central de nos travaux a consisté à rassembler l'ensemble des connaissances du domaine nécessaires et adéquates pour faire fonctionner un tel système d'aide à la reformulation de requêtes. Nous considérons les textes comme les médium dans lesquels ces connaissances sont exprimées. Notre objet devient alors de découvrir dans les textes les connaissances nécessaires au fonctionnement d'un système d'aide à la reformulation de requête.

L'ensemble de ces connaissances se structure dans une *ontologie*. Une ontologie constitue un modèle d'un domaine donné, un point de vue qui peut être pris sur ce domaine, principalement orienté par l'application visée. La construction d'ontologies relève alors de cette branche de l'intelligence artificielle appelée ingénierie des connaissances qui s'attelle, entre autres, à la

[5]Cette différence entre le langage de requête et le langage d'indexation a été mise à jour, ainsi que ses conséquences majeures, dès les années 1980 par Salton et McGill dans leur ouvrage fondateur *Introduction to Modern Information Retrieval*, [SM83].

[6]Un terme est composé d'un mot ou d'un ensemble de mots.

construction de modèles à partir de l'expression linguistique des connaissances.

Les particularités de la démarche que nous avons suivie pour élaborer notre ontologie du droit sont les suivantes, elles sont conditionnées par l'usage auquel est destinée l'ontologie :

- l'ontologie que nous visons a une large couverture thématique, suffisamment importante pour être utile dans un contexte de recherche d'informations sur le Web ;
- aucune ressource extérieure telles des glossaires, des terminologies, des ontologies préexistantes etc, n'est directement utilisée ;
- l'accent est mis sur une procédure ascendante d'acquisition de connaissances à partir de textes qui est la phase majeure de notre démarche d'élaboration d'ontologie ;
- les résultats de cette phase ascendante constituent ce que nous appelons une *ressource ontologique*, structure de connaissances intermédiaire vers la définition d'une ontologie ;
- l'ontologie résulte de la modélisation des connaissances de la *ressource ontologique* ;
- enfin, il est procédé à une évaluation par l'usage de ce système d'aide à la reformulation de requête.

L'ontologie du droit français résultant de nos travaux est installée dans un système d'aide à la reformulation de requête. La visualisation de l'ontologie ainsi que sa navigation ont également été aménagées. Le tout est disponible en ligne gratuitement à l'adresse suivante : http ://ontologie.w3sites.net.

Nous décrivons dans cette thèse la démarche qui a été la nôtre pour élaborer cette ontologie du droit dédiée à la recherche d'informations sur le Web, ainsi que l'ontologie en résultant. L'apport principal de nos travaux est une nouvelle dimension donnée à la recherche en ingénierie des connaissances et, plus spécifiquement, à l'élaboration d'ontologies. Par la couverture thématique réalisée par notre ontologie qui contient plus de 100 000 éléments, nous proposons de nouvelles pistes de recherche parallèles aux chemins déjà tracés qui conduisent à la définition d'ontologies certes précises mais peu étendues.

Une Partie préliminaire (*Positionnements*) explicite chacun des choix méthodologiques qui a été fait, ainsi que le contexte de nos travaux. Cette Partie constitue une continuation de la présente introduction.

La première Partie, *Une ontologie du droit français codifié*, s'attache à positionner nos travaux quant à la définition des ontologies (Chapitre 1, *Définition des ontologies*), aux formalismes le plus souvent utilisés (Chapitre

2, *Formalismes pour les ontologies*), et quant à la problématique particulière liant les ontologies et le droit (Chapitre 3, *Ontologie et droit*).

La seconde Partie, *Elaboration d'une ontologie à partir de corpus* décrit chacune des étapes conduisant l'élaboration de notre ontologie du droit : la définition d'un corpus de textes sur lequel se déroule la phase ascendante d'acquisition des connaissances à partir de texte (Chapitre 4, *Un corpus de référence*), la découverte du premier élément de connaissance, les termes du domaine (Chapitre 5, *Extraction des termes*), celle du deuxième élement de connaissance, les liens entre les termes (Chapitre 6, *Identification de liens entre termes*), et, enfin, l'étape de modélisation des connaissances acquises à partir des textes (Chapitre 7, *De la ressource ontologique à l'ontologie*).

La troisième Partie, *Usages et évaluations d'une ontologie dédiée à la recherche d'informations sur le Web*, s'attache à la description des utilisations qui sont faites de cette ontologie du droit français et des évaluations de ces utilisations, dans sa consultation même (Chapitre 8, *Navigation et visualisation*), et dans son intégration dans le système d'aide à la reformulation de requêtes (Chapitre 9, *Aide à la reformulation de requête*).

Enfin la Conclusion permet d'envisager les perspectives de nos travaux notamment relatives aux autres utilisations pouvant être faites de notre ontologie. Elle dévoile également les limites de la démarche adoptée, ainsi que les limites intrinsèques de l'ontologie elle-même.

Partie préliminaire : Positionnements

0.1 Contexte

Nos travaux ont pour objet la construction et l'utilisation d'une ontologie du droit dans le cadre d'un site diffusant des documents juridiques sur Internet. Ce site[7] diffuse l'édition *lois et décrets* du Journal Officiel de la République Française qui comporte les documents suivants : les lois, les décrets, les arrêtés et les avis de diverses autorités administratives telles la Commission Nationale de l'Informatique et des Libertés, le Conseil Supérieur de l'Audiovisuel etc. Le site diffuse également l'ensemble des codes en vigueur en France tel le Code civil, le Code du travail, le Code de la marine marchande etc ainsi que d'autres documents juridiques tels les directives et réglements européens. Divers moteurs de recherche plus ou moins évolués opèrent sur ces documents[8].

L'ontologie que nous construisons est destinée à être intégrée dans un système interactif d'aide à l'interrogation de moteurs de recherche. Ces moteurs de recherche sont soit des moteurs spécifiques à un site Web en particulier, tels les moteurs de droit.org, soit des moteurs généralistes paramétrés pour interroger certains sites seulement ou le Web dans sa totalité, comme par exemple Google[9].

Nos travaux font clairement partie de cette branche de l'informatique appelée ingénierie des connaissances. Cette discipline s'attache [CKZB00] à :

> «l'étude des concepts, méthodes et techniques permettant de modéliser et/ou d'acquérir les connaissances pour des systèmes réalisant ou aidant des humains à réaliser des tâches se formalisant *a priori* peu ou pas».

Cette ontologie est conçue pour un usage, globalement la recherche d'informations sur le Web. L'usage (section 0.2) conditionne la forme même de l'ontologie, sa structuration, son niveau de formalisme ; nous parlons dès lors d'une *ontologie dans l'usage* (section 0.3). Nous décrivons la démarche globale adoptée pour le développement de notre ontologie (section 0.4). Cette démarche procède en deux temps majeurs. Le premier repose sur l'acquisition de connaissances à partir de textes. Un corpus de référence (section 0.4.1) est choisi sur lequel des outils et méthodes d'acquisition de connaissances à partir de textes sont appliqués (section 0.4.2). En résulte une *ressource ontologique*, ensemble de termes du domaine et de liens non qualifiés entre ces termes. Cette *ressource ontologique* est confrontée à l'usage dans

---

[7]http ://www.droit.org, ou, http ://www.admi.net/jo

[8]Ils sont accessibles sur l'adresse *op. cit.*

[9]http ://www.google.com

le cadre d'une procédure d'évaluation (section 0.4.4). Le deuxième temps de la méthode de construction de l'ontologie consiste en une modélisation des connaissances incluses dans la *ressource ontologique* (section 0.4.5). L'objectif de cette phase est de qualifier les liens entre les termes détectés dans la première phase. Nous obtenons dès lors une ontologie en tant que telle, c'est-à-dire un ensemble de termes du domaine reliés entre eux par des relations ontologiques comme la subsumption.

0.2 Usage

L'usage est dans notre démarche une notion essentielle. Il détermine et guide la démarche d'élaboration d'ontologie que nous avons suivie et l'ontologie en elle-même.

Nous construisons une ontologie dédiée à la recherche d'informations sur le Web, notamment à l'aide interactive à la reformulation de requête. La recherche d'informations sur le Web consiste schématiquement à interroger des moteurs de recherches sur des mots dont on pense qu'ils reflètent ce qu'on cherche. Ces mots constituent la requête soumise au moteur. Les moteurs de recherche classiques s'attachent alors à donner en réponse les documents du Web contenant les mots de la requête. Cette liste de documents peut s'avérer d'une longueur inexploitable pour un humain. Pour pallier ce problème, différentes techniques existent comme, par exemple, de pondérer les documents réponse de manière à mettre en début de liste les documents les plus consultés par ailleurs, les mieux référencés etc. Ces techniques se proposent ainsi d'exploiter des connaissances intrinsèques sur Web (logs, références mutuelles etc). Un autre type de technique consiste à classer thématiquement les documents réponse de façon non supervisée dans des classes ou de façon supervisée dans des catégories prédéfinies. Nous proposons pour notre part d'insérer, dans le processus d'interrogation de moteurs de recherche, un système d'aide à la (re)formulation de requête. Ce système repose sur une ontologie du domaine pour suggérer à l'usager un ensemble de termes plus ou moins liés à ceux de sa requête. Cet ensemble de termes permet à l'usager de préciser sa requête, de la restreindre, ou de la confirmer.

Intrinsèquement à la notion d'usage transparaît la figure de l'usager. Nous pensons en effet que l'usager est un élément fondamental à prendre en compte par l'ingénieur des connaissances dans sa démarche. Notre position est que l'usager est le seul agent à même de déterminer le sens et l'objet de ses recherches, la signification exacte de ses requêtes soumises aux moteurs de recherche. De ce point de vue découle un choix technique : notre système d'aide à la (re)formulation de requête se contente uniquement de suggérer à l'usager un ensemble de termes. Ce choix méthodologique va bien au-delà des *contraintes* techniques et scientifiques qui font qu'aucun outil ou méthode de traitement automatique des langues n'est actuellement en mesure de déterminer de façon sûre qu'un terme soit synonyme d'un autre, si tant est que la synonymie parfaite existe dans le langage. Même en présence de synonymes (presque) parfaits, nous ne pouvons envisager de les imposer à un utilisateur. Nous pensons que l'intelligence est bien dans le cerveau de l'homme et notre système est là pour donner à penser à l'utilisateur et non pas pour penser par lui-même[10].

D'autre part, le fait de suggérer à l'usager un ensemble de termes plus ou moins liés à ceux de sa requête comporte un aspect pédagogique dans l'appréhension du domaine particulier envisagé par l'ontologie. La pédagogie est ainsi un usage dérivé de l'usage principal auquel est dédiée notre ontologie.

0.3 Ontologie dans l'usage

Notre ontologie constitue, dans la lignée de la définition des ontologies donnée par Thomas Gruber [Gru93], une certaine conceptualisation formalisée du domaine juridique. Cette ontologie est, nous l'avons vu, une ontologie destinée à la recherche d'informations sur le Web. Cet usage, raison même de la création de l'ontologie, conditionne sa forme même, ses éléments, sa structuration et son degré de formalisme.

L'ontologie que nous présentons est une *ontologie de domaine* [Gua98] dans le sens où elle constitue une spécification d'une conceptualisation d'un domaine bien particulier : le droit français codifié.

Cette ontologie est également une *ontologie de tâche* [Gua98] dans le sens où elle est dédiée à un certain usage.

Conditionnés par l'usage, les éléments de l'ontologie sont principalement les termes du domaine et les liens détectés entre ces termes. Cependant, chacun de ces termes a une incidence relative sur les différents thèmes abordés par le droit français codifié : le droit civil, le droit du travail, le droit des assurances etc. Ainsi, le terme *brevet* est un terme spécifique du droit de

[10]Nous faisons alors nôtre une expression de B. Bachimont, notamment développée dans sa thèse de doctorat de l'Ecole Polytechnique en épistémologie *Herméneutique matérielle et artefactuelle : des machines qui pensent aux machines qui donnent à penser* (1996).

la propriété intellectuelle. Ces informations sur la provenance thématique des termes sont conservées et exploitées dans l'interface de navigation dans l'ontologie. Les termes et les thèmes sont ainsi reliés par des relations de provenance.

L'usage guide également la particularité de nos travaux : celle d'une large couverture thématique visée par l'ontologie. De manière générale, les ontologies développées jusqu'à présent l'étaient dans des domaines relativement restreints et comprenaient alors un nombre restreint de concepts. Or, le site Web qui est le contexte de nos travaux est un site généraliste du droit, et principalement du droit français. Nous ne pouvions dès lors envisager de n'élaborer qu'une ontologie des régimes matrimoniaux par exemple. Plus notre ontologie couvre largement le droit français, plus son utilité en recherche d'informations peut être avérée. L'usage guide alors la couverture thématique de notre ontologie et, conséquemment, impose des contraintes méthodologiques pour son élaboration.

0.4 Elaboration d'une ontologie dédiée à la recherche d'informations sur le Web

Dans la lignée des travaux et résultats de la communauté de recherche TIA[11], nous avançons qu'une ontologie ne peut être conçue que dans le cadre de l'usage qui lui est destiné. L'usage guide ainsi de façon déterminante la démarche d'élaboration d'ontologie que nous avons adoptée. Cette démarche s'inscrit dans les principes généraux posés par les recherches menées en ingénierie des connaissances, comme celui de l'extraction des connaissance de corpus de documents en exploitant des résultats d'outils de traitement automatique de la langue, ou celui de l'évaluation en usage. Ces principes généraux nous ont certes guidé dans notre démarche d'élaboration d'ontologie, mais il va sans dire que toutes les questions émergeant d'une telle démarche sont loin d'être résolues. Nous considérons donc ces travaux comme un cadre théorique et parfois méthodologique. Notre démarche a néanmoins cette particularité de s'attacher à un domaine très vaste et donc à des masses de résultats inédites. En découlent des choix méthodologiques se démarquant nettement des travaux évoqués, notamment le choix de la non validation par un expert des résultats des outils. De plus, pour des raisons de complétude

[11] Le groupe de travail TIA, Terminologie et Intelligence Artificielle, travaille, au sein du PRC-I3 et de l'AFIA, à la convergence entre la terminologie, la linguistique, l'ingénierie des connaissances et l'intelligence artificielle. Les articles fondateurs de leurs travaux sont principalement [Bac00], [AGBS00a] et [BS99].

et de cohérence de la méthode que nous adoptons, nous avons choisi de ne pas utiliser directement de ressources externes tels des thesaurus existants, des glossaires, dictionnaires, index etc.

La démarche que nous avons adoptée consiste en deux étapes principales. La première, étape ascendante essentiellement automatique, consiste à utiliser au mieux les outils et techniques de traitement automatique de la langue pour obtenir ce que nous nommons une *ressource ontologique*. Nous définissons ainsi un *corpus de référence* (Section 0.4.1) sur lequel sont appliqués des outils et techniques du traitement automatique de la langue, ou TAL, (Section 0.4.2). Cette ressource rassemble les termes du domaine ainsi qu'un ensemble de relations entre ces termes pour constituer un réseau lexical. Les relations entre les termes ne sont pas qualifiées ; seuls les résultats propres aux outils et techniques de TAL sont exploités.

La deuxième étape de la méthode d'élaboration d'ontologies consiste à se baser sur cette *ressource ontologique* pour procéder à une modélisation qui permettra de qualifier les liens entre les termes établis dans la première étape. La qualification sémantique des relations entre les termes conduit, dans notre démarche, à la définition d'une ontologie sachant que d'autres tâches auraient également pu être menées, notamment la sélection des termes, leur *conceptualisation* etc. La *ressource ontologique* est alors l'unique source de connaissances. Elle définit ainsi l'ossature de l'ontologie.

0.4.1 Choix du *corpus de référence*

La phase ascendante suppose de porter une attention toute particulière quant au choix des documents constituant le *corpus de référence*. Le corpus de référence est la base de la démarche d'acquisition de connaissances à partir de textes qui inclut des outils et techniques de traitement automatique des langues. Il s'avère en effet que le droit ne fait pas partie de ces domaines ou thématiques dans lesquels un corpus fini traitant le domaine peut servir de corpus de référence. Ce peut être le cas dans des expériences de constructions d'ontologies relatives à la maintenance d'une centrale nucléaire bien particulière ou relatives à un projet bien précis de conception de véhicule[12]. La multiplicité des sources du droit au sens où l'entendent les juristes (les traités internationaux, la constitution, les lois, les règlements, la doctrine etc) ainsi que la diversité des documents traitant de questions juridiques (analyses de

[12]Cas du système SAMOVAR, fruit de la collaboration de Renault et du projet Acacia de l'INRIA.

décisions de jurisprudence, notices pratiques etc) font qu'il est impossible de considérer l'ensemble de ces documents comme constitutif d'un corpus de référence. Un choix doit donc être fait.

Le corpus de référence doit présenter conjointement les caractéristiques suivantes : être suffisamment ramassé pour être traitable par des outils de traitement automatique des langues et couvrir suffisamment largement le droit français.

Nous avons choisi d'utiliser le corpus des *codes* du droit français. Il s'avère en effet que différents thèmes du droit français, tels le droit du travail, le droit pénal etc, ont fait l'objet d'une codification. Cette codification a consisté, schématiquement, à rassembler dans un même document (le code) toutes les normes législatives et décrétales (les lois et les décrets) relatives à un même thème. Il existe ainsi en France 68 codes officiels et nous disposons de 58 d'entre eux. Ce corpus des codes est, avec les lois et les décrets en tant que tels, constitutif d'une partie du droit normatif français.

Ainsi, bien que les codes ne traitent pas de toutes les thématiques du droit français, nous pouvons affirmer que ces codes constituent un des *noyaux* du droit français, d'une part parce que les codes sont une partie de cette source du droit que sont les normes et, d'autre part, parce que les thématiques essentielles de ces normes font l'objet d'un code.

L'ontologie issue de cette méthode d'élaboration fondée en partie sur une démarche ascendante d'acquisition de connaissances à partir de textes basée sur le corpus de référence constitué par les codes du droit français est dès lors une *ontologie de référence*, d'une part parce qu'elle est thématiquement relative à un des *noyaux* du droit français, le droit français codifié, et d'autre part, parce qu'alors, un processus d'élargissement sur sa base est tout à fait envisageable, incluant d'autres sources du droit français comme la jurisprudence.

0.4.2 Outils et techniques d'acquisition de connaissances à partir de texte

Dès lors que le corpus de référence est déterminé, les outils et techniques d'acquisition de connaissances à partir de texte doivent être identifiés.

Les démarches d'acquisition de connaissances à partir de textes reposent sur le postulat que les connaissances sont exprimées dans les documents par le truchement d'expressions linguistiques. Ce sont donc bien des méthodes relevant du traitement automatique des langues qui vont permettre d'identifier les différents éléments de connaissance exprimés dans ces documents.

Nous avons donc identifié et appliqué un ensemble d'outils et de techniques permettant l'identification de ces éléments de connaissance portés par les textes : les termes du domaine et les relations lexicales et syntaxiques entre ces termes. Ces éléments sont les éléments de notre ressource ontologique.

Pour ce qui est de l'identification des termes du domaine, nous utilisons le logiciel Syntex. Ce logiciel conduit des analyses syntaxiques des textes pour proposer un ensemble de candidats termes : noms, syntagmes nominaux, adverbes, syntagmes adverbiaux etc. Syntex intègre également une plate-forme d'*analyse distributionnelle* ([Bou02]) qui, se fondant sur les résultats de l'analyse syntaxique des documents, détermine, autour des verbes, un ensemble de *termes régis* et de *termes recteurs*. Cette *analyse distributionnelle* permet d'établir un ensemble de liens syntaxiques distributionnels entre termes. Nous développons tous ces points dans la Section 6.2.

Nous avons par ailleurs utilisé différentes techniques d'identification de liens entre les termes tels qu'extraits par Syntex.

(*i*) Une première technique a consisté à considérer les relations de coordinations entre termes présentes dans le corpus. Cette technique incluant une phase de validation manuelle, nous avons limité le corpus pour ne traiter que les intitulés des titres des divisions de chaque code. Chaque code est en effet divisé de manière thématique, chaque division faisant l'objet d'un intitulé. Nous considérons ainsi appréhender une part majeure des termes importants du droit, ceux qui sont présents dans les intitulés des divisions des codes.

(*ii*) Une deuxième technique utilisée pour la détection de liens entre termes repose sur des analyses statistiques des contextes d'apparition des termes dans le corpus. Cette technique se fonde sur le principe que deux termes peuvent être lexicalement liés dès lors qu'ils apparaissent souvent dans de mêmes contextes, qu'ils sont souvent entourés des mêmes mots ou termes. Les contextes que nous considérons sont les articles de codes. Chaque code, au-delà de ses divisions en sections et sous-sections, est en effet divisé en articles. Ces articles constituent au minimum, et la plupart du temps, une seule phrase.

(*iii*) Une troisième et dernière technique d'identification de liens entre termes, bien que triviale, montre son utilité dans le contexte qui est le notre de la construction d'une ontologie dédiée à un usage de recherche d'information.

Cette technique consiste à établir un lien entre un terme donné et l'ensemble des termes contenant ce terme.

Les résultats de l'outil Syntex doublés de ceux de nos propres techniques nous permettent, par une démarche ascendante d'acquisition de connaissances à partir de textes, d'identifier une partie des éléments de l'ontologie, à savoir les termes du domaine et les liens entre ces termes.

0.4.3 Choix de la non validation

Une des particularités de notre méthode d'élaboration d'ontologies dédiées à la recherche d'informations réside dans le choix que nous faisons de ne pas faire valider les sorties de nos outils et techniques d'acquisition de connaissances à partir de textes. Ce choix s'explique par plusieurs points, comme autant de justifications.

D'une part, au-delà du choix des techniques et de la détermination de seuils pour l'analyse distributionnelle de Syntex et l'analyse statistique des contextes, une intervention humaine dans la phase d'identification de liens entre termes n'est de notre point de vue pas nécessaire. En dehors du problème de la masse difficilement gérable pour un humain, problème corollaire de celui de la masse des candidats termes, chaque lien détecté entre termes est lexicalement ou syntaxiquement *justifié*. Dans le cas des liens détectés par l'analyse distributionnelle de Syntex, chaque lien entre deux termes est en effet *justifié* par l'ensemble des termes *recteurs* partagés par ces termes. De la même façon, chaque lien détecté par l'analyse des contextes d'apparition des termes est *justifiable* par ces contextes. Il s'avère ainsi qu'il est fort difficile de faire déterminer à un expert la pertinence, ou non, de l'établissement de liens entre termes dès lors que des arguments lexicaux et syntaxiques tangibles peuvent lui être présentés.

D'autre part, la masse des candidats termes extraits par Syntex de l'ensemble de codes (plus de 300 000) est telle que nous ne pouvions envisager une phase manuelle de validation. De plus, une simple sélection sur la fréquence des termes n'était pas appropriée. Bien que de nombreux termes, environ 50 %, s'avèrent n'apparaître qu'une seule fois dans l'ensemble du corpus des codes, les caractéristiques même du corpus font que ces termes demeurent importants. En effet, le principe même de la codification des textes normatifs français répond à une demande forte de *rationalisation* des documents juridiques, rationalisation qui ne souffre guère la répétition.

En filigrane de ce choix de la non validation transparaît l'idée, avancée plus haut, que l'usager est le seul à même de déterminer le sens de ses requêtes. Nous pensons en effet que c'est à l'usager du système interactif d'aide à la reformulation de requête de déterminer l'intérêt de liens établis entre des termes. Un lien entre deux termes, pourtant identifié comme invalide par un expert, peut tout à fait être utile à un usager dans un processus d'aide à la reformulation de requête, ne serait ce que parce que ce lien lui a suggéré une nouvelle idée sur sa requête. Notre méthode et notre ontologie étant, nous l'avons déjà dit, fortement ancrées dans l'usage, nous pensons que tout ce qui peut être utile à une personne dans cet usage doit être considéré.

Nous n'envisageons donc pas de phase de validation par un expert quelconque du domaine des éléments détectés dans la phase d'acquisition de connaissances à partir de textes comme éléments futurs de l'ontologie. Une phase d'évaluation dans l'usage, détaillée dans la section 0.4.4, est par contre intégrée à notre méthode de construction d'ontologie ainsi qu'une phase de modélisation des connaissances (Section 0.4.5).

Le choix étant fait du corpus de référence, les outils et techniques d'acquisition de connaissances à partir de textes étant appliqués à ce corpus de référence, nous obtenons les éléments termes du domaine et liens entre ces termes, résultats de la phase ascendante de la méthode d'élaboration d'ontologies. Ces éléments constituent ce que nous appelons une ressource ontologique, base à un processus de modélisation. Ce processus de modélisation est cependant précédé d'une évaluation dans l'usage de la ressource ontologique.

0.4.4 Evaluation par l'usage de l'ontologie

L'usage est, nous l'avons dit, un élément essentiel dans nos travaux. C'est pourquoi nous procédons à une évaluation des résultats de nos travaux. Cette procédure d'évaluation consiste à confronter les éléments de la ressource ontologique, futurs éléments de l'ontologie, à l'usage. L'objectif est d'établir l'utilité de la suggestion de termes du système d'aide interactive à la reformulation de requête.

Plus précisément, il s'agit de faire déterminer aux évaluateurs l'utilité documentaire d'avoir lié tel terme à tel autre. Ceci revient à faire dire à l'évaluatueur si tel terme suggéré en réaction à sa requête lui a permis, ou non, de trouver plus facilement les documents qu'il attendait en réponse, et pourquoi.

Pour ce faire, et pour faciliter le processus, nous faisons appel à des professionnels de la recherche documentaire juridique. Ces personnes sont documentalistes dans des cabinets d'avocats ou dans des bibliothèques universitaires. Elles sont, de par leur activité professionnelle quotidienne, rompues à la recherche de documents juridiques, tout en ayant de fortes connaissances dans ce domaine. Nous avons prédéfini avec l'une d'entre elles un corpus qui, de part son aspect juridique fondamental, est connu de chacune. Ce corpus leur permet de connaître les documents attendus en réponse à une requête déterminée. Un moteur de recherche a été installé sur ce corpus. Ce moteur de recherche permet aux évaluateurs de poser des requêtes booléennes à leur guise, combinant plusieurs termes, ceux de la requête et ceux suggérés par le système. L'utilité de la suggestion de chaque terme à trouver plus facilement les documents attendus est ainsi vérifiable. Globalement, plus d'un terme suggéré sur deux a été jugé utile par les évaluateurs.

0.4.5 Modélisation de la ressource ontologique

La ressource ontologique telle que décrite plus haut est l'élément de base de l'étape de modélisation qui conduit à la définition de notre ontologie. Nous utilisons les données de la ressource ontologique pour procéder à une conceptualisation du domaine. La ressource ontologique rassemble les termes du domaine autour de liens lexicaux et syntaxiques. La modélisation, ou conceptualisation, consiste dès lors à donner une sémantique *ontologique* aux liens lexicaux et syntaxiques. La modélisation telle que nous l'envisageons ne consiste pas à procéder conjointement à des éliminations ou ajouts de termes, nous travaillons sur la base des connaissances de la ressource ontologique, ni plus, ni moins. Cette étape de modélisation procède d'une expertise du domaine ; elle est donc menée par un expert du domaine qui s'avère être également l'ingénieur de la connaissance dans notre expérience (voir section 0.5).

Une première étape de la modélisation consiste à identifier un ensemble fini de relations envisageables entre les termes. Cette étape a consisté à modéliser *à la main* un ensemble de termes choisis au hasard. De cette modélisation expérimentale ressort un ensemble fini de qualifications sémantiques pouvant être appliquées aux liens de la ressource ontologique. Ces qualifications des liens permettent de leur donner une sémantique proprement ontologique, ainsi le lien de subsumption entre deux concepts.

La deuxième étape de la modélisation des connaissances de la ressource ontologique consiste à la modélisation en tant que telle. Cette étape nécessite préalablement de déterminer quelles notions vont être modélisées. Nous choisissons dans un premier temps de modéliser les notions portées par les termes présents dans les intitulés des divisions du code civil. Ce choix nous permet d'envisager les termes fondamentaux du droit français, étant donné que le Code civil peut être appréhendé comme la base du droit français.

Dès lors que l'ensemble fini des qualifications des liens lexicaux est déterminé ainsi que l'ensemble des notions à modéliser, une interface facilitant la tâche du modélisateur peut être mise en place. Elle consiste à lister pour chaque terme l'ensemble des termes qui lui sont liés dans la ressource ontologique. A chacun des liens ainsi exprimés est associé le choix alternatif entre les qualifications possibles de ce lien. Le modélisateur n'a dès lors, en quelque sorte, plus qu'à choisir quelle est la qualification a apporter à chacun de ces liens. Les données de la modélisation sont récoltées dans une base de données.

0.5 L'ingénieur de la connaissance

L'ingénieur de la connaissance, *nous* dans ce document, est la personne qui guide les choix relatifs à la méthode de construction de l'ontologie et qui, finalement, conçoit cette ontologie. Nous menons ici une analyse de son rôle, de ses caractéristiques. Le travail que nous présentons ici s'est également déroulé dans des conditions particulières que nous nous devons d'exposer, dans un souci de rationalisation scientifique de la démarche d'ingénierie des connaissances.

Classiquement, les projets d'ingénierie des connaissances consistant à identifier et à exploiter les connaissances d'un contexte, d'une organisation ou d'un domaine, voient une collaboration s'établir entre une équipe chargée de l'ingénierie des connaissances et une équipe d'experts du domaine ou de l'organisation en question. Ce travail d'équipe comporte l'avantage de combiner les capacités et connaissances de chacun, l'inconvénient pouvant être un manque de compréhension entre les deux équipes dû notamment à l'usage de discours différents ou à des attentes spécifiques et différentes.

Dans notre projet d'élaboration d'une ontologie du droit dédiée à la recherche d'informations sur le Web, nous endossons alternativement, mais aussi simultanément, le rôle de l'ingénieur de connaissances et celui de spécialiste du domaine. Cette caractéristique *bicéphale* de la personne en charge de l'élaboration de l'ontologie n'est pas à négliger. Cet aspect comporte des avantages certains, notamment dans l'appréhension simultanée des différents enjeux du projet : élaborer une ontologie utile, savoir identifier les connaissances fondamentales du domaine de celles qui le sont moins, résoudre les contradictions éventuelles entre la technique et le fond etc. Un des inconvénients en est un penchant occasionnel vers la schizophrénie.

Quoi qu'il en soit, les caractéristiques que nous venons de dépeindre ne font que confirmer une tendance, celle des compétences multiples et complémentaires qu'une même personne peut exploiter. Plus largement, la démarche globale de nos travaux, plus que de relever simplement de l'ingénierie des connaissances et du droit, relève aussi de l'informatique appliquée au droit, c'est-à-dire à l'*informatique juridique*[13].

[13]Egalement connue sous le nom d'*artificial intelligence and law* ou *computer science and law*.

Première partie

Une ontologie du droit français dédiée à la recherche d'information

La notion d'ontologie a cette particularité d'être l'objet de controverses sur sa définition même. Il n'est pas une conférence, congrés ou workshop abordant les ontologies qui ne consacre une session ou un débat à la définition de la notion. Il peut ironiquement être avancé que la notion d'ontologie ne semble pas faire, dans la communauté des chercheurs, l'objet d'un *engagement ontologique*, lors même que cet engagement, reflet d'une entente collective sur le sens des notions, est posé comme le préalable essentiel par ces mêmes chercheurs à l'élaboration d'ontologies.

L'objet de cette Partie I est d'expliciter les ontologies en général et celle que nous élaborons en particulier. Nous procédons tout d'abord (Section 1) à une définition des ontologies. Ces définitions s'avèrent multiples et à géométries variables en fonction des contextes dans lesquels elles apparaissent. Différentes façons d'aborder l'élaboration d'ontologies sont ensuite décrites pour enfin tenter de circonscrire la notion d'ontologie par le rapprochement avec des notions voisines. Nous abordons dans un deuxième temps (Section 2) les différents fomalismes développés pour *exprimer* les ontologies. Enfin, nous joignons les ontologies et le droit dans une Section 3.

Chapitre 1

Définition des ontologies

Le terme même d'*ontologie* trouve son sens premier dans la philosophie. L'ontologie [All97] y fait partie de la métaphysique, science de l'être en tant qu'être :

> «[...] science de l'être en tant qu'être, c'est-à-dire de l'être en général, de ses diverses espèces, de ses propriétés et de ses relations. Les modes de l'être sur lesquels elle discute sont : le possible, le réel et l'impossible, le potentiel et l'actuel, le contingent et le nécessaire, le déterminé et l'indéterminé, le fini et l'infini, le parfait et l'imparfait, la substance et le mode, l'essence et l'accident.»

Différentes écoles philosophiques l'ont abordée. L'ontologie s'est ainsi interprétée, d'Aristote à Descartes, comme une théologie, de Descartes à Kant et aux postkantiens comme la science du fondement de la connaissance et enfin, depuis Fichte et Schelling, comme une connaissance de l'être de l'homme.

L'utilisation du terme *ontologie* en informatique n'a rien de fortuit. Les instigateurs[1] de l'*emprunt* du terme à la philosophie ont fonctionné par analogie, mode de raisonnement éminemment aristotélicien. Il apparaît ainsi clairement que l'objet métaphysique de l'ontologie est bien d'expliciter l'être dans tous ses aspects, avec toutes ses facettes, et que l'objet informatique de l'ontologie y tend également, de façon intellectuellement plus modeste.

Une ontologie se définit en informatique essentiellement par les usages qui en sont fait, nous les aborderons (section 1.1). La discipline informatique s'est donc appropriée la notion d'*ontologie* pour en faire un outil, un artefact intégré à un système informatique. Dès lors, des méthodes d'élaboration de cet outil informatique qu'est une ontologie ont été dégagées (section 1.2). Les

[1] Ou plutôt l'instigateur : John McCarthy à la fin des années 1970

ontologies en général, et celles dédiées à la recherche d'information en particulier, se rapprochent et se distinguent à la fois d'objets comme les thesaurus, terminologies ou réseaux de concepts. Nous comparons ces différents objets ainsi que leurs méthodes d'élaboration relevant de l'ingéniérie linguistique dans la Section 1.3

1.1 Les ontologies en informatique, définition par l'usage

La définition la plus fréquemment citée des ontologies en informatique est ainsi celle donnée par Thomas Gruber :

> «An ontology is a formal explicit specification of a shared conceptualization.»[2]

Une conceptualisation renvoie à une abstraction du monde, à une vision qu'on peut avoir d'une partie du monde à un moment donné et avec un point de vue donné.

Plus précisément[Kas02a], la notion de *conceptualisation* renvoie à l'*intention* d'entité conceptuelles, aux *notions*, c'est-à-dire à des règles contraignant la référence de ces entités.

Nous nous plaçons ainsi dans l'*héritage* de Thomas Gruber [Gru93], et définissons les ontologies dans notre contexte comme des descriptions structurées et formelles des concepts existant pour un ou plusieurs agents et de leurs interelations. La définition donnée par Thomas Gruber est la suivante :

> «[...]I use the term ontology to mean a specification of a conceptualization. That is, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents.»[3]

Pratiquement, tout système informatique traite des données, en utilise en entrée et en produit en sortie. Une ontologie est alors une façon de représenter et de structurer les données utilisées pour faciliter la communication interne au système informatique. Les ontologies peuvent se présenter sous différentes formes, relativement aux fonctions qui leur sont assignées dans le

[2] *Une ontologie est une spécification formelle et explicite d'une conceptualisation partagée.*

[3] *J'utilise le terme ontologie pour désigner une spécification d'une conceptualisation. Ainsi, une ontologie, de la même façon qu'une spécification formelle d'un programme, est une description des concepts et relations pouvant exister pour un agent ou une communauté d'agents.*

système informatique, elles-mêmes éminemment liées aux propres fonctions du système.

Traiter des usages des ontologies, des tâches auxquelles elles sont dédiées, c'est établir les contextes d'utilisation des ontologies et c'est aussi compléter la définition du terme. Pour finir de citer Thomas Gruber :

> «What is important is what an ontology is *for*.»[4]

On trouve ainsi des ontologies dans des systèmes informatiques réalisant du traitement de langage naturel, dans des systèmes informatiques constituant des bases de connaissances, dans des systèmes de recherche d'informations ou même de commerce électronique, ou de *knowledge management*. Les ontologies sont ainsi devenues très populaires et ceci est en partie dû à ce qu'elles promettent [Fen00b] :

> «[...] a shared and common understanding of some domain that can be communicated between people and application systems.»[5]

1.1.1 Ontologies et traitement du langage naturel

Une des premières ontologies développée dans un cadre informatique le fut dans le cadre du projet CYC du milieu des années 1980 au milieu des années 1990. L'objectif était d'élaborer un système-expert capable de comprendre et de parler un langage naturel, l'anglais. Le projet a consisté à constituer une ontologie en modélisant les *connaissances de sens commun*.

Des ontologies peuvent également être créées, comme *GENERALIZED UPPER MODEL*[6], comme base d'un système multilingue de génération de textes. Des systèmes permettant l'extraction des connaissances d'articles de journaux reposent également sur des ontologies comme *PLINIUS* ([VdVSM94]).

Wordnet®, développé à l'université de Princeton par une équipe dirigée par George Miller [Mil93], est ainsi un dictionnaire électronique[7] rassemblant les noms, verbes, adjectifs et adverbes de la langue anglaise. Ces termes sont regroupés en ensembles de synonymes appelés *synset*. Chaque ensemble représente un concept et un concept peut avoir plusieurs *synset*, comme autant d'acceptions. Des relations relient les ensembles de synonymes entre eux. WordNet® est un système de références lexicales croisées dont la conception

[4] *Ce qui compte, c'est ce à quoi sert une ontologie.*

[5] *[...] une compréhension commune et partagée d'un domaine donné qui peut être communiquée entre les personnes et les systèmes d'applications.*

[6] http ://www.darmstadt.gmd.de/publish/komet/genum/newUM.html

[7] *machine-readable dictionary*

a été inspirée par les théories actuelles de la mémoire linguistique humaine. Il est élaboré principalement à la main même si quelques travaux s'attellent à découvrir automatiquement des liens entre les termes. Contenant à la fois des données lexicales et ontologiques [Gua99], Wordnet est la base de nombreux travaux sur les moteurs de recherches lexicaux [MM00], les systèmes d'expansion de requêtes, les systèmes de *question answering* [PCCC01] etc.

Une autre ontologie fortement liée au traitement automatique des langues est l'ontologie MENELAS [ZBB+96]. MENELAS est une ontologie de domaine, celui des maladies coronaires, rassemblant des concepts et leurs relations structurés en arbre par la relation SORTE_DE. Le formalisme de représentation des connaissance utilisé est celui des graphes conceptuels. MENELAS comprend également des lexiques sémantiques et morpho-syntaxiques de mots simples et composés. Les informations sémantiques résident dans la définition des mots sous forme de description conceptuelle. L'ontologie est dédiée à l'analyse automatique de comptes-rendus d'hospitalisation liés à des maladies coronaires rédigés en français, en anglais et en néerlandais dans un but de questions-réponses.

1.1.2 Ontologies et systèmes à base de connaissances

Historiquement, les ontologies apparaissent dans les années 1990 dans le contexte global des travaux en intelligence artificielle sur les systèmes à base de connaissances (*knowledge-based systems*), nouvelle appellation des désormais révolus *systèmes-experts*. L'objectif est d'organiser au mieux les informations et les connaissances utilisées par les systèmes pour pallier les problèmes de communication et permettre le partage et la réutilisation des connaissances.

Les notions d'*information* et de *connaissance* se distinguent. L'information, constituée de données, devient connaissance à partir du moment où elle sert de fondement à une inférence, qu'elle est une des raisons du déclenchement de tel ou tel processus du système. Les connaissances sont ainsi définies par Daniel Kayser [Kay97] comme suit :

> «Les connaissances sont des données qui influencent le déroulement de processus.»

Les systèmes à base de connaissances traitent donc bien des connaissances, celles-ci étant le fondement même de ces systèmes décisionnels.

Certains[8] assimilent bases de connaissances et ontologies. Les ontologies peuvent cependant être distinguées des bases de connaissances dans le sens où les ontologies décrivent des faits admis par une communauté d'usagers par le truchement d'un vocabulaire . Les bases de connaissances décrivent, elles, des faits et assertions relatifs à un état de choses. Les ontologies peuvent tout de même être vues comme des systèmes à base de connaissances particuliers, selon [Gua98], ou, plus simplement, des éléments des systèmes à base de connaissances. Ces liens très forts qui rapprochent les ontologies des systèmes à base de connaissances expliquent le fait, que nous voyons plus loin (section 1.2.2), que les méthodes dessinées pour élaborer les uns soient également applicables aux autres, comme Common-KADS par exemple ([AGBS00a]).

Les ontologies, en fonction du rôle qu'elles ont dans le système à base de connaissances et de la complexité des connaissances qu'elles portent, présentent différents degrés de généralisation. On peut ainsi citer [Ass98] rapportant les travaux de Van Heijst de 1997 :
- les *ontologies de domaine* rassemblant les connaissances pour un domaine particulier (la biologie moléculaire, la médecine, la mécanique ou le droit) ;
- les *ontologies applicatives*, spécifiques et non réutilisables, contenant les connaissances de la couche domaine nécessaires à une application donnée ;
- les *ontologies génériques* exprimant des conceptualisations valables dans différents domaines ;
- et enfin les *ontologies de représentations* conceptualisant des primitives de langages de représentation des connaissances.

Nicola Guarino [Gua98] distingue, lui, les *top-level ontologies* à rapprocher des *ontologies génériques* de Van Heijst, les *domain and task ontologies* décrivant respectivement le vocabulaire d'un domaine et les tâches et activités, et enfin les *application ontologies* décrivant les concepts relatifs à la fois au domaine et aux tâches.

Deux types d'ontologies peuvent dès lors être distingués.

(i) Le premier regroupe les ontologies servant au stockage et à la mise en forme de connaissances d'un domaine particulier nécessaires à un système. Ce premier type, que nous nommons *ontologies de domaine* rassemble les *ontologies de domaine* et les *ontologies applicatives* de Van Heijst. Ces ontologies ont alors un rôle bien défini : porter les connaissances utiles et nécessaires et être intégrées à un système informatique donné. Ces ontologies peuvent être

[8] J. Euzenat, http ://www.inrialpes.fr/sherpa/papers/euzenat98c.html

rapprochées des terminologies [FMJ01], dès lors que toutes deux présentent un panorama d'un domaine de connaissance donné. Notre propre ontologie se place dans ce type des *ontologies de domaine*. Cette ontologie est spécifique au droit français et elle est intégrée à un système interactif d'aide à la reformulation de requêtes.

(*ii*) Le deuxième type d'ontologies, rassemblant les *ontologies génériques* et les *ontologies de représentation* de Van Heijst et les *top-level ontologies* de Guarino n'a pas d'utilité en tant que tel comme élément d'un système informatique. Elles permettent de stocker et de formaliser des connaissances sur les connaissances. Ces ontologies peuvent servir à l'élaboration d'*ontologies de domaine*. Ainsi une *ontologie générique* va expliciter les concepts d'un domaine particulier, décrire les éléments qui le composent. Une telle ontologie va par exemple, dans le domaine des courses de vélo[9], spécifier les éléments qui composent une course : les coureurs, les équipes, les sponsors, les étapes, les chutes etc, chacun avec ses caractéristiques : noms, localisation etc. Mais ce type d'ontologie ne contiendra pas les connaissances spécifiques de telle ou telle course entre Paris et Roubaix, par exemple, ayant rassemblé des coureurs tels que Virenque ou Longo. Les *ontologies de représentations* permettent elles de représenter les éléments composant le langage de représentation des connaissances. Y sont spécifiés les objets ainsi que les règles sur ces objets et les relations qui peuvent les lier. Ainsi les objets du monde peuvent être divisés en êtres animés et êtres inanimés, un être inanimé ne pouvant être en même temps un être animé, certains êtres inanimés pouvant être mangés par certains êtres animés etc. On se servira ainsi, si nécessaire, de ce genre d'ontologies *méta* pour couvrir tous les éléments d'un domaine en question et pour spécifier les relations pouvant exister entre ces éléments. Il existe au moins une de ces ontologies concernant le domaine juridique, celle développée par Valente en 1994 [Val94]. Nous y revenons plus longuement dans la section 3.2.1.

Autre rôle des ontologies dans les systèmes à base de connaissances, celui de permettre la définition d'environnements dédiés à la gestion des connaissances[10] dans le cadre d'intranet ou de sites Internet ([FHvH+01]). On-To-Knowledge[11] est ainsi un projet tendant à la définition de standards, méthodes et outils permettant une gestion des connaissances sur la base d'ontologies. Ces démarches de gestion des connaissances rencontrent plusieurs contextes comme autant de contextes d'utilisation d'ontologies :

[9]http ://opales.ina.fr/public/ontologies/Cyclisme-v1.0-OIL.html
[10]*knowledge management*
[11]http ://www.ontoknowledge.org

- la *gestion de la mémoire d'organisations.* Cette application concerne l'intranet de la société Swiss Life. OIL[12] a permi la définition d'une ontologie formant la base d'un système d'information complet reflétant l'ensemble de la mémoire de l'entreprise. Les recherches y sont ainsi facilitées dès lors que l'accès à l'information se fait selon le contenu et non plus par le biais d'un simple moteur de recherche full-text sur un ensemble de pages Web hétérogène.
- l'*intégration de sources d'informations hétérogènes dans le cadre de centres d'appel.* Dans la perspective de rationaliser la relation entre le consommateur et le service consommateurs d'une société donnée, et de diriger les premiers vers les personnes compétentes du service adéquat, une démarche de gestion des connaissances est également nécessaire. Cette démarche doit intégrer les connaissances sur le consommateur et celles sur le service. Certains centres d'appel de British Telecom[13] ont ainsi défini des ontologies formalisées en OIL permettant un accès intuitif aux données hétérogènes auxquelles les agents du centre d'appel sont susceptibles de faire appel dans leur tâche d'information du client. Ces données peuvent être celles propres à la société, les données de fichiers clientèle, l'organisation même de l'intranet ou les cas déjà traités (questions-réponses).
- *la recherche d'informations à base de connaissances.* La société EnerSearch[14] a pour principale activité la recherche d'informations R&D (recherche et développement) dans des domaines aussi divers que le marché des énergies ou les stratégies en information et télécommunication. La société a adopté une ontologie formalisée en OIL pour permettre le transfert de connaissances entre différentes disciplines, pays et spécialistes. Pratiquement, les applications basées sur l'ontologie sont utilisées pour la définition de termes de recherche appropriés (*keyword search*) et pour répondre aux questions (*query answering*).

Les ontologies sont donc un élément fondamental des systèmes à base de connaissances puisqu'elles permettent de stocker de façon centralisée et rationalisée toutes les connaissances nécessaires au système.

Ces systèmes à base de connaissances incluant des ontologies peuvent être conçus, nous venons de le voir, pour remplir différentes tâches : la gestion de la mémoire d'entreprise, l'intégration de ressources hétérogènes et aussi la recherche d'information. Ce dernier point mérite un développement particu-

[12]Le langage de spécification d'ontologies OIL est décrit dans la Section 2.3
[13]http ://www.bt.com/innovation
[14]http ://www.enersearch.se

lier.

1.1.3 Ontologies et recherche d'informations

Les ontologies ont vu leur place confirmée dans des systèmes de recherche d'informations dès lors qu'il est apparu intéressant d'aborder la recherche d'informations par le versant des connaissances. Les ontologies permettent, dans ce contexte, une représentation du contenu des documents, ce qui permet d'aller au-delà de la recherche d'informations classique. Une ontologie peut à ce titre faciliter une recherche documentaire fondée sur le contenu, comme le système OntoSeek ([Gua99], [Gua98]) ou dans le cadre du projet ACACIA de l'INRIA[15] (voir notamment [GDK01]).

L'ontologie que nous développons est dédiée à la recherche d'information, nous nous y attardons donc.

Telle que nous l'avons définie à la section précédente, l'ontologie que nous envisageons est bien une ontologie de domaine, celui du droit français. Elle est plus précisément une *ontologie applicative* telle que définie par Van Heijst dans le sens où les connaissances qu'elle rassemble ont un objet, celui de la recherche documentaire.

Nous entendons consacrer cette ontologie à être un intermédiaire entre l'utilisateur d'un moteur de recherche et ce moteur de recherche. De nombreuses études ont en effet montré, dans la mouvance des travaux de Salton et Mc Gill en 1983 [SM83] , que le *langage d'indexation* d'un moteur de recherche, ce qu'il indexe des documents qui lui sont soumis, est relativement différent de celui de l'utilisateur formulant sa requête, *language de requête*. Notre ontologie a dès lors la tâche de suggérer à l'utilisateur des termes du domaine, globalement éléments du *langage d'indexation*. Ces termes suggérés pouvant éventuellement être ajoutés par l'utilisateur à ceux de sa requête. Notre ontologie est ainsi la base à un système interactif d'aide à la reformulation de requêtes. La Figure 1.1 illustre où se situe notre système entre les requêtes des utilisateurs et l'index d'un moteur de recherche.

Notre ontologie est essentiellement composée des éléments suivants : les termes du domaine et les relations entre ces termes. Ce réseau composé de termes et des diverses relations qui les lient est porteur de connaissances sur le domaine.

[15] http ://www.inria.fr/recherche/equipes/acacia.fr.html

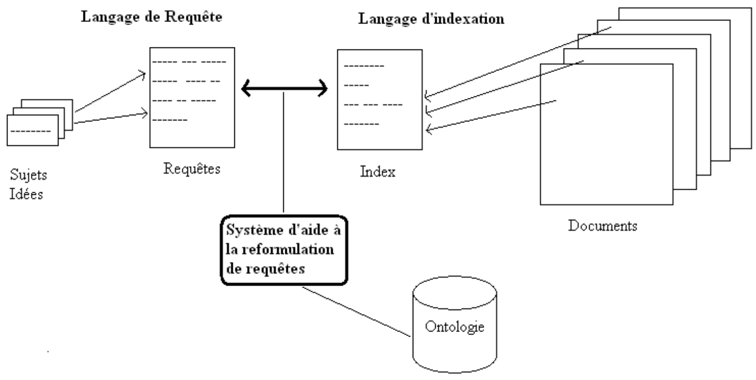

FIG. 1.1 – Le système d'aide à la reformulation de requêtes

Notre ontologie répond ainsi à la tâche généralement affectée aux ontologies, celle de la communication entre plusieurs parties, entre le système et l'utilisateur dans notre cas. L'ontologie que nous obtenons est d'une complexité peu élevée dans le sens où elle relève plus d'un vocabulaire contrôlé et structuré, une *terminologie*, que d'un système axiomatique complexe.

Les ontologies peuvent ainsi présenter divers aspects en fonction de leur *profondeur ontologique*[16]. Les ontologies pouvant dès lors relever alternativement du lexique, de la simple taxonomie, du thesaurus, du modèle relationnel jusqu'à la théorie complètement axiomatisée. Pour continuer de nous référer à Nicola Guarino, nous citons cette observation sur ce que recouvrent les ontologies [Gua98] :

> «[...] in its most prevalent use in AI, an ontology refers to an *engineering artifact*, constituted by a specific *vocabulary* used to describe a certain reality, plus a set of explicit assumptions regarding the *intended meaning* of the vocabulary words. This set of assumptions has usually the form of a first-order logical theory, where vocabulary words appear as unary or binary predicate names, respectively called concepts and relations. In the simplest case, an ontology describes a hierarchy of concepts related by subsumption relationships ; in more sophisticated cases, suitable axioms are added in order to express other relationships between concepts and and to constrain their intended interpreta-

[16] *Ontological depth* selon les termes de Nicola Guarino présentant les ontologies à l'occasion d'un séminaire du LISI à Lyon, le 5 octobre 2000

tion.»[17]

1.1.4 Ontologies et *Web semantique*

Vision de Tim Berners-Lee, le *semantic web* est un réseau de documents non plus structurés selon leur mise en forme (par des balises HTML[18]), mais selon le sens qu'ils portent, les données qu'ils contiennent (balises XML[19]). Les documents qui, en HTML, ne sont appréhendables que par les hommes qui les lisent sont, avec XML, échangeables et appréhendables par des machines. XML permet d'exprimer des informations qualifiées. Une information qualifiée sera alors de savoir, par exemple, que le chiffre 7 exprimé dans le document correspond au prix en euros d'une marchandise considérée. Différents agents peuvent dès lors communiquer entre eux en échangeant ces informations qualifiées.

Les recherches dans le domaine du Web sémantique s'intéressent de près aux ontologies à ce titre, celui de la définition préalable d'une structuration sémantique des documents ou d'un medium de communication commune entre différents agents. Ainsi, divers acteurs de la vente de billets d'avions sur Internet pourront utiliser une ontologie commune décrivant les différents éléments nécessaires à leurs propres sites (les destinations, les prix, les dates, les types d'avions etc). Ces différents éléments permettront à chaque acteur de créer ses propres documents avec des balises XML communes. Ces balises XML communes permettront d'automatiser des procédures de comparaisons de prix, de prestations etc. Nous notons à cet égard que ce type d'ontologies se créent le plus souvent *a posteriori* et que l'enjeux de ces ontologies est de rassembler sous de mêmes notions les différentes façons d'exprimer les mêmes données comme les prix, les destinations etc.

Dans cette "communauté" du Web sémantique, les ontologies sont alors vues comme ([Hen01]) :

---

[17] *Dans la plupart de ses usages en intelligence artificielle, une ontologie se rapporte à un artefact technique constitué d'un vocabulaire spécifique décrivant une certaine réalité, et d'un ensemble de postulats sur le sens intentionnel des mots du vocabulaire. Cet ensemble de postulats prend le plus souvent la forme d'une théorie logique de premier ordre où les mots du vocabulaire sont des noms de propriétés unaires ou binaires, appelés respectivement concepts et relations. Dans le cas le plus simple, une ontologie décrit une hiérarchie de concepts reliés par des relations de subsumption; dans des cas plus sophistiqués, des axiomes appropriés sont ajoutés, de façon à exprimer d'autres relations entre concepts et à contraindre l'interprétation de leurs intentions.*

[18] *HyperText Markup Language*

[19] *eXtensible Markup Language*

> «[...] a set of knowledge terms, including the vocabulary, the semantic interconnections, and some simple rules of inference and logic for some particular topics.[20]»

Le *Web sémantique* traite des informations sémantiques et non plus de simples données comme "7" ou "chat" [Kim02]. Ces informations relèvent des règles et du sens des données ("les chats sont des mammifères") et doivent être définies de façon à ce que les machines, les systèmes informatiques, puissent les interpréter et les traiter rapidement. Les ontologies dans le cadre du *Web sémantique* ont alors pour objet de codifier ces informations sémantiques. Henry Kim, dans un article intitulé *Predicting how Ontologies for the Semantic Web will evolve* [Kim02], prédit que le futur du *Web sémantique* dépend du futur des ontologies dans le sens où celles-ci sont, avec XML, le fondement de l'échange rationnel d'informations sémantiques entre systèmes informatiques.

1.2 Elaboration d'ontologies

A partir du moment où la discipline informatique au sens large, et l'ingénierie des connaissances plus précisément, s'est appropriée la notion d'ontologie, des outils ont été développés et des méthodes d'élaboration d'ontologies ont été définies.

Au-delà du rôle de l'informatique dans la construction d'ontologies et des techniques d'automatisation de certaines étapes de la construction d'ontologies, différentes méthodes ont été définies, conduisant l'élaboration d'ontologies. Ces méthodes, quelles qu'elles soient, peuvent adopter différentes approches dans la façon de conduire le processus d'élaboration de l'ontologie (Section 1.2.1). Ces méthodes sont détaillées dans la Section 1.2.2 et les outils qui ont été créés dans la Section 1.2.3.

1.2.1 Différentes approches

Approches top-down et bottom-up

L'approche *top-down*, ou descendante, consiste, comme son nom l'indique, à partir d'un niveau *méta* pour en faire découler une ontologie. On se basera ainsi sur une *ontologie générique* et/ou une *ontologie de représentation* pour la spécifier en une ontologie de domaine. On pourra aussi réunir une

[20] [...]*un ensemble de connaissances, y compris le vocabulaire, les relations sémantiques et quelques règles simples d'inférence et de logique relatifs à des sujets particuliers.*

assemblée d'experts d'un domaine donné et s'attacher à ce qu'ils s'accordent sur une conceptualisation de ce domaine. Mais, comme nous le notons dans [Lam01c], la tâche n'est pas aisée ! Pour ne parler que du domaine qui est le nôtre, de nombreux projets ont eu pour objet l'élaboration de ressources terminologiques relatives au droit. La commission de Venise[21] du Conseil de l'Europe, rassemblant des experts du domaine, a ainsi été chargée de la création d'un thesaurus du droit constitutionnel et de procédure (CODICES). Ce thesaurus, qui fonctionne, a néanmoins nécessité des années de travail avant de voir le jour.

Cette approche descendante de l'élaboration d'ontologies est inévitable lorsqu'on s'attache à la construction d'ontologies *méta* telles que nous les décrivons plus haut. La détection des éléments constitutifs d'un langage de représentation ou des supra concepts d'un domaine (les êtres animés ou inanimés par exemple) nécessite une telle approche.

Une approche *bottom-up* de la conception d'ontologies, ou approche ascendante, consiste à se fonder sur un ensemble de données (des textes) pour en extraire les éléments constitutifs de l'ontologie. Cette approche est portée en France par le groupe de recherche TIA[22] influencé par les travaux de B. Bachimont dans lesquels les textes sont considérés comme la source quasi exclusive de connaissances. L'objet de l'ingénierie des connaissances est alors la construction de modèles à partir de l'expression linguistique de connaissances [AGBS00a]. Se fondant sur des corpus, cette approche suggère que les éléments des ontologies, les concepts, leurs propriétés, leurs relations, leur regroupement, sont portés par les textes. Les principales étapes de cette approche linguistique consistent à rassembler un ensemble de documents qui définiront un corpus de référence. Diverses méthodes et techniques d'ingénierie des connaissances seront appliquées sur ce corpus pour en extraire les éléments constitutifs de l'ontologie. Cette approche permet l'utilisation d'outils logiciels de traitement automatique de la langue (ou TAL).

Néanmoins, il ne semble pas qu'une approche purement ascendante de la construction d'ontologie soit viable. Ainsi, une phase descendante de normalisation et de formalisation est envisagée dans [AGBS00a] pour organiser et formaliser les connaissances issues de la phase ascendante.

[21] Commission pour la Démocratie par le Droit

[22] Le groupe de travail TIA, Terminologie et Intelligence Artificielle, travaille, au sein du PRC-I3 et de l'AFIA, à la convergence entre la terminologie, la linguistique, l'ingénierie des connaissances et l'intelligence artificielle.

D'autres approches de constructions d'ontologies et de systèmes à base de connaissances constituent une synthèse des deux approches précédentes. Les méthodes KOD et Common-KADS, par exemple, s'attachent à récolter l'avis d'experts pour y détecter les éléments d'une ontologie. Ces avis, sous forme de textes, sont alors plus ou moins exploités pour en extraire les concepts du domaine par exemple. D'autres approches ([ZBHN97]) combinent des analyses de corpus à l'utilisation de connaissances préalables sur le domaine.

Des approches fonctionnelles

Une distinction plus *fonctionnelle* que celle précédemment exposée des méthodes d'élaboration d'ontologies est faite dans l'article *A Collaborative Approach to Ontology Design* de C. W. Hosapple et K.D. Joshi en février 2002 [HJ02]. Les auteurs se fondent sur les points de vues adoptés par les concepteurs des ontologies (*designers*), préalablement à leur création. Cinq façons d'aborder la création d'ontologies (*ontology design*), y sont définies : des approches *inspiratives*, *inductives*, *déductives*, *synthétiques* et *collaboratives* en notant que des approches hybrides peuvent également exister.

(*i*) L'approche *inspirative* de la construction d'ontologies débute par la détermination des raisons même de la construction de cette ontologie, des raisons pour lesquelles celle-ci est nécessaire, des besoins qu'elle doit combler. Utilisant ensuite son imagination, sa créativité, ses propres points de vues, la personne procède à l'élaboration d'une ontologie tendant à répondre aux besoins prédéfinis. Dans ce type d'approche, plus les engagements ontologiques[23] seront porteurs de consensus de la part des utilisateurs du système et des éventuelles personnes chargées de la maintenance de l'ontologie, plus cette ontologie aura des chances de perdurer, d'être effectivement utile, utilisée et effective.

(*ii*) L'approche *inductive* consiste à examiner et analyser un cas spécifique d'un domaine donné. L'ontologie résultant de cette caractérisation d'un cas spécifique étant alors appliquée à d'autres cas, plus ou moins similaires, du même domaine. Les auteurs donnent l'exemple des pratiques et usages

[23] Les engagements ontologiques, *ontological commitments*, portent l'accord, l'adhésion des diverses parties concernées à l'adoption d'une ontologie particulière pour ce qui est de leurs communications relatives à un domaine particulier [HJ02]. Ces parties, personnes physiques ou systèmes informatiques, n'ont par ailleurs pas nécessairement les mêmes expériences, théories ou compréhensions de ce domaine. Elles peuvent tout à fait s'accorder sur la définition générale d'un objet concept du domaine par ses éléments constitutifs par exemple, mais avoir l'expérience d'une mise en œuvre particulière. Les engagements ontologiques sont les points sur lesquels les différentes parties intéressées par l'ontologie sont d'accord.

du *knowledge management* dans une société donnée sur lesquels se fonde la définition d'une ontologie des pratiques et usages du *knowledge management* en général. La détermination des engagements ontologiques est alors une phase cruciale dont dépendra l'*adaptabilité* des observations issues d'un cas particulier à une situation générale.

(*iii*) L'approche *déductive*, corollaire inverse de l'approche *inductive*, consiste à adopter des principes généraux préalablement définis, à les adapter pour en déduire une ontologie spécifique à un cas particulier. Cette approche suppose de filtrer et adapter des notions générales de manière à ce qu'elles soient applicables à un cas particulier. Elle relève d'une approche *top-down* telle que définie plus haut, consistant par exemple à déduire d'une ontologie générale une ontologie particulière.

(*iv*) Elaborer une ontologie en adoptant une approche *synthétique* nécessite de préalablement identifier un ensemble d'ontologies ne se déduisant pas les unes des autres. Ces ontologies sont alors synthétisées en y adjoignant, le cas échéant, des concepts relatifs au phénomène en question. Une ontologie unifiée du phénomène en résulte. Les adhérents aux ontologies préexistantes sont alors plus à même d'adopter l'ontologie finale, si tant est que leurs engagements ontologiques respectifs ne soient pas trop bousculés. Cette approche peut être adoptée, par exemple, dans le cas ou plusieurs ontologies de domaine existent dans un domaine plus général, et que celles-ci s'avèrent être plus ou moins redondantes sur certains points.

(*v*) L'approche *collaborative* est l'approche la plus longuement développée par les auteurs puisque c'est celle qu'ils adoptent pour la construction d'une ontologie du *knowledge management*. Nous renvoyons à l'article pour de plus amples développements sur cette expérience. Cette approche reflète un effort commun rassemblant les expériences et points de vues de personnes ayant sciemment décidé de collaborer en vue de la construction d'une ontologie. Les auteurs notent que plus ces personnes proviennent d'horizons différents, mais sont cependant liés au phénomène en question, plus les chances d'acceptation de l'ontologie sont fortes. Ils préconisent en quelque sorte une collaboration pluridisciplinaire pour l'élaboration d'ontologies. Les auteurs notent cependant qu'un processus systématisant la définition de positions consensuelles, comme la méthode DELPHY qu'ils proposent, est nécessaire dans ce type de démarche rassemblant divers points de vues, éventuellement divergents. Une méthodologie fondée sur une approche pluridisciplinaire est également avancée par Nicola Guarino, qui note que les ontologies nécessitent une collaboration active de la philosophie et de la linguistique dans l'analyse à un haut niveau de généralité de la structure d'un phénomène donné [Gua98].

Les différentes approches que peuvent suivre les méthodes d'élaboration d'ontologies telles que nous venons de décrire semblent donc très diverses. Il s'avère cependant que c'est principalement le contexte d'élaboration de l'ontologie et l'objet même de cette ontologie qui conditionnent l'adoption de telle ou telle approche. Ainsi, si des experts du domaine ne sont pas disposés à travailler à la construction d'une ontologie, une approche strictement descendante sera inenvisageable. De la même façon, s'il n'existe pas préalablement d'ontologies dans le domaine en question ou dans un domaine suffisamment proche, une approche synthétique est également impossible.

Pour ce qui est de notre cas, le fait d'adopter une démarche essentiellement ascendante se basant sur des documents est principalement conditionnée par le fait que notre ontologie est dédiée à la recherche d'informations. Cette recherche d'informations concerne bien évidemment les documents. Sans forcément lier les documents servant de base à l'élaboration de l'ontologie et ceux dans lesquels les recherches d'informations sur la base de l'ontologie s'effectuent, nous avançons que les connaissances utiles à la recherche d'informations doivent être recherchées dans des documents plus que dans la tête d'experts du domaine. Nous utilisons pour ce faire un maximum d'outils et de techniques automatisées d'acquisition de connaissances à partir de textes.

1.2.2 Des méthodes d'élaboration d'ontologies

Dès lors que la recherche scientifique s'est intéressée aux ontologies, des méthodes d'élaboration d'ontologies ont été définies. Dans un premier temps [Kas02b], ces méthodes relevaient plutôt de cadres méthodologiques issus de projets comme *Enterprise Ontology* ou *TOVE*. Depuis, les cadres méthodologiques se sont précisés et, de recommandations en ébauches de méthodes, on est arrivé à la définition de méthodes bien plus élaborées relevant de l'ingénierie des connaissances comme *Methontology*, *Terminae*, ou *OntoSpec*. Nous notons simplement avec [JBCV98] qu'aucune méthode générale d'élaboration d'ontologies n'existe actuellement.

Quoi qu'il en soit, ces différentes méthodes sont soit globales et englobent la totalité de l'élaboration de l'ontologie (TOVE, Enterprise Model, Terminae ou Methontology) ou plus spécifiques, ne concernant qu'une phase précise de cette élaboration (KOD, CommonKADS ou Ontospec). En filigrane de chacune de ces méthodes, une approche, telle que définie dans la section précédente (Section 1.2.1), se dessine. Ainsi, une méthode comme Terminae se fonde essentiellement sur la linguistique et adopte dès lors une approche principalement ascendante alors qu'une méthode telle que celle définie par

Nicola Guarino, méthode fondée sur l'*Ontologie*, relève d'une approche exclusivement descendante.

Que ces méthodes aient été spécifiquement développées pour les ontologies ou non, elles s'inspirent à des degrés divers des méthodes d'élaboration de systèmes à base de connaissances (SBC), ce qui s'explique aisément par la filiation reconnue entre les ontologies et les SBC (voir Section 1.1.2).

La méthode Enterprise Ontology

Enterprise ontology est l'ontologie résultant du projet *Enterprise*[24] initié par le ministère britannique du travail et de l'industrie. Ce projet avait pour objectif de construire une ontologie unique permettant de modéliser les entreprises, de représenter toutes leurs activités et organisations[25]. Cette ontologie comporte ainsi des éléments *organisation*, *activités*, *stratégie*, *marketing*, *temps*, eux-mêmes incluant des éléments tels que *autorité*, *planification des tâches*, *capacité*, *partenaire financier* etc. Cette ontologie est intégrée dans un système à base de connaissance destiné à aider ces entreprises à mener des actions de conduite du changement[26].

De ce projet menant à la construction d'une ontologie, une méthode a émergé. Cette méthode est influencée par le contexte du système à base de connaissances, notamment dans le fait qu'elle distingue deux phases dans la construction de l'ontologie, une phase informelle et une phase formelle. La première phase permet d'identifier les objectifs et les raisons qui mènent à la construction d'une ontologie. Le niveau de formalisme dans lequel l'ontologie sera décrite est alors spécifié en fonction des usages et des utilisateurs de l'ontologie. Une étape de *spécification* permet, elle, de cerner le niveau d'informations que l'ontologie devra caractériser, duquel on déduit les concepts et leurs relations à inclure dans l'ontologie. Pour cette étape, les auteurs de l'*Enterprise ontology* [Usc96] nous guident dans une certaine mesure, indiquant qu'il faut élaborer des scenarii incluant un problème, et une solution telle qu'elle devrait être trouvée sur la base de l'ontologie, ces scenarii pouvant être élaborés lors de séances de *brainstorming*. De cette première phase informelle, les termes et concepts à inclure dans l'ontologie sont identifiés.

La phase formelle d'élaboration de l'ontologie va alors consister à construire un modèle formel avec tous ces éléments. Cette phase consiste à établir les

[24]http ://www.aiai.ed.ac.uk/ entprise/enterprise/
[25]*How a business works and how it is organized*
[26]*management of change*

définitions formelles et les axiomes des termes issus de la phase précédente. Plusieurs approches sont distinguées, selon le degré de formalisme requis par l'ontologie. Nous renvoyons à la référence bibliographique pour le développement de ces points [Usc96]. Enfin la phase formelle se conclu par une phase de validation formelle du modèle. Cette validation peut consister à établir la clarté, la validité et la capacité de réutilisation, l'absence de redondances etc.

La méthode d'élaboration d'ontologie déduite du projet *Enterprise* telle que nous venons de la décrire ne reste somme toute qu'un ensemble d'indications méthodologiques. Les étapes de cette élaboration restent décrites de façon totalement abstraite ([Kas02b]). Dans la phase informelle consistant à l'identification des concepts de la future ontologie, la possibilité d'utiliser des techniques d'acquisition de connaissances est évoquée mais demeure au stade de l'évocation. Aucun conseil n'est ainsi donné sur la façon d'identifier ce que seront les éléments même de la future ontologie [JBCV98].

La méthode TOVE

Dans la continuité d'indications méthodologiques résultant de projets d'élaboration d'ontologies, la méthode *TOVE* comporte de nombreuses similarités avec la méthode *Enterprise ontology* décrite ci-dessus, notamment le fait qu'elle concerne les entreprises. Le projet TOVE (Torento Virtual Enterprise)[27] consiste à élaborer une terminologie partagée sur les entreprises qui permette à différents agents de communiquer entre eux autour d'une sémantique précise et non ambiguë. Les auteurs notent à juste titre que, bien souvent, chaque système informatique d'une même entreprise a été créé de façon séparée. Ces systèmes ne partagent donc pas la même représentation de l'entreprise, ne portent pas les même connaissances sur l'entreprise. L'objectif du projet est alors de créer des ontologies portant une terminologie et une sémantique communes pour permettre, notamment, la déduction de nouvelles connaissances par des règles d'inférences.

Plusieurs étapes se distinguent dans la méthode avancée pour la construction d'ontologies dans le cadre du projet TOVE [GF95]. La première consiste à l'identification de scenarii. Ces scenarii peuvent, par exemple, décrire des problèmes spécifiques rencontrés par les partenaire industriels dans leurs entreprises respectives. Ces scenarii rendent compte de faits qui ne sont pas résolus par les applications existantes. Une deuxième étape consiste, en se basant sur les scenarii précédents, à formuler des questions informelles aux-

[27]http ://www.eil.utoronto.ca/tove/ontoTOC.html

quelles l'ontologie doit permettre de répondre. De cette phase résulte un ensemble d'engagements ontologiques. Dans une phase de *spécification terminologique*, les éléments de l'ontologie sont identifiés : objets, attributs et relations sur la base de ces questions informelles. Puis la démarche devient formelle par la spécification en langage KIF[28] des axiomes et définitions des termes de la terminologie.

Enfin, point à souligner, une phase d'évaluation de l'ontologie est envisagée. Cette évaluation consiste principalement à déterminer la complétude de l'ontologie au regard des scenarii préalablement définis. La question est alors de savoir si l'ontologie telle qu'elle est est en mesure de répondre aux problèmes soulevés dans ces scenarii.

La description de la démarche TOVE demeure sommaire et abstraite ; c'est pourquoi nous avons du mal à mieux expliciter et illustrer cette méthode. Aucune des étapes précises et techniques utilisées n'est précisément spécifiée.

Methontology

Methontology est une méthodologie structurée pour construire des ontologies [FGPJ97]. Cette méthode résulte également de l'expérience acquise lors du développement d'une ontologie particulière. Les auteurs, ayant comparé les travaux des projets *Enterprise ontology* et *TOVE* aux leurs, arrivent à la conclusion que ces méthodes ne peuvent pas être agrégées et synthétisées en une seule et même méthode. Ils proposent alors la leur.

La méthode Methontology envisage la construction d'ontologie comme un cycle[29]. L'ontologie se construit à chaque étape, le prototype s'améliorant d'étape en étape. Methontology est ainsi, de façon similaire à la méthode TOVE, divisée en plusieurs étapes distinctes mais qui peuvent être menées parallèlement, notamment la première et la deuxième.

(*i*) La première étape, étape de *spécification*, permet de produire un document de spécification de la future ontologie. Ce document comporte des indications sur l'objet de l'ontologie, les utilisateurs et utilisations envisagés, le degré de formalisme à employer, les termes à représenter, leurs caractéristiques etc.

(*ii*) Une deuxième étape d'*acquisition de connaissances* conduit l'identification des termes de l'ontologie et leur définition. Des techniques d'acquisition de connaissances sont listées : les réunions *brainstorming*, les interviews

[28] *Knowledge Interchange Format*
[29] *a life-cycle*

d'experts, les analyses formelles et informelles de textes, des outils d'acquisition de connaissances sans plus de précision. Toute technique d'acquisition de connaissances semble donc viable et, sur ce point, on peut noter avec [JBCV98] que la méthode Méthontology est peu prescriptive.

(*iii*) Une étape de conceptualisation va ensuite consister à structurer les connaissances du domaine en un modèle conceptuel décrivant le problème et ses solutions telles qu'envisagées dans l'ontologie. A ce stade, la représentation demeure informelle.

(*iv*) L'étape suivante, étape d'*intégration*, permet d'envisager dans quelles mesures d'autres ontologies préexistantes peuvent être utilisées dans la construction de l'ontologie donnée. L'objectif des auteurs est de *situer* leurs travaux par rapport à des travaux similaires de construction d'ontologies et également de tenter d'unifier la recherche dans ce domaine [FGPJ97]. Ainsi des *ontologies méta* (ou top ontologies) peuvent être utilisées, donnant la structure de base des concepts d'une nouvelle ontologie. Les auteurs notent que ces ontologies méta préexistantes peuvent ne pas convenir et il faut dans ce cas développer une nouvelle méta-ontologie dans un langage formel. D'autres ontologies peuvent également être utilisées pour la définition de termes communs. De cette phase d'intégration résulte un document spécifiant les ontologies réutilisées, les termes qui y sont communs, les façons de traduire ces différents termes etc.

(*v*) La phase suivante est celle d'*implémentation* ; elle consiste à représenter formellement l'ontologie dans un langage formel tel LOOM, Ontolingua mais aussi Prolog ou C++. Un environnement de développement d'ontologies est envisageable tel Ontolingua.

(*vi*) Une phase d'*évaluation* consiste alors à vérifier et à valider l'ontologie en question ainsi que son environnement logiciel et sa documentation. Les problèmes d'incomplétude, d'inconsistance et de répétition sont alors identifiés. Les techniques utilisées s'inspirent largement de celles afférentes aux systèmes à base de connaissances.

(*vii*) Enfin, la dernière étape de la construction de l'ontologie repose sur la *documentation* de celle-ci. Ce point semble très important aux auteurs qui indiquent qu'il n'existe pas de manière unifiée de documenter les ontologies, et ce principalement car il n'existe pas de méthode unifiée pour construire ces ontologies. En résulte le plus souvent une documentation incomplète, insuffisante à la compréhension globale de l'ontologie et/ou à sa réutilisation. Et donc une impossibilité à élaborer des standards gouvernant la construction d'ontologies. Methontology se propose de pallier ces insuffisances, notamment par le fait que, de chaque étape de la construction, résulte un document de spécification de l'étape considérée.

La particularité de la méthode Methontology est de s'attacher fortement à la maintenance de l'ontologie, dont son évaluation. Les auteurs [FGPJ97] insistent également fortement sur le fait que tout concepteur de nouvelle ontologie doit s'efforcer le plus possible de réutiliser des ontologies existantes.

La méthode IDEF5

Nous n'avons trouvé sur cette méthode que des articles généraux ne la traitant qu'en partie ([JBCV98], [gra00]) mais elle y est suffisamment citée pour que nous la prenions en compte[30]. Cette méthode a les même caractéristiques que les précédentes. La construction d'une ontologie y est ainsi vue comme un cycle, un processus itératif en plusieurs étapes. Cette méthode suggère le fait qu'une solution *clé en main*[31] pour l'élaboration d'ontologies n'est pas envisageable du fait notamment que les analyses ontologiques sur des domaines trouvent rarement une fin définitive. La méthodologie IDEF5 est donc une procédure générale soulignant néanmoins quelques étapes majeures.

(*i*) La première (*organising and scoping*) consiste, de la même façon que les autres premières étapes des autres méthodes, à établir l'objet, le but de l'ontologie, ses objectifs et ses exigences (*requirements*), son point de vue sur le domaine et son contexte.

(*ii*) La deuxième étape (*data collection*) consiste à rassembler les données brutes[32] nécessaires au développement de l'ontologie. Ces données sont obtenues en utilisant des techniques classiques d'acquisition de connaissances comme les interviews d'experts.

(*iii*) L'étape suivante (*data analysis*) consiste à exploiter les données issues de l'étape *data collection*. De ces données sont ainsi distingués les objets du domaine de ceux qui n'en sont pas et qui, à ce titre, ne seront pas inclus dans l'ontologie. L'étape des premiers développements de l'ontologie peut alors commencer, durant laquelle une première ontologie est élaborée, constituée de concepts *prototypes*[33]. L'ontologie prototype est ensuite affinée et validée par un processus itératif d'instanciation des concepts sur la base des données. Pour faciliter le travail entre les experts du domaine et les constructeurs de l'ontologie, les premiers prototypes de l'ontologie sont schématisés en utilisant une notation graphique relativement intuitive (*schematic language*),

[30] La référence mentionnée dans les papiers que nous citons est, à notre connaissance, introuvable sur le Web *KBSI, "The IDEF5 Ontology description Capture Method Overview", Technical Report, KBSI, Texas, USA, 1994*

[31] A cookbook approach [JBCV98]

[32] "Raw data"

[33] proto-concepts

le modèle final étant exprimé dans un langage plus structuré (*elaboration language*), extension de KIF.

Les auteurs de IDEF5 notent eux-mêmes qu'une méthode *clé en main* n'est pas envisageable dans le domaine de la construction d'ontologies. La description qu'ils font de leur démarche reste sommaire, se contentant de donner de grandes orientations.

La méthode CommonKADS et le projet KACTUS

CommonKADS[34], continuation de la méthode KADS[35], est une méthode d'acquisition de connaissances définie pour le développement de systèmes à base de connaissances (ou SBC). Cette méthode est ainsi utilisée lors d'élaboration de systèmes réalisant de la gestion des connaissances[36] dans des organisations (entreprises etc). CommonKADS a été développée dans le cadre des programmes européens ESPRIT (I et II) et tend à devenir un standard, du moins en Europe, de l'analyse des connaissances et du développement de systèmes utilisant les connaissances. Cette méthode permet de mettre à jour la façon dont une organisation donnée produit, échange, utilise et applique ses connaissances propres. CommonKADS est une approche dirigée par les modèles, et non pas par l'implémentation, qui découpe le processus d'élaboration d'un SBC en deux étapes : d'une part la formulation et le développement de modèles (la phase d'analyse), d'autre part, la conception et l'implantation du logiciel (la phase de conception). Une des particularités de CommonKADS est de séparer la représentation (ou modélisation) du domaine de celle de la résolution de problème. Ainsi, il existe dans CommonKADS plusieurs modèles pour représenter et analyser la connaissance et notamment le *modèle d'organisation* décrivant l'organisation en question (l'entreprise) et les connaissances qu'elle porte, et le *modèle de tâches* décrivant les tâches réalisées par l'organisation.

Plus précisément, et pour ce qui nous intéresse, la démarche d'acquisition des connaissances telle que définie dans CommonKADS propose de combiner l'utilisation d'ontologies du domaine, quand elles existent, à l'exploitation de textes. Ces textes, des retranscriptions d'entretiens avec des experts ou des documents techniques propres au système en question, ne tiennent néanmoins qu'une place marginale dans le processus d'acquisition des connaissances [AGBS00a]. Ils peuvent néanmoins aider à l'identification

[34] http ://www.commonkads.uva.nl/

[35] *Knowledge Acquisition and Design System*

[36] knowledge management

des concepts du domaine.

Le projet KACTUS[37], *modelling Knowledge About Complex Technical systems for multiple USe*, est la continuation d'un projet ESPRIT dédié au développement de méthodes relatives à la réutilisation par un système technique de ses connaissances propres tout au long de son cycle de vie. Le projet KACTUS inclut la construction d'une ontologie. L'ontologie en question est développée sur la base d'une librairie rassemblant des ontologies préexistantes [SWJ95]. Une série d'outils a été développée (VOID[38]) définissant un environnement d'édition, de consultation et d'organisation de cette librairie d'ontologies. Ces ontologies utilisées pour l'élaboration d'une nouvelle ontologie doivent donc être articulées, mises en perspective les unes des autres, cartographiées[39]. L'identification des ontologies adaptées à la construction d'une nouvelle est rendue possible par la définition d'un schéma d'indexation des ontologies existantes. Ce schéma d'indexation caractérise les ontologies selon les usages qu'on peut en faire [SWJ95] : les types de tâches concernées (diagnostics, prédiction, planification etc), les méthodes de résolution de problèmes, les types de modèles du domaine (modèle structurel, modèle fonctionnel etc) et le type de domaine (réseaux électriques, ascenseurs etc). Le formalisme utilisé dans KACTUS est CML[40], un langage de modélisation et de représentation des connaissances développé dans le cadre de la méthode CommonKADS.

Il est à noter, pour le contexte qui est le nôtre, qu'André Valente a proposé l'utilisation de la méthode CommonKADS, dont il a par ailleurs participé au développement avec Joost Breuker, pour le développement de systèmes à base de connaissances juridiques. Cette proposition l'a conduit à définir une représentation du domaine juridique, son *ontologie fonctionnelle du droit*[41] que nous détaillons par ailleurs.

Le principal apport de la méthode issue du projet KACTUS est de rationaliser la démarche consistant à réutiliser des ontologies pour en développer de nouvelles. Cette rationalisation passe par l'utilisation d'outils qui permettent de consulter une librairie d'ontologies existantes.

---

[37] http ://www.swi.psy.uva.nl/projects/Kactus/home.html
[38] http ://www.swi.psy.uva.nl/projects/Kactus/toolkit/intro.html
[39] mapping
[40] *Conceptual Modeling Language*
[41] Functional ontology of law

La méthode Terminae

Terminae est une méthode qui se fonde essentiellement sur l'analyse linguistique de textes pour conduire l'élaboration d'ontologies. Cette méthode, développée à l'université de Villetaneuse[42], propose ainsi l'élaboration d'ontologies principalement à partir de textes.

Cette méthode tente de répondre à des questions relatives à la construction d'ontologies qui ne sont le plus souvent pas réglées dans les autres méthodes, par exemple, le fait de sélectionner les concepts, le choix de leurs propriétés et de leurs relations, leur regroupement, l'influence de l'application dans ces choix etc.

Cette méthode repose sur le principe que les textes comprennent des termes et des relations lexicales qui sont plus ou moins identifiables à l'aide de différents outils de traitement automatique des langues. Ces termes sont regroupés selon leurs contextes et, le cas échéant, donnent lieu à la création de concepts et de relations sémantiques. Ces concepts et relations sont ensuite formalisés dans un modèle.

Plusieurs étapes rythment l'élaboration d'ontologies par cette méthode [AGBS00a, AGBS00b].

(*i*) L'étape initiale consiste, classiquement au regard des autres méthodes décrites ci-dessus, en une description des besoins.

(*ii*) L'étape suivante, étape de *constitution du corpus*, conduit à constituer un corpus sur lequel les outils de traitement automatique de la langue seront appliqués. Il est spécifié que ce corpus doit couvrir complètement le domaine requis par l'application et que celui-ci peut contenir des documents de types différents : notices techniques, textes didactiques, comptes-rendus d'expérience etc. Cette phase nécessite une expertise sur le domaine. Nous notons, à propos du corpus, que les auteurs envisagent leur méthode dans le contexte de textes techniques.

(*iii*) L'étape suivante, l'*étude linguistique*, consiste à utiliser les outils de traitement automatique des langues sur ce corpus pour en extraire les termes et leurs relations lexicales. Ces outils sont des extracteurs de candidats termes, des extracteurs de relations à partir de marqueurs linguistiques, des outils de regroupement conceptuel. La plupart de ces outils, quelle que soit la technique sur laquelle ils reposent (analyse syntaxique, analyse statistique, définition préalable de marqueurs de relation etc), nécessitent un travail humain fastidieux d'analyse et de sélection des résultats. Ce travail fastidieux se

[42] http ://www-lipn.univ-paris13.fr/ szulman/TERMINAE.html

double d'une nécessité d'expertise sur le domaine. De cette étape, on obtient néanmoins un ensemble de termes, de relations lexicales entre ces termes et de regroupements de ces termes.

(*iv*) La phase suivante, phase de *normalisation*, revient à conceptualiser les résultats de l'étape précédente en se fondant sur l'analyse de corpus. Un premier travail reste dans le domaine lexical pour sélectionner les termes à conserver par la suite, ceux qui font l'objet d'une acception différente en fonction du contexte et, enfin, donner à ces termes une définition en langage naturel. Le second travail définit les concepts et relations sémantiques sur la base des termes et de leurs relations lexicales. On obtient alors une structuration du domaine sous forme de réseau sémantique.

(*v*) Enfin, la dernière étape, celle de *formalisation*, utilise des ontologies ou terminologies existantes pour structurer les connaissances. Les concepts et relations sémantiques de l'étape précédente sont traduits en concepts et rôles dans le langage de la base de connaissances. Les auteurs évoquent alors une validation du modèle qui est envisageable pour établir sa cohérence.

Les auteurs de cette méthode notent eux-mêmes que leur méthode reste générale [AGBS00a], que les choix méthodologiques et techniques restent ouverts, dépendant de plusieurs facteurs comme la description des besoins relatifs à l'application, la documentation technique à disposition, l'expertise disponible etc. Cette méthode est utilisée pour le développement de diverses ontologies. Ainsi, le projet de construire une ontologie du domaine de l'ingénierie des connaissances qui permette aux chercheurs de décrire leurs propres outils et de les comparer à ceux qui existent déjà. Les textes techniques auxquels les auteurs se réfèrent pour la constitution du corpus sont alors principalement des articles scientifiques et des descriptifs techniques de leurs propres outils, y compris ceux utilisés dans cette méthode. Nous notons à cet égard une dérive possible vers le nombrilisme de la recherche dans ces domaines. Quoi qu'il en soit, la méthode telle qu'elle est décrite répond à un certain nombre de questions relatives à la construction d'ontologies, par exemple le fait de donner des définitions aux concepts. Elle ne répond cependant pas à des questions soulevées par les auteurs eux-mêmes comme irrésolues comme, par exemple, les façons de sélectionner les concepts, d'identifier les propriétés des concepts qui soient adéquates etc. Nous pensons que ces questions sont éminemment dépendantes du contexte, du domaine et peut-être encore plus des applications de l'ontologie. Une réponse claire et définitive ne peut y être donnée. Nous proposons, pour notre cas, de répondre à certaines de ces questions en fonction d'un élément principal : notre ontologie est dédiée à la recherche d'informations. Les principaux choix se font ainsi à ce titre, dans cette perspective particulière.

Une méthode sur corpus

Certains ([MS00]) proposent d'induire des ontologies de textes[43]. Les auteurs proposent une architecture générale permettant d'élaborer des structure conceptuelles et par là de construire des ontologies de façon semi-automatique dans un processus cyclique. Trois étapes majeures sont envisagées. La première conduit l'acquisition de concepts, la seconde l'établissement d'une taxinomie de concepts et la troisième la découverte de relations conceptuelles non taxinomiques. Les ressources textuelles utilisées par les auteurs dans leur méthode sont des dictionnaires techniques et des textes relatifs au domaine.

Des outils menant des analyses syntaxiques sont utilisés. Ces outils comportent des modules d'extraction de termes, un lexique contenant des termes et des catégories, ce lexique étant la base de l'analyse lexicale sur les textes : analyses morphologiques, reconnaissances de noms propres etc. Des analyses syntaxiques sont également faites sur les textes, permettant de découvrir des dépendances syntaxiques entre termes. Des analyses fondées sur les calculs d'information mutuelle entre paires de termes sont menées, permettant d'établir des relations supplémentaires aux relations taxinomiques.

Les auteurs illustrent leur approche d'un exemple. Il consiste à établir une taxinomie de concepts, un *noyau* à partir duquel les méthodes précédentes seront utilisées dans le but de l'élargir dans un processus cyclique vers une ontologie. L'établissement de cette taxinomie de concepts se fait en utilisant un dictionnaire technique du domaine en question. Chaque entrée du dictionnaire est envisagée comme un concept et les relations taxinomiques établies dans le dictionnaires sont utilisées. Schématiquement, l'élargissement de cette première taxinomie de concepts consiste à ajouter de nouveaux termes ou concepts issus de l'analyse linguistique de textes du domaine et d'établir de nouvelles relations non taxinomiques entre ces termes et concepts sur la base d'analyses statistiques utilisant elles-mêmes les données de la taxinomie de concepts.

La description et l'exemple donnés par les auteurs dans leur description de leur méthode d'élaboration d'ontologie nous semble trop vague pour pouvoir en faire une description plus complète. Nous pouvons cependant nous positionner vis-à-vis de ce que nous avons pu établir de leur méthode. Notre méthode, bien qu'également fondée sur des textes, ne se base pas sur des dictionnaires de la matière pour définir un réseau taxinomique de concepts. Par ailleurs, nous adoptons une démarche beaucoup plus globale que le processus cyclique décrit par les auteurs. Nous notons cependant que les grands

[43] Le titre de leur article est *Mining ontologies from text.*

principes de leur méthode d'élaboration d'ontologies sont aussi les nôtres : utilisation de textes du domaine, d'outils de traitement automatique des langues.

Conclusion sur les méthodes d'élaboration d'ontologies

De l'analyse de toutes ces méthodes tendant à l'élaboration d'ontologies émergent, malgré des différences notables, quelques grands principes [JBCV98]. Ainsi, la plupart de ces méthodes envisagent une application comme point de départ. C'est alors que la première question à laquelle doit répondre le concepteur d'une ontologie est : pour quoi faire ? Cette question permet d'envisager une évaluation de cette ontologie ainsi qu'une description de ses capacités à répondre à des questions données. Notre propre méthode, telle que nous la décrivons plus bas, s'inscrit dans cette mouvance, somme toute assez logique, pour la prolonger. Nous avançons en effet que les étapes mêmes de l'élaboration de l'ontologie, ainsi que sa structure propre, sont éminemment dépendantes de l'application visée.

Cette focalisation sur l'application de l'ontologie a cependant un inconvénient, le fait que sa réutilisation pour d'autres applications semble dès lors condamnée [JBCV98]. Ce point est indéniable. La conclusion à laquelle il amène est cependant contestable dans la mesure où l'application en question peut faire l'objet d'une définition plus ou moins précise. Nous proposons, par exemple, une méthode d'élaboration d'ontologies dédiées à la recherche d'information. Cette application se diversifie en plusieurs applications particulières comme la reformulation de requêtes, la catégorisation de documents ou la constitution de résumés de documents. Toutes ces applications particulières sont alors autant de nouvelles applications pour l'ontologie en question. Concernant cette question de la réutilisation d'ontologies, deux cas principaux se distinguent. Le premier est celui où les ontologies sont spécialement conçues en vue d'une réutilisation. Il en va ainsi des ontologies méta telles que nous les décrivons plus haut comme les *top ontologies*. Ces ontologies ont pour objet de décrire les primitives d'un domaine particulier ou d'un langage de spécification. Elles ont ainsi pour vocation première d'être utilisées comme fondement à des ontologie plus précises. Leur objet est donc la réutilisation. Le deuxième cas relatif à la réutilisation des ontologies concerne les ontologies résultant de projets dont l'objet est bien de répondre à un problème précis et non pas de tendre à la réutilisation de l'ontologie en résultant. Dans ce cas, des extrapolations et même des expérimentations peuvent être menées permettant de déterminer si l'ontologie en question est utilisable pour une autre application. Nous avançons cependant que de telles réutilisations ne sont envisageables que dans des applications relevant d'une même logique. Ainsi toutes les applications relatives à une démarche de gestion des connaissances dans une organisation comme la catégorisation thématique de documents, la gestion de forums thématiques de discussion, la création d'équipes dans des domaines particuliers etc, peuvent relever d'une même ontologie. Mais on ne peut *a contrario* envisager qu'une telle ontologie soit utilisée, par exemple dans un système multilingue de génération de documents si elle n'a pas été initialement prévue à cet effet.

Autre principe commun à une majorité de méthodes d'élaboration d'ontologies, celui d'un processus en deux étapes. Une première étape est alors dédiée à la définition d'une ontologie informelle, ou *ontologie conceptuelle*, ou *ontologie pré-computationnelle*, la deuxième étape consistant à formaliser cette ontologie dans une ontologie formelle computationnelle. Cette dernière est alors exprimée dans un langage de représentation approprié, c'est-à-dire opérationnel. Le passage de cette ontologie informelle portant les connaissances du domaine à l'ontologie formelle telle que ces connaissances soient utilisables par un système est ainsi un des enjeux de la construction d'ontologies. L'objet des ontologies en général est en effet de permettre le lien entre des systèmes informatiques et le monde réel qu'ils modélisent [JBCV98].

Dans chacune des méthodes décrites ci-dessus, une attente est plus ou moins formulée explicitement, celle de la création de bibliothèques d'ontologies déjà existantes, ce genre de bibliothèque facilitant des développements futurs d'ontologies dans la mesure où la réutilisation de certaines d'entre elles puisse être envisagée. Un effort de rationalisation et de centralisation est en effet certainement nécessaire dans ce domaine de la construction d'ontologies. Cet effort est amorcé par diverses initiatives, notamment le développement d'ontologies d'ontologies (!), ou la normalisation de descriptions et d'indexations d'ontologies [AVGPLTP98].

La plupart des méthodes d'élaboration d'ontologies ont fait l'objet d'une étude comparative [FL99] selon le standard IEEE 1074-1995[44]. La principale conclusion de cette étude porte sur la viabilité de ces méthodes pour avancer qu'aucune ne s'avère être complètement mature. Une échelle de degrés de maturité est établie qui classe les méthodes par ordre décroissant de maturité. La première est Methontology, puis TOVE, puis Enterprise Ontology, puis SENSUS et KACTUS.

[44] IEEE Standard 1074-1995, standard IEEE sur la qualité des logiciels, plus spécifiquement sur le cycle de vie du logiciel

Toutes ces méthodes telles que nous les décrivons s'attachent à répondre à un ensemble de questions que rencontre(nt) inévitablement le (ou les) concepteur(s) d'une ontologie. Ainsi la question de savoir si d'autres ontologies préalables doivent être réutilisées, le formalisme à adopter, comment identifier les connaissances c'est-à-dire les éléments de la future ontologie et les relations entre ces éléments, comment les modéliser etc. Toutes ces méthodes d'élaboration d'ontologies ont donc largement inspiré nos travaux, mais essentiellement du fait qu'elles répondent aux questions basiques de l'élaboration d'ontologies. Les réponses apportées à ces questions sont, il nous semble, principalement conditionnées par le contexte de l'élaboration de l'ontologie et par l'objet même de cette ontologie, ses applications.

Nous pensons qu'une méthode générique d'élaboration d'ontologies n'est pas envisageable et que la méthode adoptée dépend essentiellement du contexte d'élaboration de l'ontologie et des tâches auxquelles cette ontologie est dédiée. Nous proposons ainsi pour notre part (Partie II) une méthode d'élaboration d'ontologies dédiées à la recherche d'informations.

Notre objet est de constituer une ontologie du droit français dédiée à la recherche d'informations sur le Web. Notre domaine est alors extrêmement étendu : plus celui-ci sera bien couvert, plus l'ontologie sera utile. La méthode que nous adoptons pour l'élaboration de cette ontologie doit alors intégrer un maximum de techniques automatiques facilitant la tâche du concepteur.

1.2.3 Des outils dédiés aux ontologies

Dans la perspective de construction d'ontologies, des outils informatiques peuvent être développés, intervenant dans différentes phases de l'élaboration de l'ontologie. Ont ainsi été conçus des outils informatiques de travail collaboratif facilitant la communication entre les personnes chargées de la construction de l'ontologie. Ces outils s'inscrivent dans le contexte de l'approche d'élaboration d'ontologies consistant à rassembler des ingénieurs de la connaissance et des experts du domaine pour, par un travail de coopération, arriver à élaborer une ontologie par tous acceptable. Cette méthode d'élaboration a donné lieu à la définition de démarches d'élaboration principalement axées sur le travail coopératif et à distance. Les systèmes Tadzebao et Ontolingua ont été développés dans cette optique, permettant à plusieurs personnes distantes de travailler ensemble à la définition d'une même ontologie. Le système Tadzebao gère ainsi les problèmes d'échanges de messages (discussion) asynchrones et synchrones [Dom98] ainsi que l'affichage simultané d'images, de texte et la possibilité d'utiliser des outils graphiques pour la maintenance de l'ontologie en construction. Le système Ontolingua permet lui de publier, consulter, créer et éditer des ontologies stockées sur un *serveur d'ontologie*[45] dans le but d'atteindre un consensus entre plusieurs personnes impliquées dans l'élaboration d'une ontologie [FFR97].

D'autres outils ont été conçus pour être utilisés dans la phase même de construction de l'ontologie : interfaces de visualisation des premiers prototypes de l'ontologie, outils graphiques de création d'éléments de l'ontologie (objets et relations), outils d'inférence déterminant la validité des évolutions de l'ontologie etc.

Un ensemble d'outils a été créé, principalement sur la base du langage de spécification d'ontologies OIL, que nous décrivons section 2.3. Ces outils se différencient selon leurs tâches (pour plus d'information sur chacun d'entre eux, voir [FHvH$^+$01]) :

Les éditeurs d'ontologies pour la création de nouvelles ontologies. Ces outils sont principalement des interfaces aidant le concepteur d'ontologie. Ils intègrent les possibilités de définir des hiérarchies de concepts, les attributs des concepts, des axiomes et contraintes. Il permettent la création, la navigation et la modification d'ontologies. On peut citer OntoEdit[46], OILed[47] et Protégé[48].

Les outils d'annotation fondés sur des ontologies permettant de lier des sources structurées et semi-structurées d'informations avec des ontologies. Des DTD XML et des schémas XML peuvent ainsi être inférés d'ontologies exprimées en OIL. Et des définitions d'instances en RDF ou en RDFS peuvent également être inférées d'OIL.

Les outils de *raisonnement* sur ontologies aidant à la création d'ontologies et les utilisant, en accédant à l'ontologie elle-même. De tels outils *raisonnent* sur des ontologies dans le sens où ils peuvent établir la position adéquate d'un nouveau concept dans une hiérarchie de concepts donnée. Ainsi OIL utilise le système FaCT comme support pour la création, l'intégration et la vérification d'ontologies. FaCT (*Fast Classification of Terminologies* [49]) repose sur la logique des descriptions.

[45] *ontology server*

[46] développé par le Knowledge Management Group de l'Université de Karlsruhe, http ://ontoserver.aifb.uni-karlsruhe.de/ontoedit

[47] développé par l'Université de Manchester en collaboration avec Vrije Universiteit d'Amsterdam et SemanticEdge (http ://img.cs.man.ac.uk/oil)

[48] développé à l'Université de Stanford (http ://www.smi.stanford.edu/projets/protege)

[49] http ://www.cs.man.ac.uk/ horrocks/FaCT

Notre expérience de développement d'ontologie ne nécessitait pas d'outils de travail coopératif puisque le concepteur de l'ontologie, l'ingénieur de la connaissance, est également expert du domaine. Notre démarche ne fait de plus pas appel à d'autres experts du domaine dans la phase de conception. Pour ce qui est des outils d'aide à la création d'ontologies comme OntoEdit, OILed ou Protégé, nous les avons également écartés. D'une part, la plupart de ces outils sont conçus pour la construction manuelle d'ontologies, ce qui n'est pas notre problématique. D'autre part, quand les outils permettent l'importation d'ontologies préexistantes, celles-ci ne peuvent être d'une taille trop importante. Ainsi, notre ontologie est composée de plus de 100 000 termes.

Notre expérience de construction d'ontologie et la méthode que nous dégageons ne prévoient donc pas l'utilisation d'outils préexistants. Des outils propres ont cependant été développés dédiés à la partie manuelle de la construction de l'ontologie, à son évaluation pour la reformulation de requête etc.

1.3 Ontologies et analyse de texte

Comme il est posé dans l'introduction de cette étude, la démarche que nous adoptons pour la construction d'ontologie est fondée sur les textes. Nous nous plaçons ainsi dans la continuation des travaux menés conjointement en terminologie et intelligence artificielle. Ces travaux proposent, entre autre, une méthode de construction d'ontologies qui se fonde sur l'analyse de corpus en utilisant ses principes linguistiques et ses logiciels de traitement automatique de la langue (voir notamment [AGBS00a]). Nous adoptons ces principes dans notre démarche d'élaboration d'ontologie et nous attardons conséquemment dans cette section à l'ingéniérie linguistique.

De plus, notre ontologie est dédiée à la recherche d'informations et s'inscrit alors dans les problématiques des sciences de l'information que nous abordons également.

Par la démarche d'élaboration adoptée et par l'usage même auquel elle est destinée, notre ontologie ressemble beaucoup à un thesaurus mais aussi à une terminologie et un réseau sémantique. Nous développons ces notions dans une première section pour ensuite aborder des méthodes plus ou moins automatiques de construction.

1.3.1 Les thesaurus, terminologies et réseaux sémantiques

Les thesaurus

Aux débuts des bases documentaires informatiques, l'accès aux documents se faisait par le biais d'index, listes de termes avec lesquels ils avaient été manuellement indexés. C'était alors le travail préalable d'attribution de termes descripteurs aux documents, l'indexation manuelle, qui permettait, par le biais de requêtes sur ces descripteurs, d'identifier les documents en question. Les thesaurus ont alors été créés, avec pour principal objet de normaliser ce langage d'indexation manuelle et d'en structurer les éléments pour faciliter la recherche d'information ([Hea78] p. 31, [Jon95]). Les thesaurus sont ainsi définis (Norme ISO 2788) :

> «[...] vocabulaire d'un langage d'indexation contrôlé organisé formellement de façon à expliciter les relations a priori entre les notions.»

Sachant qu'un langage d'indexation se défini comme suit :

> «Ensemble contrôlé de termes choisis dans une langue naturelle et utilisés pour représenter, sous une forme condensée, le contenu des documents.»

Les thesaurus rassemblent alors les termes (ou descripteurs) autour de concepts par le biais de relations telles que *terme générique*, *terme spécifique*, *terme associé*, *terme équivalent*. Les thesaurus étaient alors, et le sont encore souvent, élaborés manuellement par des comités experts en documentation et du domaine en question.

Peu à peu, les capacités de stockage des ordinateurs allant sans cesse croissant, les bases et la recherche *full text*, en texte intégral, se sont développées à côté des systèmes d'indexation manuelle basée sur thesaurus[50]. Parallèlement, sont apparus les traitements statistiques et linguistiques des langues naturelles et la mise au point de système d'information *intelligents*[51]. Les thesaurus se sont alors vu attribuer de nouvelles tâches comme la reformulation plus ou moins automatique de requêtes.

[50] Nous notons que dans ce domaine juridique à été l'un des premiers à aménager l'accès en texte intégral aux documents

[51] Comme le système SPIRIT (*Système Syntaxique et Probabiliste d'Indexation et de Recherche d'Information Textuelle*) créé en 1979 dans la continuité de travaux menés conjointement au CEA et au CNRS notamment par C. Fluhr

Les réseaux sémantiques

Un réseau sémantique, ou graphe de concepts, ([Lef00], p. 136) est un graphe formé de nœuds réprésentant des concepts, reliés par des arcs orientés et étiquetés, qui représentent les relations sémantiques entre les concepts. WordNet[52] est ainsi un réseau sémantique organisant l'ensemble des termes du vocabulaire anglais courant autour de relations comme la synonymie, l'hyponymie-hyperonymie et l'holonymie-méronymie.

Dans les réseaux sémantiques, il n'y a pas de standards d'élaboration comme dans les thesaurus ni sur le choix des termes employés, ni sur celui des relations entre ceux-ci.

Les terminologies

Les terminologies sont ([Lef00], p. 58) des listes de termes d'un domaine ou sujet donné représentant les concepts ou notions les plus fréquemment utilisés ou les plus caractéristiques, cette liste étant ou non structurée.

Ces termes peuvent servir de descripteurs pour indexer le contenu de documents. Les terminologies sont alors, de la même façon que les thesaurus, des langages contrôlés. Pour être utilisées en recherche d'information, les terminologies doivent être structurées et inclure des relations, comme la synonymie, entre les termes. Les terminologies regroupent ([Lef00]) les *glossaires* (explicitant les définitions de termes choisis), les *terminologies multilingues spécialisées* (inventoriant et donnant la traduction de termes techniques d'un domaine), les *index terminologiques* (recensant les termes marquant d'un ouvrage). Elles rassemblent également ([BS99]) les *index structurés* (pour les systèmes d'indexation automatique), les *terminologies de référence* (pour les systèmes d'aide à la rédaction), les *référentiels terminologiques* (pour les systèmes de gestion de données techniques) etc.

Conclusion

P. Lebevre dans son ouvrage sur la recherche d'informations ([Lef00], p. 140) distingue les terminologies des thesaurus. L'auteur avance que les thesaurus sont des terminologies particulières possédant une structure de réseau sémantique tout en notant que ces notions sont différentes. Cette différence reposant essentiellement sur le fait qu'un thesaurus est défini avec une extrême rigueur selon des normes précises, chaque concept étant exprimé avec

[52]http ://www.cogsci.princeton.edu/ wn/

un terme préférentiel unique et les relations sémantiques étant normalisées.

Nous notons que d'autres ne distinguent pas les thesaurus des terminologies incluant les premiers dans ces dernières ([BS99]). De ce fait, la difficulté de distinguer les thesaurus des réseaux sémantiques et des terminologies n'en est que plus grande, aucune définition uniforme et partagée de chacun de ces objet n'étant, à notre connaissance, dégagée. En conséquence de quoi, la démarche qui voudrait que nous positionnions notre ontologie par rapport à ces objet n'en est que plus délicate. Ce que nous pouvons avancer, c'est que, du fait de son objet, de sa structure et des éléments qu'elle rassemble, notre ontologie relève à la fois du thesaurus, du réseau sémantique et de la terminologie. De ce fait, la méthode de construction d'ontologie que nous avons adoptée relève à la fois des méthodes développées pour la construction plus ou moins automatique de thesaurus, de réseaux sémantiques et de terminologies.

1.3.2 Méthodes d'élaboration (semi)automatiques

Historiquement, les travaux sur la construction automatique de telles ressources ont débuté à la fin des années 1980 et ont d'abord porté sur les *thesaurus*[53]. Dans les années 1990, les travaux portant sur les thesaurus se poursuivent (ainsi, les travaux de G. Greffenstette en 1994 [Gre94]) et sont parallèlement apparus des travaux sur la construction automatique de *terminologies* et de *réseaux sémantiques* (notamment avec le développement du logiciel Termino [DP90] ou les travaux sur la construction automatique de graphes de concepts [BB92]). Nous choisissons de ne pas distinguer ces différentes notions pour mener une étude qui, ne prétendant pas à l'exhaustivité, se veut illustrer les différents travaux menés dans ce domaine de la construction (semi)automatique de telles ressources.

Schématiquement, en considérant donc indifféremment les thesaurus, terminologies et réseaux sémantiques, la construction plus ou moins automatique de ces ressources se base sur des textes et procède en deux temps :
- l'identification des éléments de base (les *termes*, *concepts*, *mots* ...)
- et l'établissement de relations entre ces éléments (relations lexicales, syntaxiques, sémantiques ...).

Ces démarches s'appuient sur un ensemble d'outils de traitement automatique de la langue comme étiquetteurs syntaxiques, extracteurs de termes

[53]Notamment en France dans le laboratoire CLIPS (*Communication Langagière et Interaction Personne-Système*) de l'Institut d'Informatique et de Mathématiques appliquées de Grenoble http ://www-clips.imag.fr/

complexes, modules d'acquisition de relations entre termes, de classification de termes etc. et combinent, ce qui n'est pas nouveau[54], analyse linguistique et analyse statistique. Le système IOTA[55] (voir notamment [Bru89]) utilise ainsi un analyseur syntaxique de surface combiné à une analyse de la fréquence des termes et leurs cooccurrences pour extraire automatiquement un thesaurus des textes qui lui sont soumis.

L'identification de termes

A partir du début des années 1990, plusieurs logiciels d'extraction de canditats termes ont été développées se fondant sur des analyses, combinées ou non, statistiques et syntaxiques des textes (nous abordons également ce point de façon introductive dans la section sur les outils extracteurs de termes que nous avons testés Section 5.1). Nous en citons quelques uns.

Le logiciel ANA ([Eng92]) a été développé au Centre d'Etudes Atomiques pour l'enrichissement de réseaux lexicaux exploités par un système de gestion des connaissances. Il extrait les termes candidats sans recourir à des analyses syntaxiques mais en établissant des égalités approximatives entre mots et en menant l'observation de répétition de patrons.

Le logiciel ACABIT ([Dai94]) a été développé au sein de la société IBM pour la construction de lexiques terminologiques multilingues dans le cadre d'un projet de traduction automatique. Le logiciel travaille sur des corpus de textes préalablement étiquettés et désambiguïsés. La première étape repose sur une analyse linguistique permettant d'extraire des séquences nominales qui sont ensuite ramenées à des candidats termes binaires. Un filtrage statistique est ensuite opéré sur la liste des candidats termes.

Le logiciel LEXTER (ancètre de Syntex que nous utilisons et détaillons dans la Section 5.1) a été développé à Electricité de France pour la mise à jour d'un thesaurus utilisé par un système d'indexation automatique. Le logiciel travaille lui aussi sur des corpus de textes préalablement étiquettés et désambiguïsés. Lexter mène l'acquisition de termes par repérage de frontières syntaxiques combiné à une méthode d'apprentissage endogène sur corpus permettant de résoudre les ambiguïtés de rattachement prépositionnel, les syntagmes nominaux maximaux sont alors identifiés. Ces syntagmes

[54] Cette combinaison est implémentée dès 1979 dans le système d'indexation et de recherche d'information SPIRIT (voir [ABD$^+$81] et [DFR89])

[55] http ://www-mrim.imag.fr/realisations/

sont ensuite décomposés pour produire une liste de candidats termes.

Les relations entre termes

De la même façon que pour l'identification des termes, les relations peuvent être détectée sur la base de deux approches, combinées ou non : l'analyse statistique des termes dans les textes et l'analyse syntaxique des phrases des textes.

La statistique est utilisée dans les outils de classification de termes sur la base de cooccurrences de ceux-ci dans des textes ou des fenêtres de mots et dans les outils de classification de termes sur la base de la distribution syntaxique.

Les outils se fondant sur la cooccurrence des termes reposent sur le principe, dégagé très tôt en recherche documentaire ([SJ71]) selon lequel deux termes appaissant souvent ensembles dans un même contexte peut être rapprochés. Ainsi le projet ILIAD ([TND$^+$98]) dans lequel les candidats termes extraits par des outils d'extraction sont regroupés en fonction de leurs cooccurrences. Les classes ainsi constituées sont ensuite soumises à interprétation par des experts.

L'outil LEXICLASS ([Ass98]) regroupe les termes à l'aide de méthodes de classification automatique hiérarchique ascendante en se basant sur la distribution syntaxique des termes donnée par LEXTER.

D'autres méthodes statistiques de détection de relation entre termes consistent à mener une classification automatique de ceux-ci sur la base de leur apparition conjointe dans un même contexte. Ainsi, le système CONTERM ([MN96]) se fonde sur les algorithmes de classification par réseaux de neurones pour rassembler des termes en classes comme autant d'indices de régularités d'associations lexicales.

D'autres méthodes consistent à détecter des relations entre termes dans des textes en utilisant des listes de marqueurs syntaxiques de relations préalablement établies. Ainsi les travaux de Hearst sur l'extraction automatique de liens d'hyperonymie ([Hea92]).

Ces outils peuvent reposer sur une analyse syntaxique des textes plus ou moins précise. Ainsi, le logiciel SEEK ([Jou93]) se contente d'une «analyse syntaxique approximative» pour proposer des relations sémantiques entre termes. Le système se fonde sur une base de marqueurs linguistiques préalablement établie (savoir linguistique indépendant d'un quelconque domaine [Jou95]) qui, couplée à un moteur d'inférence et à un ensemble de règles d'exploration contextuelle, détecte des indices linguistiques pertinents orientant la construction de représentations sémantiques.

Une analyse syntaxique plus précise est menée par le module Upery ([BL02]) d'analyse distributionnelle de l'outil Syntex. Ce module détecte les relations de dépendance syntaxique entre termes. Nous avons utilisé, entre autre, Syntex et son module Upery pour l'identification de relations entre termes. Pour une description plus détaillée de l'outil, voir Section 6.2.

Les méthodes syntaxiques et statistiques peuvent être combinées de façon à combiner les avantages de chacunes. Par exemple, les systèmes CONTERM et SEEK ont été associés ([JBD+97]) dans le but de permettre le traitement de larges corpus de textes par l'approche statistique, tout en y associant une analyse sémantique.

La construction plus ou moins automatique de thesaurus peut également conduire à l'intégration d'éléments divers comme des thesaurus préexistants élaborés manuellement (voir [Che02] par exemple). Ces approches tentent de résoudre des questions de chevauchement, d'articulation de fond et de forme entre sources parfois hétérogènes.

La méthode que nous adoptons pour la construction de notre ontologie s'inspire plus ou moins des méthodes ci-dessus décrites. Nous choisissons de conjuguer les analyses syntaxiques et statistiques des textes pour à la fois extraire les termes et trouver des relations entre ceux-ci. Cependant, dans notre démarche, une analyse syntaxique relativement fine est menées sur la totalité du corpus (avec l'outil Syntex) et ce sont ensuite les méthodes statistiques qui viennent améliorer ses résultats (voir la Partie II).

Chapitre 2

Formalismes pour les ontologies

La problématique majeure des ontologies en informatique est de savoir comment structurer au mieux les éléments de l'ontologie pour que l'utilisation par un système informatique en soit la plus rationnelle possible [Cas02]. Dès lors, la définition de formalismes semble nécessaire.

La notion de *formalisme* recouvre dans notre propos celle de modélisation ainsi que celles de *normalisation* et de *formalisation* ([AGBS00a], [Bac00]) conduisant à un modèle de plus en plus formel traduit en langages de plus en plus formels.

Il faut cependant distinguer deux types de formalismes relativement aux ontologies : le formalisme pour donner du sens, qui consiste à exprimer de façon plus ou moins formelle les connaissances, et le formalisme pour implanter un système, qui consiste à exprimer l'ontologie dans un langage informatique opérationnel.

Les ontologies constituant des conceptualisations partagées de domaines, elles "portent" une hiérarchie de concepts et une description des propriétés majeures de chacun de ces concepts par un mécanisme d'attributs et de valeurs. Les liens entre les concepts sont également explicités [DMVH+00]. A titre d'exemple, une ontologie des animaux est envisageable, constituée de la classe des carnivores, qui ont certains attributs, et incluant les lions, les tigres etc, la classe des herbivores qui ne sont pas des carnivores et qui inclut les girafes etc.

Dès lors, les langages de représentation ou de spécification d'ontologies doivent permettre l'expression de toutes ces caractéristiques, ou du moins de celles nécessaires à la représentation d'une ontologie donnée pour une (ou plusieurs) tâche(s) donnée(s).

Les langages de représentation d'ontologies s'inscrivent dans la continuité des langages de représentation des connaissances, plus génériques reposant chacun sur des paradigmes de représentation différents et utilisant des primitives de représentation différentes. Ces paradigmes de représentation sont généralement distingués comme suit ([BCKN01]) :

- les logiques de description, ou *logiques terminologiques*, utilisant les *concepts* et les *rôles*, qui sont des relations binaires ;
- les langages de frames, ou *langages de schémas*, utilisant les *frames* et les *attributs* ;
- les graphes conceptuels, utilisant les *types*, les *relations* et les *graphes* ;
- les langages fondés sur la logique de premier ordre, utilisant les *prédicats* comme primitives de représentation.

Ces différents langages de représentation ont chacun leurs spécificités et leurs avantages et inconvénients. L'objet de nos travaux n'est cependant pas de développer ces points ni d'expliciter les différentes filiations ayant conduit à la définition des différents langages de représentation d'ontologies. Ce que nous notons, c'est que les langages de représentation d'ontologies combinent parfois plusieurs de ces langages de représentation pour en tirer le plus d'avantages possibles, comme OIL par exemple. L'objet de cette section est de décrire les principaux langages pouvant être utilisés pour la spécification d'ontologies, que ce soit leur finalité propre ou pas, et que ce soit au niveau du sens ou au niveau de l'implémentation.

2.1 XML

XML (*eXtensible Markup Language*) est un métalanguage dérivé de SGML (*Standart General Markup Language*) utilisé pour définir des langages de marquages comme XHTML. Ces langages de marquage permettent la structuration des documents du Web non plus sur la base d'une structure figée, comme le permettait HTML, mais en laissant la possibilité au concepteur de distinguer les données selon leur sens et leur contenu. Un document XML va alors se présenter sous la forme d'un ensemble de données *taggé* par un ensemble de balises, chacune pouvant comporter des attributs et des valeurs.

Il n'y a pas de définition figée des balises, et leur *grammaire* peut être consignée dans un document séparé des documents XML, la DTD (*Document Type Definition*). Il est important de noter avec [DMVH+00] que la DTD n'a aucune vocation à traiter de la sémantique ; la DTD ne sert qu'à définir des conventions de syntaxe.

Les schémas XML, normalisé par le W3C en mai 2001, permettent, de la même façon que les DTD, de modéliser et de valider les documents et leurs données. Les schémas XML sont cependant beaucoup plus précis que les DTD ([BKL+02], p. 47), des contraintes plus strictes pouvant être données aux valeurs.

Une des contradictions fondamentales entre XML et les ontologies est la suivante : XML ne traduit finalement que des grammaires et les ontologies nécessitent de représenter la sémantique. Les DTD, par exemple, ne permettent pas de déterminer les concepts et relations d'un domaine donné, éléments fondamentaux d'une ontologie. De la même façon concernant les schémas XML, la logique de classe, nécessaire à la représentation d'une ontologie, ne peut être traduite. Seul un programme effectuant des tâches sur des données XML est en mesure d'en extraire une quelconque sémantique. XML est un bon moyen de décrire des données et de les stocker mais ne permet pas d'interpréter les données mêmes d'un document [DMVH+00] et ne peut donc être utilisé en tant que tel pour formaliser une ontologie.

Un language a pourtant été défini sur la base de XML avec pour principal objet l'échange d'ontologies. XOL (*XML-based Ontology exchange Language*) tente de répondre à deux impératifs : la nécessité d'avoir un language de représentation des connaissances orienté objet, une variante de OKBC[1] (OKBC-lite) est alors utilisée, et la nécessité d'une syntaxe XML permettant l'interopérabilité et intégrant les capacités des parsers XML. Nous notons cependant que l'objet même de XOL n'est pas la représentation d'ontologies mais bien leur échange. XOL peut cependant être utilisé comme langage intermédiaire lors du transfert d'ontologies dans différents systèmes de bases de données [CGP00].

Nous utilisons XML pour stocker les données de notre ontologie que nous laissons à disposition sur le Web pour téléchargement. Un schéma XML a été défini (Annexe schéma XML). Mais il faut bien rappeler que XML est utilisé uniquement pour le stockage des données de notre ontologie : termes, liens, types de liens etc.

[1] *Open Knowledge Base Connectivity*

2.2 RDF

RDF (*Ressource Description Framework*) est un modèle de représentation sémantique des informations du Web qui utilise la syntaxe d'XML. Ces représentations comportent des *méta-données* sur les *ressources* du Web comme les auteurs de pages Web, leur date de création etc, les *ressources* du Web étant l'élément de base en RDF, toute entité imaginable accessible ou non par le Web : une page HTML, une personne, une sculpture etc. Chaque *ressource* est pourvue d'une URI[2]. Initialement recommandé par le W3C dans le but de standardiser les définitions et les usages des méta-données, RDF est également utile à la représentation de données elles-mêmes. Les éléments principaux de RDF sont les objets, leurs attributs, et les valeurs de ces attributs. L'objet principal de RDF est de définir un mécanisme pour décrire des données indépendamment de tout domaine et spécificité. De la même façon qu'avec XML, RDF ne permet pas la déclaration de propriétés particulières ; leur définition est totalement libre.

Les schémas RDF (RDFS) permettent ainsi, de la même manière qu'en XML, la définition du vocabulaire utilisé dans les descriptions RDF. Les schémas RDF confèrent un formalisme de représentation riche et adapté aux ontologies, incluant des classes, des sous-classes, des propriétés, sous-propriétés, des règles d'héritages de propriétés etc. Dans les schémas RDF, des types sont ainsi prédéfinis tels *Class*, *subPropertyOf*, *subClassOf* etc. Le langage RDFS se fonde ainsi sur le langage RDF. Ces types permettent de spécifier des caractéristiques hiérarchiques de classes et de propriétés. La structure objet-classe des RDFS permet de représenter un modèle du domaine définissant des objets du domaine et leurs relations tel qu'est une ontologie.

Des aménagements du langage RDF semblent cependant nécessaires pour en faire un véritable langage de représentation d'ontologie. Ainsi, le W3C, dans ses toutes premières recommandations tendant à la définition d'un language d'ontologie (OWL 1.0 en avril 2002 *Ontology Web Language*), spécifie bien que ce language doit aller au-delà de schémas RDF.

> «The next element for the Semantic Web is a Web ontology language which can formally describe the semantics of classes and properties used in the Web documents. In order for machines to perform useful reasoning tasks on these documents, the language must go beyong the basic semantics of RDF Schema.[3]»

[2]Universal Resource Identifier. ex : une URL comme http ://www.organisation.org/personnel/paul.durand.html

[3]*Le prochain élément pour le Web sémantique est un langage spécifique pour les ontolo-*

De la même façon, une équipe de recherche de l'université de Karlsruhe ([SEMD00]) se fonde sur les RDFS pour définir un language de description d'ontologies. Notant que RDF en tant que tel présente quelques limites dans la description des ontologies, notamment dans le fait que seuls les concepts et relations issus de mécanismes d'héritages sont pris en compte, l'équipe propose d'adapter RDF en y adjoignant la possibilité de définir des axiomes en tant qu'objets.

2.3 OIL

OIL[4] (*Ontology Inference Layer*) est un language de représentation d'ontologie qui dérive de RDF. Les principaux fondements du langage OIL sont les langages de frame (tels OKBC, XOL ou RDF) et les logiques de descriptions. OIL a été défini dans l'objectif de permettre la spécification et l'échange d'ontologies. Le premier objectif prévalant à la création du language OIL, projet financé par la Communauté Européenne [5], était d'en faire un langage largement intuitif pour l'homme. Le langage devait également avoir une sémantique formelle bien définie et enfin, pour des raisons d'interopérabilité, être lié aux langages existants tels XML ou RDF. Pour résumer, le défit des concepteurs du language OIL ([FHvH+01]) était de réunir dans un même langage la richesse épistémologique de la modélisation objet (*frame-based logic*), la sémantique formelle et un support de raisonnement efficient porté par la logique des description et enfin les standards d'échanges de la communauté Web.

OIL permet la description d'ontologies sur la base des éléments de la logique de frame ([RC01]) : les classes, les propriétés de classes, les relations entre les classes etc. Les classes s'organisent en hiérarchie incluant des classes et des sous-classes. Certaines classes ([FHvH+01]) peuvent être définies de façon à ce que leurs propriétés soient des conditions nécessaires et suffisantes à l'appartenance à ces classes (*defined classes*). Les classes peuvent être combinées par le biais d'expressions logiques : intersection, union etc. Les propriétés se déclarent de la même façon que les axiomes logiques. Elles peuvent être fonctionnelles (*functional*), c'est-à-dire n'avoir qu'une seule et même valeur,

gie du Web qui permette la description formelle des sémantiques de classes et de propriétés utilisées dans les documents du Web. De façon à ce que les machines puissent opérer des tâches utiles de raisonnement sur ces documents, ce langage doit aller au delà de la sémantique basique des schémas RDF.

[4]http ://www.ontoknowledge.org/oil/index.shtml

[5]Les projets Ibrow (*an Intelligent Brokering Service for Knowledge-Component Reuse on the World-Wide Web*) et *On-to-Knowledge*.

transitives ou symétriques. Des restrictions peuvent être apportées sur le type des propriétés ainsi que sur un nombre restreint de valeurs qu'elles peuvent, le cas échéant, prendre.

Un des aspects fondamentaux de OIL ([FHvH+01]) est sa sémantique formelle. Ainsi, chaque classe est organisée en ensemble d'objets et chaque propriété en ensemble de paires d'objets, cette organisation devant bien évidemment se conformer aux contraintes spécifiées lors de la définition des classes et des propriétés.

OIL est organisé en *couches hiérarchisées* (*ever-increasing layers*) de différents sous-langages. Chaque couche, ou niveau, ajoute des fonctionnalités et, partant, de la complexité au niveau précédent. Un agent, humain ou machine, traitant spécifiquement d'une ontologie définie par un sous-langage d'un certain niveau est ainsi en mesure de *comprendre* partiellement une ontologie définie dans un sous-langage de niveau supérieur.

Les différents sous-langages de OIL sont les suivants ([FHvH+01]) :
- Standard OIL : permet d'exprimer les principales primitives de modélisation de façon suffisamment expressive et compréhensible. La sémantique est spécifiée, ainsi que les règles d'inférence.
- Instance OIL : permet de spécifier une modélisation plus profonde. Instance OIL rend possible la spécification complète d'une base de données.
- Heavy OIL : devrait posséder (il est en cours de développement) des capacités encore supérieures de représentation et de raisonnement, la définition de règles et de *méta-classes* devant être possible.

Le choix entre les différents sous-langages OIL de description d'ontologies dépend de la complexité même de l'ontologie en question. Il ne sert à rien d'exprimer une ontologie dans un sous-langage OIL comportant plus d'expressivité et de complexité que nécessaire [FHvH+01].

Une ontologie exprimée en OIL est elle-même annotée avec des données propres telles le titre de l'ontologie, son ou ses créateurs, sa date de création, sa date de réactualisation etc. OIL suit les spécifications du W3C Dublin Core Standard pour exprimer ces méta-données sur l'ontologie elle-même.

L'objectif d'OIL est ([Fen00a]) de permettre la définition d'environnements complets de gestion des connaissances (*knowledge management*) dans le cadre d'intranets.

2.4 OntoSpec

Ontospec [Kas02a] propose une méthode de spécification *semi-informelle* d'ontologies conceptuelles ou *ontologies pré-computationnelles.* OntoSpec se situe ainsi dans la première étape de l'élaboration d'ontologies, celle de la modélisation des connaissances ontologiques, ou comment leur donner du sens. OntoSpec n'est donc pas une méthode complète d'élaboration d'ontologie mais sa réponse, même partielle, à ce problème n'en demeure pas moins à noter.

La distinction entre une modélisation pour donner du sens, modélisation des connaissances ontologiques conduisant à la définition d'une *ontologie conceptuelle*, et une modélisation pour implémenter un système conduisant à une *ontologie computationnelle*, est issue de la méthode CommonKADS. Les méthodes d'élaboration d'ontologies Methontology et Terminae distinguent également ces deux phases.

OntoSpec propose une méthode de spécification d'ontologies conceptuelles en langue naturelle fortement structurée et contrôlée ([Kas02a]). Cette méthode conduit à définir les entités conceptuelles, les concepts et leurs relations, comme des ensembles structurés de propriétés. Chaque propriété est présentée comme une proposition logique et des règles typographiques sont posées.

L'objectif d'OntoSpec est de pousser le concepteur à passer par une phase initiale de modélisation des connaissances. L'ontologie conceptuelle en résultant, spécifiée en langue naturelle structurée et contrôlée, sert de base à un travail collaboratif entre le cogniticien et l'expert. OntoSpec conduit à une modélisation précise des entités conceptuelles, ce qui tend à faciliter le travail du cogniticien, ainsi que la compréhension de l'ontologie par un acteur extérieur et sa réutilisabilité. L'auteur [Kas02a] note dès lors que l'introduction d'une ontologie formelle spécifiée en langage de représentation comme KIF ou RDFS est optionnelle pour le développement d'une ontologie computationnelle.

La suite logique de la définition d'une ontologie conceptuelle est la définition d'une ontologie computationnelle qui puisse être utilisée en tant que telle dans un système informatique. Cette ontologie opérationelle est définie, dans la suite d'OntoSpec, dans un langage de représentation d'ontologies qui est un langage opérationnel Def-* [Kas99].

OntoSpec s'attache donc à ce que l'élaboration d'une ontologie computationnelle passe par celle d'une ontologie conceptuelle porteuse d'une modélisation précise et rigoureuse des connaissances. Cette ontologie conceptuelle est spécifiée en langue naturelle structurée et contrôlée, ce qui la rend lisible par un non cogniticien. L'objectif d'OntoSpec est donc de prévoir à la fois une modélisation précise des connaissances de l'ontologie, et une lisibilité de l'ontologie conceptuelle en résultant.

2.5 Autres langages

Nous évoquons dans cette section quelques langages utilisés et utilisables pour la représentation d'ontologies, que ceux-ci soient spécifiquement dédiés à cette finalité ou pas.

CoGITo [6] se fonde sur le formalisme des graphes conceptuels proposés par Sowa en 1984. Pour l'équipe qui a développé CoGITo, les travaux réalisés en théorie des graphes sont applicables aux graphes conceptuels.

DAML [7] est un langage permettant la représentation d'ontologies développé dans le cadre d'un programme DARPA[8] sur le Web sémantique. DAML est une combinaison de XML et de RDF permettant de spécifier des objets mais aussi les relations entre ces objets. La dernière version de DAML se combine avec OIL (DAML+OIL) et cette association préfigure une future large utilisation du fait de la coopération internationale entre un projet américain et un projet européen qu'elle induit.

DefOnto [9], développé à l'Université de Picardie, est un langage spécifiquement dédié à la représentation d'ontologies. DefOnto se fonde en partie sur le langage de schémas (ou *langage de frames*) KL-ONE. Il se différencie des logiques de description dans le fait qu'un concept est obligatoirement défini par rapport à au moins un autre concept le subsumant et en permettant de définir des relations en tant qu'entités autonomes [BCKN01].

FLogic (pour *Frame Logic* [KL89]) est un langage combinant les langages de frame (ou *langage de schémas*) et les prédicats de premier ordre. Ses primitives de représentation sont les objets, les objets complexes, les héritages etc.

[6] http ://www.lirmm.fr/c̃ogito
[7] http ://www.daml.org
[8] Agence américaine *Defense Advanced Research Projects Agency*
[9] http ://www.laria.u-picardie.fr/EQUIPES/ic/DefOnto/

LOOM [10], développé à partir du milieu des années 1980 à l'Université de Californie du Sud, permet de définir des applications composées d'une partie déclarative et d'une partie procédurale. C'est la partie déclarative de ces applications qui correspond à la représentation de connaissances ontologiques et de modèles de domaine. LOOM se fonde à la fois sur les logiques de descriptions et sur des expressions de calcul des prédicats.

OCML [11] (*Operational Conceptual Modeling Language*) permet de spécifier des modèles de résolution de problèmes au niveau des connaissances et de les opérationnaliser. Son architecture distingue les tâches, les méthodes de résolution, le domaine et l'application. La couche domaine permet de définir des connaissance de la même façon qu'avec Ontolingua. OCML est utilisé pour représenter les ontologies dans WebOnto[12], interface d'édition et de navigation pour les ontologies [Dom98].

OKBC [13] (*Open Knowledge Base Connectivity*) spécifie un protocole d'accès aux bases de connaissances organisées selon le paradigme des langages de frame. Ce protocole consiste en un modèle uniforme de systèmes de représentation des connaissances[14], le modèle objet, fondé sur une conception commune des classes, propriétés, individus, valeurs [RC01]. OKBC combiné à XML a permi de définir le langage XOL, lui-même inspirateur de OIL.

Ontolingua [15] est un langage fondé sur KIF et sur *Frame Ontology*. KIF (*Knowledge Interchange Format*) a été développé dans l'objectif de résoudre les problèmes d'hétérogénéité des langages de représentation des connaissances. *Frame Ontology* permet la spécification d'ontologies selon la logique des langages de frame. Ontolingua est le langage de spécification d'ontologies du serveur Ontolingua de l'université de Stanford dédié à la création coopérative d'ontologies [FFR97].

SHOE [16] (*Simple HTML Ontology Extension*) a été développé à l'université du Maryland aux Etats-Unis. Ce langage est à la base une extension de HTML permettant d'annoter des pages Web avec des connaissances traitables par les machines (*machine readable*), ces connaissances étant elles-mêmes décrites dans des ontologies [LSR96]. SHOE a été rendu

[10] http ://www.isi.edu/isd/LOOM
[11] http ://kmi.open.ac.uk/projects/ocml/
[12] http ://kmi.open.ac.uk/projects/webonto/
[13] http ://www.ksl.stanford.edu/software/OKBC/
[14] KRS *Knowledge Representation Systems*
[15] http ://ksl.stanford.edu/software/ontolingua
[16] http ://www.cs.umd.edu/projects/plus/SHOE/

compatible avec XML. Le principe de SHOE est de permettre à des agents de collecter des informations sur le sens des pages Web. Les classes, les relations et les règles d'inférences peuvent être définies.

2.6 Conclusion sur les langages

Comme nous le montrons dans la section précédente, l'éventail des choix possibles de langages de représentation et de spécification d'ontologies est très large. Un choix a du être fait parmi cet ensemble de langages. Ce choix est fondé principalement sur des critères techniques et de portabilité, critères qui peuvent également s'avérer *politiques* dès lors que le choix se porte sur un langage largement utilisé et reconnu dans la communauté internationale ou sur un langage aux répercussions plus discrètes.

Le choix entre les différents langages de représentations d'ontologies se fonde tout d'abord sur des comparaisons menées sur ceux-ci. L'objet de nos travaux n'étant pas de mener nous-mêmes ces études comparatives, nous nous contentons d'exploiter les résultats déjà obtenus.

Dans une étude menée en 2001 et présentée au cours de la conférence IC'2001 à Grenoble, des chercheurs se sont attachés à l'*Evaluation de langages opérationnels de représentation d'ontologies* [BCKN01]. Cette étude se concentre sur des *langages opérationnels de représentation*, c'est-à-dire des langages formels fournissant des spécifications directement exécutables par une machine. Un protocole de comparaison a été établi qui permet de comparer les puissances d'expression, le mode d'opérationnalisation ainsi que l'existence d'outils associés d'aide au développement. Les langages comparés sont les suivants : CoGITo, DefOnto, LOOM, OCML et OIL. Ces langages de représentation d'ontologies sont des langages de programmation de haut niveau qui sont soit compilés, pour CoGITo et DefOnto, soit interprétés dans un langage hôte tel LISP, C++ ou JAVA, pour LOOM, OCML et OIL. Des critères de comparaison ont donc été établis tels le fait de savoir si l'implantation de l'ontologie nécessite de développer dans le langage hôte, de savoir de quels services inférentiels, permettant de déduire des connaissances nouvelles à partir des connaissances déjà représentées, le langage est doté etc. Les langages sont positionnés les uns par rapport aux autres selon trois dimensions : la puissance d'expression, le mode d'opérationnalisation et l'existence d'outils associés d'aide au développement.

Une étude plus précise des capacités de représentation et de raisonnement des langages permettant de spécifier des ontologies est présentée dans [CGP00]. Cette étude distingue les langages traditionnels de spécification d'ontologies (*traditional ontology specification languages*) comme Ontolingua, OKBC, OCML, FLogic et LOOM des langages orientés Web pour la construction d'ontologies (*web languages for building ontologies*) tels XOL, SHOE et OIL. Cette étude se fonde sur le modèle proposé dans CommonKADS distinguant les connaissances de domaine des connaissances d'inférence pour définir des critères de comparaison entre chacun de ces langages. Ainsi, sont analysées les capacités de chaque langage à exprimer toutes les caractéristiques des concepts, des attributs, des relations entre les concepts, des instances, des règles ainsi que des mécanismes d'inférences sur ces connaissances du domaine tels que la classification automatique, les héritages, les exceptions etc. La conclusion globale de cette étude est que les langages traditionnels de spécification d'ontologies (*traditional ontology specification languages*) tels qu'Ontolingua, OKBC, OCML, FLogic ou LOOM ont globalement une expressivité, c'est-à-dire une capacité à représenter un ensemble d'éléments constitutifs d'ontologies, plus large que les langages orientés Web pour la construction d'ontologies (*web languages for building ontologies*) tels XOL, SHOE et OIL.

Du fait de la taille de notre ontologie (plus de 100 000 termes) et du fait que celle-ci est consultée en ligne, le paramètre essentiel est pour nous l'accessibilité des données de l'ontologie. De plus, l'usage même de notre ontologie ne nécessite qu'une consultation de celle-ci, aucun raisonnement n'est opéré sur les données de l'ontologie.

Les langages de représentation d'ontologies décrits précédemment ont donc été écartés, du fait que leur objet est différent du nôtre, par exemple le langage OntoSpec qui sert à la spécification d'ontologies pré-computationnelles, et du fait que les services d'inférence qu'ils proposent ne nous sont pas nécessaires. Il était dès lors inutile de spécifier notre ontologie dans un de ces langages. Nous utilisons seulement XML pour permettre l'échange de notre ontologie laissée à disposition pour téléchargement sur le Web[17] (voir Section 2.1).

Les données de notre ontologie sont simplement stockées dans une base de donnée[18] créée à cet effet. Les programmes Web dynamiques y accèdent alors directement et rapidement par le biais de tables de hashage. Dans la base de données, une entrée par terme est aménagée. A chacun de ces termes

[17]http ://ontologie.w3sites.net
[18]*DB Files* : DataBase Files.

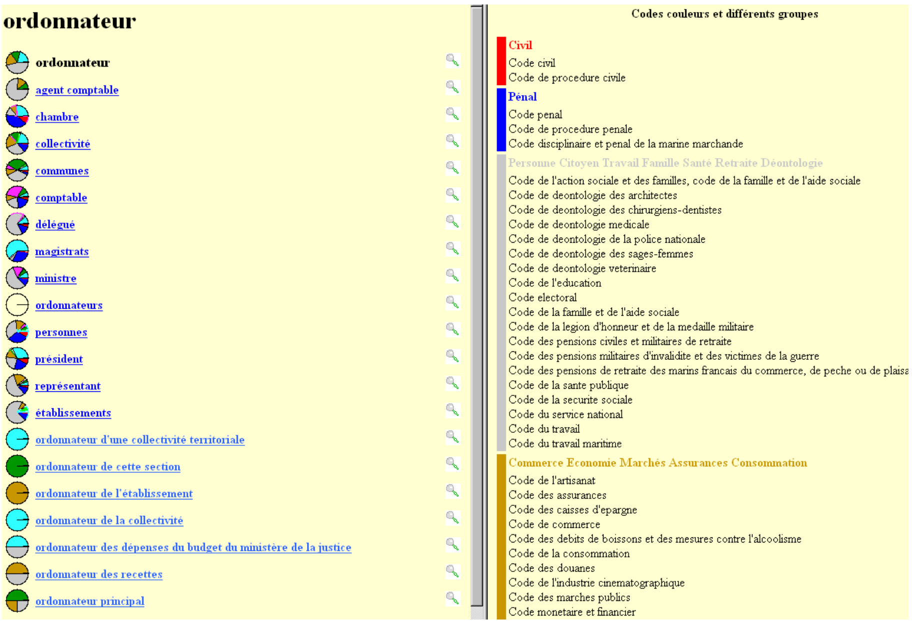

FIG. 2.1 – Les termes liés dans notre ontologie au terme *ordonnateur*

clés correspond les termes qui lui sont associés (Figure 2.1). Sont également associées à chaque terme les informations sur ses provenances thématiques, les codes, et fréquences. Ainsi, chaque clé de la base de données correspond à un concept de l'ontologie et à une classe. Les sous-classes sont représentées par les termes liés au terme clé qui se trouvent être eux-mêmes des termes clés. Les relations entre les classes et entre les classes et les sous-classes sont spécifiées par la sémantique des relations liant les termes-clé entre eux et les termes clés à leurs termes associés.

Chapitre 3

Ontologie et droit

Le projet d'élaboration d'une ontologie du droit ne peut se concevoir sans considérer un tant soit peu la discipline juridique. Nous nous attardons à la problématique conjointe des ontologies et du droit dans ce Chapitre 3.

Le droit et l'informatique se rencontrent dans plusieurs cas, tous relevant de cette discipline nouvelle née du traitement automatique de l'information juridique : l'informatique juridique. L'informatique juridique implique en effet à la fois l'étude de la logique du droit, de l'argumentation juridique, l'analyse cognitive de la décision juridique ainsi qu'une approche raisonnée du langage et de la structure du texte juridique ([BHR00]). Une démarche relevant, comme la nôtre, de l'informatique juridique, doit impliquer, à des degrés divers, une approche à la fois épistémologique et théorique sur le droit.

Le recours aux formalismes et aux modèles pour tenter de comprendre le droit remontent à Aristote et à Platon [Bou99]. Le premier analysa l'éloquence judiciaire *via* ses raisonnements dialectiques. Le second s'intéressa aux lois pour conclure qu'elles ne peuvent décrire le monde dans sa totalité et que leur application, et donc interprétation, par des *fonctionnaires* est alors nécessaire. En 1588, Abraham Faunce [Bou99] écrivait un texte sur la logique du droit dont il représentait les éléments sous forme d'arbre binaire. Leibniz ensuite introduisit le rapprochement entre le raisonnement juridique et la logique mathématique. Ces idées ont ressurgi par la suite au XX siècle, avec la naissance de la cybernétique.

Tous ces modèles révèlent les réflexions théoriques sur le droit qui sont le préalable au développement de tout système informatique juridique. Ainsi, l'élaboration d'une ontologie du droit quelle qu'elle soit est porteuse d'une certaine *vision*, ou *conceptualisation*, du droit.

Nous présentons dans la Section 3.1 les particularités théoriques et pratiques du droit dont tout ingénieur des connaissances traitant de la matière doit avoir conscience. Nous présentons ensuite les principales ontologies du droit existantes dans la Section 3.2. Enfin, sachant qu'une ontologie du droit rassemble les objets juridiques et que ces objets sont portés par des termes, nous présentons dans la Section 3.3 des éléments importants de linguistique juridique.

3.1 Particularités juridiques

Une démarche relevant de l'informatique juridique implique une approche à la fois épistémologique et théorique sur le droit ([BHR00]). Nous notons cependant que l'importance de cette approche théorique varie en fonction de l'objectif poursuivi. Il est clair que le développement d'un système expert d'aide à la décision ne soulève pas les mêmes questions théoriques sur le droit que le développement d'une ontologie documentaire du droit. Cependant, la question majeure soulevée est celle-ci : qu'est-ce que le droit ?

Nous évoquons tout d'abord (Section 3.1.1) les caractéristiques théoriques et philosophiques de la matière juridique qui sont abordées lors du développement d'ontologies ou, plus largement, lors de modélisations du droit. Dans un deuxième temps (Section 3.1.2), un point de vue plus pratique est envisagé.

3.1.1 Caractéristiques théoriques de la discipline juridique

L'ontologie en philosophie traite de la nature de l'existence et tente de déterminer l'existence réelle des objets, indépendamment de leur perception par un agent cognitif. Elle se distingue alors de l'épistémologie qui traite des relations entre les agents cognitifs et les objets de la connaissance, la réalité. Elle se distingue également de la sémiologie qui étudie les systèmes symboliques et leurs significations. Une vision ontologique sur un domaine des sciences humaine comme le droit est particulière, comparée à une vision ontologique sur un domaine purement scientifique. Dans les disciplines appelées les *sciences exactes*, la connaissance consiste à faire un ensemble d'observations sur le monde et à construire des théories sur cette réalité, tentant de l'expliquer. Ces théories sont des constructions. Dès lors des interrogations émergent : les observations sont-elles correctes ?, les théories scientifiques ont-elles réellement à voir avec la réalité ? etc. Cependant, si un point de vue relevant du *réalisme philosophique* est pris, combiné à une vision épistémologique, un ensemble de vérités peuvent être établies sur la réalité du monde. Ces points de vues sont alors de solides fondements à la pratique des *sciences exactes*. Dans le cas où le terrain d'investigation concerne des institutions humaines, la réalité sociale, la situation est différente [MSO97]. En droit, ce qu'on peut assimiler aux *théories scientifiques* traite d'une réalité construite par les hommes eux-mêmes, par les agents cognitifs. Cette réalité n'est donc pas indépendante de leur perception. En conséquence, dans cette perspective, les points de vue ontologiques et épistémologiques sont difficiles à différencier [MSO97].

Quoi qu'il en soit, la rencontre des notions d'ontologie et de droit ne peut faire abstraction de deux controverses majeures de la philosophie juridique [MSO97] : le débat entre le droit naturel et le positivisme et la question de savoir si le droit énonce ce qui devrait être ou ce qui est en tant que tel.

(*i*) En ce qui concerne la première controverse, si le droit est une pure construction de l'homme, son existence peut être affirmée en tant que telle, et des théories explicatives fondées dessus. Cependant, pour adopter un point de vue proprement ontologique sur le droit, la question de savoir si le droit ou les normes que nous avons créés ne l'ont pas été de façon trop arbitraire doit trouver une réponse. En effet, l'existence du droit dépend largement de l'observation par l'homme [MSO97]. Dans la doctrine du droit naturel, une partie du droit est une chose, existe en tant que telle ; c'est le droit naturel. Le droit positif, les normes juridiques édictées par les hommes, sont alors des projections plus ou moins parfaites de ce droit naturel. La doctrine du droit naturel se rapproche alors du *réalisme* philosophique. Les partisans du positivisme considèrent, eux, que les normes juridiques sont des pures constructions. Cette position relève du *relativisme* philosophique. Dans chacune de ces deux positions, les normes positives, comme les lois émanant du Parlement français, sont des artefacts. On peut affirmer leur existence et adopter un point de vue ontologique à leur égard. Cependant, aucun critère n'est disponible pour affirmer l'existence et justifier la véracité des classifications faites par ces normes juridiques (les *personnes*, les *biens*, les *institutions* etc). Ces classifications étant les éléments fondamentaux de l'ontologie philosophique et de l'ontologie informatique, le problème majeur de la notion d'ontologie du droit en informatique est mis à jour. Les auteurs ([MSO97]) notent dès lors qu'on ne devrait pas parler d'*ontologie du droit* dans la sphère informatique mais simplement de *schémas conceptuels de concepts juridiques*[1] relevant uniquement de points de vue épistémologiques.

(*ii*) La deuxième controverse concerne la question de savoir si le droit

[1] *conceptual schema of legal concepts*

établit ce qui devrait être ou ce qui est. Souvent, le droit est envisagé comme édictant les choses de la façon dont elles devraient être. Dès lors, on ne traite pas d'ontologie, mais d'éthique. Cependant, les normes juridiques peuvent également être appréhendées comme dérivant de ce que nous observons du monde. Les choses, les personnes, les institutions évoquées dans les normes étant également observées dans la réalité. Un point de vue ontologique est alors parfaitement envisageable, tant sur l'existence des choses que sur l'existence des catégories juridiques.

Traiter des ontologies du droit doit se faire en pleine conscience de ces problématiques propres à la philosophie juridique. La question *qu'est-ce que le droit ?* doit trouver une réponse. Nous adoptons à leur égard un point de vue proprement pragmatique considérant les normes juridiques comme existant en tant que telles. De la même façon, les objets et les catégories créées par elles sont supposés existant en tant que tels, dès lors qu'ils sont édictés par une norme. Ces prises de position sont les fondements théoriques essentiels de notre démarche qui consiste à dériver une ontologie de textes normatifs du droit français.

Pour certains, et [SL97] en particulier, le fondement des ontologies du droit doit se trouver dans la théorie juridique. A ce titre, la *théorie pure du droit* de H. Kelsen peut servir de base au développement d'une ontologie du droit intégrée dans un système informatique. Les ontologies traitent des conceptualisations des objets, appelés respectivement *idées* et *réalisations* chez Kelsen. Kelsen distingue ainsi dans sa théorie *être* (*sein*) et *devoir être* (*sollen*). La première notion (les idées) regroupe les substrats factuels d'un acte (un papier, un comportement humain) et les résultats interprétatifs de l'acte (le contenu même du texte par exemple). La deuxième notion (les réalisations ou concepts) regroupe les qualifications juridiques de l'acte. Chez Kelsen, le fondement du système juridique est la *norme basique* (*Grundnorm*) qui fonde la validité juridique des normes, par exemple le fait que la signature apposée en bas d'un document le transforme de texte quelconque en contrat valide. Ces quelques éléments de la *théorie pure du droit* de H. Kelsen, avec d'autres (voir [SL97], peuvent alors servir de fondements théoriques à l'élaboration d'ontologies du droit[2].

Le droit peut également être envisagé, dans des tentatives de modélisations [HV99], comme un système dynamique d'états de fait connectés entre

[2]Les ontologies par là envisagées sont alors ce que nous appelons des ontologies *méta*, ou *top ontologies*. Elles envisagent l'ensemble du domaine juridique pour en porter une conceptualisation globale. Nos travaux se situent en deçà de ces considérations générales.

eux par des règles et des évènements[3]. Les *états de fait* sont des éléments comme le fait qu'un contrat en question a été conclu entre des parties. Un *événement* modifie l'*état de fait*, par exemple, la signature en elle-même du contrat. Enfin, les *règles* font la connexion entre les *états de fait*. Par exemple, la règle selon laquelle la signature d'un contrat porte la naissance d'obligations entre les parties au contrat. Dès lors, plusieurs caractéristiques du droit peuvent être modélisées. Une règle par exemple est constituée de conditions et de conclusions. Les conditions ainsi que les conclusions sont des états de fait. Les auteurs donnent comme exemple la règle selon laquelle tout voleur doit être puni. Les règles peuvent alors être formulées logiquement selon la forme *si ..., alors ...*. Néanmoins, les caractéristiques du système juridique font qu'il n'y a aucune garantie que la satisfaction des conditions d'une règle donnée entraîne la réalisation de ses conclusions. Ce point, appelé *defeasibility*, est la conséquence, par exemple, d'une contradiction entre règles elles-mêmes (par exemple la règle selon laquelle tout criminel doit être puni peut se trouver en contradiction avec la règle selon laquelle les personnes non conscientes de leurs actes ne peuvent être considérées comme responsables de ceux-ci).

Autre caractéristique du droit relevée par les auteurs ([HV99]), le fait qu'un des premiers raisonnements juridiques consiste à procéder à des classifications. Nous retrouvons une des problématiques majeures des ontologies. Cette classification consiste à attribuer à une situation factuelle donnée les conditions telles que définies dans les règles. La classification relève alors de la qualification juridique : à une situation de fait est associé un statut juridique particulier. Cette classification consiste soit à subsumer un objet concret d'une catégorie abstraite (établir, par exemple, si tel moyen motorisé de locomotion peut être envisagé comme un *véhicule terrestre à moteur* tel que défini par la règle), soit à établir un lien de causalité entre une situation factuelle et une qualification (par exemple, dans le cas de la responsabilité civile délictuelle, doit être établi le lien de causalité entre les faits en question et le dommage constaté).

Enfin, le système juridique inclut des règles plus générales que d'autres dans le sens où elles dictent la validité même de ces dernières. Ainsi, une norme ne sera *valide* que si la personne qui la signe est légitimée à le faire, cette légitimité étant conférée par une norme. Ces règles sont alors des règles supérieures aux autres et leur prévalant. Dans le même ordre d'idées, les systèmes juridiques incluent des normes plus ou moins générales, celles-ci

[3][...] *law is a dynamic system of states of affairs which are connected by means of rules and events.*

s'organisant entre elles dans ce qui est appelé la *hiérarchie des normes*. En haut de cette hiérarchie, symbolisée en pyramide, se trouvent les traités internationaux, puis les normes constitutionnelles, les lois, les décrets, les arrêtés etc. Ces points de vues sur la matière juridique portent les prémisses d'une conception ontologique du droit. La *théorie générale des normes* de H. Kelsen est de ceux-là.

Des caractéristiques théoriques de la discipline juridique conduisant notamment à des modélisations de la matière sont déduites des caractéristiques pratiques, en particulier quant à l'informatisation du droit.

3.1.2 Caractéristiques pratiques de la discipline juridique

Pour ce qui est des bases documentaires, le droit est un domaine qui a eu un rôle moteur à la fois dans les systèmes à indexation ou en texte intégral. En témoigne la création en 1970 du CEDIJ (Centre de Recherche et de Développement en Informatique Juridique) sous l'édige de la Chancellerie qui deviendra le CNIJ (Centre National d'Informatique Juridique) en 1984. Ces centres initieront la création de bases documentaires juridiques et promouvront leur accès par voies télématiques. Autre illustration de cette importance du droit dans l'informatique documentaire, les travaux de Andreewsky et al. (voir notamment [ABD$^+$81]) portant sur des bases juridiques ou les travaux de l'IRETIJ (Institut de Recherche et d'Etudes pour le Traitement de l'Information Juridique).

Par ailleurs, dès les débuts de la recherche sur les systèmes experts informatiques, l'intérêt pour la matière juridique a été fort. Le présupposé était, et demeure parfois encore, le suivant : le droit est une discipline dont les connaissances et les raisonnements sur ces connaissances sont tels qu'ils peuvent être facilement automatisés. De façon schématique, le droit peut en effet être appréhendé comme un ensemble de règles applicables à un ensemble de faits. Ces règles sont telles qu'on peut les traduire en propositions logiques et les inclure dans un système informatique. Dès lors, une explicitation suffisante des faits suffit à leur appliquer les règles et, par là même, à produire un système d'aide à la décision. On a, par là, élaboré des systèmes informatiques basés sur les règles[4].

Un tel système est, par exemple, incorporé dans le démonstrateur MILE

[4] *rule-based reasoning systems*

élément du projet CLIME[5]. MILE ([Win00]) est un système d'informations maritimes. Il incorpore toutes les réglementations afférentes à la classification des bateaux et à leur contrôle. Un des modules de MILE est chargé de la tâche d'*évaluation normative*. Celle-ci consiste à *évaluer* un cas donné, par exemple un type particulier de tanker présentant certaines caractéristiques à la lumière d'un groupe de normes afin de déterminer si ce cas est conforme. Le système MILE incorpore un ensemble de normes, les normes maritimes relatives à certains navires, sous la forme d'objets structurés. Pratiquement, les normes sont divisés en objets texte ayant un nom et un contenu ; ainsi, la plupart des sources juridiques sont disponibles au format SGML. Le processus d'*évaluation* des cas va tout d'abord consister en une *abstraction* : l'affaire va être décrite de façon structurée en éliminant les informations inutiles. L'enjeu de l'évaluation réside dans la *généralisation* du cas particulier. En effet, les normes se réfèrent rarement à des situations spécifiques comme le cas du navire X, propriété de Z, mais plutôt à des situations générales comme un type W de tanker. Les termes du cas en question sont donc transformés en termes des conditions d'application de la norme. Dès lors, un ensemble de normes est dégagé, applicable au cas en question. Des problèmes peuvent alors survenir comme le résultat de conflits de normes. Le système incorpore donc un module de résolution de conflits de normes fondé sur des connaissances méta-juridiques, ce qui permet de produire une qualification définitive pour le cas en question.

Dans un autre point de vue, le droit peut être appréhendé comme un ensemble de cas. Une description appropriée de ces cas et leur comparaison permet de raisonner sur ceux-ci. Ont alors été construits des systèmes de raisonnement à partir de cas[6]. Ces systèmes sont particulièrement appropriés dans l'étude de la jurisprudence, et plus particulièrement dans les pays de système juridique anglo-saxon où le précédent jurisprudentiel a une grande importance. Les principes fondamentaux de ces systèmes sont la représentation des différents aspects des cas, l'expressivité de ces représentations ainsi que les comparaisons entre ces représentations. Ces problématiques sont largement traitées dans les domaines relatifs à l'informatique juridique, ou en anglais, *artificial intelligence and law*[7].

[5]CLIME, *Computerized Legal Information Management and Explanation*, est un projet européen ESPRIT lancé en 1998 ayant pour objectif général d'améliorer l'accès et la récupération de l'information juridique.

[6]*case-based reasoning systems*

[7]Nous renvoyons aux actes des conférences ICAIL (*International Conference on Artificial Intelligence and Law*), notamment [Pra01] et JURIX (*International Conference on Legal Knowledge and Information Systems*), notamment [VLLM01].

Autre dimension du droit qui fait l'objet de tentatives d'automatisation : l'argumentation juridique ou plus généralement le raisonnement juridique. Nous ne nous attardons pas sur cet aspect qui s'éloigne de nos propres préoccupations[8]. Notre ontologie n'entend en effet pas contenir les normes en tant que telles et les raisonnements qu'elles peuvent mener, mais simplement les objets du droit par elles créés. Ces objets sont cependant reliés entre eux, et des rôles juridiques peuvent être identifiés entre eux. Ces *rôles* ne sont cependant pas comparables avec ceux représentés dans des systèmes décisionnels [Bou99].

Les réflexions théoriques sur la discipline juridique, comme celles que nous exposons plus haut, ont cependant permis de mettre au jour la complexité de tout système juridique. Une telle complexité ne peut être ignorée dans le cadre d'un projet d'automatisation, et, par exemple, de système d'aide à la décision. L'automatisation du droit n'est donc pas tâche plus aisée que pour une autre discipline. Tout système informatique juridique doit inclure des connaissances de la discipline, ces connaissances étant rassemblées dans des ontologies, ou plus généralement, dans des modélisations abstraites du domaine.

Nous constatons qu'une des caractéristiques principales des ontologies, celle de classer les objets dans des catégories, et par là même de permettre de généraliser les raisonnements, rejoint une des principales activités juridiques, la qualification. La qualication consiste en droit à attribuer à une situation de fait une catégorie juridique préétablie, c'est-à-dire exprimée dans la norme[9]. L'ingénieur de la connaissance procédant à l'élaboration d'une ontologie fait dès lors le même type de raisonnement que le juriste faisant de l'analyse juridique. Les catégories définies dans une ontologie du droit ne sont alors ni plus ni moins que les catégories définies par le droit lui-même (les *biens*, les *personnes*, les *véhicules terrestres à moteur* ou les *établissements publics*). Nous adoptons alors, dans notre démarche de construction d'ontologie du droit, un point de vue relevant du *réalisme* philosophique[10] et justifions par là même notre double rôle d'ingénieur de la connaissance chargé de l'élaboration de l'ontologie et d'expert du domaine, juriste capable de valider cette même ontologie.

Nous utilisons, dans notre démarche d'élaboration d'une ontologie du

[8] *ibid.*

[9] Cette tâche de qualification est la fonction principale incombant au juge.

[10] Voir [Mom98] citant les travaux de Devitt et de Searle

droit[11], un corpus rassemblant les codes existant en droit français. Les codes du droit français sont le résultat d'un processus appelé codification. La codification consiste à regrouper des textes dans des ensembles cohérents [Bra99]. Les textes concernés sont les textes normatifs et en aucun cas des jurisprudences[12]. La codification peut être distinguée selon qu'elle réforme le droit, c'est alors le cas des grands codes de Napoléon, le Code civil et le Code pénal, qui ont modifié les règles préexistantes et la codification qui reforme, comme celle à laquelle nous assistons depuis la fin des années 1980 et qui est le fait de la Commission supérieure de codification. Le rôle de cette dernière est de rassembler toutes les normes, législatives et réglementaires, pour un thème donné et à les rationaliser, sans les modifier, dans un seul et même recueil, le code. D'après G. Braibant, président de la Commission supérieure de codification, environ la moitié du droit français serait codifié. Les codes sont donc des recueils thématiques de normes applicables. Le choix de ces différentes thématiques constitue une sorte de conceptualisation du droit dans le sens où celui-ci est alors appréhendé comme un ensemble se divisant entre les divers thèmes couverts par les codes. Le choix des différents thèmes faisant l'objet d'un code est le fait des membres de la Commission. Ce choix est foncièrement politique, et du fait que c'est au Parlement d'entériner tout nouveau code, il peut faire l'objet de revirements et de vives oppositions. Quoi qu'il en soit, en choisissant les grandes thématiques du droit, le codificateur procède en quelque sorte à une modélisation du domaine. De la même façon, le codificateur structure les codes en Parties, Chapitres, Sections et Sous-sections, chacun faisant l'objet d'une thématique particulière. Ceci participe de la modélisation de la matière. En utilisant ces structurations thématiques du droit dans notre démarche d'élaboration d'une ontologie du droit, et en les exploitant, nous nous appuyons de façon plus ou moins explicite sur la modélisation du domaine qu'elles incarnent. Cette modélisation est une manière de voir le droit, il y en a bien d'autres, mais celle-ci a l'avantage indéniable d'être validée par le législateur lui-même, le Parlement. Notre ontologie est donc porteuse de cette *vision* du droit incarnée dans les codes.

[11]Ce point fait l'objet de la Section 4.

[12]Les versions d'éditeurs privés de codes comportent pourtant des dispositions jurisprudentielles. Ceci est cependant dû au fait que ces codes sont *commentés*, c'est-à-dire que chacune des dispositions codifiées fait l'objet d'un développement notamment dans ses applications par le juge.

3.2 Les ontologies juridiques existantes

Plusieurs ontologies juridiques, ou plus généralement des conceptualisations du domaine juridique, ont été développées depuis le milieu des années 1990. Ces ontologies sont de différents niveaux, et principalement de niveau *méta* : ontologies génériques, *top ontologies* et *core ontologies* comme l'ontologie fonctionnelle de Valente [VB94]. D'autres sont sous-jacentes à la définition de langages de représentation de la réalité juridique comme le *Language for Legal Discourse* de McCarty ([McC89]) ou le langage LEGOL[13] et le formalisme NORMA[14] développés par Stamper ([Sta80], [Sta91]). Ces ontologies ne sont pas considérées dans cette section, du fait des trop grandes différences d'approches avec nos propres travaux qu'elles présentent. Une étude comparative a été menée entre certaines de ces ontologies dans [VBC98].

Nous décrivons l'ontologie fonctionnelle du droit de Valente et l'ontologie de Van Kralingen et Visser, sachant que notre point de vue à leur égard a été d'envisager, ou non, leur réutilisation. Notre objet est en effet bien d'élaborer une ontologie du domaine juridique, et ce dans un but bien précis, celui de l'inclure dans un système de recherche d'informations juridiques sur le Web. Nous ne développons donc pas une ontologie générique mais bien une ontologie *applicative*. Dès lors, une ontologie générique peut nous être utile dans la phase d'élaboration de notre ontologie applicative pour la définition des catégories et concepts supérieurs. Chacune des ontologies décrites cidessous a été développée dans le but d'être intégrée dans des systèmes à base de connaissances juridiques. Toutes ont été spécifiées dans des langages formels ou semi-formels.

3.2.1 L'ontologie fonctionnelle du droit de Valente

L'ontologie fonctionnelle du droit de Valente [VB94], ou *functional ontology of law* (FOLaw), est qualifiée de telle du fait qu'elle est fondée sur une perspective fonctionnelle du système juridique. Le système juridique est en effet appréhendé comme un instrument modifiant et influençant la société en fonction d'objectifs sociaux[15]. Sa principale fonction est ainsi de réagir aux comportements sociaux.

---

[13] *Legally Oriented Language*

[14] Signifiant *logique de normes et permissions.*

[15] [VBC98] *The legal system is considered an instrument to change or influence society in specific drections, determined by social goals.*

FOLaw comporte un ensemble de catégories primitives et sous-catégories des connaissances juridiques, chacune de ces connaissances correspondant à une façon de réagir aux comportements sociaux. Ces catégories portent une interprétation de la connaissance juridique. FOLaw spécifie également les structures de ces catégories ainsi que leurs interrelations.

FOLaw est une ontologie de niveau supérieur, c'est une *core ontology* dans le sens où elle définit les catégories supérieures du droit.

Les principales catégories des connaissances juridiques dans la FOLaw sont les suivantes, les trois premières étant les plus importantes :

les connaissances normatives[16] . Elles définissent, de manière générale, des standards, elles prescrivent les comportement des personnes vivant en société. Les normes portent en effet la définition d'une sorte de monde idéal auquel le monde réel est comparé. Les normes peuvent être distinguées[17] selon qu'elles sont des normes de commandement[18], des normes dérogatoires[19] ou des normes conférant pouvoir[20]. Cette distinction se fait selon les fonctions portées par ces différentes normes dans le système juridique. Les normes de commandement sont celles au cœur du système juridique : elles prescrivent un certain comportement social, que ce soit l'obligation de faire ou l'interdiction de faire. Les normes conférant pouvoir font référence au pouvoir normatif dont peuvent être pourvus des agents. Ces normes ont pour fonction de conférer le pouvoir normatif à certains agents, comme le Parlement par exemple. Enfin, les normes dérogatoires sont des normes sur les normes dans le sens où elles agissent sur et se réfèrent à d'autres normes, par exemple, une loi en annulant de précédentes.

les connaissances de responsabilité[21] . Elles définissent les conditions de la responsabilité d'un agent du fait, ou non, de son comportement qui seules justifient l'établissement de la culpabilité de cet agent. Ces connaissances comprennent tous les raisonnements juridiques qui attribuent ou limitent la responsabilité d'un agent au regard d'une ou plusieurs situations de fait ; ainsi, les principes de la responsabilité civile délictuelle du droit français, mais aussi ceux de la responsabilité du fait des choses sont inclus dans les connaissances de responsabilité. Les connaissances de responsabilité connectent ce que les auteurs appellent la *responsabilité causale* et la *responsabilité juridique*. La première permet d'établir un lien de cause à effet entre une série de faits (un com-

[17] en suivant une conception simplifiée de la dernière typologie des normes de Kelsen
[18] *commanding norms*
[19] *derogative norms*
[20] *empowering norms*

portement et un dommage par exemple) et la seconde de déterminer qu'un agent a dérogé à la norme. Cette catégorie des connaissances juridique est liée à une sous-catégorie de la catégorie des connaissances du monde, celle des connaissances causales.

les connaissances du monde[22] . Les connaissances du monde sont les connaissances juridiques permettant de décrire le monde objet des régulations. Ces connaissances sont celles utilisées dans la loi pour décrire le monde qu'elle régule ; elles comportent donc les concepts juridiques et les relations entre ces concepts. Ces connaissances sont principalement des connaissances *terminologiques* dans le sens où leur objet est d'établir quel concept juridique correspond à telle entité du monde. Cependant, les auteurs notent qu'une des principales difficultés réside dans le fait que ces concepts juridiques ne sont définis que de façon imprécise et large, laissant toute latitude à l'interprétation, notamment à celle du juge[23]. Ce processus d'interprétation concerne la relation classe-sous-classe (savoir si les *bicyclettes* sont des *véhicules*) et la relation classe-instance (savoir si tel engin muni de deux roues et d'un pédalier est une *bicyclette*). Les auteurs notent également que les connaissances du monde exprimées dans les lois donnent une large part, du fait des définitions relativement vagues qui sont faites, à des connaissances de sens commun[24] comme celles exprimées dans des ontologies généralistes comme CYC par exemple[25]. Les connaissances du monde comprennent également ce que les auteurs appellent les *connaissances causales* qui font le lien entre des agents et des états de fait.

les connaissances réactives[26] **et les connaissances créatives**[27] . Ces connaissances sont spécifiées par les auteurs qui notent cependant que leur importance demeure toute relative au fait qu'elles ne prennent pas part au raisonnement juridique. Ces connaissances ont pourtant un rôle important dans la structure et le fonctionnement du système juridique. Les connaissances réactives définissent les réactions, en général des punitions, prévalant au constat de l'illégalité d'une situation au regard des connaissances normatives et à l'établissement de la responsabilité d'un agent, conformément aux connaissances de responsabilité. Les connaissances créatives sont celles qui permettent au législateur de créer une nouvelle entité qui n'existait pas préalablement dans le monde. Les au-

[23]Les concepts juridiques ont une *texture ouverte*, référence faite aux *open-texture legal concepts* de Hart (voir [VB94] citant les travaux de Hart, dont son ouvrage *The Concept of Law* [Har61]).

[24]*commonsense knowledge*

[25]CYC est partiellement décrite dans la Section 1.1.1

teurs donnent comme exemple la création d'un nouveau département ministériel.

enfin, les connaissances méta-juridiques concernent les relations s'établissant entre les normes elles-mêmes : ainsi les conditions d'application des normes, les règles permettant d'ordonnancer les normes entre elles et, par là, de régler les conflits de normes etc.

L'ensemble de ces connaissances juridiques permet de conceptualiser le système juridique, ce système juridique étant envisagé par les auteurs dans son principal objectif de faire en sorte que le monde soit le plus conforme possible au monde idéal décrit par les connaissances normatives. Les tâches du système juridique sont alors de contrôler cette conformité et de réguler par là les comportements sociaux. Ces tâches sont elles-mêmes divisées en sous-tâches, le tout constituant, selon les auteurs, une ontologie du domaine juridique.

Pratiquement, les auteurs envisagent plusieurs usages à leur ontologie, comme la résolution de problème, l'élaboration de systèmes à base de connaissances juridiques. FOLaw est aussi vue comme un moyen de représenter les sources juridiques et comme pouvant être intégrée dans un système de base de données juridiques procédant à des recherches d'information au niveau conceptuel. La principale utilisation que nous connaissons de la FOLaw de Valente est son intégration dans un système interactif d'apprentissage PROSA, un système qui aide les étudiants à résoudre des questions (ou cas) relevant du droit administratif néerlandais [BM99]. Un tel système permet de mettre l'accent non plus tant sur l'acquisition mais sur la compréhension de la matière. Les ontologies, et en particulier les *core ontologies* comme la FOLAw, permettent de donner une vision d'ensemble sur la matière, et ce même si le système d'apprentissage en question ne concerne qu'un aspect particulier de la matière. Elles permettent également de refléter, par leur structure même, les structures de raisonnement propres au domaine.

Les connaissances telles que décrites plus haut sont donc les éléments de l'ontologie sur lesquels des raisonnements pourront être réalisés.

L'ontologie de Valente se présente donc comme une ontologie d'un niveau supérieur de celle que nous élaborons. Notre point de vue à son égard est d'établir dans quelle mesure nous pouvons la réutiliser. Notre objet étant d'élaborer une ontologie du droit dédiée à la recherche d'informations juridiques sur le Web, nous traitons des connaissances juridiques identifiées par Valente comme les *connaissances du monde*. Celles-ci servent en effet à décrire le monde régulé par le droit et comporte, à cet égard, les concepts juridiques. Or, c'est bien à ces éléments là de la connaissance juridique que nous nous intéressons, indépendamment des types de raisonnement pouvant prévaloir avec les autres éléments de connaissance. Dès lors, une réutilisation en tant que telle de l'ontologie de Valente n'est pas pertinente. Nous nous contentons simplement de nous positionner vis-à-vis d'elle.

3.2.2 L'ontologie de Van Kralingen et Visser

Une autre ontologie bien connue du domaine juridique est l'ontologie développée par Van Kralingen et Visser. Plus connue sous le nom de FBO[28], elle est souvent envisagée comme une seule et même ontologie ([VBC98], [HV99] et même [VKVBCVDH99]) alors qu'elle regroupe en réalité une ontologie conceptuelle développée par Van Kralingen et une ontologie formelle développée par Visser. L'ontologie globale est proposée dans le cadre du développement de systèmes à base de connaissances juridiques. Elle comporte deux approches : une proprement ontologique (*legal ontologie*) qui se veut générale, dont les éléments sont envisageables dans chacun des sous domaines du droit, et une autre (*statute-specific ontology*) comportant des éléments spécifiques à certains sous domaines du droit. Dans la première, les connaissances juridiques sont divisées entre les normes, les actes et les descriptions de concepts, chacun de ces éléments étant commun à tous les sous domaines du droit. Chacun de ces éléments est défini dans l'ontologie par un modèle appelé *structure de cadre*[29] qui établit la liste de tous les attributs relatifs aux éléments. Les normes sont vues comme les règles générales, les standards et les principes s'appliquant aux sujets de droit. Les *normes* sont définies dans l'ontologie à l'aide de plusieurs éléments : un identifiant de la norme, son type, son objet, ses conditions d'application etc. Les *actes* représentent les aspects dynamiques modifiant les états du monde. Ils regroupent les événements, changements instantanés entre deux états et les processus qui ont une durée. Les actes sont définis dans l'ontologie par un ensemble d'attributs : un identifiant, une source, un objet, des aspects temporels d'application de l'acte etc. Enfin, les *descriptions des concepts* traitent de la signification des concepts du domaine. Elles permettent de les définir, d'énoncer leurs conditions nécessaires et suffisantes etc. Les descriptions de concepts comportent alors les éléments suivants : le concept en lui-même, son type, son objet, les conditions dans lesquelles il est applicable etc. La seconde approche (*statute-specific ontology*) comporte les éléments spécifiques complétant ceux de la première approche. Cette ontologie doit être, selon les auteurs, créée pour

[28] *frame-based ontology*
[29] *frame-based model*

chaque sous domaine juridique en question. Elle comporte les éléments nécessaires à l'interprétation de ce sous domaine, par exemple, la distinction entre le fait que l'agent en question soit un homme ou une femme qui peut être importante dans le cas d'une loi spécifique sur les conditions de travail. Cette ontologie inclut donc les éléments de vocabulaire spécifiques au sous domaine en question.

De la même façon que la FOLaw de Valente, l'ontologie proposée par Van Kralingen et Visser adopte un point de vue général sur le domaine juridique. Notre objet est bien plus spécifique, celui d'appréhender les concepts tels que définis par le droit, quel qu'en soit le sous domaine. Notre ontologie n'est de ce point de vie ni générale, ni spécifique, mais plutôt transversale puisqu'elle propose d'embrasser les concepts relevant de divers sous domaines du droit. Nous nous rapprochons en partie de la (*statute-specific ontology*) de Van Kralingen et Visser tout en n'en adoptant pas tous les principes puisque l'ensemble des connaissances sur les concepts qui nous aurons pu appréhender font l'objet d'une seule et même ontologie. L'ontologie proposée par Van Kralingen et Visser ne peut donc être réutilisée dans notre perspective.

3.3 Linguistique juridique

Notre démarche de construction d'ontologie se fonde donc essentiellement sur des textes pour en extraire les éléments de notre ontologie, les termes du domaine juridique. Dès lors, en amorce de la présentation de notre démarche, des précisions doivent être données sur une discipline se revendiquant en tant que telle : la linguistique juridique. Cette discipline fait l'objet d'un ouvrage de G. Cornu ([Cor00]). Nous en décrivons dans cette section les principales caractéristiques en reprenant le déroulement de l'ouvrage lui-même : le vocabulaire juridique (Section 3.3.1) et les discours du droit, plus principalement celui qui nous intéresse, le discours législatif (Section 3.3.2).

3.3.1 Le vocabulaire juridique

Le langage du droit est, note l'auteur, un langage de spécialité au sein de la langue commune. Simplement, son entendement nécessite plus que la maîtrise de la langue commune. Ainsi, la communication du droit se heurte à ce que l'auteur appelle un *écran linguistique*. Nous nous plaçons dans cette perspective pour avancer que l'objet de notre système de recherche d'informations sur le Web incluant une ontologie du droit français tente, à son échelle et dans la mesure de ses capacités, de réduire cet *écran linguistique*.

Cet objectif se couple et se place dans la continuité d'un mouvement de mise à disposition du droit à tous sur le Web. Ce mouvement, œuvre de quelques irréductibles acteurs du Web dans ses débuts, a été officialisé en 2000 par le Premier Ministre par la création du *service public d'accès au droit* dont l'objet principal est de mettre à disposition l'ensemble des sources du droit pouvant être utile au citoyen de façon libre sur le Web.

Notre ontologie est élaborée dans le but d'être intégrée à un système d'aide à la reformulation de requête, qui permette aux usagers du Web d'accéder plus facilement aux ressources juridiques du Web. Pour élaborer cette ontologie, notre travail s'oriente principalement sur les termes juridiques. C'est à ceux-ci que nous nous intéressons plus particulièrement dans cette section. Certains termes de la langue française n'ont de sens que juridique comme *antichrèse*, *nantissement* ou *synallagmatique*. La plupart sont cependant le siège d'une sorte de polysémie : le même terme ayant un sens donné, voire plusieurs, dans la langue commune et un autre, voire plusieurs, souvent plus précis, dans le langage du droit. Le vocabulaire juridique français est ainsi défini comme *l'ensemble des termes de la langue française qui reçoivent du droit un ou plusieurs sens*. Il regroupe d'après l'auteur un ensemble d'environ[30] 10 000 termes. Parmi ces termes, une partie ont leur sens primordial dans le domaine juridique, ils ont en quelque sorte été empruntés par le langage commun du fait de leur usage, de leur assise sociale. Leur sens a alors dérivé, de façon plus ou moins forte, du sens juridique initial. On peut citer les termes *justice*, *permission*, *légitime*, *transaction* ou *tutelle*. L'auteur note que ces termes, porteurs de notions fondamentales du droit, se trouvent au cœur du système juridique dont ils expriment les catégories et les forces essentielles. Un autre ensemble notable de termes juridiques sont porteurs d'une rupture de sens entre le sens juridique et le sens extrajuridique. On peut citer *distraction* qui désigne en droit un prélèvement, un recouvrement direct s'agissant des dépens, de même le terme *minute* qui désigne dans son sens juridique l'original d'un acte authentique demeurant dans les archives de l'autorité en étant dépositaire. Des phénomènes de polysémie interne sont également observables dans le vocabulaire juridique, lorsque deux termes homographes se trouvent avoir des sens juridiques différents. L'auteur note que ce phénomène de polysémie interne n'est aucunement un accident du langage du droit mais bien plus une marque essentielle du vocabulaire juridique. Ce sont les termes les plus représentatifs du langage qui présentent le plus ces aspects polysémiques (par exemple, *conseil* qui désigne tout à la fois un avis

[30]G. Cornu note que cette évaluation n'est qu'une estimation. Elle se fonde sur le comptage des termes et sous termes définis par lui dans son ouvrage Vocabulaire Juridique [Cor01].

sur ce qu'il convient de faire, la personne qui donne cet avis, l'assemblée de personnes chargées de donner un avis, le lieu où siège l'organe consultatif ou la séance même de consultation).

Ces phénomènes de polysémie, qu'ils soient internes (plusieurs sens juridiques) ou externes (plusieurs sens juridiques et extrajuridiques), nous concernent au premier chef, ils sont même un des fondements de l'élaboration de notre ontologie. L'idée sous-jacente est que, plus une requête comporte des termes précis, plus sont évincés ces problèmes de polysémie. D'un autre point de vue, en procédant à l'élaboration même de cette ontologie, nous faisons face à ces phénomènes de polysémie, que nous devons envisager et présenter à l'utilisateur final. Ces points sont précisés dans les Parties 2 et 3 de cette thèse.

Le langage juridique, et son vocabulaire, est donc le vecteur d'expression du droit. A ce titre, le vocabulaire juridique évolue, principalement au travers des néologismes. Ainsi, l'influence prégnante du droit communautaire dans le droit national est sensible au travers de nouveaux termes apparaissant dans le vocabulaire juridique, porteurs de nouvelles notions, comme *producteur de bases de données* par exemple. De la même façon, l'évolution du droit régulant les rapports entre enfants et parents a fait passer du terme *puissance paternelle* au terme *autorité parentale*. Les termes juridiques sont alors chargés de tout leur sens. Cette évolution du langage juridique doit être considérée dans le développement d'une ontologie telle que nous l'envisageons. Les nouveaux termes apparaissant, comme *pacte civil de solidarité*, par exemple, doivent être intégrés. L'ontologie que nous avons élaborée l'a été à un moment donné, en prenant en compte le droit tel qu'il était à ce moment-là[31]. Les éléments de notre ontologie sont alors tous ceux existant à ce moment-là. Un mécanisme de réactualisation de l'ontologie doit être prévu ; nous en proposons les principes généraux en fin de thèse.

Certains des termes juridiques ont ce que Cornu appelle une *valeur fondamentale*. Nous les appelons pour notre part les *termes fondamentaux du domaine*, notamment dans la Section 5.2 où nous partons à leur découverte. Ces termes sont *fondamentaux* car porteurs de catégories fondamentales du droit ([Cor00], p.133), ces catégories fondamentales étant communes à toutes les disciplines juridiques. Ainsi l'auteur distingue trois types de termes fondamentaux : ceux relevant du langage de l'établissement, ceux relevant du vocabulaire de l'action et ceux relevant du langage du dénouement, chacune

[31]Nous utilisons les codes tels que diffusés sur le site officiel Légifrance en mai 2001.

de ces notions participant de l'établissement du droit. Le droit établit, il érige. Des termes comme *norme*, *compétence*, *interdiction*, *contrat*, *don* ou *sûreté* en sont l'illustration linguistique. Le droit a des effets, matérialisés par des actes et des faits, ainsi que par des réactions. Des termes comme *infraction*, *faute*, *dommage*, *déni*, *lésion* ou *corruption* en sont le siège. Enfin, le droit tente de rétablir la paix ou du moins l'ordre. Les termes *demande*, *fin de non recevoir*, *requérant*, *procédure*, *enquête* ou *sentence* en sont l'illustration.

Le langage du droit assure une pluralité de fonctions, à chacune desquelles correspond un sous vocabulaire juridique. Le langage juridique est ainsi le langage d'élaboration de la loi, du jugement, des conventions. Il participe alors à la fonction législative, à la fonction juridictionnelle, à l'activité contractuelle, à la création doctrinale etc. Nous nous intéressons dans notre perspective d'élaboration d'une ontologie du droit, précisément au langage d'élaboration de la loi. L'attention est donc portée sur le vocabulaire législatif. Ce choix est fait non seulement d'un point de vue méthodologique, nous l'expliquons dans la Section 4, mais aussi structurel. Il nous a paru logique d'envisager dans un premier temps les concepts tels que définis par la loi. Ce point étant posé, il est certain que la considération des concepts tels que définis dans les jugements ou dans les conventions, respectivement le vocabulaire juridictionnel et le vocabulaire contractuel, peut être une perspective de nos travaux, une manière d'élargir le champ de compétence de notre ontologie.

Cornu note également que la linguistique juridique est un auxiliaire de la science fondamentale du droit, notamment par le fait qu'elle procède aux définitions des notions juridiques, catégories et concepts juridiques. Elle procède, de plus, à la recherche des relations unissant les catégories juridiques. Cette recherche s'appuie sur des études lexicales des textes juridiques. Sans prétendre participer par quelque manière que ce soit à la science fondamentale du droit, l'élaboration d'une ontologie du droit, quelle qu'elle soit, participe cependant, à sa mesure, à une meilleure connaissance du droit. Ce faisant, et par les techniques que nous utilisons, décrites dans la Partie II, nous procédons à une étude relevant, en partie, de la linguistique juridique. C'est ainsi que l'auteur note que la linguistique juridique participe aussi de l'informatique juridique, à tout le moins à cette partie de l'informatique juridique relative à la recherche d'informations et s'intéressant aux moteurs de recherche, ainsi la recherche de mots clés, l'analyse de contenus, les appels aux champs sémantiques etc. L'auteur remarque simplement que les points de vue du linguiste, de l'informaticien et du législateur, ne seront pas nécessairement les mêmes. Nous avons, dans nos travaux, adopté alternativement chacun de ces points de vue. L'objectif a cependant été celui de les concilier en un seul et unique, matérialisé par l'ontologie finale. Or concilier, c'est en partie renoncer. En conséquence, chacun des spécialistes précédemment cités[32] ne pourra retrouver la totalité de son point de vue dans cette ontologie du droit dédiée à la recherche d'information. Notre objectif ne consiste pas à tenter de satisfaire chacun de ces spécialistes dans leur vision qu'ils ont de la problématique d'une ontologie du droit. Il est simplement de tenter d'élaborer un système utile à l'usager[33] du Web recherchant des informations juridiques. Nous traitons alors avec le vocabulaire juridique, en utilisant des techniques relevant en partie de la linguistique juridique.

3.3.2 Les discours du droit

Le langage du droit réside également dans les textes qu'il produit, les textes du droit. Nous mettons de côté, dans notre perspective, l'expression orale du discours du droit. Le caractère proprement juridique du discours réside alors dans les finalités même du discours qui sont, schématiquement, la création (discours législatif) ou la réalisation du droit (discours juridictionnel). Ces finalités confèrent une certaine structure et un certain style au discours juridique et, pour de ce qui nous intéresse, aux textes juridiques.

Le premier à parler le langage du droit, et par là, à adopter un discours juridique, est le législateur. Les autres acteurs à parler le droit sont le juge, le gouvernement et l'administration, les professionnels et même les simples particuliers. Nous nous intéressons dans nos développements aux textes tels que présents dans les codes. Ces textes codifiés sont à la fois des textes législatifs pour ce qui est, schématiquement, des parties législatives des codes, et des textes réglementaires, pour ce qui est des parties décrets et arrêtés des codes.

Le rôle premier du discours législatif est l'énoncé de la règle. De cette fonction première et essentielle, naît un certain style de discours et des caractéristiques qu'il nous semble important de noter dès lors que des traitements automatiques sont prévus sur de tels textes.

Le discours législatif, de par sa fonction même, est un discours normatif, [Cor00], p. 268 :

[32]Auxquels s'ajoutent les ingénieurs de la connaissance, les cogniticiens, les juristes en général etc...

[33]L'usager du Web recouvre à la fois, dans notre problématique, le spécialiste du domaine, juriste, et le non spécialiste. Toute personne est ainsi confrontée de près ou de loin au droit, et peut en conséquence être amenée à rechercher du droit sur le Web.

> «Il porte les marques de la souveraineté, de la généralité et de la détermination de la loi.»

Ces caractéristiques sont ainsi prégnantes dans les textes de lois, dans l'énoncé même de la règle. Elles se retrouvent dans l'emploi de certains verbes (*devoir*[34], *obliger*[35], *défendre*[36]), dans l'emploi du présent de l'indicatif[37] etc.

Le discours législatif, incarné dans le texte de la loi tel que publié au Journal Officiel, est encadré de formules consacrées comme autant de marqueurs de souveraineté. Ces formules sont des formules de promulgation. Ainsi toute loi commence et se termine avec ces formules :

> «L'Assemblée Nationale et le Sénat ont adopté, le Président de la République promulgue la loi dont la teneur suit [...]»

> «La présente loi sera exécutée comme loi de l'Etat.»

Les dispositions législatives et réglementaires qui se trouvent être codifiées présentent également certaines caractéristiques de forme. Ainsi, tout article de code fait référence à la loi ou au règlement dont il est issu. Cette référence se fait par le rappel du numéro et de la date de la loi ou du règlement ainsi que de l'article en question.

Le discours législatif a cette particularité de séparer ses phrases sous forme d'articles. Ainsi, à chaque disposition correspond un article. Chacun de ces articles est numéroté. Pour ce qui est des codes, ceux qui sont le résultat des dernières campagnes de codification sont numérotés comme suit : une lettre spécifiant la partie dans laquelle se trouve l'article en question : L pour la partie législative, R pour la partie réglementaire en général, une distinction pouvant être faite entre les décrets en Conseil d'Etat (R), les décrets simples (D) et les arrêtés (A). Chaque chiffre du numéro de code correspond à une division du code dans laquelle se trouve l'article en question, le chiffre après le tiret correspondant au numéro de l'alinéa[38]. Ainsi, l'article L.241-1 du Code de la sécurité sociale se trouve être la première phrase de la Section 1 du Chapitre 4 du Livre 2 du code. Les autres codes, comme le Code civil ou le Code des impôts, sont numérotés à partir de 1. C'est par ce numéro que les renvois aux articles sont faits à l'intérieur d'un même code et entre différents

[34] *Les enfants doivent des aliments à leurs père et mère ou autres ascendants qui sont dans le besoin* (C. civ., art. 205)

[35] *Les époux s'obligent mutuellement à une communauté de vie* (C. civ., art. 215, al. 1)

[36] *Il est défendu aux juges de prononcer par voie de disposition générale et réglementaire sur les causes qui leurs sont soumises* (C. civ., art. 5)

[37] *Si l'autre époux ne reconnaît pas les faits, le juge ne prononce pas le divorce* (C. civ., art. 235)

[38] Un alinéa correspond à une phrase.

codes.

Cornu note que la concision et l'unité intellectuelle sont les deux caractéristiques principales des articles de codes. Leur longueur n'est cependant pas uniforme : de quatre mots à plus de 400 pour le seul Code civil. Le plus grand nombre d'articles de codes se trouvent cependant dans une moyenne de 10 à 100 mots.

Chaque article se trouve placé dans un ensemble, une loi ou un code. Pour ce qui est des codes, chacun est divisé en diverses sections et sous-sections (livres, titres, chapitres, parties, sections, sous-sections, paragraphes) mais de façon différente, chacun ne disposant pas forcément de toutes ces divisions. A chaque division correspond un titre, nous parlons pour notre part d'*intitulés de divisions*. Les articles sont ainsi thématiquement rassemblés. Ces divisions thématiques du droit sont, de la même façon que la division en différents codes, porteuses d'une certaine conceptualisation de la matière. Nous utilisons dans nos développements ces divisions et principalement leurs intitulés avec le présupposé suivant : les termes utilisés dans les intitulés des codes ne l'ont pas été de manière fortuite. On peut présumer à leur égard une importance plus grande qui leur est attribuée au regard des autres termes présents dans le corps même des articles de code.

Le domaine couvert par notre ontologie est donc envisagé en tant que tel, avec ses caractéristiques de forme et de fond. En utilisant les codes du droit français, nous faisons reposer notre ontologie sur un socle conceptuel de la matière. Les éléments de notre ontologie sont les termes du domaine, étiquettes des concepts juridiques exprimés dans les codes. Nous verrons par ailleurs plus loin (Section 7.3.1) que certaines des relations que nous dégageons entre ces éléments sont spécifiques à la matière. Nous entendons ainsi répondre à une interpellation de T. Bench Capon lors du *Workshop on Legal Ontologies* de la conférence Jurix 2001. Celui-ci nous demandait en effet dans quelle mesure la méthode d'élaboration d'ontologie que nous développions à l'époque était spécifique au droit et à l'informatique juridique. Nous répondons en avançant que beaucoup des choix méthodologiques sont guidés par la matière juridique. De plus, dans l'ontologie finale, des relations établies entre éléments découlent de spécificités du droit (voir Section 7.3).

Deuxième partie

Elaboration d'une ontologie à partir de textes

Nous décrivons dans cette Partie la démarche que nous avons adoptée pour conduire l'élaboration de notre ontologie à partir de corpus. De la même façon que pour les méthodes d'élaboration d'ontologies telles que TOVE ou Methontology décrites plus haut (Section 1.2.2), nous n'avançons pas avoir dégagé une méthode générique d'élaboration d'ontologies. Nous nous attachons à décrire les différentes étapes que nous avons réalisées pour élaborer notre ontologie. Le contexte dans lequel nous nous plaçons est celui d'ontologies dédiées à la recherche d'information, ce qui justifie la grande place qui y est faite aux documents, notamment dans la phase d'acquisition de connaissances à partir de textes. Nous pensons ainsi avoir dégagé un cadre méthodologique pour la construction d'ontologies dédiées à la recherche d'information. Ce cadre méthodologique pouvant être utilisé pour le développement d'autres ontologies.

La démarche que nous adoptons est ainsi principalement une démarche ascendante, déduisant des textes l'essentiel des connaissances portées par l'ontologie : les termes du domaine et les relations entre ces termes. Ce genre de démarche ascendante a cependant des limites [Mor99], en autres, celle de ne permettre d'identifier que des liens *neutres* entre les termes, c'est-à-dire non étiquetés par des prédicats sémantiques ; nous parlons de liens *non qualifiés*. Or, de telles qualifications sémantiques des liens entre termes sont nécessaires pour constituer une ontologie en tant que telle. Nous procédons donc *a posteriori* à la qualification sémantique des liens détectés dans la phase ascendante d'acquisition de connaissances à partir de textes.

La démarche de construction d'ontologie que nous présentons est centralisée dans le sens où elle n'a pas requis de travail d'équipe pluridisciplinaire. Ainsi, l'ingénieur de la connaissance s'avère être en même temps un expert du domaine.

Par ailleurs, la démarche que nous décrivons est essentiellement basée sur des textes du domaine qui tiennent lieu, en quelque sorte, de rapports d'experts sur le domaine. Le rôle de l'expert du domaine se limite alors, au-delà de son rôle dans la modélisation des connaissances acquises, à faire des choix quant aux textes à utiliser et non pas à produire des connaissances sur le domaine.

La démarche de construction d'ontologies dédiées à la recherche d'informations que nous adoptons procède en deux temps majeurs. Le premier temps repose sur une phase ascendante d'acquisition de connaissances à partir de texte qui rassemble plusieurs étapes. Il faut tout d'abord définir un corpus de référence (Chapitre 4). Sur ce corpus de référence, nous procédons à l'extraction des termes du domaine (Chapitre 5) et à l'identification des liens entre ces termes (Chapitre 6). Le second temps consiste à la modélisation des connaissances acquises à partir des textes (Chapitre 7) pour obtenir une ontologie en tant que telle.

Chapitre 4

Un corpus de référence

Nous développons dans la Section 4.1 la définition d'un corpus de référence. Ce corpus, dès lors qu'il est choisi (Section 4.2), servira de base à une démarche d'acquisition de connaissances à partir de textes.

4.1 Définition du corpus de référence

Le corpus de référence est un ensemble de textes rassemblés dans un objectif précis, celui d'être la base d'une démarche d'acquisition de connaissances. C'est sur ce corpus que seront utilisés des outils et techniques de traitement automatique de la langue avec pour objectif d'obtenir les deux éléments majeurs de la future ontologie : les termes du domaine et les relations entre ces termes.

Nous parlons de corpus de référence tandis que d'autres préfèrent parler de *corpus dédié* ou *special purpose corpus* [MP01a]. Certains distinguent le corpus de référence du corpus dédié [Pea98] arguant que la caractéristique principale du premier est la représentativité et que le corpus dédié est un corpus dont la composition est déterminée par un objet précis pour lequel il est élaboré[1]. Dans notre expérience, le corpus dédié et le corpus de référence se confondent dans un seul et même corpus ; nous parlons alors de corpus de référence.

Plusieurs cas peuvent être distingués quant à l'élaboration de ce corpus de référence.

(i) Soit ce corpus est spécialement élaboré à cet effet et les textes qui le composent n'existaient pas auparavant. Ce cas regroupe les démarches d'ac-

[1] *A corpus whose composition is determined by the precise purpose for which it is to be used*

quisition de connaissances à partir de textes qui consistent à récolter préalablement des avis d'experts sur les questions diverses pouvant concerner une même ontologie. Les spécifications sur les connaissances d'un domaine sont alors réunies dans des documents qui, rassemblés, constituent le corpus de référence, fondement de l'élaboration de l'ontologie. Un tel corpus est constitué, par exemple, quand l'ontologie traite de la mémoire d'une entreprise, ce qui implique la prise en compte de connaissances organisationnelles sur l'entreprise. Ainsi, une méthode de gestion des connaissances impliquant une phase d'acquisition prévoit un tel recueil des savoir des experts dans des documents textuels par des procédures d'interviews. Les savoir des experts sont en effet souvent opposés aux connaissances explicites exprimées en tant que telles et stockées sur un support physique. Les connaissances des experts sont, elles, dites tacites, dans le sens où elles sont difficilement exprimables. Ces connaissances des experts, et surtout les connaissances organisationnelles, sont ainsi difficilement communicables donc accessibles. L'enjeu de systèmes de gestion des connaissances organisationnelles, et des ontologies qui y sont incluses, est alors de faire passer ces connaissances du tacite vers l'explicite. La constitution de corpus de référence incluant des interviews d'experts est fonction du domaine en question, une organisation en particulier, et de l'application visée. Notre contexte d'une ontologie du droit français dédiée à la recherche d'informations ne nécessite ainsi pas particulièrement de recueils de connaissances tacites d'experts. Nous n'avons à faire qu'à des connaissances explicites exprimées dans des textes.

(*ii*) Un deuxième cas consiste à élaborer ce corpus de référence en rassemblant des documents épars préexistants dont l'objet premier n'était pas de servir de fondement à l'élaboration d'une quelconque ontologie. Ainsi, des travaux s'attachent à la constitution de corpus de référence en déterminant des mots clés du domaine, en cherchant des documents correspondants sur le Web et en sélectionnant parmi eux les documents les plus représentatifs ([GB01]). Un corpus de référence est alors élaboré, réunissant un ensemble de documents représentatifs du domaine. L'enjeu est en effet bien celui-là : sélectionner des documents représentatifs du domaine. Cette sélection peut se faire soit en lisant ces documents, soit en décidant de ne garder que ceux présentant un nombre de termes du domaine considéré comme suffisamment élevé. L'utilisation de documents préexistants à l'expérience pour servir de base à l'acquisition de connaissances à partir de texte comporte l'avantage indéniable de la rapidité et de la facilité. Des interviews d'experts ne sont ainsi pas nécessaires. Pour adopter cette méthode, il faut bien sûr que le domaine s'y prête et que les documents existent. Notre cas du domaine juridique français est typique de ce point de vue. Le droit est ainsi un domaine présidé par un ensemble de normes, ces normes étant l'objet même du domaine. Le droit français n'étant par ailleurs pas un droit coutumier, les normes qui le compose, les connaissances qu'il exprime, le sont sous la forme de textes. Ainsi, l'un des fondements du droit français est le Code civil, document créé sous Napoléon. Les sources des normes du droit français sont donc bien textuelles ; l'essence même du droit français transparaît dans des textes.

(*iii*) Enfin, dernier cas, le nôtre, la définition du corpus de référence consiste à identifier un corpus préexistant et valide à servir de corpus de référence. La totalité du droit français faisant l'objet de textes, à quelques exceptions près de quelques pratiques coutumières, notre corpus de référence est constitué de ces textes. Notre objet étant d'élaborer une ontologie du droit français dédiée à la recherche d'informations sur le Web, la question de sa couverture devait être résolue. Pour un tel objet, plus l'ontologie en question couvre de larges pans du droit français, plus son utilité est grande puisqu'elle appréhende alors la plupart des concepts juridiques. Nous ne pouvions cependant envisager de considérer tous les documents du droit français, de tous les inclure dans le corpus de référence. Les sources du droit français, bien que toutes textuelles, sont de divers type : les textes constitutionnels, les lois organiques, les lois ordinaires, les normes réglementaires, la jurisprudence etc. A cet ensemble déjà vaste et quasi-infini, dans le sens où un recueil centralisé de tous ces documents n'existe pas à notre connaissance, s'ajoute les analyses d'experts qui constituent la doctrine. Cette doctrine est de la même façon difficilement quantifiable, même à considérer toutes les revues autorisées traitant du droit français. Ainsi, les sources textuelles du droit français sont quasiment infinies donc indéfinissables. De plus, sachant que notre démarche consiste à utiliser des outils de traitement automatique de la langue sur les documents constituant le corpus de référence, nous ne pouvions en conséquence envisager d'embrasser toutes les sources du droit. Il fallait donc trouver un équilibre entre un corpus de référence couvrant suffisamment le droit français tout en ayant une taille raisonnable, c'est-à-dire traitable par un logiciel de traitement automatique des langues. Le corpus de référence devait dès lors présenter une densité terminologique relativement importante.

Au regard des documents dont nous disposions et des expériences que nous explicitons ci-dessous, nous en sommes arrivé à envisager l'ensemble des codes du droit français comme notre corpus de référence. Cet ensemble constitue un corpus en tant que tel dans le sens où il rassemble tous les codes qui sont eux-mêmes une des sources du domaine, celle de droit codifié. Nous explicitons dans la section 4.2 les expériences qui ont mené à ce choix.

4.2 Choix du corpus de référence

Nous disposons au Centre de recherche en informatique de l'Ecole des mines de Paris d'un ensemble de documents juridiques du droit français ainsi qu'un ensemble de documents de droit communautaire. Ces documents sont ceux qui se trouvent diffusés par les sites officiels tels Légifrance[2] ou Europa[3]. Nous disposons ainsi de documents issus du Journal Officiel de la République française édition lois et décrets (ci après JO), des codes du droit français et des directives et réglements européens. Tenant à nous limiter au droit français dans un premier temps[4], un corpus devait être choisi et/ou élaboré à partir de celui du JO et de celui des codes.

4.2.1 Le Journal Officiel de la République française

Chronologiquement, notre démarche a tout d'abord consisté à tenter d'exploiter les documents du Journal Officiel [Lam00]. La démarche ne s'est cependant pas révélée fructueuse comme nous le notons dans [Lam01b] et avec plus de détails dans [Lam01c]. Il s'avère en effet que le corpus du Journal Officiel se révèle être inadapté à une démarche d'acquisition de connaissances à partir de textes, ou à tout le moins, il est moins adapté à cette démarche que le corpus des codes.

Le corpus des JO dont nous disposons est tel que celui diffusé par le site officiel Légifrance. Le Journal Officiel de la République française édition lois et décrets rassemble un ensemble de documents propres au droit français : les lois, les décrets, les arrêtés émanant des différents ministères du gouvernement et des avis d'autorités telles l'Autorité de Régulation des Télécoms ou le Conseil Supérieur de l'Audiovisuel. Nous disposons de l'ensemble de ces documents depuis le 1er janvier 1998 et des textes relativement importants pour les dates antérieures. Ce corpus de JO avoisine (avril 2002) les cent mille documents, chaque loi, décret, arrêté ou avis constituant un seul et même document.

Ce corpus représente un ensemble d'environ 600 000 mots différents et trois à quatre millions de groupes de mots différents. Ces chiffres élevés sont expliqués, ou du moins explicables, par le fait que de nombreux noms propres

---

[2]http ://www.legifrance.gouv.fr/jo

[3]http ://www.europa.eu.int

[4]Des travaux peuvent tout à fait être envisagés, adoptant une méthode similaire de la nôtre pour développer une ontologie du droit communautaire ou, à tout le moins, élargir la nôtre avec les concepts du droit communautaire.

apparaissent dans ces documents, des noms de personnes dans le cas des décrets de nomination par exemple, ou des noms de localités.

Pour donner un ordre d'idées, l'ensemble des documents du JO correspondant à l'année 2000, soit un total de 24 178 documents, comportent 139 410 mots et 447 324 groupes de mots différents[5]. Ce sous corpus du JO représente 172 Moctets de données au format HTML et 132 Moctets au format TXT.

Nous ne pouvions raisonnablement envisager de traiter l'ensemble des quelques cent mille documents du JO. Des tentatives ont donc été menées sur l'ensemble des lois publiées au JO durant l'année 2000 ainsi que celles publiées en 2001, jusqu'en février. Cet ensemble représente 87 documents. Le logiciel Lexter d'extraction de candidats termes (décrit dans la section 5) a été lancé sur ces 87 documents. Un ensemble de plus de 500 000 candidats termes différents a ainsi été détecté. Parmi ceux-ci, de nombreux noms propres, et un constat : celui que les termes propres du droit se trouvaient "noyés" dans un ensemble de termes de la langue naturelle, la densité terminologique juridique est trop faible. Du fait de la masse des sorties de l'outil, nous ne pouvions envisager une phase de validation pourtant classique dans ce type de travaux[6]. Cette phase de validation consiste à faire déterminer par un expert du domaine, parfois aux vues de l'application visée, les termes pertinent du domaine dans un ensemble de candidats termes sorties d'extracteurs. Une autre solution devait être envisagée quant au corpus de référence.

Le corpus de référence doit en effet être suffisamment homogène pour constituer un ensemble cohérent, ce que ne constituent pas, finalement, les lois du Journal Officiel dont nous disposons. Il doit par ailleurs être d'une taille "raisonnable" et de préférence avoir la particularité de contenir une majorité de termes du domaine.

4.2.2 Les codes du droit français

Nous utilisons comme base de l'élaboration de notre ontologie les codes du droit français tels que présents sur le site officiel Légifrance au courant de l'année 2001 (Mai 2001). En réalisant quelques regroupements, par exemple assembler le Code Rural Ancien et le Code Rural, ou bien le Code de l'action sociale et des familles avec ce qui reste du Code des familles, nous obtenons un ensemble de cinquante-huit codes, indépendamment du fait que certains codes comportent des parties législatives, décrets ou arrêtés. La liste des

[5]données fournies par le moteur Pertimm

[6]voir par exemple la méthode sous-tendant l'utilisation de l'outil de structuration de terminologie Terminae [SBAG01]

codes est consultable en Annexe.

Nous notons que chacun de ces codes a une importance juridique relative au thème qu'il traite. Il va ainsi de soit que le Code civil ou le Code pénal ont une importance majeure au regard de codes beaucoup plus particuliers tels le Code de déontologie des sages-femmes ou le Code de la médaille militaire et de la légion d'honneur.

Les codes ont par ailleurs cette particularité d'avoir un vocabulaire relativement ramassé sur le domaine juridique. Ainsi, la fonction première des codes est de rassembler thématiquement les normes. Ils sont actuellement créés par la Commission supérieure de codification dont l'objet est de rassembler thématiquement les normes, de les rationaliser et d'organiser un code avec ces normes. Toutes les normes n'ont pas encore fait l'objet d'une codification. Les codes ne couvrent donc pas l'intégralité des sources normatives du droit français. Ils en couvrent cependant une bonne partie et, à tout le moins, les éléments les plus importants comme le code civil. Recueils de normes, les codes utilisent donc les termes fondamentaux du droit normatif. L'expérience[7] a également montré que, globalement, le vocabulaire utilisé dans les codes est relativement ramassé autour du droit. Ce fait est expliqué par l'objet même des codes : rassembler les normes de façon rationnelle. Ainsi, nous ne disposons pas dans les codes d'indications sur les mesures d'applications de la norme en question, ou des ministres chargés de son application comme c'est le cas dans les lois publiées au JO. Le corpus des codes s'avère donc, pour toutes ces raisons, être un corpus adapté pour constituer en tant que tel un corpus de référence. De plus, le corpus des codes a l'avantage de constituer par lui-même un corpus divisé thématiquement, ce qui est exploité dans notre ontologie.

Pratiquement, nous disposons de ces codes sous la forme d'un document HTML par article de codes tels que spécifiés dans les codes eux-même, par exemple, les articles L.123-7 ou R.613-46 du Code de la propriété intellectuelle. Un total de 65 536 articles correspondait aux 58 codes à la date du 4 avril 2002. L'ensemble de ces fichiers HTML des articles des codes correspond à 194 Mo de données. La totalité des codes rassemble un total de 6 403 216 mots.

Dans une démarche comme celle que nous adoptons pour la construction

[7] Cette expérience est celle de l'extraction des termes du domaine que nous décrivons dans la section qui suit (section 5).

d'ontologies essentiellement basée sur des textes, le choix du corpus de référence tel que nous le décrivons est un choix fondamental. De ce bon choix va dépendre une partie de la réussite de l'opération. La démarche telle que nous la décrivons, sur la base du corpus de référence des codes, a été opérationnelle pour élaborer une ontologie du droit français dédiée à la recherche d'informations sur le Web. La construction d'ontologies fondée sur cette démarche devra particulièrement s'attacher à cette question préalable du choix du corpus de référence. Celui-ci pourra être, en fonction du domaine et des circonstances, élaboré de toutes pièces, rassemblé, ou simplement identifié, comme nous le décrivons plus haut.

Chapitre 5

Extraction des termes

Notre objet est, dans ce Chapitre, d'identifier les termes du domaine sur la base du corpus de référence.

Le fondamental est ainsi celui-là : les connaissances sont exprimées dans les textes au travers de structures linguistiques. C'est un travail sur ces structures linguistiques qui va permettre d'acquérir les connaissances exprimées dans les textes. Un travail sur les termes permet d'appréhender les notions, donc les objets, d'un domaine.

L'extraction des termes du domaine consiste à utiliser des outils d'extraction de termes sur le corpus de référence (Section 5.1). Ces outils fournissent des listes de candidats termes ; encore faut-il identifier parmi eux les termes du domaine qui nous intéressent en tant qu'étiquettes de concepts du domaine. Ces termes du domaine que nous recherchons sont définis dans la Section 5.2. Pour les identifier, nous procédons à un filtrage sur la liste des candidats termes (Section 5.3). Nous avons ainsi tenté d'utiliser les techniques statistiques classiques de pondération de termes pour identifier les termes du domaine. En conclusion de cette expérience, nous avançons que ces techniques statistiques ne sont pas appropriées. Cette conclusion fonde un choix méthodologique fondamental dans notre démarche, celui de la non-validation de la liste des candidats termes (Section 5.4).

5.1 Outils d'extraction de termes

Divers outils existent permettant d'identifier les termes présents dans les textes qui leurs sont soumis. Les termes sont les mots et les groupes de mots lexicalement pertinents tels les groupes nominaux, les groupes adverbiaux, les groupes adjectivaux etc.

Ces outils peuvent reposer sur deux méthodes permettant d'identifier ces termes : une méthode statistique et une méthode syntaxique. Les deux méthodes peuvent être utilisées simultanément ou alternativement.

5.1.1 Méthodes d'extraction de termes

(i) La méthode statistique d'identification des termes consiste à déterminer les suites de mots qui, apparaissant suffisamment souvent dans les textes, sont susceptibles d'avoir une cohérence lexicale. L'extraction des mots ne pose ainsi aucun problème à ce type d'outils. C'est la détermination des groupes de mots qui est plus complexe, nécessitant des procédures itératives. Les outils basés sur de telles méthodes sont donc très gourmands en ressources informatiques et les traitements qu'ils opèrent sont relativement longs. De plus, de tels outils nécessitent certaines caractéristiques du corpus du référence. Si celui-ci ne contient que peu de répétitions et si les termes n'apparaissant qu'une seule fois ont aussi leur importance, l'outil, basé sur des données statistiques, risque de ne pas les considérer. Or c'est justement le cas du corpus de référence que nous utilisons dans notre expérience. Les outils statistiques d'extraction de termes fonctionnent ainsi d'autant mieux que le corpus est important et comporte des répétitions. Ils ont cependant l'avantage d'être multilingues dans le sens où la langue dans laquelle les textes sont exprimés ne leur importe finalement que peu. Des méthodes statistiques, notamment de filtrage de termes, peuvent également être utilisées dans un même outil en complément d'une méthode syntaxique.

(ii) Un deuxième type d'outils procédant à l'identification de termes dans des textes repose sur des analyses syntaxiques menées sur ces textes. Ces outils fonctionnent avec des documents "étiquetés" sur lesquels ils appliquent des règles de grammaires pour identifier les syntagmes pertinent. Ces outils comportent donc un module d'étiquetage des textes qui, sur la base de dictionnaires ou pas, déterminent la nature syntaxique de chaque élément des documents ; chacun est alors "étiqueté". Dès lors que le texte est "étiqueté", les règles de grammaires sont appliquées sur les phrases. Ces outils intègrent ainsi des modules contenant, de façon plus ou moins simplifiée, les règles grammaticales d'une ou plusieurs langues ; ils sont alors dépendants de la langue dont ils connaissent la grammaire. La détermination des termes syntaxiquement pertinents se fera, entre autre, par l'utilisation de *mots frontière*. Une liste est ainsi incluse dans l'outil, parfois paramétrable par l'utilisateur, indiquant la liste de ces mots déterminant la fin d'un terme et le commencement d'un nouveau. Ces mots frontière sont le plus souvent des verbes.

En sortie de ces outils, on obtient une liste des candidats termes pour les textes qui lui ont été soumis. Des indications syntaxiques sur ces termes peuvent être données : savoir si c'est un nom, un verbe, un adjectif, un syntagme nominal, un syntagme adverbial etc. Du fait des caractéristiques du corpus de référence utilisé dans notre expérience, nous avons préféré utiliser un extracteur de termes se fondant sur des études syntaxiques. De plus, nous utilisons les données syntaxiques fournies sur les sorties de tels outils.

5.1.2 Les outils Sylex et Syntex

Nous avons testé deux outils se fondant sur des analyses syntaxiques : Sylex et Syntex, pour choisir ce dernier.

(*i*) Le logiciel Sylex [Con95] est présenté comme un analyseur linguistique, cherchant à "comprendre et analyser un texte quelconque". L'outil présente une large part d'analyse syntaxique des textes, sur un "axe syntagmatique". L'analyse du texte sur un "axe paradigmatique" s'attelle à la sémantique, s'enrichissant de chaque étape de l'analyse syntaxique. Cette dernière part de l'analyse des textes n'apparaît cependant que peu dans les sorties de l'outil. Nous avons donc envisagé l'outil Sylex pour ses capacités d'analyses syntaxiques des documents. L'outil Sylex incorpore les règles de grammaire de la langue française (et d'autres mais qui ne nous intéressent pas dans notre contexte). Il se fonde alors sur la théorie de Tesnière, arguant que celle-ci permet de minimiser la création de concepts linguistiques, ce qui réduit d'autant plus les arbitraires potentiels. Ainsi, toute ambiguïté lexicale (le cas, par exemple, des mots qui sont à la fois des noms et des verbes) est partiellement résolue par le choix arbitraire d'une des interprétations possibles, l'avancement de l'analyse permettant de revenir sur ces choix arbitraires. L'analyse syntaxique du texte se fait donc par "couches", la structuration syntaxique du texte évoluant au fur et à mesure du passage des différents modules grammaticaux. L'objet premier de Sylex est donc l'analyse de textes et pas l'extraction de termes.

(*ii*) Le logiciel Syntex[1] est un analyseur syntaxique. Il fonctionne sur des textes préalablement étiquetés et lemmatisés dans lesquels sont indiquées,

[1] Le logiciel Syntex a été développé au sein de l'Equipe de Recherche en Syntaxe et Sémantique de Toulouse, par Didier Bourigault et Cécile Fabre. Il fait l'objet d'un contrat de valorisation passé entre le CNRS, l'Université Toulouse Le Mirail et la société Synomia (www.synomia.fr). Syntex fonctionne actuellement pour le français et pour l'anglais.

pour chaque mot, des informations sur la catégorie grammaticale, les traits morphologiques, comme le genre et le nombre, et la forme lemmatisée. L'outil repère dans les textes les relation de dépendance syntaxiques suivantes : sujet de verbe, objet de verbe, complément prépositionnel de verbe, nom ou adjectif, épithète et attribut de nom. Grâce à des procédures d'apprentissage endogène, le système acquiert lui-même les informations syntaxiques de sous catégorisation qui lui sont nécessaires pour effectuer un repérage précis des termes complexes [Bou93]. Sur la base des relations de dépendance établies, le module d'extraction de syntagmes construit un réseau de mots et syntagmes. De ce réseau, on obtient l'ensemble des candidats termes, quelles que soient leurs catégories grammaticales et les contextes syntaxiques de chacun de ces termes exploités dans le cadre de l'analyse distributionnelle. Nous nous intéressons dans cette section au premier résultat de l'outil : la liste des candidats termes. Des sorties de l'outil sur notre corpus de référence, l'objectif est d'identifier les termes du domaine.

5.2 Les termes du domaine

Utilisant une technique automatique d'extraction des candidats termes, il nous faut veiller à circonscrire, ou filtrer, les termes spécifiques du domaine du droit de ceux qui ne le sont pas. Nous appelons ces termes qui nous intéresse les *termes du domaine*.

Les termes du domaine sont des termes qui représentent des artefacts juridiques. Les termes du domaine concernent (voir Section 3.3) d'une part les termes spécifiques à la matière juridique, ainsi *licéité*, *stellionataire*, *colon partiaire*, *répétition de l'indu* ou *tontine*. Les termes du domaine sont également des termes ayant un sens juridique et d'utilisation commune dans le langage commun : *contrat*, *mariage*, *testament*, *société à responsabilité limitée* ou *avocat* etc, preuve que ces artefacts juridiques sont entrés dans la vie quotidienne. Enfin, les termes du domaine rassemblent également des termes polysèmes ayant un sens juridique propre et un sens autre dans la langue naturelle, par exemple *absence*, *minute* ou *grosse* [Cor00]. Les *termes du domaine* réunissent alors un ensemble de termes qui présentent au minimum un sens juridique.

Par ailleurs, le droit est une discipline qui interprète le monde. Les objets du monde sont à ce titre appréhendés par le droit et sont exprimés, de notre point de vue, par des termes du domaine, ainsi *machine*, *médicament*, *passager* ou *clôture*. Nous entendons donc comme termes du domaine à la fois les termes représentant des artefacts juridiques et des objets du monde appréhendés par le droit.

5.3 Filtrage des candidats termes

Un ensemble très hétérogène de termes complexes (ou syntagmes) peut être extrait des textes (cf le Tableau 5.1). Nous avons donc choisi de n'envisager que les termes s'avérant être des noms et des syntagmes nominaux (des groupes nominaux). Nous pensons ainsi embrasser une partie essentielle des objets du domaine juridique. En effet, les objets du droit sont des entités matérielles (la *voirie routière*) ou immatérielles (une *société commerciale*, la *capacité juridique*) nommée dans les textes même du droit par des noms ou syntagmes nominaux. Du fait que nous tentons d'identifier ces objets du domaine, nous ne nous intéressons en conséquence pas à la sémantique portée par les verbes et les mots d'autres catégories grammaticales. Par ailleurs, les noms et les syntagmes nominaux sont les catégories grammaticales de termes les plus souvent saisies dans des moteurs de recherche.

Abréviation	Type	Exemple
Adj	Adjectif	*budgétaire*
Adv	Adverbe	*éventuellement*
NomPr	Nom propre	*Hauts-de-Seine*
Nom	Nom	*avis*
Ppa	Participe passé	*élaboré*
SAdj	Syntagme adjectival	*accessible aux personnes handicapées*
SNom	Syntagme nominal	*accueil thérapeutique*
SPpa	Syntagme participial	*notifié à l'intéressé*
SV	Syntagme verbal	*arriver à échéance*
V	Verbe	*rétracter*

TAB. 5.1 – Types et exemples de syntagmes extraits par Syntex du Code de l'action sociale et des familles

Le logiciel Syntex[2] a été lancé sur le corpus de référence tel que nous l'avons préalablement défini. Au moment de cette expérience, nous disposions d'un ensemble de 57 codes, sachant que nous n'avions pas encore rassemblé le Code de l'action sociale et des familles avec le Code des familles, ni le Code rural avec le Code rural ancien. Nous notons simplement que le Code

[2]Nos premières expériences sur l'extraction de termes ont cependant été menées avec le logiciel Lexter, l'"ancêtre" de Syntex.

des caisses d'épargne n'était pas en notre possession à ce moment là. L'outil extrait un ensemble de plus de 500 000 termes de notre corpus de référence. La quantité est énorme ; elle interdit toute velléité de traiter manuellement ces termes. Quoi qu'il en soit, un premier tri s'impose ; il va consister à éliminer un ensemble de termes non appropriés.

Ainsi, un certain nombre de termes peuvent être ignorés, selon leurs caractéristiques propres :

Les termes contenant des indications chiffrées : ces indications chiffrées peuvent être propres au corpus même. Ainsi, de nombreuses références internes aux articles des codes par leurs numéros sont présentes dans notre corpus. Nous avançons que toute indication chiffrée peut être, dans notre contexte, ignorée, celle-ci n'étant pas caractéristique d'un objet du droit. Ainsi, notre objet n'est pas de connaître la valeur d'un taux d'imposition en particulier, mais plutôt de déterminer qu'un *taux d'imposition* est un objet du domaine juridique.

Les termes contenant des caractères non alphabétiques : sont également ignorés selon la même logique que précédemment (par exemple, %).

Les adjectifs, les adverbes etc : ayant choisi de ne considérer les termes uniquement noms ou syntagmes nominaux, nous ignorons en conséquences les autres. Ainsi les adverbes, les adjectifs, les verbes et les syntagmes verbaux, les participes passés, les syntagmes participiaux et les noms propres sont retirés de la liste des termes extraits par Syntex.

Les termes comportant des lettres majuscules : ces termes sont soit des débuts de phrases, soit contiennent des noms propres. Ainsi, on ne gardera pas le terme *tribunal de grande instance de Paris* mais le terme *tribunal de grande instance*.

Un ensemble de termes comme listés dans le tableau 5.2 sont ignorés des sorties de Syntex sur notre corpus de référence.

Nous obtenons dès lors une liste d'environ 300 000 termes. Des expérimentations ont été menées sur cette liste de termes, dans l'objectif de différentier parmi eux ceux susceptibles d'être des termes du domaine de ceux qui ne le seraient pas. L'expérience décrite ci-dessous s'est soldée par un échec, nous en dégageons la conclusion que la détection des termes du domaine par des techniques statistiques fondées sur les indices classiques de la pondération de termes n'est pas envisageable.

5.3.1 Pondération des candidats termes

Une première expérience a été menée sur la liste de 300 000 termes dans le but de détecter dans cette liste, les termes du domaine des autres. Cette expérience a consisté à établir un certain nombre de caractéristiques statistiques et d'indices classiques de pondération des termes. Pour chacun de ces indices, le corpus considéré est le corpus de référence tel que défini précédemment, l'ensemble des codes. Ce corpus est envisagé comme un ensemble de documents que sont les articles de codes, 59 275 au total.

Les caractéristiques établies pour chacun des termes de la liste des 300 000 sont les suivantes :

Le nombre de mots composant le terme : un mot est alors considéré comme une suite de caractères séparés par des espaces typographiques. Ainsi, le terme *résiliation de l'abonnement* contient 3 mots.

La fréquence totale du terme , ou *tf*, représentant le nombre total d'occurrences du terme dans la totalité du corpus,

L'*inverse document frequency* , ou *idf* ([SJ73], [Moe00]), établit la répartition d'un terme dans un corpus, partant du principe que l'importance d'un terme est inversement proportionnelle au nombre de documents du corpus dans lesquels il apparaît ; *idf* est le logarithme du nombre de documents dans le corpus N divisé par le nombre de documents n du corpus contenant le terme en question i,

$$\log\left(\frac{N}{n_i}\right) \tag{5.1}$$

***Tf.idf* :** ([SJ73], [Moe00]) l'*inverse document frequency*, ou *idf*, peut être combinée à la fréquence totale *tf* du terme i dans le corpus, ce qui

Syntagme	Observation
l'objet du litige peut être modifié	SV
relatives à l'organisation judiciaire	SAdj
ultérieurement prononcée	SAdj
statué sur le fond	SPpa
rendue sur le fondement de l'article 808	caractère non alphabétique
alinéa 2	caractère non alphabétique
Le jugement est réputé	lettre majuscule
tribunal de grande instance de Paris	lettre majuscule

TAB. 5.2 – Exemple de termes ignorés extraits pas Syntex du code de procédure civile

permet de prendre en compte les termes n'apparaissant pas dans tous les documents du corpus (*idf*) et qui, malgré tout, apparaissent souvent dans le corpus (*tf*) :

$$tf_i \cdot \log\left(\frac{N}{n_i}\right) \tag{5.2}$$

L'entropie de la répartition : la répartition du terme i dans le corpus s'établit comme suit, le nombre d'occurrences *tf* du terme i dans le document x divisé par le nombre total d'occurrence *TF* du terme i dans le corpus total :

$$r(i)_x = \frac{tf_i(x)}{TF_i} \tag{5.3}$$

L'entropie S de la répartition s'établit alors comme suit :

$$S = -\sum_{i,x} r(i)_x \log r(i)_x \tag{5.4}$$

Les résultats de chaque indice que nous présentons ci-dessous ont été établis avec le logiciel libre de statistiques R[3]. Le nombre de mots des termes se répartit comme suit dans nos 300 000 termes (Tableau 5.3) :

Minimum	1
1er quartile	3
Médiane	4
Moyenne	4.494
3eme quartile	6
Maximum	29

TAB. 5.3 – Nombre de mots par termes

La moyenne du nombre de mots par termes s'établit à 4.5. Des termes comportant un seul mot sont par exemple *cargaison*, *enveloppe*, *juin* ou *non-respect*. Le terme comportant le maximum de mots (29) est *titre de le régime de assurance vieillesse complémentaire de le travailleur non-salarié de le profession artisanal organisation autonome de assurance vieillesse de le travailleur non-salarié de le profession artisanal*[4]. Ce terme est le résultat d'une approximation de découpage de l'extracteur de termes. Une des façons de sélectionner les termes du domaine pourrait consister à ne considérer que les termes comprenant un nombre raisonnable de termes. Cependant, déterminer ce seuil raisonnable n'est pas chose aisée et relève forcément d'heuristiques. Une telle démarche sur une liste de 300 000 termes s'avère difficilement

[3]http ://www.r-project.org/

[4]Le terme est sous sa forme lemmatisée par Lexter.

réalisable.

Le tableau 5.4 décrit la façon dont les fréquences des termes se comportent dans notre liste de 300 000 termes. Nous notons que le nombre de termes

Minimum	1
1er quartile	1
Médiane	1
Moyenne	16.6
3eme quartile	2
Maximum	106386

TAB. 5.4 – Fréquences des termes

n'apparaissant qu'une fois, les *hapax*, est relativement élevé (188 158), ce qui représente 63 % de nos termes. Nous ne pouvons cependant envisager de ne pas les considérer en tant que termes du domaine.

La moyenne d'apparition des termes dans notre corpus est de 16.6, le terme apparaissant 106 386 fois dans notre corpus est, sans surprise, le terme *article*[5]. Des exemple de termes n'apparaissant qu'une fois dans le corpus sont listés dans le Tableau 5.5 ainsi que des termes ayant une fréquence élevée (supérieure à 50 000). Le nombre moyen de mots des termes n'apparaissant qu'une fois dans le corpus est de 4.9. Ce qui correspond, à peu près, au nombre moyen de mots de tous les termes. Les termes n'apparaissant qu'une fois ne sont donc pas que des termes contenant beaucoup de mots, même si plus les termes comportent un nombre élevé de mots, plus leur fréquence est faible.

Termes freq=1	Termes freq > 50 000
activité économique taxable	*article*
activités agricoles à prédominance pastorale	*chapitre*
actualisation des données scientifiques	*code*
acheminement de stupéfiant	*disposition*
acheteur potentiel	*décret*
achèvement du risque	*journal*

TAB. 5.5 – Exemples de termes selon leur fréquence

L'indice de fréquence sur les termes peut, à la limite, permettre d'identifier les termes n'étant pas des termes du domaine au sens où on l'entend, comme

[5]Ce constat est *sans surprise* car les codes sont divisés en articles, chacun étant intitulé comme tel. De plus, les références internes aux articles sont fréquentes.

article ou *chapitre*. Cependant, des termes apparaissant souvent peuvent néanmoins se révéler être des termes du domaine comme par exemple *décret*. L'indice de fréquence des termes ne permet donc pas en tant que tel d'identifier les termes du domaine.

L'*inverse document frequency* des termes donne une indication de leur répartition dans un corpus. Cet indice est conséquemment dépendant du corpus utilisé [Moe00]. Il permet d'identifier des descripteurs des documents et, parfois [SJ73], les termes ayant une faible *idf* sont identifiés comme termes vides.

Minimum	0.6932
1er quartile	4.2767
Médiane	4.6540
Moyenne	4.7049
3eme quartile	5.0690
Maximum	7.8136

TAB. 5.6 – *Inverse document frequency*

La moyenne de l'*idf* de nos termes s'établit à 4.7 (voir Tableau 5.6). Le terme qui a l'*idf* la plus faible est *firme* ; celui qui a l'*idf* la plus élevée est *candidatures à la préfecture*[6]. Classiquement, plus un terme a une *idf* faible, moins il peut être considéré comme un descripteur intéressant puisqu'il apparaît dans beaucoup des documents du corpus. Il s'avère cependant que, dans notre expérience, une telle conclusion ne soit guère envisageable. Nous nous sommes ainsi attelés, en tant que spécialiste du domaine juridique, à étudier de façon expérimentale les listes de termes ayant une *idf* faible et une *idf* élevée. Leur parcours n'a pas permit de dégager un indice d'*idf* particulier auquel correspondrait un nombre important de termes juridiques. A titre d'illustration, le Tableau 5.7 liste un ensemble de termes ayant une *idf* faible (idf < 2.5) et un ensemble de termes ayant une *idf* élevée (idf $>$ 7.7).

L'indice *idf* est largement dépendant du corpus et de la façon dont il est découpé. Nous notons à cet égard qu'il semble, d'un point de vue expérimental, que les termes ayant un indice d'*idf* élevé relèvent principalement du Code électoral.

La démarche demeure expérimentale mais nous ne pouvons envisager de ne pas prendre en compte les termes ayant un indice d'*idf* faible.

[6]Si on omet le terme *agrave*, issu d'erreurs typographiques dans les documents.

Termes idf $<$ 2.5	Termes idf $>$ 7.7
assistance morale	*agrave*
assuré naître	*candidatures à la préfecture*
caution judiciaire	*consultation par voie d'affichage*
délibéré du juge	*dépôt des candidatures à la préfecture*
essieux	*jour du scrutin au bureau de vote intéressé*
firme	*souscription de titres*
ministre	*liste en nombre*
obligations du dépositaire	*période de consultation par voie d'affichage*
obligations scolaire	*période de dépôt des candidatures à la préfecture*
prison de rattachement	*quantité de bulletin de vote*
revendication indépendante	*scrutin au bureau de vote intéressé*

TAB. 5.7 – Exemple de termes selon leur *inverse document frequency*

Le fait de combiner l'*inverse document frequency* avec la fréquence permet de mettre en valeur les termes qui, apparaissant souvent dans le corpus, n'apparaissent cependant pas dans tous les documents. L'indice *tf.idf* [SJ73] est ainsi souvent utilisé pour pondérer des termes [Moe00]. Classiquement, plus un terme a un indice *tf.idf* élevé, plus il est intéressant.

Minimum	2.398
1er quartile	4.522
Médiane	5.159
Moyenne	72.25
3eme quartile	10.33
Maximum	423100

TAB. 5.8 – *Tf.idf*

La moyenne de l'indice *tf.idf* s'établit pour nos termes à 72.25 tandis que la médiane se situe à 5.159. Le terme dont l'indice de *tf.idf* est le plus élevé (423 000) est le terme *article* ; celui dont l'indice de *tf.idf* est le plus faible est *délibéré du juge*. Le Tableau 5.9 liste un ensemble de termes dont les indices de *tf.idf* sont faibles (tf.idf $<$ 2.6) et élevés (tf.idf $>$ 220 000).

De la même façon que pour la fréquence, l'indice *tf.idf* pourrait être utilisé pour identifier des termes qui ne sont pas des termes du domaine dans le sens où on l'entend, comme *article* ou *chapitre*. Cependant, le terme *décret* est bien, lui, un terme du domaine. De la même façon, des méthodes heuristiques seraient alors nécessaires pour fixer un seuil de l'indice *tf.idf* à partir

Termes tf.idf < 2.6	Termes tf.idf $> 220\,000$
Lorsqu'un vétérinaire	*art*
critique ouverte	*article*
délibéré du juge	*chapitre*
fonction publique temporaire	*code*
intention contraire	*décret*
intention de l'humain	*journal*
même domicile	*journal officiel*
obligations scolaire	*livre*

TAB. 5.9 – Exemple de termes selon leur *Tf.idf*

duquel on ne trouverait que des termes du domaine.

L'entropie est une mesure de *désordre*. Nous établissons l'entropie de la répartition de chaque terme dans le corpus, ce qui revient à mesurer le désordre provoqué par certains termes dans la répartition globale des termes dans le corpus. Ainsi, un terme largement répandu, c'est-à-dire apparaissant dans un nombre important de documents du corpus, aura une entropie élevée, signifiant que ce terme n'apporte pas de désordre à la répartition globale des termes dans le corpus, qu'il n'apporte que peu d'informations. Nous avons calculé l'entropie pour chacun des 300 000 termes de notre liste.

Minimum	~ 0.0000
1er quartile	~ 0.0000
Médiane	~ 0.0000
Moyenne	0.6103
3eme quartile	0.3466
Maximum	60.2408

TAB. 5.10 – Entropie

La moyenne des entropies de nos termes s'établit à 0.619 tandis que la médiane est de l'ordre de 0. Le terme dont l'entropie est la plus élevée est *partie*. Le Tableau 5.11 liste des exemples de termes dont l'entropie est faible (entropie $\simeq 0$) et dont l'entropie est élevée (entropie > 52).

Les termes dont l'entropie est élevée peuvent être mis de côté de la même façon qu'avec la fréquence ou *tf.idf*. Le même problème demeure. De la même façon que pour les autres indices, la détermination d'un ou plusieurs seuils qui permettraient d'identifier les termes du domaine est illusoire. Cette démarche nécessiterait de définir des heuristiques sur la liste des 300 000 termes.

Termes dont entropie $\simeq 0$	Termes dont entropie > 52
abandon anticipé	*article*
abattage de bois	*chapitre*
abattement de salaire	*code*
abattoir en régie	*dispositions*
abolition	*décret*
acquisitions de bien	*livre*
acte de prévention	*partie*
adversité	*section*
aedificandi	*titre*

TAB. 5.11 – Exemple de termes selon leur entropie

La conclusion de cette première expérience de tentative d'identification des termes du domaine en utilisant des méthodes classiques de pondération de termes est que ces méthodes sont inadéquates. Nous avançons que des méthodes statistiques ne peuvent en aucun cas être menées automatiquement sur une liste de candidats termes puisque la détermination de seuils est nécessaire et semble difficile à réaliser. Cette détermination de seuils exigerait de définir des heuristiques inenvisageables dès lors que la liste de candidats termes est un tant soit peu importante.

5.3.2 La liste de termes juridiques

Pour conforter la conclusion résultant de la tentative de détection par des méthodes statistiques de termes du domaine, une seconde expérience a été menée. Elle a consisté à tenter d'établir si des termes connus comme juridiques, des candidats termes du domaine donc, avaient des particularités statistiques par rapport aux autres termes. Cette deuxième expérience a été menée sur la liste des 300 000 termes, utilisant une liste préalablement établie de termes connus comme étant des termes juridiques. Le préalable était, en effet, d'établir une sous liste contenant des termes connus comme des termes du droit de notre liste de candidats termes. L'idée sous-jacente était de déterminer les caractéristiques statistiques de ces termes dans notre corpus de référence. L'expérience montre qu'aucune caractéristique statistique discriminante n'a pu être dégagée pour les termes juridiques par rapport aux autres, ce qui conforte l'idée que les méthodes statistiques ne sont pas appropriées pour l'identification de termes du domaine.

Pour identifier les termes juridiques (candidats termes du domaine), nous avons préalablement établi une liste de termes issus de dictionnaires de droit. Ces termes ont alors forcément au moins une signification juridique. Cette liste est établie automatiquement à partir des dictionnaires juridiques identifiés comme tels *à la main* sur Internet. Elle rassemble un total de 1 490 termes consultable en Annexe. Ces termes sont principalement des noms et des syntagmes nominaux. Nous utilisons cette liste pour détecter dans notre liste de 300 000 termes ceux correspondant exactement à ces termes du droit ou les contenant. C'est ainsi une procédure de *pattern matching* qui est implémentée, grâce à un programme Perl. Les termes de notre liste de 300 000 sont alors distingués selon qu'ils sont des termes juridiques de façon certaine ou qu'ils sont autres. Par exemple, le terme *contrat* fait partie de la liste des termes connus comme juridiques. Par le processus de *pattern matching*, l'ensemble des termes de notre liste de 300 000 contenant le mot *contrat* (2 092) seront identifiés comme des termes juridiques. Parmi ceux-ci : *caractéristiques du contrat*, *date de conclusion du contrat à durée indéterminée*, *donation par contrat de mariage* ou *exécution du contrat*. Nous extrayons ainsi de notre liste initiale une sous liste de 111 202 termes juridiques, ce qui représente environ à 37% du total. La liste des termes restant (appelés *termes autres*) est de 185 478.

Les termes juridiques et les termes autres ont été comparés selon les indices établis dans la première expérience. Toutes les données ainsi que les graphes ont été réalisés avec le logiciel de statistique *R*.

Nombre de mots des termes : Si l'on établit le rapport entre le nombre de termes juridiques et le nombre de termes autres pour chacune des valeurs possibles du nombre de mots par termes (Figure 5.1), nous constatons que, globalement, plus le nombre de mots est élevé, plus la proportion de termes juridiques est importante et dépasse la part de termes autres.

Ce point s'explique par la méthode qui a conduit à l'identification des termes juridiques. Celle-ci repose sur le principe du *pattern matching* entre une liste de termes du droit et notre liste de 300 000 termes. De ce fait, plus un terme de la liste des 300 000 a un nombre élevé de mots, plus il a de chances de contenir des termes de la liste des termes du droit.

L'établissement du nombre de mots ne permet cependant pas de distinguer une caractéristique particulière des termes juridiques par rapport aux autres. Nous constatons ainsi que, globalement, Figure 5.2, les termes juridiques et les termes autres montrent les mêmes caractéristiques quant au nombre de mots les composant (maximum entre 2 et

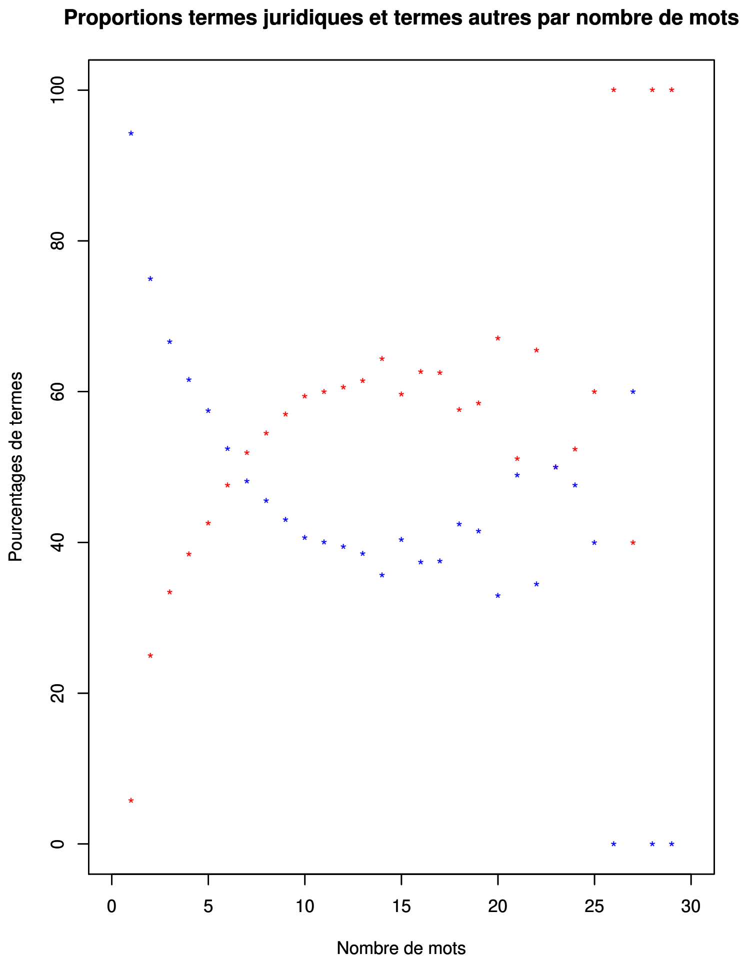

FIG. 5.1 – Proportion des termes juridiques (en rouge) et des termes autres (en bleu) par nombre de mots

4 mots, plus le nombre de mots augmente, moins le nombre de termes correspondant est élevé). Ainsi, les termes juridiques et les autres ont le même comportement statistique en terme de nombre de mots.

La fréquence des termes : De la même façon, l'établissement de la fréquence d'apparition des termes dans le corpus de référence ne permet pas de distinguer les termes juridiques des autres.

La Figure 5.3 illustre le fait que les termes juridiques et les autres ont le même comportement en ce qui concerne leurs fréquences dans le corpus. Le graphique représente, pour des questions des lisibilité, les termes dont la fréquence est comprise entre 1 et 30, ce qui correspond à 97 % des termes (97 % des termes juridiques et 96.9 % des termes autres). Les points représentent le nombre de termes juridiques en fonction du nombre de termes autres pour chaque niveau de fréquence entre 1 et 30. La corrélation entre les termes juridiques et les termes autres pour ce qui est de leur fréquence d'apparition dans le corpus est très forte : le coefficient de corrélation linéaire est de 0.9998108.

On ne peut donc distinguer les termes juridiques des autres sur la base de leurs fréquences d'apparition dans le corpus.

L'*inverse document frequency* : Pour représenter les indices d'*idf* de nos termes dans la Figure 5.4, nous avons choisi d'établir pour chacune des valeurs d'*idf* fixées, le nombre de termes juridiques et de termes autres dont l'*idf* est comprise entre la valeur précédente et celle en question. Nous avons ainsi fixé arbitrairement les valeurs d'*idf* par intervalles de 0.25 entre 0.65 (le minimum d'*idf*) et 7.85 (le maximum).

Le Graphique 5.5 établit le nombre de termes juridiques en fonction du nombre de termes autres pour chaque valeur d'*idf* telles qu'établies précédemment. Le coefficient de corrélation linéaire entre les termes juridiques et les termes autres du point de vue de l'*idf* est de 0.9979602.

La détermination de l'*idf* de chaque terme ne permet également pas de distinguer les termes juridiques des autres.

L'indice *tf.idf* : De la même façon que pour les indices établis préalablement, la considération de *tf.idf* ne permet pas de distinguer particulièrement les termes juridiques des autres.

Le Graphique 5.6 représente le nombre de termes juridiques en fonction du nombre de termes autres pour l'indice *tf.idf*. Pour élaborer ce graphique, nous avons considéré les termes dont l'indice *tf.idf* est compris entre 2 et 60, ce qui couvre 94 % des termes. Les nombres de termes juridiques et de termes autres ont été établis pour chaque valeur de *tf.idf* comprise entre 2 et 60 avec un pas de 2.

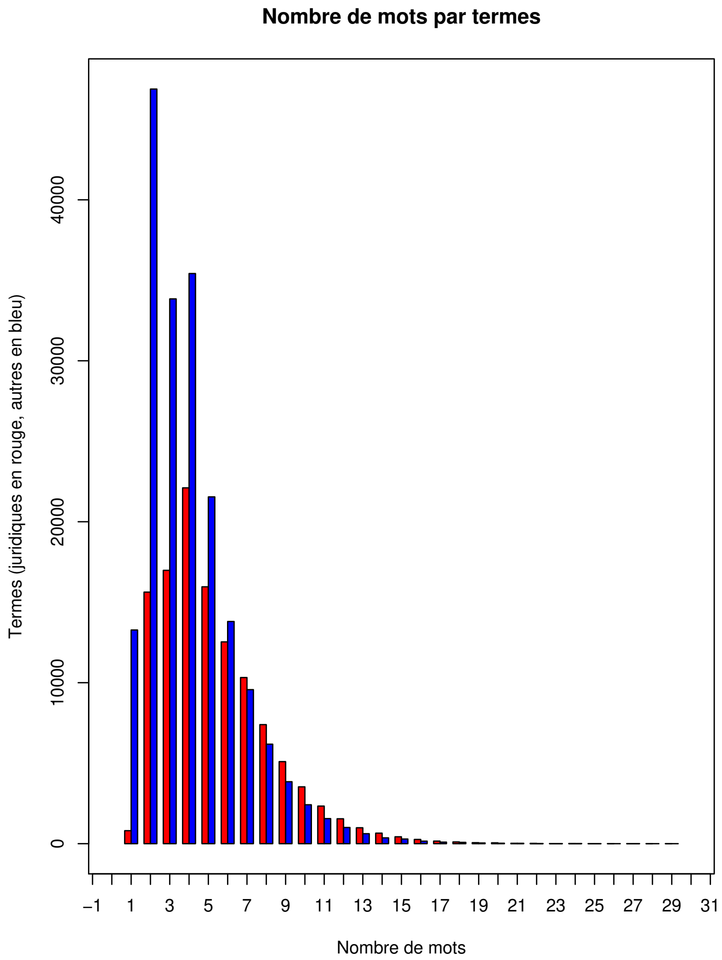

FIG. 5.2 – Termes juridiques (enrouge) et termes autres (en bleu) par nombres de mots

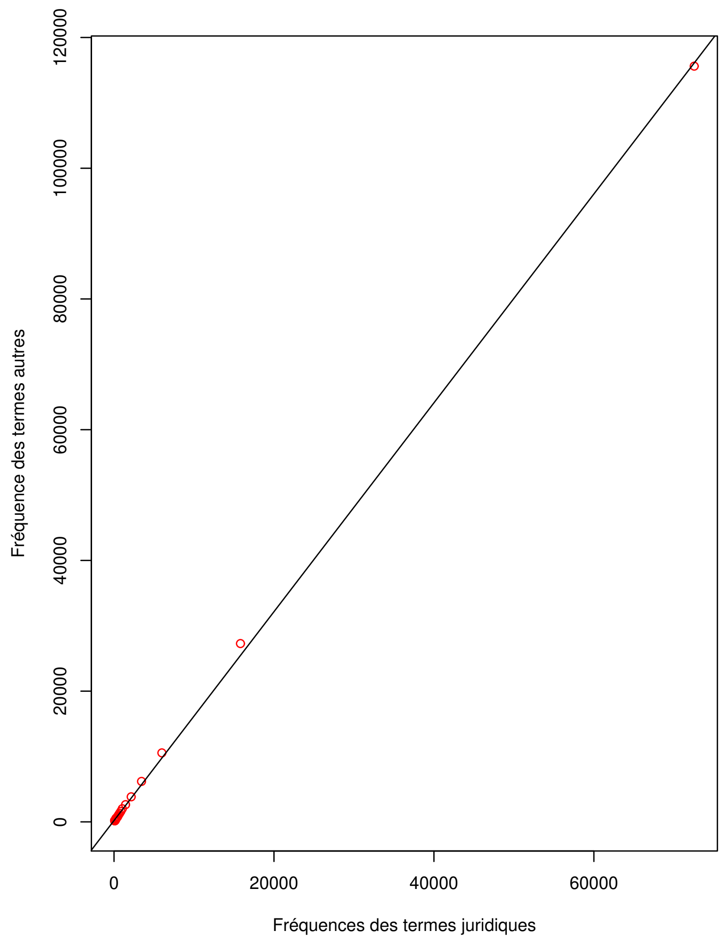

FIG. 5.3 – Nuage de corrélation entre les termes juridiques et les termes autres par fréquences

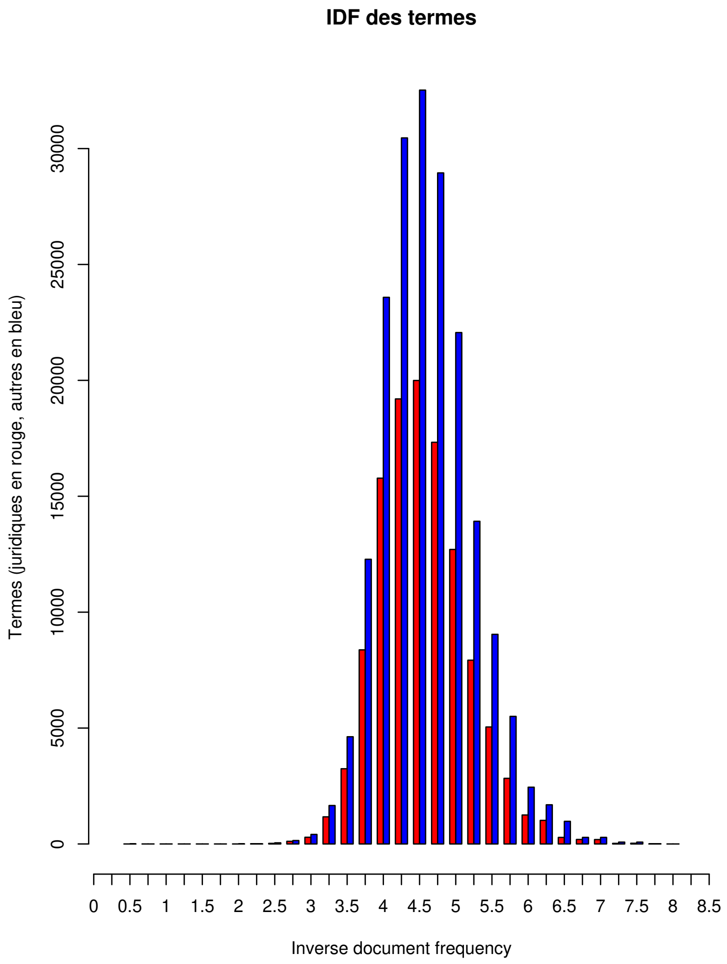

FIG. 5.4 – *Inverse document frequency* des termes juridiques (en rouge) et des termes autres (en bleu)

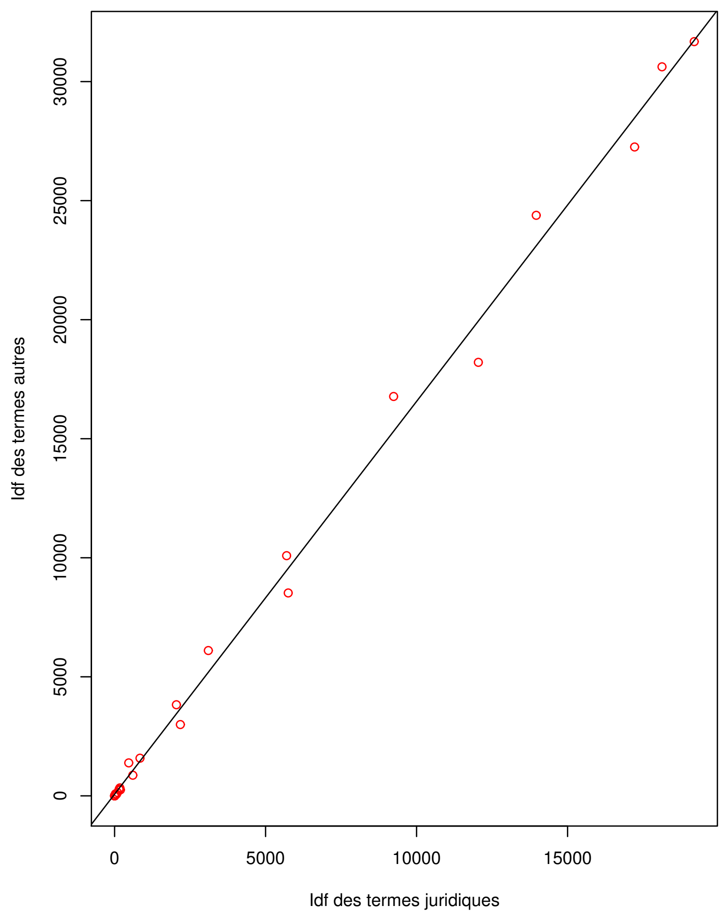

FIG. 5.5 – Nuage de corrélation entre les termes juridiques et des termes autres par indices d'*idf*

La corrélation entre les termes juridiques et les termes autres pour ce qui est de l'indice *tf.idf* est très bonne, le coefficient de corrélation linéaire s'établissant à 0.9996807.

L'entropie : Le calcul de l'entropie de la répartition des termes juridiques dans le corpus n'a également pas permis de dégager des caractéristiques spéciales attribuables aux termes juridiques. Le Graphique 5.7 représente le nombre de termes juridiques et le nombre de termes autres quant à l'entropie de leur répartition. Le graphique concerne les termes dont l'entropie est comprise entre 0 et 2, ce qui correspond à 93 % des termes. Les nombres de termes juridique et de termes autres ont été établis pour chaque valeur de *tf.idf* comprise entre 2 et 60 avec un pas de 0.1.

La corrélation entre les termes juridiques et les termes autres pour ce qui est de l'entropie est très bonne, le coefficient de corrélation linéaire s'établissant à : 0.9999307.

Cette expérience tendant à l'identification de caractéristiques statistiques particulières des termes juridiques par rapport aux autres conduit à la conclusion qu'il ne semble pas y en avoir. Les termes juridiques adoptent les mêmes comportements statistiques que les autres termes. De ce fait, la tentative d'identification de ces termes juridiques par des méthodes statistiques semble vouée à l'échec. Nous ne pouvons donc pas envisager une telle solution pour l'identification des termes du domaine parmi les résultats de l'outil d'extraction de termes.

5.4 Choix de la non validation

Suite à ces expériences visant à l'identification des termes du domaine par des techniques statistiques, nous arrivons à plusieurs constats.

(i) Le premier est celui que les termes du domaine ne sont pas des descripteurs, ou *index terms*, en tous cas, ne peuvent pas être envisagés comme tels. La plupart des méthodes tendant à l'identification de termes descripteurs, comme *idf* ou *tf.idf*, s'attelent en effet à l'identification des termes indicateurs du contenu de documents. Ainsi, l'*inverse document frequency* repose sur le principe que les termes les plus communs apparaissent dans la plupart des documents d'un corpus [Moe00]. De la même façon, le calcul de l'entropie de la répartition des termes permet d'identifier les termes qui créent un désordre dans cette répartition, ce qui revient à identifier les termes discriminants.

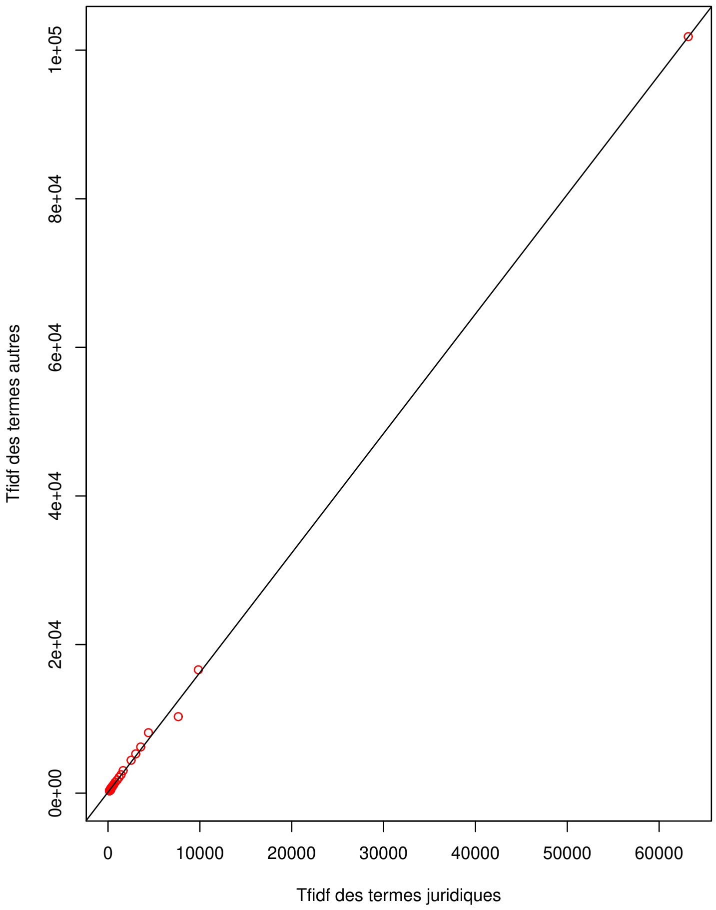

FIG. 5.6 – Nuage de corrélation entre les termes juridiques et des termes autres par indices *tf.idf*

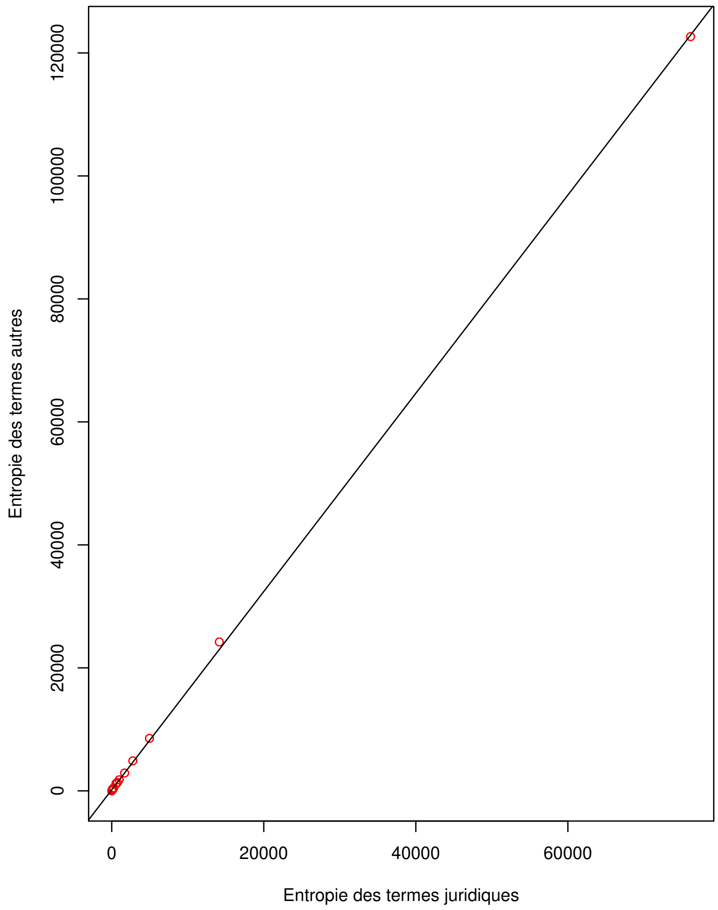

FIG. 5.7 – Nuage de corrélation entre les termes juridiques et des termes autres par indices d'entropie

(*ii*) Or, deuxième constat, les termes du domaine incluent les termes apparaissant souvent dans un corpus. Ainsi, le terme *contrat* ne peut être envisagé comme un bon descripteur pour notre corpus dans le sens où il n'est pas discriminant et apparaît dans beaucoup de documents[7] (voir Tableau 5.12[8]). On ne peut cependant pas l'ignorer en tant que terme du domaine. Les méthodes classiques d'identification des descripteurs, reposant sur des analyses statistiques, ne sont donc pas appropriées à l'identification des termes du domaine dans un corpus en question.

Corpus	Fréquence	Corpus	Fréquence
total	1790	CTRAVAI	1121
CCIVILL	187	CCOMMER	184
CACTSOCL	72	CJUSADM	20
CTRAVMA	4	CAVIACI	5
CCONSOM	6	CASSURA	11
CGLIVPF	4	CPROCPE	9
CEDUCAT	4	CORGJUD	9
CSECSOC	4	CRURAL	10
CPROCIV	22	CENVIRO	15
CPENSIM	2	CCONSTR	10
CPROINT	86	CPENALL	5

Tab. 5.12 – Répartition du terme *contrat* dans le corpus de référence

(*iii*) Enfin, dernier constat, cette expérience d'identification des termes du domaine sur notre corpus de référence nous a menée vers la conclusion que ce corpus est extrêmement ramassé d'un point de vue thématique et que la grande majorité des termes extraits par Syntex de celui-ci peut être envisagée comme des termes du domaine. Nous pourrions d'ailleurs appuyer ce point par le résultat de l'expérience précédente : nous ne pouvons distinguer dans notre corpus de référence les termes juridiques des autres, peut-être est-ce dû au fait que tous les termes extraits de ce corpus par les outils extracteurs peuvent être considérés comme des termes juridiques.

La conclusion à laquelle nous arrivons, suite à cette expérience, est donc que l'ensemble des 300 000 termes extraits de notre corpus de référence peut être envisagé comme l'ensemble des termes du domaine. Nous notons néanmoins que cette conclusion est toute relative à notre expérience et à notre corpus de référence. La reproduction d'une telle conclusion dans un autre cas

[7]Nous envisageons dans ce cas chaque code comme un document du corpus.

[8]La signification des abréviations pour chaque code est consultable en Annexe.

peut certes être envisagée, mais nous ne pouvons présumer de sa pertinence générale.

(*iv*) De cette conclusion résulte une conséquence majeure : celle de la non sélection manuelle dans les sorties de l'outil d'extraction de termes. Cette non sélection induit une non validation des résultats de l'outil automatique. Une phase de validation par un expert des termes est pourtant prévue dans toutes les expériences conduisant à l'utilisation d'un outil automatique d'extraction de termes ([AGBS00b], [AB96], ...). Le choix que nous faisons est guidé principalement par les résultats de l'expérience que nous détaillons ci-dessus. Il est également le résultat d'une tentative de validation menée sur les résultats de Syntex sur notre corpus de référence. Cette tentative, totalement expérimentale, a permis de constater plusieurs points. Le premier et principal enseignement est que la décision de considérer tel terme comme un terme du domaine ou pas est une décision extrêmement complexe dans la matière juridique. Comme nous l'expliquons dans [Lam01c], du fait que le droit appréhende le monde, l'interprète, tout objet du monde peut être un objet du droit. Ainsi, le terme *couteau à pain* peut désigner un objet du droit dès lors qu'un jugement considère le couteau en question comme une arme par destination[9]. Il devient dès lors extrêmement complexe pour un expert de décider si tel terme est effectivement un terme du domaine ou pas. Ainsi, dans l'expérience de validation des résultats de Syntex sur notre corpus de référence, la plupart des termes que nous estimions n'être pas des termes du domaine étaient de fait des termes issus d'erreurs de l'outil. Ces erreurs peuvent être induites par le texte lui-même comportant des erreurs typographiques. Elles peuvent également être le fait de l'outil lui-même qui ne découpe pas le terme au bon endroit. Des exemples d'erreurs telles que détectées dans les sorties Syntex est présenté dans le tableau 5.13.

Syntagme	Remarque
application des articlesL	erreur dans le texte
membres équipage	erreur dans le texte
télécommunications sous réserve	erreur de l'outil
bureaux de poste au nombre minimum	erreur de l'outil

TAB. 5.13 – Exemple de termes erronés identifiés dans les sorties Lexter sur le Code des postes et télécommunications

Les erreurs que nous décrivons apparaissent relativement rarement dans la liste des candidats termes. Nous faisons à leurs propos le choix de ne pas

[9]Tout objet peut être considéré comme une arme dès lors que l'usage qui en est fait en fait un objet dangereux ; on parle alors d'*arme par destination*.

chercher à les identifier pour les éliminer. Ce travail nécessiterait en effet d'une personne, pas nécessairement experte du domaine, une tâche longue, répétitive et peu gratifiante. Ce choix reflète également un point de vue que nous prenons, en tant que non-spécialiste de la discipline, sur les outils de traitement automatique des langues que nous utilisons : celui d'utiliser leurs résultats en tant que tels. Les termes erronés dans la liste des 300 000 termes ne sont donc pas identifiés ni éliminés.

L'ensemble des 300 000 termes extraits par Syntex de notre corpus de référence peut donc être considéré comme l'ensemble des termes du domaine. Nous proposons cependant une sous-liste de ces 300 000 termes rassemblant ce que nous nommons les *termes fondamentaux du domaine* (Section 5.4.1). De cette liste de 300 000 sont exclus les termes contenant des mots déterminés vides de sens (Section 5.4.2), ce qui permet de circonscrire la liste initiale. Enfin, nous abordons dans une Section 6 la question des formes plurielles des termes.

5.4.1 Les *termes fondamentaux du domaine*

Notre démarche consiste donc à considérer dans un premier temps l'ensemble des 300 000 termes tels qu'extraits par Syntex de notre corpus de référence. Une distinction doit cependant être menée parmi eux. Ainsi, lors de la phase de détermination des liens entre les termes, certains calculs ne peuvent être envisagés sur la totalité d'une liste de 300 000 termes. De plus, la phase de modélisation, qui est manuelle, ne peut être envisagée sur l'ensemble des 300 000 termes. Pour ces deux raisons, nous distinguons, parmi la liste des 300 000 termes, certains, considérés comme plus fondamentaux.

Ces *termes fondamentaux du domaine* sont identifiés en utilisant la structure même du corpus de référence, ce qui est appelé communément *discourse structure* ([Moe00]). Ainsi, l'importance de termes peut être présumée au vu de la place spéciale qu'ils occupent dans la structure du document en question. Un terme présent dans le titre d'un document, dans son résumé ou dans son premier paragraphe peut être identifié comme un terme plus important que les autres.

Nous identifions pour notre part les termes présents dans les intitulés des divisions des codes de notre corpus de référence. Il s'avère en effet que chaque code est thématiquement divisé en sections, chacune de ces sections faisant l'objet d'un intitulé. Nous avons donc rassemblé, dans un même fichier, l'ensemble des intitulés des sections des codes constituant notre corpus de référence, ce qui représente un total de 5 560 intitulés. Ce fichier a été soumis au traitement de Syntex. De la même façon que précédemment, nous n'avons retenu que les noms et les syntagmes nominaux, en ignorant les termes comprenant des caractères non alphabétiques. Nous obtenons alors une liste de 16 681 termes que nous nommons *termes fondamentaux du domaine*.

5.4.2 Les termes vides

Sur la base de la liste des 16 000 *termes fondamentaux du domaine*, est rapidement élaborée une liste de termes vides ([Lam01c]). Un simple calcul de la fréquence globale des termes fondamentaux du domaine dans les intitulés nous a permis d'identifier un ensemble de 22 termes n'étant pas, à proprement parler, des termes du domaine. Ces termes sont ainsi le reflet de la propre organisation des codes constituant notre corpus de référence. Ces codes sont divisés, comme nous l'avons dit, en sections qui sont des *livres*, des *titres*, des *chapitres* et des *sections*. Chacun de ces termes apparaît conséquemment dans les intitulés même des codes. De plus, la plupart des codes sont organisés de telle manière que sont d'abord traitées des dispositions générales relatives au code et ensuite des dispositions particulières. De ce fait, ces termes apparaissent dans les intitulés. Les termes identifiés comme termes vides sont listés dans le Tableau 5.14.

Terme	Fréquence	Terme	Fréquence
code	111	*relatives*	67
titre	111	*applicables*	59
dispositions	109	*ier*	57
chapitre	107	*dispositions relatives*	55
générales	96	*paragraphe*	45
section	95	*décret*	41
partie	94	*sous-section*	41
livre	91	*bis*	39
particulières	89	*législative*	39
dispositions générales	87	*partie législative*	36
dispositions communes	80	*règles générales*	34

TAB. 5.14 – Termes vides

Ces termes identifés et validés vides de sens sont alors considérés dans notre approche comme n'étant pas des termes du domaine. Il sont à ce titre exclus de la liste des 300 000 termes. Un processus de reconnaissance de formes, ou *pattern matching* est alors opéré sur la liste des 300 000 termes, il consiste à en éliminer les termes correspondant exactement à un de ces termes vides ou en contenant un. Ce processus réduit le nombre de termes à 118 000. Nous considérons donc cette liste de 118 000 termes comme les termes du domaine.

5.4.3 Les formes plurielles

Nous nous attardons dans cette Section sur les formes au pluriel des termes. Les noms et les syntagmes nominaux que nous envisageons comme des termes du domaine se trouvent, la plupart du temps, dans les textes, à la fois dans leur forme au singulier et dans leur forme au pluriel. Nous n'entendons pas régler toutes les questions afférentes au traitement automatique de la langue relativement aux formes plurielles et singulières des termes. Ces points font l'objet d'études spécialisées. Nous devons simplement envisager le problème et lui trouver des solutions acceptables dans notre contexte et en fonction de nos capacités.

Dans notre objet qui est celui de détecter les termes étiquettes de concepts juridiques, nous simplifions la problématique en avançant qu'un terme, quel que soit son nombre, désigne un seul et même concept[10]. Par exemple, les termes *acte* et *actes* désignent bien le même concept. Nous considérons alors un seul et même terme, en retenant sa forme au singulier. Il s'avère cependant que certains termes n'apparaissent dans notre corpus de référence que dans leur forme plurielle ou, à tout le moins, le plus fréquemment. Il en va ainsi des termes *acquêts* ou *servitudes*. Nous conservons dès lors le terme dans sa forme au pluriel.

Nous utilisons les formes lemmatisées des termes fournies par Syntex. Pour ce qui est des syntagmes nominaux, à une seule forme lemmatisée peut correspondre plusieurs syntagmes tels que trouvés dans les textes. Par exemple, à la forme lemmatisée *acte de huissier de justice* correspond les formes suivantes : *actes des huissiers de justice*, *acte d'huissier de justice*, *actes d'huissier de justice*. La difficulté réside alors dans le choix de la forme à afficher à l'usager[11]. Pour l'instant, le choix dans la liste des formes se fait au hasard, ce qui, nous en convenons, est une solution guère satisfaisante. Le problème resurgit lors de l'interfaçage de l'ontologie, nous l'évoquons dans la Section 8.1.1.

---

[10]Nous notons que cette approche est bien une simplification puisque cette position n'est pas toujours exacte et certains termes n'ont pas la même signification dans leur forme au singulier et dans leur forme au pluriel.

[11]Nous ne souhaitons en effet pas afficher la forme lemmatisée des termes à l'usager.

Nous avons donc procédé à l'extraction des termes du domaine. Ces termes du domaine constituent un des éléments fondamentaux de la ressource ontologique et donc de l'ontologie. L'autre élément fondamental en est l'ensemble des relations que l'on peut établir entre ces termes du domaine. Nous décrivons la démarche d'identification des liens entre termes dans le Chapitre suivant.

Chapitre 6

Identification de liens entre termes

Dès lors nous avons défini les termes du domaine, les relations entre ceux-ci peuvent être identifiées. L'identification des liens entre les termes procède également d'un processus d'acquisition de connaissances à partir de textes. Nous nous basons sur le corpus de référence pour identifier les relations pouvant exister entre les termes du domaine. De la même façon que pour l'extraction des termes du domaine, l'idée sous-jacente à notre démarche est que les relations entre les termes du domaine sont exprimées dans les textes par des structures linguistiques particulières. L'étude de ces structurations linguistiques permet donc de dégager des liens entre ces termes. Des étude statistiques peuvent également être menées, qui permettent de mettre à jour des relations entre termes. Ces études statistiques reposent, elles, sur le postulat principal que des termes apparaissant dans des contextes similaires peuvent être rapprochés.

Les liens que nous obtenons dans cette phase sont non qualifiés sémantiquement ; ils sont les éléments de la ressource ontologique.

Plusieurs méthodes tendant à l'identification de liens ont été appliquées sur le corpus de référence tel que défini dans la Section 4.

Nous entendons par relations un couple de termes liés par une des méthodes ci-dessous explicitées. Un quelconque sens dans la relation n'est pas considéré et la relation Terme1 Terme2 équivaut à la relation Terme2 Terme1. Une seule des relations Terme1 Terme2, Terme2 Terme1 est comptabilisée lorsque nous parlons de nombre de relations.

6.1 Analyse des relations de coordinations

6.1.1 La méthode

Une des méthodes permettant l'identification de liens entre termes consiste à repérer entre ces termes, dans un corpus donné, des marqueurs lexico-syntaxiques susceptibles d'établir une relation entre ces termes. Une telle démarche se rapproche, de façon très simplifiée, de la démarche décrite dans [Mor99] d'identification de *schémas lexico-syntaxiques* caractéristiques de relations sémantiques données. Notre objet n'est cependant pas, comme dans le système PROMETHE ([Mor99]), de définir à la main un ensemble de termes liés par une relation sémantique donnée pour ensuite en déduire plus ou moins automatiquement des schémas lexico-syntaxiques permettant d'extraire de nouveaux couples de termes liés par cette relation. Nous nous sommes, pour notre part, contenté de considérer *a priori* que les conjonctions de coordinations *et* et *ou* entre des termes dans un corpus permettaient d'établir des liens entre ces termes [Lam00].

Notre démarche a consisté à détecter, grâce à un programme Perl parcourant notre corpus, l'ensemble des séquences correspondant à ces patrons :

SN1 *et* SN2
SN1 *ou* SN2

SN correspondant à un syntagme nominal connu comme tel, élément de notre liste de termes du domaine.

6.1.2 La pratique, résultats

Nous obtenons d'un tel processus un ensemble de séquences susceptibles d'héberger une relation entre les termes SN1 et SN2. Une étape de validation est alors nécessaire ; elle doit être opérée par une personne experte du domaine.

Nous ne pouvions en conséquence envisager de considérer toutes les séquences correspondant à ces patrons dans la totalité de notre corpus. Nous avons donc, dans cette démarche, uniquement considéré les intitulés des divisions des codes. Ainsi ([BL02]), sur l'ensemble des intitulés des codes, 5 660 séquences hébergeant de tels patrons lexico-syntaxiques ont été identifiées. La validation manuelle de ces résultats, qui a nécessité une quinzaine d'heures, permet d'identifier 2 580 relations entre 3 762 termes différents.

Un filtre a été opéré sur ces résultats présentés dans [BL02]. En effet, notre intérêt se porte, nous l'avons dit, sur les noms et les syntagmes nominaux. Or la détection automatique des relations par conjonctions de coordination et sa validation, établissent des relation entre termes qui ne sont pas forcément des noms ou syntagmes nominaux. Parmi les relations détectées de cette manière, nous avons donc procédé à une sélection qui consiste à ne conserver la relation détectée et validée que si elle concerne des noms ou syntagmes nominaux. Nous avons donc comparé ces résultats à la liste de tous les termes noms et syntagmes nominaux tels qu'extraits par Syntex du corpus de référence. Certaines de ces relations peuvent également s'établir entre termes non connus en tant que termes du domaine. Nous avons fait le choix de les conserver cependant, c'est alors une façon de prendre en considération des termes non sélectionnés en tant que termes du domaine et de, par là, élargir cette liste de termes du domaine.

On obtient alors un ensemble de 2 356 relations entre 2 974 termes différents.

Le Tableau 6.1 liste un ensemble de liens établis entre termes par la détection et la validation de relations de coordination.

SN1	SN2	Code
aérodromes	*installation à usage aéronautique*	CAVIACIL
absence	*insuffisance*	CGCTERRM
accident du travail	*accident du trajet*	CSECSOCA
artisans	*commerçants*	CCOMMERL
camping	*stationnement des caravanes*	CURBANIA
déductions fiscales	*réductions d'impôt*	CGIMPX20
enseignement	*recherche*	CSANPUBD
funérailles	*lieux de sépulture*	CGCTERRL
magistrats	*greffiers*	CPROCPEA

TAB. 6.1 – Exemples de liens entre termes détectés par les relations de conjonctions

L'analyse de la coordination des termes dans les codes permet de détecter des polysémies internes au domaine. En effet, d'un code à l'autre, un même terme peut être coordonné à différents autres termes. Certaines relations de coordination sont stables selon les codes : c'est ainsi qu'*action publique* et *instruction* sont coordonnés à la fois dans les intitulés du Code de justice militaire et dans ceux du Code de procédure pénale, et qu'*actes de terrorisme* et *infractions* sont coordonnés à la fois dans les intitulés du Code des assurances et dans ceux du Code des impôts. Par contre, l'analyse des contextes de coordination des mots comme *liquidation* ou *planification* fait apparaître aisément la polysémie de ces termes dans le domaine (Tableau 6.2).

SN1	SN2	Code
liquidation	*acquittement*	CDYANES
liquidation	*paiement*	CTRAVMA
liquidation	*recouvrement*	CPROCPE
liquidation	*recouvrement*	CURBANI
liquidations	*ventes au déballage*	CCOMMER
planification	*conseil familial*	CSANPUBN
planification	*intervention économique*	CCOMMUN

TAB. 6.2 – Exemples de polysémies détectées

6.2 Analyse distributionnelle

Des relations entre termes peuvent être établies sur la base d'analyses distributionnelles menées sur les termes dans le corpus de référence. Nous utilisons pour ce faire les résultats de l'analyse distributionnelle menée par le module Upery de Syntex [Bou02]. L'analyse distributionnelle consiste à rapprocher les termes en fonction des contextes syntaxiques qu'ils partagent (on parle de lien *régi* entre termes *voisins*) et, respectivement, les contextes syntaxiques en fonction des termes qu'ils partagent (on parle alors de lien *recteur* entre contextes voisins).

6.2.1 la méthode

Comme nous le décrivons dans la section 5, Syntex établit un réseau de dépendances qui, à chaque candidat terme associe ses contextes syntaxiques. Ainsi, comme nous le montrons dans [BL02], à partir de l'analyse syntaxique de la phrase "*Les agents bénéficient de la formation professionnelle*", Syntex établit un ensemble de couples termes-contextes (voir Tableau 6.3[1])

Ces données sont ensuite utilisées par le module Upery d'analyse distributionnelle pour effectuer un calcul des proximités distributionnelles entre les termes du réseau de dépendances, qu'ils soient des mots ou des syntagmes nominaux[2].

[1] *SUJ* pour sujet du verbe, *ADJ* pour adjectif du nom et *DE* pour complément en *de*

[2] On parle alors d'*analyse distributionnelle étendue* car prenant en compte non seulement les mots mais aussi les syntagmes.

Terme	Contexte syntaxique
agent	*bénéficier* _SUJ
agent	*bénéficier de formation professionnelle* _SUJ
agent	*bénéficier de formation* _SUJ
professionnel	*formation* _ADJ
formation professionnelle	*bénéficier* _DE
formation	*bénéficier* _DE
formation professionnelle	*agent bénéficier* _DE
formation	*agent bénéficier* _DE

TAB. 6.3 – Données extraites par l'analyse distributionnelle à partir de l'analyse syntaxique de la phrase *Les agents bénéficient de la formation professionnelle*

Deux termes peuvent être voisins par un lien régi ou par un lien recteur. Ainsi ([BL02]) les noms *enfant* et *mineur* sont voisins de type régi dans le code civil du fait qu'ils partagent les contextes suivants : (*atteindre l'âge*, SUJ), (*représentant légal*, DE), (*émanciper*, OBJ), (*moralité*, DE), (*tuteur*, DE), (*adoption*, DE), (*placer*, OBJ), (*recueillir*, OBJ) etc. Les noms *rémunération* et *salaire* sont voisins de type recteur dans le code du travail car ils font partie de contextes de type adjectival avec les adjectifs *perçu*, *versé*, *offert*, *horaire*, *net*, *inférieur* et *dû*.

L'identification des syntagmes dans les textes et les données utilisées dans l'analyse distributionnelle sont liées. Chaque couple terme-contexte correspond à un syntagme, et réciproquement. Ainsi le syntagme *représentant légal de l'enfant* génère, entre autre, le couple (*enfant*, *représentant légal* _DE). On peut cependant distinguer l'extraction des termes complexes et la génération des couples terme-contexte utilisés pour l'analyse distributionnelle. Une restriction de fréquence des syntagmes dans le corpus en question peut être posée à la génération des couples terme-contexte correspondant, on parle de *fréquence de syntagmes générateurs*. Cette restriction n'étant pas forcément valable pour l'extraction du syntagme en tant que tel.

Le module Upery calcule alors pour chaque terme et contexte les coefficients suivants :

- la productivité d'un terme z $prod_z(c)$ qui correspond au nombre de contextes avec lesquels ce terme apparaît, et respectivement, la productivité d'un contexte w $prod_w(t)$ qui correspond au nombre de termes avec lesquels ce contexte apparaît ;

- le coefficient $a_{(x,y)}$ qui établit, pour chaque paire *(x,y)* de termes (ou de contextes), le nombre de contextes (ou respectivement de termes) partagés ;
- le coefficient *prox* permet de tenir compte du fait que, plus un contexte partagé par deux termes est productif, moins sa contribution au rapprochement des deux termes est importante, et respectivement pour les termes partagés par les contextes ; *prox* est établi comme suit (c est le contexte partagé) :

$$prox = \sum_{i=1}^{n} \frac{1}{prod\,(c_i)^{\frac{1}{2}}} \tag{6.1}$$

- les coefficients j_1 et j_2 permettent de prendre en compte, lors de l'évaluation de la proximité entre deux termes, les contextes partagés (a) mais aussi les contextes que chacun des termes a en propre ; j_1 et j_2 s'établissent donc comme suit pour les termes t_1 et t_2 :

$$j_1 = \frac{a_{t_1,t_2}}{prod\,(t_1)} \tag{6.2}$$

$$j_2 = \frac{a_{t_1,t_2}}{prod\,(t_2)} \tag{6.3}$$

Pour chacun des couples de termes et des couples de contextes de notre corpus de référence, le module Upery a donc établi l'ensemble de ces coefficients.

6.2.2 En pratique, l'établissement de liens

Ces coefficients ont pour fonction de donner la possibilité de *resserrer*, ou non, les résultats de l'analyse distributionnelle. Ceci repose sur l'*hypothèse distributionnelle* de Harris [Har71], fondement de l'analyse distributionnelle, qui lie la distribution syntaxique des mots (ou termes) à leur signification. De façon schématique, plus deux termes partagent des contextes syntaxiques, plus ils sont proches dans leurs significations. Ainsi, plus les coefficients a, *j1*, *j2* et *prox* seront élevés pour une relation établie entre deux termes, plus cette relation a des chances d'être pertinente, et inversement.

A titre d'exemple, le tableau 6.4 classe par ordre décroissant des indices a, *prox* et *j1*, *j2* les termes liés au terme *action sociale* par le module Upery dans le Code de l'action sociale et des familles. Le classement tient compte en premier lieu du coefficient a, puis du coefficient *prox* et enfin des coefficients *j1* et *j2*.

Terme1	Terme2	*a*	*j1*	*j2*	*prox*
action sociale	*éducation*	5	0.28	0.25	0.99
action sociale	*personnes*	4	0.22	0.02	0.97
action sociale	*établissements*	4	0.22	0.03	0.96
action sociale	*aide sociale*	4	0.22	0.08	0.93
action sociale	*sécurité sociale*	4	0.22	0.31	0.89
action sociale	*budget*	3	0.17	0.27	0.68
action sociale	*affaires*	3	0.17	0.38	0.59
action sociale	*emploi*	3	0.17	0.30	0.59

TAB. 6.4 – Relations établies avec le terme *action sociale* en fonction des coefficients de l'analyse distributionnelle

En se basant sur l'*hypothèse distributionnelle* de Harris, on peut établir que les termes *action sociale* et *éducation* ont une signification plus proche que les termes *action sociale* et *affaires* ou les termes *action sociale* et *emploi*. L'évaluation de ce genre de résultats ne peut être que suggestive, de la même façon que l'interprétation qu'elle nécessite. Nous pouvons simplement avancer que le Code de l'action sociale et des familles associe plus souvent la notion d'*action sociale* à celle d'*éducation* qu'à celle d'*emploi*. La sélection des relations entre termes établies par l'analyse distributionnelle de Syntex sur la base des coefficients *a*, *prox*, *j1* et *j2* permet donc de *resserrer* la précision de ces relations.

Reste, dès lors, à déterminer des seuils pour chacun de ces coefficients, dans le but de ne détecter que des relations *pertinentes* entre les termes[3]. Une telle démarche sur les coefficients est envisageable dans ce cas, contrairement à celui de la sélection des termes du domaine. Dans ce dernier cas, aucune hypothèse ne peut en effet être établie sur les caractéristiques statistiques des termes permettant de distinguer ceux qui seraient des termes du domaine des autres (Section 5.3).

La détection de relations *pertinentes* suggère que les relations détectées par l'outil à partir des seuils fixés sont interprêtées et validées par une personne. Logiquement, plus les coefficients tels que décrits ci-dessus sont élevés, plus les relations détectées sont valides mais moins leur nombre est élevé. Il faut donc établir des seuils établissant un équilibre entre le nombre de rela-

[3]Sachant que notre objet est de détecter des liens entre des termes et que nous n'envisageons alors que les noms et les syntagmes nominaux, nous ne tenons compte des résultats du module Upery que sur ceux-ci.

tions détectées et la qualité de celles-ci.

La détermination de ces seuils se fait, et ne peut se faire, qu'avec une démarche plus ou moins empirique.

Nous avons cependant mené une expérience, tentant de rationaliser quelque peu l'établissement des seuils des coefficients *a*, *j1*, *j2* et *prox* de détection de relations entre termes par l'analyse distributionnelle étendue de Syntex.

Nous avons utilisé les relations détectées par l'analyse et la validation des coordinations dans les intitulés des codes telles que décrite dans la Section 6.1. Ces relations ont été comparées à celles détectées par l'analyse distributionnelle de Syntex. Nous avons ainsi établi la moyenne des coefficients *a*, *j1*, *j2* et *prox* pour l'ensemble des relations communes aux deux méthodes de détection de relations entre termes.

Cette expérience a été conduite sur un ensemble de 12 codes dont la liste est disponible en Annexe. La fréquence des syntagmes générateurs était fixée à 1. Les seuils des coefficients étaient fixés comme suit : *a* à 3, *prox* à 0.5 et *j1* ou *j2* devait être supérieur à 0.25. L'analyse distributionnelle de Syntex détecte alors un ensemble de 27 397 relations entre termes, incluant les relations détectées entre les adjectifs, les syntagmes verbaux etc. Le nombre de relations communes à l'analyse des coordinations et à l'analyse distributionnelle de Syntex est de 126 ; quelques unes de celles-ci sont listées dans le Tableau 6.5.

Terme1	Terme2
action civile	*action publique*
apprentissage	*formation professionnelle continue*
contributions indirectes	*taxes*
contrôle judiciaire	*détention provisoire*
directoire	*conseil de surveillance*
employeur	*travailleurs*
impôts directs	*taxes*
lois	*réglements*
redressement	*liquidation*

TAB. 6.5 – Exemples de relations entre termes communes à l'analyse des coordinations et à l'analyse distributionnelle de Syntex

Nous notons que le nombre de relations entre termes détectées par l'analyse des coordinations dans les 12 codes en question est de 533. Toutes les relations détectées par l'analyse des coordinations ne sont donc pas *confir-*

mées par l'analyse distributionnelle de Syntex. Les deux méthodes sont alors complémentaires quand à la détection de relations entre termes.

Pour ces 126 relations détectées par l'analyse des coordinations et *confirmées* par l'analyse distributionnelle de Syntex, la moyenne des coefficients s'établit comme suit (Tableau 6.6) :

Coefficient	Moyenne	Maximum	Minimum
a	9.34	54	3
j1	0.24	1	0.04
j2	0.28	1	0.04
prox	2.48	13.8	0.5

TAB. 6.6 – Moyenne, valeurs maximales et minimales des coefficients *a*, *j1*, *j2* et *prox* de l'analyse distributionnelle pour les relations communes avec l'analyse des coordinations

Si une sélection est faite, parmi les relations détectées par l'analyse distributionnelle de Syntex, sur la base des coefficients *a*, *j1*, *j2* et *prox* fixés à leurs moyennes respectives établies dans l'expérience, le nombre de relations est de 105. Ce nombre, inférieur au nombre de relations communes à l'analyse des coordinations et à l'analyse distributionnelle de Syntex, s'explique par le fait que nous avons utilisé les moyennes des coefficients. Quoi qu'il en soit, les seuils ainsi fixés le sont de façon bien trop drastique, notamment les coefficients *a* et *prox*.

6.2.3 Des seuils empiriques

Nous nous sommes donc résolus à adopter une démarche totalement empirique conduisant à la détermination des seuils des coefficients. Cette démarche a consisté à isoler des relations détectées par l'analyse distributionnelle de Syntex selon les valeurs de coefficients. Les coefficients *a*, *j1* et *j2* et *prox* ont été laissés au même niveau que pour l'expérience sur les 12 codes. Le coefficient *a* est donc fixé à 3, signifiant que les contextes partagés doivent au minimum être de 3, un des coefficients *j1* ou *j2* doit être supérieur à 0.25, signifiant que l'un des deux termes doit partager au moins le quart de ses contextes avec l'autre, et le coefficient *prox* doit être supérieur à 0.5. Seule, la fréquence des syntagmes générateurs a été restreinte pour être fixée à 2, ce qui signifie que les couples terme-contexte ne sont déduits que des syntagmes apparaissant au moins deux fois dans le corpus. C'est la fixation de ce dernier seuil à 2 qui est la plus restrictive et qui resserre qualitativement les résultats de l'analyse distributionnelle.

Syntex a été lancé sur l'ensemble des codes dont nous disposons. Chacun des codes a fait l'objet d'un traitement particulier ce qui signifie que les relations que nous établissons sur la base de l'analyse distributionnelle de Syntex le sont par code ; nous avons donc une indication sur la provenance des relations qui s'établissent entre termes d'un même code.

L'analyse distributionnelle de Syntex sur l'ensemble de notre corpus de référence[4] donne un total de 16 404 relations entre 3 283 termes différents, sachant que les même règles prévalant à la sélection des termes du domaine ont été appliquées[5]. Nous rappelons que si la relation du type Terme1 Terme2 est trouvée, ainsi que la relation Terme2 Terme1, cela est comptabilisé comme une seule relation.

Le Tableau 6.7 liste, à titre d'exemple, des relations établies par l'analyse distributionnelle de Syntex sur notre corpus de référence.

Terme1	Terme2	Code
faute	*crime*	CPROCPE
faute	*infraction*	CPROCPE
identité	*adresse*	CSANPUB
identité	*domicile*	CGIMPOX
identité	*domicile fiscal*	CGIMPOX
identité	*infraction*	CPROCPE
matériel	*entretien*	CORGJUD
matériel	*mobilier*	CSECSOC
matériels	*véhicule*	CROUTE

TAB. 6.7 – Exemples de relations entre termes détectées grâce à l'analyse distributionnelle de Syntex

Une étude empirique et non quantifiée sur les relations établies par l'analyse distributionnelle a été menée. L'ensemble de ces relations a ainsi été estimée *acceptable* dans le sens où une grande majorité a été identifiée comme sémantiquement valide.

Le nombre de relations établies par l'analyse distributionnelle est relativement élevé. Le nombre de relations moyennes par termes s'établit à 11.8[6].

[4]A l'exception notable du code de la caisse d'épargne et du code du travail maritime. Le premier n'était pas en notre possession à la date de l'expérience et le deuxième ne peut être analysé par Syntex ; il provoque une erreur de l'outil d'étiquetage préalable à Syntex.

[5]Seuls les noms et syntagmes nominaux sont pris en compte. Sont de plus éliminés les termes contenant des caractères non alphabétiques et des lettres majuscules.

[6]L'établissement de cette moyenne de relations attribuées par termes se fait sur la base

Cette moyenne cache de fortes disparités comme le montre le Tableau 6.8 associant des termes aux nombres de relations qui leur sont attribuées par l'analyse distributionnelle de Syntex.

Terme	Nbre relations
attestation	4
commission permanente	4
conception	12
convocation	6
déclaration	330
délégué	2
demande	178
droit	94
électeurs	4
modalités de fonctionnement	2
personne	474
utilisation	2
valeur	244

TAB. 6.8 – Exemples de nombres de relations détectées grâce à l'analyse distributionnelle de Syntex par termes

6.2.4 La sélection du nombre de relations

L'usage auquel l'ontologie finale est dédié prend alors toute sa place dans cette tache d'ingénierie des connaissances consistant à établir des liens entre des termes. Notre ontologie est dédiée à la recherche d'informations sur le Web, et plus précisement, à l'aide à la reformulation de requêtes. Il ne peut dès lors être envisagé de suggérer un nombre trop important de termes relatifs à une requête donnée. L'utilité d'un tel système étant bien de suggérer un nombre de termes raisonnables sans *perdre* l'usager avec un trop grand nombre d'informations. On ne peut ainsi envisager de suggérer plus de 330 termes à un usager ayant saisi une requête comprenant le mot *déclaration*.

Sachant que l'ensemble des relations détectées par l'analyse distributionnelle sur le corpus de référence est estimé sémantiquement acceptable, le problème à résoudre se trouve alors être celui de limiter le nombre de relations s'avérant trop élevé pour certains termes, tout en conservant les relations

de nombre total de relations, celles du type Terme1 Terme2 et celles du type Terme2 Terme1.

établies en nombre raisonnable pour d'autres. L'usage commande de limiter le nombre de relations pour certains termes, ceux dont le nombre de relations est trop élevé.

Nous ne pouvons, pour ce faire, envisager de faire varier les coefficients *a*, *prox*, *j1* et *j2*. Il s'avère en effet qu'en étant plus restrictif sur ces coefficients, le nombre de relations ainsi que le nombre de termes concernés se trouve trop grandement affecté. A titre d'exemple, le Tableau 6.9 liste le nombre de relations établies et le nombre de termes différents qu'elles concernent en fonction des coefficients *a*, *prox*, *j1* et *j2*. Ces données sont établies à partir des résultats de Syntex sur le corpus d'expérience de 12 codes précédemment évoqué.

Coefficients	Nbre relations	Nbre termes
$a \geq 3$, $prox \geq 0.5$, $j1 \| j2 \geq 0.25$	27 397	3 283
$a \geq 4$, $prox \geq 0.6$, $j1 \| j2 \geq 0.3$	8 296	1 888
$a \geq 5$, $prox \geq 0.7$, $j1 \| j2 \geq 0.4$	2 608	948
$a \geq 6$, $prox \geq 0.8$, $j1 \| j2 \geq 0.5$	912	433
$a \geq 7$, $prox \geq 0.9$, $j1 \| j2 \geq 0.6$	262	168
$a \geq 8$, $prox \geq 1$, $j1 \| j2 \geq 0.7$	72	54

Tab. 6.9 – Nombre de relations et de termes différents en fonction des coefficients *a*, *prox*, *j1* et *j2* de l'analyse distributionnelle

L'approche que nous adoptons pour limiter le nombre de relations pour les termes dont ce nombre est trop élevé consiste à fixer un seuil de nombre acceptable de relations par termes et, au-delà de ce seuil, à limiter le nombre de relations en fonction du nombre de codes dans lesquels elles apparaissent. L'analyse distributionnelle est en effet réalisée code par code. En conséquence, chaque relation détectée entre deux termes peut l'avoir été dans plusieurs codes. A titre d'exemple, le terme *personnes morales* est lié au terme *sociétés* dans le Code des impôts et dans le Code de la construction. On peut à ce titre établir que plus une relation est détectée dans des codes différents, plus cette relation est *importante*[7]

Les différentes étapes de notre approche sont les suivantes :

- la fixation d'un seuil établissant le nombre acceptable de relations établies pour chaque terme. Sachant qu'aux relations issues de l'analyse distributionnelle seront ajoutées, le cas échéant, les relations détectées

[7] L'*importance* s'entend alors relativement aux relations qui ne sont détectées que dans un code.

par ailleurs, nous avons fixé de façon totalement empirique ce seuil à 10, les termes dont le nombre de relations issues de l'analyse distributionnelle est inférieur à 10 ne sont donc pas concernés par les étapes suivantes ;

- pour les termes dont le nombre de relations est supérieur à 10, nous éliminons les relations n'apparaissant que dans un seul code ;
- si à la suite de l'étape précédente, le nombre de relations reste supérieur à 10, nous ne prenons alors en considérations que les 10 relations apparaissant dans le plus de codes ;
- les termes dont aucunes des relations ne sont détectées dans plus d'un code sont rejetés dans le sens où toutes les relations qui leur sont attribuées sont ignorées. Cette option peut paraître par trop radicale. Nous n'avons identifié aucun critère permettant de sélectionner les relations. Nous aurions cependant pu choisir de sélectionner 10 relations par un procédé aléatoire. Le nombre de termes concernés est cependant relativement faible (96) ce qui correspond à 0.25% du total[8]. Le Tableau 6.10 liste un ensemble de termes dont les relations ne sont détectées que dans un seul code.

Terme	Nbre relations
activités d'assistance médicale	28
créancier	49
habitants	30
infirmier	25
ministre chargé du logement	19
médicaments	88

Tab. 6.10 – Termes dont les relations issues de l'analyse distributionnelle ne sont détectées que dans un seul code

En appliquant les étapes précédentes sur les résultats de l'analyse distributionnelle de Syntex (16 404 relations entre 3 283 terme différents), nous obtenons un total de 8 167 relations entre 2 987 termes différents. La moyenne des relations par terme s'établit alors à 3.3[9]. Le Tableau 6.11 donne un échantillon de relations détectées par l'analyse distributionnelle sur le corpus de référence, après la phase de sélection explicitée ci-dessus. L'ensemble des relations détectées par l'analyse distributionnelle pour le terme *création* est listé

[8]Le total correspond à l'ensemble des termes différents concernés par les 16 404 relations détectées par Syntex dans le corpus de référence

[9]L'établissement de cette moyenne se faisant en comptabilisant chacune des relations du type Terme1 Terme2 et Terme2 Terme1.

en Annexe –, ainsi que les relations sélectionnées par le procédé qui nous explicitons ci-dessus.

Terme1	Terme2
juridiction d'instruction	*chambre*
juridiction d'instruction	*tribunal*
juridiction d'instruction	*cour*
juridiction d'instruction	*cour d'appel*
juridiction d'instruction	*juge*
juridiction d'instruction	*chambre de l'instruction*
kilogrammes	*tonnes*
location	*vente*
location	*cession*

Tab. 6.11 – Exemple de termes reliés par l'analyse distributionnelle, après sélection

6.3 Analyse statistique

6.3.1 La méthode

Une autre méthode pour établir des relations entre les termes d'un corpus repose sur des analyses statistiques des textes. Elle consiste à lier les termes qui partagent des contextes identiques, quelles que soient les catégories syntaxiques de ces contextes. Le contexte des termes utilisé dans cette analyse statistique se différentie de celui utilisé lors de l'analyse distributionnelle. Le contexte statistique, souvent appelé *fenêtre de mots* est composé de l'ensemble des termes entourant le terme en question dans les textes en question. L'analyse distributionnelle et l'analyse statistique des contextes des termes ne reposent donc pas sur les mêmes principes et peuvent conséquemment être envisagées comme complémentaires l'une de l'autre.

L'idée sous-jacente de l'analyse statistique est que peuvent être rapprochés des termes qui, non pas cooccurrent souvent, mais qui apparaissent souvent avec des contextes similaires. Ainsi deux termes apparaissant souvent dans les fenêtres de mots, ou contextes, proches seront reliés, ces deux termes n'apparaissant pas forcément dans les mêmes documents. Une telle méthode est décrite notamment dans [GWR99]. Le choix de la longueur de la fenêtre de mot est fondamental dans cette méthode. Plus ce contexte, une phrase, un paragraphe ou un document, est spécifique, plus les relations établies entre les termes seront précises ([Moe00], p. 110). L'auteur ([GWR99]), se basant sur des études antérieures, note également que ce genre de méthodes donne des résultats acceptables si tant est qu'elles sont appliquées sur de gros corpus de textes comprenant un vocabulaire spécialisé. Enfin, l'auteur note que ce genre de techniques permet effectivement de détecter des relations de divers degrés entre des termes (synonymie, quasi synonymie, termes spécifiques, termes génériques etc) mais qu'en aucun cas la nature de ces relations n'est identifiée.

L'analyse des contextes communs d'apparition des termes a donc été menée sur notre corpus de référence, dans le but de compléter la liste des relations entre termes établies par les méthodes précédemment décrites.

Nous nous sommes inspirés de la méthode décrite dans [GWR99], tout en l'adaptant notamment du fait que nous n'avons pas d'information sur la position des termes. Cette méthode se déroule selon les étapes suivantes :

- identification des *termes cibles* w. Tout d'abord, doit être définie une liste de termes, *termes cibles*[10], dont les contextes d'apparition sont comparés. Nous utilisons pour ce faire la liste des 18 000 termes fondamentaux du domaine tels que nous les avons identifiés précédemment (section 5.4). Nous appliquons donc cette méthode de détection de relation entre termes sur les mots mais aussi sur des syntagmes, noms et syntagmes nominaux en l'occurrence.
- définition du contexte. Pour chacun des *termes cibles* identifié, son vecteur de contexte doit être établi. Nous définissons les contextes des termes comme les articles de code. La liste des termes détectés comme tels par notre outil d'extraction de termes apparaissant dans le même article que le terme cible en question constitue le vecteur des termes contextuels c (*context words*).
- pondération des termes contextuels. Chaque terme contextuel est pondéré sur la base de l'indice d'*information mutuelle*[11] entre le terme contextuel et le terme cible en question. L'indice d'*information mutuelle* entre un terme cible et un terme contextuel est élevé quand leur fréquence conjointe f_{cw} est élevée relativement à leurs fréquences individuelles dans le corpus, f_c et f_w. La formule de l'*information mutuelle* ajoute 1 au ratio des fréquence de façon à ce que, à des occurrences de zéro correspondent des indices d'*information mutuelle* de zéro. La

[10] *target words*
[11] *Mutual Information, MI*

formule de l'indice d'*information mutuelle* [GWR99] s'établit comme suit :

$$MI_{(cw)} = log_2 \left(\frac{f_{cw}}{f_c f_w} + 1 \right) \tag{6.4}$$

- comparaison des contextes. Les vecteurs des termes contextuels pondérés sont alors comparés deux à deux pour chacun des termes cibles. La comparaison entre les vecteurs se fait par le biais de la mesure de similarité cosinus [GWR99] qui peut varier entre -1 et 1. Plus la mesure de similarité cosinus est proche de 1 en valeur absolue, plus les vecteurs sont proches. La formule de la mesure de similarité cosinus entre les vecteurs de termes contextuels des termes a et b s'établit comme suit (p_a correspond aux poids des termes contextuels du terme cible a, p_b aux poids des termes contextuels du terme cible b, ab est le nombre de termes contextuels communs à a et à b) :

$$Sim_{a,b} = \frac{\sum_{ab} p_a p_b}{\sqrt{\sum_a p_a^2 \sum_b p_b^2}} \tag{6.5}$$

- détermination d'un seuil de similarité entre termes cibles. Dès lors que les indices de similarité entre chaque paire de termes cibles définis sont calculés, il faut déterminer le seuil à partir duquel les relations sont jugées valides. De la même façon que pour la méthode précédente, cette détermination ne peut se faire qu'en adoptant une méthode empirique. Cette méthode consiste à interpréter et à valider les relations entre termes détectées par divers niveaux des indices de similarité. Le seuil de similarité a été dans notre contexte fixé à 0.8.

6.3.2 La pratique, résultats

Nous avons appliqué la méthode statistique des contextes sur les 16 000 termes fondamentaux du domaine. L'analyse s'est faite code par code du corpus de référence, de telle sorte que chaque relation établie entre deux termes est assortie d'une indication sur sa provenance, le code en question.

Cette méthode statistique de comparaison des contextes d'apparition de termes dans des corpus permet d'établir des relations entre ces termes. Nous rappelons que les termes concernés dans cette étude sont les *termes fondamentaux du domaine* tels que définis précédemment[12]. Ces termes sont des noms ou des syntagmes nominaux. Il peut dès lors y avoir redondance entre

[12] A titre de rappel, ces *termes fondamentaux du domaine* sont ceux rencontrés dans les intitulés des divisions thématiques de l'ensemble des codes.

un lien trouvé entre deux termes, et notamment deux noms, et un syntagme connu en tant que tel. Dans ce cas, il est utile d'éliminer une telle relation détectée. Par exemple, la méthode statistique de comparaison des contextes permet d'établir une relation entre les termes *limite* et *responsabilité*. Or le syntagme *limites de la responsabilité* est connu en tant que tel dans les sorties de notre outil d'extraction de termes. Les relations détectées par la comparaison statistique des contextes des termes correspondant à des syntagmes connus sont donc éliminées. Un programme Perl considère les résultats de l'analyse statistique des contextes et les compare un par un aux syntagmes issus de l'outil d'extraction de termes. Les relations concernant deux termes présents dans un même syntagme sont éliminées. On obtient, par ce filtre sur les résultats de l'analyse statistique des contextes, un ensemble de 6 534 relations établies entre 2 971 termes différents.

Le Tableau 6.12 liste un ensemble de relations telles que détectées par l'analyse statistique des contextes des termes.

Terme1	Terme2	Code
abattages d'arbres	*défrichements*	CURBANI
abattages d'arbres	*régime des coupes*	CURBANI
abattages d'arbres	*parties de communes*	CURBANI
abattages d'arbres	*territoire de communes*	CURBANI
abattages d'arbres	*pos*	CURBANI
absence d'usufruitier	*conventions*	CCIVILL
absence d'usufruitier	*exercice des droits indivis*	CCIVILL
enlèvement	*séquestration*	CPENALL
enlèvement	*détention des choux-palmistes*	CFOREST
enlèvement	*résidus*	CGCTERR

TAB. 6.12 – Exemples de relations entre termes détectées par à l'analyse statistique des contextes

6.4 Analyse des formes

6.4.1 La méthode

Une dernière méthode a été utilisée pour établir des liens entre les termes du domaine sur la base de notre corpus de référence. Cette méthode consiste à lier à un terme donné, tous les termes contenant ce terme donné. La relation est donc établie entre termes sur la base du principe du *pattern matching*[13] que nous retrouvons notamment dans la commande Unix *grep*. Nous convenons que cette méthode est triviale. Nous avançons cependant que dans le contexte qui est le nôtre, celui de la construction d'une ontologie dédiée à la recherche d'information, cette méthode permet de détecter des relations entre termes qui peuvent être utiles pour une aide à la reformulation de requêtes.

6.4.2 La pratique, résultats

Nous avons appliqué cette méthode à l'ensemble des quelques 118 000 termes extraits par Syntex de notre corpus de référence et sélectionnés selon les méthodes que nous présentons dans la Section 5. La méthode a été implémentée par un programme Perl qui considère un à un ces termes et cherche parmi le reste de la liste ceux le contenant. De cette façon, ne sont liés à un terme donné que des termes le contenant, donc plus longs. Les liens détectés entre termes par cette méthode sont donc des liens qui permettent de préciser un terme en question. Par exemple, le terme *acte de naissance* va être, de cette façon, lié aux termes *acte de naissance d'un enfant majeur*, *acte de naissance de chaque époux* ou *acte de naissance de la personne décédée*. Il ne sera pas lié aux termes *acte* et *naissance*. Cependant, *acte* va être lié à *acte de naissance*.

A chaque relation établie entre un terme et un autre le contenant, l'indication de la (ou des) provenance(s) de ce dernier est conservée. En effet, chaque terme est connu pour avoir été extrait par le module d'extraction de syntagmes de Syntex d'un ou plusieurs codes donnés. Cette information est conservée et sert d'attribut à la relation établie. Par exemple, le terme *acquisition* est lié au terme *acquisitions de l'Etat* dans le Code de l'environnement et au terme *bourse des acquisitions* dans le Code de commerce.

L'application de cette méthode de détection de liens entre termes permet d'établir un ensemble de 179 938 relations entre 90 830 termes différents. Une moyenne de 10 relations par termes est trouvée par ce que nous appelons l'*analyse des formes*. De la même façon que pour les relations détectées par l'analyse distributionnelle de Syntex, cette moyenne cache de fortes disparités, certains termes se voyant établir plus de 25 relations par cette seule méthode alors que d'autres n'en ont aucune.

Pour les même raisons que celles exposées lors de la description des sé-

[13] reconnaissance de forme

lections opérées dans les résultats de l'analyse distributionnelle, on ne peut envisager de conserver toutes les relations établies par l'*analyse des formes*. C'est donc l'usage final auquel est dédié notre ontologie qui justifie la procédure de sélection que nous décrivons :

- un seuil de nombre *raisonnable* de relations par termes a été fixé, au-dessus duquel la sélection est réalisée. Ce seuil est établi sachant que d'autres relations sont susceptibles d'être établies pour chacun des termes en question par les autres méthodes que nous exposons précédemment. Ce seuil a été fixé de façon empirique tout en tentant de prendre en considération des appréciations d'*accessibilité* à l'outil final. Ce seuil est fixé à 12. Au-dessus de ce seuil, les relations sont sélectionnées.
- la sélection des relations établies par la méthode de l'*analyse de forme* se fait sur la base des codes dans lesquels ces relations sont établies, ou, pour être plus précis, sur la base des codes dans lesquels les termes liés sont trouvés. Nous ne prenons en compte que les relations dont le terme lié se trouve présent dans plus d'un code. Ainsi, la relation établie entre les termes *action* et *action en justice* est conservée car le terme *action en justice* se trouve à la fois dans le Code civil et le Code du travail. Par contre la relation établie entre *action* et *action en dissolution des associations* sera ignorée, ce dernier terme n'étant trouvé que dans le Code de l'organisation judiciaire.

La procédure de sélection des relations établies par *analyse des formes* a été appliquée. Nous obtenons alors une moyenne de 6.8 relations par termes.

Le Tableau 6.13 liste un ensemble de relations établies entre termes par *analyse des formes*.

Terme1	Terme2	Code
cession totale	*cas de cession totale*	CCOMMER
cession totale	*cession totale de l'entreprise*	CCOMMER
cession totale	*cession totale de l'entreprise*	CTRAVAIL
cession totale	*cession totale des actifs de la société*	CCIVILL
cession totale	*cession totale partielle*	CPROINT
cession totale	*cession totale partielle*	CCOMMER

TAB. 6.13 – Exemples de relations entre termes détectées par *analyse des formes*

6.5 Ensemble des relations entre termes

Si nous regroupons l'ensemble des termes faisant l'objet de relations par au moins une des méthodes explicitées plus haut, nous obtenons un ensemble de 103 994 termes. Sur l'ensemble de ces termes, 17 688 se trouvent être liés à plus d'un terme. La différence, plus de 80 000 termes, sont des termes qui ne sont liés qu'à un seul terme, ces relations sont principalement issues de la méthode d'analyse des formes.

Si nous comparons pour chacun des termes les relations qui lui sont attribuées avec d'autres termes, nous pouvons établir les constats suivants. D'une part, aucune des relations n'est établie par chacune des quatre méthodes différentes d'établissement de liens entre termes. Seules 24 d'entre elles sont confirmée par trois méthodes. Ces trois méthodes n'incluant jamais l'*analyse des formes*. Ce point s'explique aisément par le fait que cette dernière méthode ne peut établir de liens qu'avec des termes contenant le terme en question. Ce type de liens ne sont quasiment jamais établis par les autres méthodes. Le Tableau 6.14 liste certaines de ces relations établies par trois des méthodes d'établissement de liens entre termes.

Terme1	Terme2
chirurgiens-dentistes	*médecin*
dépens	*frais*
déportés	*prisonniers*
déportés	*réfractaires*
entreprises d'investissement	*établissements de crédit*
perquisitions	*saisie*

TAB. 6.14 – Exemples de relations entre termes détectées par trois méthodes de détection de liens entre terme

Le nombre de relations détectées par plus d'une méthode d'établissement de liens entre termes est de 928.

Ces derniers points justifient le fait que les méthodes d'établissement de liens entre termes telles que nous les décrivons ci-dessus sont complémentaires les unes des autres et que chacune permet de découvrir des relations différentes.

L'exploitation des informations sur les provenances des relations établies entre termes, quelles que soient les méthodes utilisées, permet de détecter des phénomènes de polysémies.

A titre d'illustration le Tableau 6.15 liste un ensemble de termes liés au terme *cadre* et les codes dans lesquelles ces relations ont été détectées. Des regroupements sémantiques peuvent être opérés. Ces regroupements ne peuvent, de notre point de vue, être établis qu'à la main. Une présomption sur le sens induit par le code en question ne pouvant être établie. Ainsi le terme *cadre* pris dans son sens de statut de salarié est représenté par des relations détectées dans le Code du travail. Cependant, le terme pris dans son sens général de circonscription fonctionnelle est également représenté par des relations détectées dans le Code du travail.

Terme1	Code
rénovation	CCONSTR
amélioration du cadre de vie	CURBANI, CENVIRO
cadre supérieur	CTRAVAI
qualité de cadre	CTRAVAI
cadre des accords	CTRAVAI
cadre des opérations	CCONSTR
cadre de conventions	CFOREST

TAB. 6.15 – Exemples de relations établies pour le terme *cadre* et leurs provenances

L'exploitation des données issues de nos méthodes de détection de relations entre termes vers l'étude des polysèmes juridiques n'est pas une voie que nous avons décidé d'approfondir, principalement par manque de temps. Notre interface de visualisation et d'utilisation de notre ontologie ne présente donc pas ces phénomènes de polysémie à l'utilisateur. Nous avançons cependant à cet égard que nous plaçons toute notre confiance dans l'utilisateur. Celui-ci nous semble en effet le plus à même de détecter les polysémies d'un terme, notamment lors de la consultation d'une liste de termes liés à celui-ci. D'une part, l'identification des divers champs sémantiques d'un même terme nécessiterait un long travail manuel. D'autre part, les choix qu'il engendrerait, éminemment subjectifs car relevant de l'interprétation, seraient forcément sujets à caution.

Chapitre 7

De la ressource ontologique à l'ontologie

La phase ascendante d'élaboration d'ontologies à partir de corpus telle que nous la décrivons dans les Chapitres 4, 5 et 6 permet d'extraire deux des éléments principaux de l'ontologie future : les termes du domaine et des relations entre ces termes. Ces éléments sont regroupés dans ce que nous appelons la ressource ontologique. L'ontologie n'est cependant que *future* ; une étape supplémentaire est nécessaire permettant, sur les bases de la ressource ontologique, d'obtenir une ontologie en tant que telle.

Les termes du domaine et les relations entre ces termes, éléments de la ressource ontologique, ont été trouvés par des analyses menées sur corpus. Ces analyses, combinant des approches syntaxiques à des approches statistiques, permettent de découvrir des relations purement linguistiques entre termes. Aucune signification n'est en effet attachée à chacune des relations que nous avons établies entre termes ; ces relations ne sont pas *qualifiées*. Un travail de qualification de ces relations, travail s'assimilant à une modélisation, permet alors de faire un pas de plus vers l'ontologie. Cette qualification des relations entre les termes permet de fournir une information supplémentaire à l'utilisateur de notre ontologie, que ce soit lors de sa consultation ou lors de son utilisation dans le cadre du système d'aide à la reformulation de requête.

La démarche de construction d'ontologie que nous décrivons propose de se fonder sur les éléments de la ressource ontologique pour faire le travail nécessaire à la définition d'une ontologie en tant que telle. C'est pourquoi nous avons choisi l'appellation de *ressource ontologique* au réseau constitué des termes du domaine et des relations entre ces termes. La phase permettant de passer de relations purement linguistiques entre termes à des relations sémantiquement validées est appelée *phase de modélisation des connaissances acquises*. Elle permet de passer de la *ressource ontologique* à l'ontologie. Nous choisissons de n'utiliser que les éléments de la ressource ontologique pour élaborer l'ontologie, nous refusant à utiliser d'autres éléments. Ce choix est fort ; il diffère des autres démarches d'élaboration d'ontologies reposant en partie sur une approche linguistique. Notre démarche consiste certes à incorporer des connaissances aux éléments constitutifs de la ressource ontologique. Cependant, la structure même de la ressource ontologique demeure l'ossature de l'ontologie, et l'apport de connaissance ne se fait qu'à la qualification des relations. Ce choix méthodologique est principalement guidé par l'approche pragmatique que nous adoptons quant à la construction de cette ontologie. Nous n'entendons pas réaliser un modèle parfait de la matière. Nous entendons simplement organiser les éléments de la matière de manière à ce que cette *organisation*, l'ontologie, soit utile dans le cadre d'un système d'aide à la reformulation de requête.

Nous décrivons dans un premier temps la ressource ontologique (Section 7.1). Le passage de la ressource ontologique à l'ontologie proprement dite est décrit dans un deuxième temps (Section 7).

7.1 La ressource ontologique

La ressource ontologique rassemble donc les éléments résultats de la phase ascendante de construction d'ontologie sur corpus relevant d'un procédé d'acquisitions de connaissances à partir de textes. Nous définissons la ressource ontologique, notamment au regard d'autres concepts comme les *bases de connaissances terminologiques* (Section 7.1.1). Nous décrivons ensuite notre point de vue majeur relatif à cette *ressource ontologique*, celui selon lequel nous ne procédons à aucune validation humaine de ses éléments (Section 7.1.3). Par contre, une phase d'évaluation en usage de la ressource ontologique a été réalisée (Section 7.2).

7.1.1 Définition

Ce que nous appelons ressource ontologique est la structure, ou l'objet, qui rassemble les résultats de la phase ascendante d'acquisition des connaissances à partir de textes de notre méthode d'élaboration d'ontologies. Notre ressource ontologique rassemble donc les termes du domaine tels que nous les avons définis et les relations établies entre ces termes sur la base des analyses également décrites. La ressource ontologique se présente donc comme un réseau de termes liés entre eux.

A ces termes et relations sont ajoutées des informations thématiques sur les termes. Nous exploitons pour ce faire les données sur la provenance des termes. Les termes sont issus des codes, et chaque code traite d'un domaine particulier du droit. Par conséquent, chaque terme se rattache à une ou plusieurs thématiques juridiques. Par exemple, le termes *legs à titre universel* n'apparaît que dans le Code civil. Nous pouvons donc lui rattacher la thématique *droit civil*. Nous faisons ainsi correspondre à chaque code la thématique qu'il traite.

L'ensemble des 103 994 termes qui se trouvent reliés entre eux par une ou plusieurs relations constitue donc la base de la *ressource ontologique*. La ressource ontologique se présente alors comme un graphe, les nœuds étant les termes et les arcs les relations entre les termes.

Nous notons que les arcs présentent entre eux une différence relative à la méthode ayant permis l'établissement du lien entre les deux termes. Ainsi, les relations établies par analyse des formes ne sont que dans un sens. Par exemple, le terme *nationalisation* se trouve être lié par cette méthode au terme *dispositions spéciales applicables aux opérations de nationalisation*. Par contre, l'établissement du lien inverse entre *dispositions spéciales applicables aux opérations de nationalisation* et *nationalisation* n'a pas été fait. *A contrario*, les relations établies par chacune des trois autres méthodes (analyse distributionnelle, analyse statistique des contextes et analyse des coordinations) le sont dans les deux sens. Par exemple, le terme *vice de forme* se trouve être lié au terme *nullité* par l'analyse statistique des contextes. Dans ce cas, le lien entre le terme *nullité* et le terme *vice de forme* est également établi.

Une distinction pouvant être faite entre les termes de la ressource ontologique est celle qui sépare les termes *nœuds*, ceux qui se trouvent liés à plusieurs termes, eux même liés à d'autres termes, et les termes *feuilles* qui ne se trouvent être reliés que par une seule relation avec un seul terme. Du fait que les liens établis entre termes sur la base de l'analyse de forme ne sont établis que dans un sens allant du terme contenu au terme contenant, les termes *feuilles* sont les termes les plus longs en nombre de mots qui ne se trouvent par ailleurs pas reliés à d'autres termes par d'autres méthodes. Ainsi, en quelque sorte, une plus grande valeur est attribuée aux relations établies par d'autres méthodes que l'analyse de forme, c'est-à-dire aux méthodes statistiques et syntaxiques. Cette distinction entre termes *nœuds* et termes *feuilles* peut en effet être interprétée comme une différence d'importance, les termes *nœuds* pouvant être vus comme plus importants que les termes *feuilles*. En effet, lors de la navigation dans l'ontologie, les termes *feuilles* apparaissent comme des "impasses" de navigation, aucun termes ne leur sont liés, tandis que les termes *nœuds* se trouvent eux liés à d'autres termes. A ce titre, des termes comme *capitalisation* ou *assurance sur la vie* sont des termes plus importants, car reliés à d'autres termes, que des termes comme *contrat de capitalisation* ou *épargne populaire* qui ne s'avèrent liés à aucun terme[1]. Cette distinction entre termes sera utilisée par la suite, dans la phase de modélisation des connaissances acquises de la ressource ontologique. Cette distinction ne permet cependant pas de distinguer parmi les termes ceux qui pourraient être qualifiés de *concepts* au sens ontologique.

Le Tableau 7.1 liste les termes liés dans la ressource ontologique au terme *capitalisation*. Les termes en gras sont ceux pouvant être qualifiés de plus importants que les autres du fait qu'ils sont reliés à d'autres termes.

assurance
assurance sur la vie
participation des entreprises d'assurance
épargne
bons de capitalisation
contrat de capitalisation
opération de capitalisation

TAB. 7.1 – Termes reliés dans la ressource ontologique au terme *capitalisation*

Les relations constitutives de ce réseau ont été détectées par des analyses linguistiques sur des textes. Dès lors, ces relations entre termes ne sont pas interprétées sémantiquement. Ces relations permettent simplement de relier des termes entre eux. Par exemple, les éléments de la ressource ontologique nous permettent d'avancer que le terme *mandataire* est notamment lié aux termes *experts en diagnostic d'entreprise*, *personnes*, *représentant*, *mandataires judiciaires* ou *qualité de mandataire*. Ces liens établis entre termes ne nous permettent en aucune façon d'établir qu'un *mandataire* est une *personne*, ni que la *qualité de mandataire* est un attribut juridique des *mandataires*. C'est ce type de qualification de liens entre termes que nous tentons d'établir dans la phase de modélisation des connaissances de la ressource ontologique.

[1]Le terme *épargne populaire* est une *feuille* du terme *épargne*.

7.1.2 La ressource ontologique et les BCT

Au regard de tous ces éléments descriptifs de la *ressource ontologique*, il nous a semblé nécessaire de la positionner relativement à des notions proches et plus précisément de la notion de *base de connaissances terminologiques* ou BCT.

Les *bases de connaissances terminologiques* ont ce point commun avec notre *ressource ontologique* qu'elles sont une étape de la construction de modèles formels de domaines, ou plus précisément d'ontologies [AGBS00a]. Elles se situent en effet dans la phase de structuration des connaissances du domaine dans un modèle conceptuel non computationnel élaboré à partir du vocabulaire repéré dans le domaine [JS00]. Ainsi, une BCT rassemble ce que Bachimont ([Bac00]) appelle des *concepts sémantiques* par opposition aux *concepts formels*, nœuds de l'arbre ontologique.

De plus, les BCT sont issues d'une approche purement linguistique [CAG00], de la même façon que notre ressource ontologique ; elles constituent ainsi une représentation structurée des connaissances issues d'un corpus à partir de son analyse linguistique.

Enfin, les BCT, telles qu'elles sont définies dans [AC00], sont envisagées comme :

> [...]« un modèle du texte, c'est-à-dire comme un produit dont la construction ne fait pas appel à une connaissance extérieure à celle du corpus.»

C'est également le point de vue que nous adoptions sur les *ressources ontologiques*, dès lors que le corpus de référence qui a servi de base à leur élaboration est constitué à bon escient. Des connaissances autres que celles des textes sont utilisées dans la phase suivante de modélisation des connaissances acquises qui permet de passer de la ressource ontologique à l'ontologie en tant que telle.

Les deux notions ne se recoupent cependant pas, loin de là, et principalement en raison du fait que les BCT sont élaborées essentiellement manuellement et par des linguistes. L'idée sous-jacente à l'élaboration des BCT est que les termes et leurs usages sont révélateurs de la conceptualisation propre au domaine et que l'étude de la terminologie est le moyen d'identifier les objets concrets et abstraits valides dans le domaine et les opérations sur ces objets [JS00]. L'élaboration que nous faisons de notre ressource ontologique répond en partie à ces idées, notamment celle selon laquelle l'étude de la terminologie permet d'identifier les objets du domaine. Simplement, nous ne nous attachons pas, à ce stade de l'élaboration de l'ontologie, à l'identification des concepts du domaine, éléments des ontologies, ni à la définition d'opérations entre les objets du domaine puisque les relations établies entre les termes, étiquettes des objets, ne sont pas sémantiquement qualifiées.

De plus, les BCT ([JS00], [SBAG01]) comprennent l'accès aux occurrences des termes dans les textes, ainsi qu'une description en langage naturel des concepts, ce que ne comprend pas notre ressource ontologique.

Enfin, nous n'envisageons pas notre ressource ontologique de la même façon que le sont les BCT. Ainsi, il a été avancé que des expérience ont montré que les éléments d'une BCT ne suffisaient pas pour constituer un modèle du domaine [AGBS00a]. Certains éléments linguistiquement pertinents peuvent s'avérer inadéquats pour l'application alors que d'autres doivent être ajoutés. Nous envisageons pour notre part la ressource ontologique comme constituant l'ossature principale de l'ontologie. Des connaissances sont certes ajoutées à ces éléments, mais celles-ci résident simplement dans la qualification de liens déjà éléments de la ressource ontologique. Les éléments de la *ressource ontologique* sont donc ceux de l'ontologie ; aucun n'est ajouté. Certains peuvent cependant être minimisés dans la qualification des liens ; nous le voyons plus loin, Section 7.3.2.

Notre ressource ontologique n'est donc pas une *base de connaissances terminologiques*, bien qu'étant également une étape dans la construction d'ontologies.

7.1.3 Le choix de la non validation

Les éléments constitutifs de notre ressource ontologique sont donc les termes du domaine identifiés lors de la phase ascendante d'acquisition de connaissances à partir de textes. Ces termes sont liés entre eux par des relations établies sur la base de méthodes d'analyses sur corpus.

La démarche utilisée pour l'identification des termes du domaine est décrite dans la Section 5.2. Elle conduit à un choix méthodologique majeur, celui de la non validation humaine de la liste des termes définis comme les termes du domaine. Cette orientation méthodologique est importante car elle démarque notre approche de construction d'ontologies de toutes les autres. Elle est justifiée dans la Section 5.4.

Pour ce qui est de l'identification des relations entre les termes, nous adoptons une démarche également fondée sur le corpus. Les différentes analyses conduisant à l'identification de liens entre termes sont décrites dans la Section 6. L'intervention humaine dans cette phase d'identification de relations entre termes réside tout d'abord dans le choix des méthodes utilisées, mais essentiellement dans l'ajustement de ces méthodes. Seule l'analyse des conjonctions de coordination a fait l'objet d'une validation humaine. Pour ce qui concerne l'analyse distributionnelle, l'analyse des formes et l'analyse des contextes, l'établissement de seuils est nécessaire pour chacun des indices utilisés. Ces démarches sont largement empiriques. L'intervention humaine de l'ingénieur de la connaissance est alors fondamentale ; elle fonde la justesse et la finesse des résultats obtenus de chacune de ces méthodes.

L'établissement d'heuristiques permettant de limiter dans certains cas le nombre de relations établies par terme (l'analyse distributionnelle et l'analyse de forme) nécessite également une intervention humaine. Ces interventions de l'ingénieur des connaissances se traduisent par des mesures d'appréciations sur ce qu'est un nombre raisonnable de termes suggérés à un utilisateur de système d'aide à la reformulation de requêtes. Dans ce cas comme dans le précédent, ces interventions humaines sont un des paramètres d'une automatisation de processus.

Dans la mesure de ces interventions de l'ingénieur des connaissances, l'incorporation des relations établies sur la base d'analyses de corpus ne fait pas l'objet de validation humaine, de la même façon que précédemment et essentiellement pour la raison suivante.

Notre ontologie est destinée à être intégrée à un système d'aide à la reformulation de requêtes. C'est à ce titre que les relations établies entre les termes sont utilisées. Elles permettent de lister les termes plus ou moins liés à ceux de la requête de l'utilisateur. Le but de la suggestion de ces termes est principalement de donner à penser à l'utilisateur. Des termes suggérés lui permettent de restreindre sa requête initiale par un terme plus spécifique. D'autres lui permettent de procéder à un élargissement de sa requête avec un terme plus général. D'autres encore lui permettent de formuler une nouvelle requête en utilisant des termes proches de sa requête initiale. Enfin, d'autres termes peuvent lui évoquer une nouvelle requête plus ou moins liée à la première. L'utilisateur a toute latitude, dans le système d'aide à la reformulation de requête, pour choisir les termes qui lui semblent appropriés pour une reformulation. Par ce choix, l'utilisateur procède à une interprétation et à une comparaison des termes suggérés au regard du sens qu'il attribue à sa requête initiale. Ce processus est ainsi totalement subjectif, empreint des interprétations de l'utilisateur. Dès lors, nous ne pouvons concevoir une phase de validation manuelle des relations établies entre les termes. Quelle serait la légitimité d'une personne (éminente experte du domaine, ou non) à établir que telle relation entre deux termes est susceptible de donner correctement à penser à l'utilisateur ? Nous pensons qu'elle n'en aurait aucune. Ce point de vue repose sur l'idée sous-jacente que c'est bien l'utilisateur qui est le plus à même d'établir la signification de sa requête, et le plus à même de pouvoir la reformuler.

Il est également suggéré par le fait que les relations que nous établissons le sont par des méthodes fondées sur un corpus et sont à ce titre justifiées. D'une part, ces méthodes se fondent sur un corpus de référence défini de telle sorte que nous l'envisageons comme définissant les principaux objets du domaine et les relations entre ces objets. De plus, les relations établies entre termes par l'analyse distributionnelle sont justifiées par le fait que deux termes partagent un ensemble de contextes syntaxiques. Par exemple[2], les termes *enfant* et *mineur* sont liés du fait qu'ils partagent les contextes suivants : (*atteindre l'âge*, SUJ), (*représentant légal*, DE), (*émanciper*, OBJ), (*moralité*, DE), (*tuteur*, DE), (*adoption*, DE), (*placer*, OBJ), (*recueillir*, OBJ). D'autre part, les relations établies entre termes par l'analyse statistique des contextes sont justifiées à chaque fois par le fait que les deux termes en question apparaissent effectivement dans le corpus de référence, souvent dans des contextes similaires. Comment la personne chargée de la validation manuelle de ces relations fonderait-elle ses décisions qui conduiraient à écarter une relation plus qu'une autre, sachant que les deux sont justifiées et justifiables de la même façon et dans les mêmes mesures ?

Dès lors, et pour toutes les raisons exposées ci-dessus, nous n'incluons pas de phase manuelle de validation des relations établies entre les termes. Nous ne pouvons cependant, dans notre double rôle d'ingénieur de la connaissance et d'expert du domaine, ignorer que certaines sont de meilleure qualité que d'autres. A titre d'exemple, le Tableau 7.2 liste l'ensemble des termes liés au terme *divorce* dans la *ressource ontologique*. Il va sans dire que le lien entre *divorce* et *réduction* est plus obscur sémantiquement que le lien entre *divorce* et *séparation*. Nous n'envisageons cependant pas de considérations de ce genre dans la ressource ontologique.

[2]Nous reprenons ici un exemple donné dans la Section 6.2, également cité dans notre article [BL02].

adoption
nullité
rescision
réduction
révocation
séparation
séparation de corps
cas de divorce
divorce demandé par un époux
divorce pour faute
divorce pour rupture de la vie commune
divorce sur demande conjointe des époux

TAB. 7.2 – Termes liés dans la ressource ontologique au terme *divorce*

7.2 Evaluation de la ressource ontologique

Les éléments termes et relations de la ressource ontologique ne font donc l'objet d'aucune validation manuelle quant à leur pertinence. Une phase d'évaluation de la ressource ontologique est cependant prévue. Cette phase d'évaluation prévoit de la confronter à l'usage auquel est destinée l'ontologie. Le but est de se faire une idée de l'utilité d'intégrer dans un système d'aide à la reformulation de requête un tel réseau de termes établis dans les conditions définies précédemment. L'évaluation en usage de l'ontologie elle-même n'a pas été menée par manque de temps.

Le principe fondamental de ce processus d'évaluation est de mettre la ressource ontologique en contexte, c'est-à-dire de l'intégrer dans un système d'aide à la reformulation de requête. Il repose sur les principes de ce qui est appelé l'*évaluation en usage*. C'est alors aux usagers eux-mêmes d'établir l'utilité d'un tel système, et par là d'évaluer en quelque sorte sa validité. Cette idée de faire évaluer en contexte par des usagers un système de recherche d'informations n'est pas inédite mais est relativement nouvelle dans le domaine. Par exemple, au sein des conférences TREC[3], l'étude des comportements et réactions des usagers se fait dans le cadre de *Interactive TREC* et date de 1997.

Les principes de notre processus d'évaluation en usage s'inspirent en partie de l'article *Interactive query expansion : a user-based evaluation in a re-*

[3]Text Retrieval Conference

levance feedback environment ([Eft00]). Les bases en sont posées dans notre article *Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du droit* ([BL02]).

Un protocole a été défini, établissant les principales étapes de l'évaluation en usage de la ressource ontologique. Chacune d'elle est décrite dans les lignes qui suivent.

7.2.1 Mise en place du système d'aide à la reformulation de requête

Installation du système d'aide à la reformulation de requêtes : L'étape préliminaire à l'évaluation en usage de la *ressource ontologique* consiste à l'installer dans un système d'aide à la reformulation de requête. Nous détaillons dans le Chapitre 9 les principes de la reformulation de requêtes. Nous rappelons simplement ici que l'idée force de notre système d'aide à la reformulation de requête est de suggérer à l'utilisateur d'un moteur de recherche, en réaction à sa requête, des termes plus ou moins liés à ceux de sa requête. L'utilisateur est alors libre de choisir un ou plusieurs de ces termes pour reformuler, ou non, une requête au moteur. Dans la perspective qui est la nôtre ici, les termes liés sont ceux de la ressource ontologique. Notre système d'aide à la reformulation de requête consiste donc à identifier les termes de la requête connus dans la ressource ontologique et, pour chacun de ces termes, à suggérer à l'utilisateur les termes qui s'y trouvent liés. Le principe de notre système est donc, schématiquement[4], de découper la requête saisie par l'utilisateur, y identifier les termes qui s'avèrent être connus dans la *ressource*, lister ces termes et y associer les termes liés. Les données de la ressource ontologique sont préalablement stockées dans un tableau associatif rémanent[5] dont l'accès par les programmes Web dynamiques[6] est optimisé par son chargement dans un tableau. Un programme Perl récupère donc la requête de l'utilisateur et la découpe en termes (mots et groupes de mots). Chacun des termes identifiés y est comparé aux entrées du tableau de hashage. Si des termes sont dans la ressource ontologique, ceux-ci sont renvoyés à l'utilisateur, ainsi que les termes qui leurs sont liés. Sur la page sur laquelle ces termes sont listés, un

[4]Nous renvoyons à la Section 9.1.2 pour une explication plus détaillée du système d'aide à la reformulation de requêtes. Un schéma y illustre le procédé.

[5]Un *DataBase_File*, pour base de donnée.

[6]Programmes *Common Gateway Interface*.

nouveau formulaire de recherche est disponible, qui permet de saisir une nouvelle requête. Ce formulaire permet d'interroger un moteur de recherche indexant les documents visés.

Quels évaluateurs ? La première question à se poser dans notre perspective est sans doute de savoir quels sont les usagers de notre système d'aide à la reformulation de requêtes. Nous ne l'abordons cependant pas ici mais dans la Partie III. Les deux questions sont certainement liées, mais notre point de vue est que des personnes au fait de la matière juridique sont peut-être plus à même d'évaluer un tel système, notamment sur sa qualité. Les évaluateurs sont donc des juristes. Pour rester dans notre perspective d'évaluation dans l'usage, nous avons également pensé que des évaluateurs connaissant les enjeux de la recherche documentaire seraient également bien placés pour procéder à l'évaluation de notre système. Nous avons donc fait appel aux membres de l'association Juriconnexion[7]. Cette association regroupe des personnes qui, professionnelles du droit, n'en sont pas moins des spécialistes de la recherche documentaire. La plupart de nos évaluateurs travaillent ainsi dans les services documentation de cabinets d'avocats ou dans des bibliothèques universitaires de droit. Nous avons fait appel à eux dans le but qu'ils testent le système, qu'ils en perçoivent les limites, notamment en lui soumettant des requêtes qu'ils connaissent comme ambiguës, complexes ou difficiles pour des moteurs de recherche classiques.

Quelle évaluation ? L'idée de notre processus d'évaluation en usage de la ressource ontologique est, nous l'avons dit, de déterminer l'utilité d'un tel système d'aide à la reformulation de requête. Or le principe d'un système d'aide à la reformulation de requête, comparé à un système de moteur de recherche classique, est de permettre à l'utilisateur de mieux formuler sa requête et, par là, de trouver plus facilement des documents y répondant. C'est donc bien ce critère qui doit être pris en compte. Par ailleurs, l'évaluation d'un tel système ne peut se faire d'un bloc, et c'est requête par requête et terme suggéré par terme suggéré qu'une évaluation peut être établie. L'évaluation en contexte de la ressource ontologique se formule alors ainsi : *étant donnée une requête que j'ai adressée, quelle a été l'utilité de la suggestion de tel terme pour reformuler une question qui me permette de trouver plus facilement des documents répondant à ma requête ?* Le préalable, pour que les utilisateurs puissent répondre à cette question, est qu'ils soient en mesure de déterminer quels sont les documents qu'ils attendent en réponse à leur

[7]http ://www.juriconnexion.org, «*Association ouverte à tous ceux qui interrogent et forment à l'interrogation des banques de données juridiques sous tous supports.*»

requête. Nous dévoilons là tout l'intérêt d'avoir des évaluateurs professionnels du droit. Nous avons ainsi défini, en coopération avec l'un d'eux[8], un sous-ensemble de codes qui, de par les domaines juridiques couverts, sont bien connus des juristes qui les *pratiquent* tous les jours. Du fait de leur connaissances de ces codes, nos évaluateurs sont donc en mesure de savoir quels articles ils estiment être des réponses pertinentes à leurs requêtes. Ces codes sont listés dans le Tableau 7.3. Le système d'aide à la reformulation de requête basé sur la ressource ontologique est installé sur ces onze codes[9]. Ainsi, la requête que les évaluateurs auront, ou non, reformulée sur la base des termes suggérés est lancée sur ces seuls onze codes. Un moteur de recherche, spécialement dédié à l'évaluation, a donc été installé sur ces onze codes. C'est un moteur de recherche *classique* dans le sens où aucun algorithme de classement des documents réponses[10] n'est installé. Le moteur renvoit les articles de codes dans lesquels le ou les termes de la requête sont présents. Les requêtes booléennes sont possibles avec les opérateurs logiques *AND* et *OR*. L'opérateur *AND* est présumé entre deux mots de la requête. Ce moteur permet aux évaluateurs de soumettre des requêtes sur le corpus de l'évaluation et, en soumettant des requêtes reformulées, d'être en mesure de déterminer si une requête reformulée permet de trouver *plus facilement*[11] les documents attendus en réponse. Nous avons choisi de développer nous-mêmes ce moteur en Perl, ce qui nous permet de contrôler son fonctionnement. Nous disposons ainsi d'un moteur de recherche dans sa plus grande simplicité, n'incluant ni algorithme de *ranking* des documents réponses, ni algorithme de découpage de requête.

Le principe de l'évaluation en usage de la ressource ontologique a prévalu au fait que les évaluateurs ont toute latitude quant au choix de leur requête. Cette situation est bien celle d'une utilisation *normale* d'un moteur de recherche et, plus spécialement, d'un système d'aide à la reformulation de requête. Il induit cependant une plus grande difficulté dans l'analyse des résultats de l'évaluation puisque les requêtes ne peuvent être comparées entre elles, elles sont toutes différentes.

Il est demandé aux évaluateurs de déterminer l'utilité de la suggestion de chaque terme, relativement à leur requête. Pour ce faire, nous avons

[8] Je tiens, par ces lignes, à remercier E. Barthe.

[9] La ressource ontologique a été incluse dans sa totalité.

[10] Plus connu sous le nom d'algorithme de *ranking*.

[11] Des documents trouvés *plus facilement*, sont des documents apparaissent dans une liste de documents réponses de taille raisonnable, c'est-à-dire pas trop longue. Ce point a été expliqué aux évaluateurs.

Code de l'action sociale et des familles
Code civil
Code de commerce
Code de l'environnement
Code de la justice administrative
Code de l'organisation judiciaire
Code pénal
Nouveau code de procédure civile
Code de procédure pénale
Code de la propriété intellectuelle
Code du travail

TAB. 7.3 – Liste des codes utilisés pour le moteur de l'évaluation

définit une nomenclature. Pour chacun des termes suggérés, il leur est demandé de choisir entre diverses solutions :
- soit le terme est inutile à la reformulation d'une requête qui permette de trouver plus facilement les documents attendus en réponse ;
- soit le terme est utile car c'est un terme synonyme d'un des termes de la requête ;
- soit le terme est utile car c'est un terme, non pas synonyme, mais proche d'un des termes de la requête ;
- soit enfin le terme est utile car il apporte une nouvelle idée[12].

Les évaluateurs ont donc un choix exclusif à faire entre chacune de ces qualifications pour chaque terme suggéré.

Installation technique : Techniquement, le procédé d'évaluation a été conçu pour qu'il se déroule en ligne. Nous avons crée des pages HTML, destinées à cette seule évaluation. Les premières pages consistent à expliquer le but de la démarche d'évaluation et ses différentes étapes. Les personnes intéressées pour évaluer sont invitées à s'inscrire dans un formulaire HTML destiné à cet effet. Quand l'inscription est achevée, les personnes ont accès à un formulaire de recherche qui leur permet de saisir leur requête. A la validation de cette requête, notre programme de découpage de la requête et de consultation de la ressource ontologique est activé. Ce programme renvoi alors une nouvelle page listant les termes découpés de la requêtes connus dans la ressource ontologique

[12] C. Fluhr, rapporteur de nos travaux, nous a fait remarquer qu'une autre alternative aurait pu être donnée aux évaluateurs, celle permettant de désigner un terme comme *nuisible* dans le sens où il induirait des déformations de sens par rapport à la requete initiale. Il est vrai que nous aurions pu donner cette alternative aux évaluateurs ...

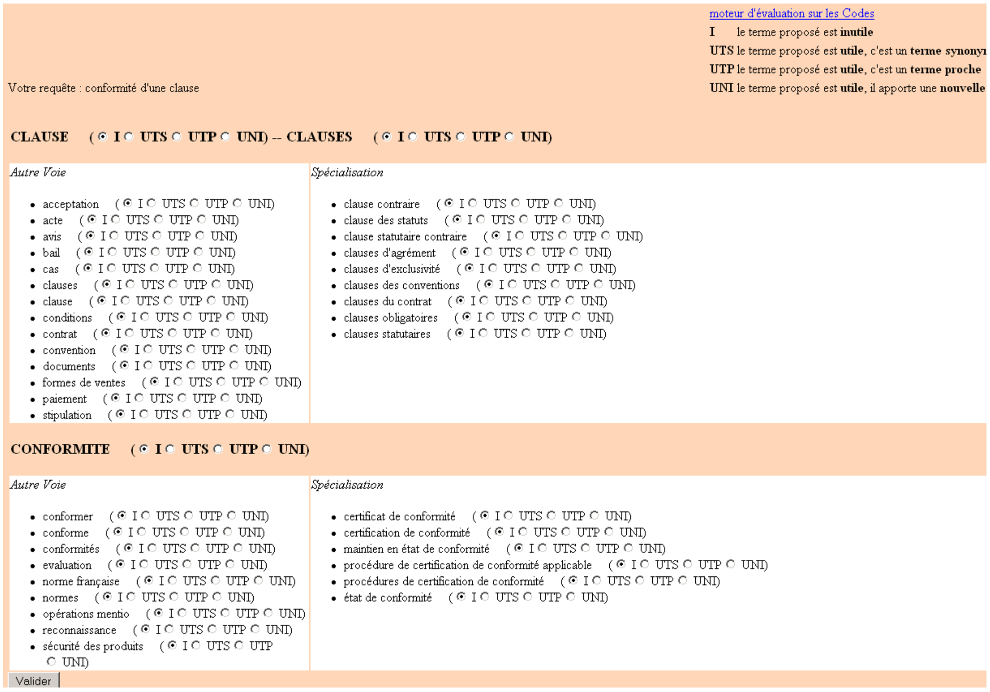

FIG. 7.1 – Ecran d'évaluation des termes suggérés pour la requête *conformité d'une clause*

et, sous chacun d'eux, les termes qui lui sont liés. En face de chacun de ces termes, des boutons radios sont aménagés, permettant à l'évaluateur de choisir une qualification qui lui semble convenir pour le terme en question. Un lien hypertexte aménage un accès au moteur de recherche installé sur les 11 codes de l'évaluation pour aider les évaluateurs de déterminer l'utilité de la suggestion de chaque terme à la reformulation d'une requête. La Figure 7.1 montre un écran de validation des termes suggérés pour une requête. A chacune des étapes de l'évaluation en ligne, les données sont stockées dans une base de donnée MySQL créée à cet effet. Ainsi, les renseignements sur l'identité des évaluatueurs, la formulation de leurs requêtes, les termes suggérés pour chacune de ces requêtes, et enfin, la qualification choisie pour chacun des termes suggérés.

Nous avons lancé notre procédure d'évaluation en usage de la ressource ontologique le 6 juin 2002 par un e-mail envoyé sur la liste Juriconnexion. 18 personnes différentes ont participé à l'évaluation[13] ; elles ont posé un total de

[13]Je tiens à leur signifier toute ma gratitude d'avoir bien voulu participer à une telle procédure d'évaluation.

55 questions.

Ces questions abordent des sujets très divers ; le Tableau 7.4 liste un ensemble de questions posées par les évaluateurs, significatif de la diversité des thèmes abordés.

abus de position dominante
bases de données juridiques
clause d'exclusivité
défricher un terrain
discothèque et troubles de voisinage
la pêche en rivière
travail à temps partiel

TAB. 7.4 – Quelques questions posées par les évaluateurs

Ces questions sont d'une longueur variable, entre un[14] et neuf mots[15].

A chaque question, des termes sont suggérés sur la base de ceux identifiés dans la ressource ontologique. Le nombre de termes suggérés est variable[16]. Certaines questions se voient suggérer deux termes[17], et d'autres 97 termes[18]. La moyenne de termes suggérés par question s'établit à 25.

7.2.2 Résultats de l'évaluation

Parmi les quatre qualifications possibles de chaque terme suggéré, nous estimons que les qualifications *utile car terme synonyme*, *utile car terme proche* et *utile car terme suggérant une nouvelle idée* peuvent être rassemblées sous le qualificatif général *terme utile*. Les termes estimés utiles sont alors comparés à ceux qualifiés d'*inutiles*. La figure 7.2 établit pour chacune des 55 questions, le nombre de termes suggérés qui ont été estimés utiles et le nombre de termes suggérés qui ont été estimés inutiles. Elle montre que de grandes disparités prévalent dans les évaluations de chaque question. Une majorité des questions (32 sur un total de 55) voit néanmoins le nombre de

[14] La question *ententes* par exemple.

[15] La question *monuments historiques obligations du propriétaire et de l'Etat*.

[16] Nous tenons à cet égard à signaler que les méthodes permettant de limiter le nombre de termes suggérés par terme, telles qu'explicitées dans le Chapitre 6, n'étaient pas encore implémentées au début de la phase d'évaluation. A ce titre, les chiffres ainsi qu'une partie des résultats que nous présentons dans cette section ne sont pas représentatifs du système à l'état final. L'analyse des résultats de l'évaluation permet cependant de se faire une idée sur la perception qu'ont eu du système les évaluateurs.

[17] La question *viol*

[18] La question *exploitation d'une hypermarché et petits commerces*

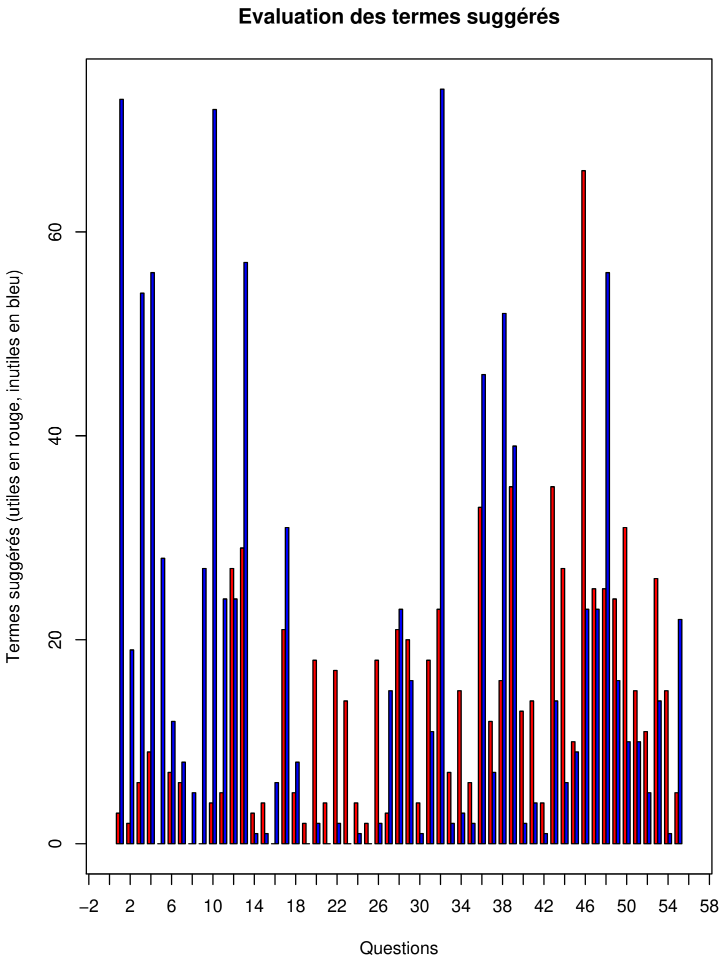

FIG. 7.2 – Evaluation des termes suggérés pour chacune des 55 questions

termes suggérés évalués *utiles* supérieur au nombre de termes suggérés évalués *inutiles*.

Contrairement à ce qu'il aurait pu être intuitivement établi, les évaluateurs ne semblent pas se trouver trop *lassés* par un nombre trop important de termes suggérés. On aurait en effet pu croire que plus le nombre de termes suggérés est élevé, moins le nombre de termes évalués *utiles* le soit, du fait du découragement des évaluateurs face à une liste trop importante de termes à considérer. Nous avons établi dans la Figure 7.3 le nombre de termes évalués *utiles* par rapport au nombre de termes suggérés. La droite de régression est également tracée. Celle-ci montre que le nombre de termes évalués *utiles* ne chute pas quand le nombre de termes suggérés augmente. La tendance est même à la hausse, même si cette hausse n'est pas très importante.

Nos évaluateurs n'ont donc pas été découragés par un nombre important de termes suggérés. Nous avons cependant préféré réduire ce nombre de termes suggérés par question[19], de façon à ce que les utilisateurs finaux de notre système d'aide à la reformulation de requête ne soient pas, eux, découragés.

La distinction que nous avons faite entre les termes *utiles car synonymes*, *utiles car proches* et *utiles car donnant une nouvelle idée* permet d'évaluer la qualité des liens détectés entre termes. En réponse aux 55 questions posées par les utilisateurs, notre système d'aide à la reformulation de requête a suggéré globalement un ensemble de 1 381 termes, soit une moyenne de 25 termes suggérés par requête. Seuls 120 termes suggérés se trouvent être évalués plus d'une fois[20]. Sur ces 120 termes, 49 ont été à chaque fois évalués avec le même qualificatif. Ainsi, 71 termes, soit 5 % du total, font l'objet d'un désaccord quant à leur validation[21]. Nous ne prenons pas ces termes en compte. Le Tableau 7.5 distingue les termes suggérés en fonction des évaluations dont ils ont fait l'objet.

Evaluation	Nbre termes
terme synonyme	152
terme proche	218
nouvelle idée	119
terme inutile	821

TAB. 7.5 – Nombre des termes en fonction de leur qualification

Nous notons que, sur la globalité des termes, 821 ont été évalués *inutiles*, ce qui représente 63 % du total. Nous relativisons ce résultat en en précisant la formulation : sur les 55 questions qui ont servi à l'évaluation, 63 % des termes suggérés par notre système sur la base de la ressource ontologique

[19]Nous faisons ici référence aux méthodes de filtrage de liens établis entre termes par l'analyse distributionnelle, l'analyse de forme et l'analyse statistique des contextes (Section 6).

[20]Ces termes sont suggérés pour plus d'une question.

[21]Nous notons bien évidemment que ce faible chiffre est essentiellement expliqué par le fait que 90 % des termes ne sont évalués qu'une seule fois et ne peuvent donc pas faire l'objet d'un désaccord quant à leur évaluation.

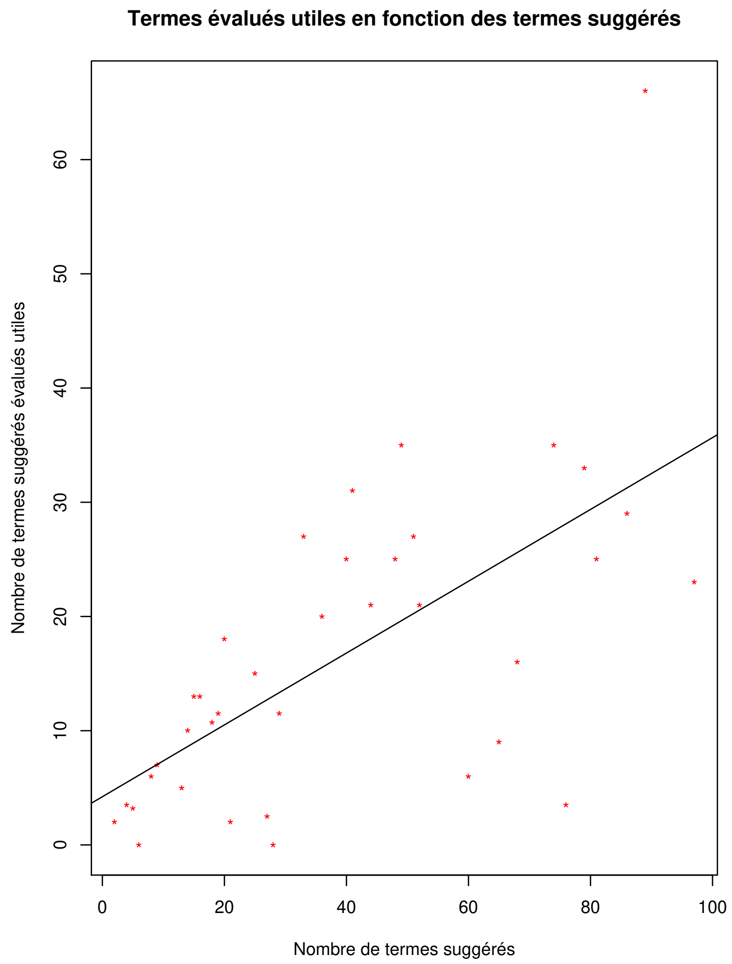

FIG. 7.3 – Nombre de termes évalués *utiles* par rapport au nombre de termes suggérés

n'ont pas été estimés utiles à la reformulation d'une requête. Ce résultat ne présume pas de l'utilité de ces termes suggérés pour d'autres requêtes, et présume encore moins de la validité, dans l'absolu, de la relation établie entre les deux termes.

Dans la globalité des termes évalués *utiles* (soit 489), 31% l'ont été du fait d'une relation de synonymie prévalant entre le terme de la requête en question et le terme suggéré, 45% du fait d'un lien de proximité, et 24% du fait que le terme suggéré apporte une nouvelle idée par rapport à la requête initiale, cette nouvelle idée permettant de reformuler une requête utile à obtenir plus facilement du moteur les documents réponses attendus. Ainsi, une part significative des termes évalués *utiles* le sont de par ce qu'ils donnent à penser à l'utilisateur, leur suggérant une idée qu'ils n'auraient pas eu au départ.

Les résultats de l'évaluation en usage de la *ressource ontologique* nous permettent également d'évaluer chacune des méthodes qui ont été utilisées pour établir des liens entre termes. Nous traitons chacune de ces méthodes sur la base des termes suggérés issus de ces méthodes et en fonction des évaluations dont ils ont fait l'objet. Le Tableau 7.6 liste le nombre de termes concernés par chacune des méthodes et leurs différentes qualifications. Nous rappelons que ces méthodes, telles qu'explicitées dans la Section 6, sont les suivantes : analyse des conjonctions de coordinations (*coordination*), analyse distributionnelle (*distribution*), analyse statistique des contextes (*contexte*) et analyse des formes (*formes*).

–	*coordination*	*distribution*	*contexte*	*formes*
Nbre Termes	68	26	78	1138
inutile	22	12	50	737
utile	46	14	28	301
synonyme	10	1	5	136
proche	20	7	11	180
nouvelle idée	16	6	12	85

Tab. 7.6 – Evaluation des méthodes d'établissement de liens entre termes

Les deux méthodes dont la majorité des termes sont validés *utiles* sont l'analyse des conjonctions de coordinations et l'analyse distributionnelle de Syntex (respectivement 68% et 54% des termes évalués *utiles*). La méthode de l'analyse statistique des contextes présente des performances moindres mais tout de même intéressantes (36% des termes évalués *utiles*), tandis que la méthode de l'analyse des formes obtient les moins bonnes évaluations (26% des termes évalués *utiles*). Le fait que l'analyse des conjonctions de coordination soit la méthode qui présente les meilleurs résultats n'est guère étonnant du fait que la signification de l'usage des coordinations *et* et *ou* dans un texte est bien d'établir le rapprochement entre plusieurs éléments donc termes sémantiquement proches. L'analyse distributionnelle obtient également de bons résultats. L'utilisation d'informations syntaxiques sur les textes, comme c'est le cas pour l'analyse distributionnelle, doublée d'ajustements humains des coefficients obtenus, permet d'obtenir de bons résultats. Puis, une méthode utilisant des données purement statistiques doublée d'ajustements humains des coefficients obtenus, l'analyse des contextes, donne des résultats très moyens mais non négligeables. Enfin, une méthode triviale consistant à relier à un terme tous les termes le contenant, l'analyse des formes, engendre des résultats médiocres, contrairement à ce que nous présupposions.

Les résultats d'évaluation que nous venons de présenter sont relatifs à la procédure même d'évaluation. Ce qui a été estimé par les évaluateurs, c'est bien l'utilité de l'établissement de liens entre termes pour reformuler des requêtes permettant de trouver plus facilement les documents réponses attendus. Cette évaluation n'a concerné que 55 questions. Le pas est alors grand qui mènerait à déduire de cette évaluation la nécessité de l'abandon de l'une ou l'autre méthode d'établissement de relations entre termes. Tout au plus pouvons nous les situer, d'un point de vue qualitatif, les unes par rapport aux autres.

La procédure d'évaluation en usage de la ressource ontologique avait pour but de donner une idée sur la façon dont cette ressource ontologique est perçue en usage et d'établir son utilité à être intégrée dans un système d'aide à la reformulation de requête. Nous estimons que la perception est globalement bonne et en déduisons l'utilité d'un système d'aide à la reformulation de requête basé sur une structure telle que notre ressource ontologique.

La ressource ontologique rassemble donc les éléments issus de la phase ascendante d'acquisition de connaissances à partir de textes. Ces éléments sont des termes du domaine et des liens entre ces termes. La ressource ontologique a fait l'objet d'une procédure d'évaluation en usage. Elle est vue, dans notre méthode d'élaboration d'ontologies dédiées à la recherche d'informations sur le Web, comme un intermédiaire. Des connaissances sont ainsi ajoutées à celles contenues dans la ressource ontologique. Cet ajout se fait durant ce que nous appelons la *phase de modélisation des connaissances acquises*.

7.3 Phase de modélisation des connaissances

La phase de modélisation des connaissances acquises consiste à qualifier les liens entre les termes identifiés lors de la phase ascendante d'acquisition de connaissances à partir de textes. Cette qualification a pour objet de conférer un sens à des liens déjà identifiés. Aucune nouvelle relation n'est ajoutée à celles déjà identifiées ; la ressource ontologique définit alors l'ossature principale de l'ontologie. La qualification sémantique des liens établis entre les termes permet, de notre point de vue, de ne plus considérer les seuls termes du domaine, mais aussi les concepts. Nous obtenons alors une ontologie en tant que telle. Cette étape est également appelée *ontologisation* chez [Kas02a] ; elle est commune à d'autres méthodes d'élaboration d'ontologies comme celles définies dans [FGPJ97] ou dans [AGBS00a]. La modélisation qu'elle implique est une modélisation *pour donner du sens* [Kas02a], en comparaison à la modélisation *pour implémanter un système* consistant à coder l'ontologie dans un langage de programmation dans le but d'obtenir une ontologie computationnelle.

La procédure de qualification sémantique des liens entre les termes de la ressource ontologique se déroule en deux étapes. La première consiste, par une analyse de la tâche doublée d'une expertise du domaine, à identifier un ensemble fini de qualifications envisageables des relations (Section 7.3.1). La seconde consiste à la modélisation en elle-même (Section 7.3.2).

7.3.1 L'ensemble des qualifications possibles

Notre position conjointe d'ingénieur de la connaissance et d'expert du domaine nous a permis d'avoir une vision globale sur la tâche que nous avons définie de qualification sémantique des relations linguistiques établies entre les termes de la ressource ontologique.

Dans la ressource ontologique, des liens sont établis entre des termes, sans que ces liens soient porteurs de sens. Nous avons en effet choisi de ne pas utiliser des méthodes linguistiques consistant à identifier un ensemble de patrons lexico-syntaxiques relatifs à une relation sémantique donnée. De telles méthodes sont décrites dans [Mor99]. Comme exemple de ces méthodes, nous citons le travail sur les contextes définitoires[22] de [Reb00]. Les raisons de ce choix reposent principalement sur le fait que l'identification des patrons lexico-syntaxiques se fait essentiellement de façon manuelle, bien qu'une par-

[22]Du type *est un*, *peut être vu comme* etc.

tie de cette identification puisse être automatisée [Mor99]. De plus, ce travail nécessite des compétences de linguistes que nous n'avons pas et, enfin, ce genre de méthodes, du dire même des personnes qui les pratiquent, sont génératrices de beaucoup de bruit [Reb00]. Notre démarche consiste à se fonder sur les résultats de méthodes automatiques[23]. C'est ainsi que les résultats de la phase ascendante d'acquisition de connaissances à partir de textes rassemblés dans la *ressource ontologique* sont envisagés comme le fondamental de notre démarche d'élaboration d'ontologie. Des connaissances du domaine sont ajoutées aux éléments de la ressource ontologique par la qualification sémantique des relations linguistiques établies entre les termes.

Nous qualifions cette tâche de *modélisation*. Nous pensons ainsi que, par l'ajout de sens au réseau de termes du domaine que constitue la *ressource ontologique*, nous sommes amenés à définir des concepts du domaine dans ce réseau, et par là, obtenons un modèle de ce domaine. C'est en effet bien conceptualiser le domaine, le modéliser, que d'établir des réseaux sémantiques autour de notions juridiques. Dans notre cas, les notions, ou les concepts, sont portés, incarnés, par des termes. La qualification des relations entre les termes permet d'identifier des termes plus importants que d'autres, aux confluents de diverses relations. Ces termes importants sont appréhendés comme des concepts du domaine, les éléments de l'ontologie.

Nous basant sur le réseau lexical de la ressource ontologique pour en inférer, en quelque sorte, un modèle du domaine, nous procédons à une sorte de modélisation ascendante. Une telle modélisation *ascendante* s'oppose à une modélisation *descendante* qui consisterait à considérer le droit comme rassemblant, par exemple, des personnes, des actes, des institutions et des biens. A partir de là, les concepts de *citoyens*, d'*époux*, d'*actionnaires*, de *contrat* ou d'*immeubles* seraient affectés à chacun de ces concepts supérieurs. La démarche se poursuivrait alors, considérant la *carte électorale* comme un attribut nécessaire mais non suffisant d'un *citoyen* ou le *contrat de nantissement* comme une sorte de contrat. Nous entendons, pour notre part, établir une telle modélisation du domaine en partant du lien linguistique détecté par exemple entre *mariage*, *charges du mariage*, *contrat de mariage* et *procréation* pour dégager des concepts du domaine juridique, les relations entre ces concepts et établir par là un modèle du domaine. Ce modèle, incarné dans l'ontologie, est un modèle parmi d'autres, de la même façon que chaque point de vue sur un domaine engendre un modèle différent. Notre souci est simple-

[23]Même si celles-ci nécessitent, nous l'avons vu dans la Section 6, d'indispensables ajustements par l'humain.

ment de veiller à ce que le modèle que nous établissons remplisse les fonctions qui lui sont attribuées lors de son intégration dans un système d'aide à la reformulation de requête.

Nous envisageons la tâche de *modélisation* de la façon suivante. Chaque terme élément de la ressource ontologique se trouve lié en un ensemble d'autres. Si la ressource ontologique est envisagée comme un graphe, certains des termes se trouvent être des "nœuds" - ils sont liés à plusieurs termes - et d'autres se trouvent être des "feuilles", ils ne sont liés qu'à un seul terme. Les données de la ressource ontologique sont telles que l'arbre qu'elle forme est peu profond et très étendu ; beaucoup de termes sont des termes "feuilles".

Par exemple, le Tableau 7.7 liste l'ensemble des termes liés au terme *contrat de travail* ; les termes liés en caractère gras, sont des *nœuds*, les autres sont des *feuilles*.

contrat de travail d'un salarié
convention
règlement des créances
contrat de travail des salariés
contrat de travail temporaire
contrat de travail à durée déterminée
contrat de travail à durée indéterminée
fait de la rupture du contrat de travail
intention de rompre le contrat de travail
périodes de suspension du contrat de travail
rupture du contrat de travail
salariés liés par un contrat de travail temporaire
suspension du contrat de travail

TAB. 7.7 – Termes liés dans la ressource ontologique au terme *contrat de travail*

La tâche de modélisation consiste à considérer les termes *nœuds*. Pour chacun des termes qui se trouvent liés à ce terme, la qualification du lien qui unit le terme lié au terme *nœud* est opérée. Dans l'exemple du terme *contrat de travail*, la modélisation consiste à établir la sémantique du lien existant entre *convention* et *contrat de travail* ou entre *contrat de travail des salariés* et *contrat de travail*, le terme *convention* faisant lui aussi l'objet d'une *modélisation* en tant que terme *nœud*. Nous procédons donc à une qualification des liens dans un sens donné : la relation qualifiée est celle reliant le terme lié au terme *nœuds* en question. Du fait que la plupart des relations établies entre termes dans la ressource ontologique le sont dans deux sens, Terme 1 - Terme 2 et Terme2 - Terme 1, la procédure de modélisation conduit à une double modélisation de nombreux liens. Cette double modélisation conduit à la définition d'une règle quant à l'importance relative des qualifications (voir Section 7.3.2). Chacun des termes *nœud* de la ressource ontologique est donc envisagé comme un futur concept de l'ontologie, qu'il deviendra dès lors qu'il aura été modélisé. Les termes *feuilles* de la ressource ontologique seront des *termes* dans l'ontologie.

Nous procédons ainsi à une *conceptualisation.* Celle-ci renvoie ([GG95]) à l'intention des entités conceptuelles, c'est-à-dire aux *notions* qui sont des règles contraignant la référence des entités. Peuvent être modélisés les concepts, les relations et les intentions de ces concepts et relations. Nous nous attachons ici à expliciter les relations que nous établissons, ces relations permettant de définir les concepts, c'est-à-dire de modéliser.

Une étude de la ressource ontologique et de la tâche de *modélisation* a permis de dégager un ensemble fini de qualifications sémantiques pouvant être établies entre termes.

(i) Le lien de subsumption.

La relation la plus fréquente que l'on trouve dans les ontologies est le lien de subsumption. Pour simplifier, ce lien correspond à la relation *est un* ou *ISA*. Par exemple, la *vente de marchandise* est une *vente*, de même que la *vente forcée* est une *vente.* Simplement, dans notre domaine, deux types de liens de subsumption peuvent être distingués. Le premier est un lien de subsumption proprement juridique ; le second est un lien de subsumption plus général. Ceci est dû au fait que la matière juridique définit elle-même des catégories pour ses artefacts, chacune de ces catégories conduisant l'application de règles différentes. Pour reprendre l'exemple de la *vente*, la *vente forcée* est une catégorie juridique de *vente* alors qu'une *vente de marchandise* est une sorte générale de vente. Ainsi une *vente de marchandise* particulière pourra être amenée à être qualifiée de *vente forcée* dès lors que les conditions définies dans la loi seront réunies. Nous distinguons donc les deux relations que nous nommons respectivement EST_UNE_SORTE_JURIDIQUE_DE et EST_UNE_SORTE_GÉNÉRALE_DE.

Un troisième type de lien de subsumption existe dans notre ontologie : c'est celui qui lie un concept à une des formes qu'il peut prendre dans certaines conditions, de temps notamment. Nous définissons ainsi la relation EST_UN_ÉTAT_DE s'établissant, par exemple, entre *legs réclamé* et *legs* ou

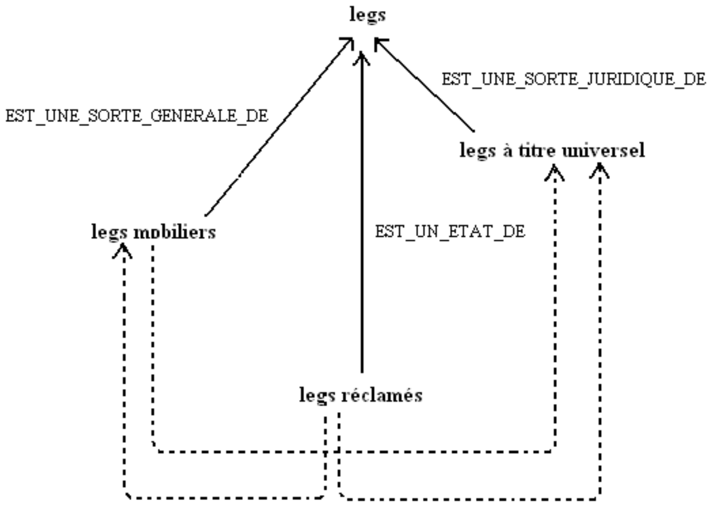

FIG. 7.4 – Articulation des liens de susbsumption pour le terme *legs*

entre *mandat détenu* et *mandat*.

Des règles peuvent être dégagées sur la façon dont les trois liens de subsumption s'articulent entre eux. Ainsi, la relation EST_UNE_SORTE_GÉNÉRALE_DE n'exclut pas la relations EST_UNE_SORTE_JURIDIQUE_DE. Par exemple (Figure 7.4), sachant qu'un *legs mobilier* EST_UNE_SORTE_GÉNÉRALE_DE *legs* et qu'un *legs à titre universel* EST_UNE_SORTE_JURIDIQUE_DE *legs*, l'établissement du premier lien n'exclut pas qu'un *legs mobilier* particulier puisse être qualifié de *legs à titre universel*. De la même façon, sachant qu'un *legs réclamé* EST_UN_ÉTAT_DE *legs*, le legs en question toutefois qualifiable de *legs à titre universel* ou de *legs mobilier*.

(*ii*) Le lien *composant*.

Le deuxième type de relations fréquentes dans une ontologie est le lien s'établissant entre les composants d'un tout et celui-ci. Ce lien permet en quelque sorte de produire une définition du tout. A ce titre, nous définissons le lien EST_UN_ÉLÉMENT_DE. Ce lien est par exemple établi entre *acte de vente* et *vente* ou entre *prix de la vente* et *vente*. Nous notons simplement que l'établissement de tels liens ne présume pas de conditions nécessaires et suffisantes. Par exemple, il est établi qu'une *résolution de vente* est un élément d'une *vente* qui permet de définir le concept de *vente*, sachant qu'une vente est constituée d'un acte et d'un prix. Simplement, toute vente ne se solde pas par une résolution[24]. Pratiquement, nous ne distinguons pas les conditions nécessaires et suffisantes des autres.

(*iii*) Le lien *est lié à*.

Nous définissons pour notre ontologie un troisième type de relation qui permet de lier entre eux des concepts sémantiquement proches. Par exemple, le concept *legs* est sémantiquement lié au concept *donation*, mais aucun ne permet de définir l'autre par les relations que nous venons de définir. Un legs se définit ainsi ([Cor01]) :

> «Acte unilatéral de disposition à cause de mort et à titre gratuit contenu dans un testament [...] par lequel le testateur laisse tout ou partie de ses biens en pleine propriété, en usufruit ou en nue-propriété à un légataire ou lui transmet un autre droit (usage, habitation, servitude, créance, action).»

Une donation se définit comme suit ([Cor01]) :

> «[...] contrat par lequel une personne - le donateur - se dépouille actuellement et irrévocablement sans contrepartie et dans une intention libérale d'un bien présent lui appartenant en faveur d'un autre personne - le donataire - qui l'accepte.»

Nous avons donc défini la relation EST_LIÉ_À qui s'établit par exemple entre *legs* et *donation*.

(*iv*) Les autres acceptions.

Enfin, dernier type de relation possible, celui permettant de signifier une autre acception que la définition du concept donnée par les autres relations le concernant. Par exemple, le concept *échange* se trouve défini, après modélisation, par les relations suivantes[25] :

- *échange international* EST_UNE_SORTE_JURIDIQUE_DE *échange*
- *échange multilatéral* EST_UNE_SORTE_JURIDIQUE_DE *échange*
- *échange de biens* EST_UNE_SORTE_GÉNÉRALE_DE *échange*
- *échange de parts* EST_UNE_SORTE_GÉNÉRALE_DE *échange*
- *vice rédhibitoire* EST_UN_ÉLÉMENT_DE *échange*
- *échange fixé* EST_UN_ÉTAT_DE *échange*

D'autres termes se trouvent être liés à *échange* dans la ressource ontologique comme *échanges de vues*, *échanges de renseignements*, *échanges de télégrammes*. Ces termes connotent l'idée de communication et pourraient

[24]La résolution d'une vente consiste dans la restitution au vendeur de la chose vendue en cas de non paiement du prix.

[25]Nous ne les avons pas toutes reportées.

se trouver liés à un concept *communication*. Il est clair que ces termes ne correspondent pas à l'acception telle qu'entendue en droit où un échange se définit de la sorte ([Cor01]) :

> «Contrat (translatif de propriété) par lequel les parties (copermutants) se donnent respectivement une chose pour un autre.»

Nous avons donc défini la relation AUTRE_ACCEPTION qui permet de lier, notamment, *échanges de vues* et *échange*. Cette relation permet de mettre à jour les phénomènes de polysémie. Ainsi, les termes liés à un concept mais présentant une acception différente que les autres termes liés sont mis de côté. Nous nous contentons donc de distinguer les termes et concepts liés à une acception donnée d'un concept, des autres. Nous ne prenons dès lors pas le parti de la *sémantique différentielle* telle que définie par Bachimont [Bac00]. Celle-ci veut qu'à chaque acception différente d'un concept corresponde autant de concepts différents. Cela suppose l'insertion de nouveaux concepts dans l'ontologie. Nous avançons que, dans l'usage qui est celui de notre ontologie, la création de nouveaux concepts pour chaque acception n'est pas nécessaire. Dans le cadre d'un système d'aide à la reformulation de requête, il nous semble suffisant de signifier à l'utilisateur lesquels des termes qui lui sont suggérés correspondent à une autre acception que celle, principale, portée par les autres termes. Le préalable d'un tel choix est de déterminer quelle acception d'un concept, quand il y en a plusieurs, est considérée comme principale par rapport à une autre. L'acception choisie est, logiquement, l'acception juridique. Cependant, les éléments de linguistique juridique que nous notons dans le Chapitre 3.3 traitent, entre autre, de phénomènes de polysémies internes, c'est-à-dire de termes juridiques présentant plusieurs sens dans le domaine juridique même. Ces termes sont alors porteurs de concepts juridiques différents. De notre point de vue, deux cas se distinguent alors. Dans le premier, les différents concepts auxquels se rapportent chacune des acceptions peuvent être rassemblés en un seul et unique concept. Par exemple le terme *dépôt* qui désigne à la fois le contrat de dépôt, l'action de déposer la chose confiée et le lieu du dépôt. Chacun de ces sens correspond bien à un concept différent. Nous prenons cependant le parti de rassembler ces concepts sous un seul et unique, étiqueté par le terme *dépôt*. Une modélisation fine du domaine nécessiterait, elle, de différentier ces différents concepts. Cependant, notre objet n'est pas d'établir un modèle inutilement détaillé du domaine, mais un modèle qui soit utile dans la tâche d'aide à la reformulation de requête.

Dans le deuxième cas, qui concerne notamment les polysèmes homonymes, le regroupement des différents concepts est plus problématique. Par exemple

([Cor00]), le terme *police* désigne à la fois l'ordre dans la cité ou la force publique et l'écrit constatant à fin de preuve certains contrats, comme la police d'assurance. Le terme *police* correspond alors à deux concepts différents. Autre exemple, le terme *forfait* qui renvoi à la fois au concept de *marché* puisqu'il peut désigner un marché fait d'avance, et à la fois au concept d'*infraction* quand il désigne une *transgression*. Les distinctions de ce genre ne sont que rarement détectées[26] par les termes et liens éléments de la ressource ontologique. Nous dévoilons alors une des limites de notre démarche d'élaboration d'ontologie. Ce genre de distinction de sens peut néanmoins être appréhendée par une analyse appropriée des autres données de la ressource ontologique, les données sur la provenance thématique des termes. Cette analyse ne sera cependant le fait que d'un spécialiste du domaine connaissant déjà le phénomène de polysémie en question. Quoi qu'il en soit, le choix donné au modélisateur de qualifier des termes liés comme correspondant à un AUTRE_SENS permet alors de mettre de côté plusieurs acceptions d'un terme. Le sens choisi sera alors celui qui se trouve le plus représenté dans les termes liés au terme en question. La pratique de la modélisation nous a conduit à limiter à une le nombre d'autre acception que l'acception principale d'un concept.

L'ensemble des qualifications sémantiques que nous pouvons apporter aux relations établies entre termes dans la ressource ontologique sont donc les suivantes :

- EST_UNE_SORTE_JURIDIQUE_DE ;
- EST_UNE_SORTE_GÉNÉRALE_DE ;
- EST_UN_ÉTAT_DE ;
- EST_UN_ÉLÉMENT_DE ;
- EST_LIÉ_À ;
- AUTRE_ACCEPTION.

Nous notons que cet ensemble de qualifications fait la particularité de notre ontologie. Ainsi les éléments de notre ontologie, chacun des concepts et des termes, est lié avec un autre par une de ces relations. L'établissement de ces relations permet de mener une modélisation des connaissances de la ressource ontologique vers la définition d'une ontologie. Ces relations sont également spécifiques du domaine juridique, principalement la distinction entre les liens de subsumption EST_UNE_SORTE_JURIDIQUE_DE et EST_UNE_SORTE_GÉNÉRALE_DE. La sémantique des relations est explicitée dans le nom qui leur est donné.

[26]Les termes liés dans la *ressource ontologique* aux termes concernés par la phase de modélisation à laquelle nous nous sommes attelés n'ont jamais permis la détection d'une telle polysémie.

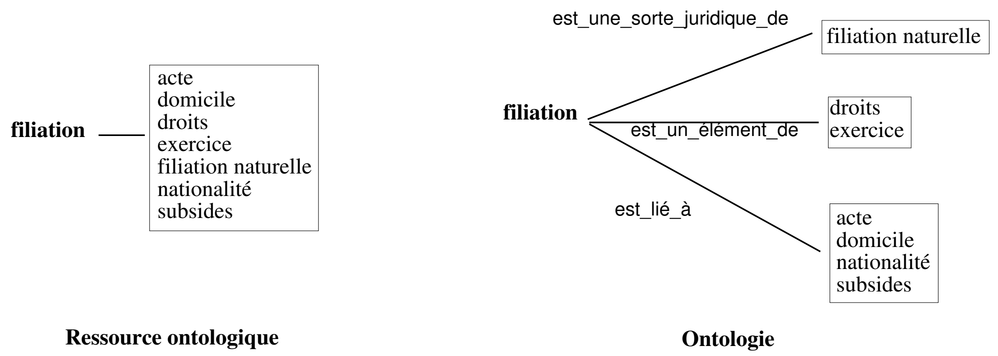

FIG. 7.5 – Modélisation du terme *filiation*

La Figure **??** représente le passage des données de la ressource ontologique à l'ontologie. D'un terme *nœud* et ses termes liés dans la ressource ontologique, on obtient un concept et ses relations sémantiques en constituant une définition dans l'ontologie.

7.3.2 La modélisation

La définition d'un ensemble fini de qualifications des relations permet la rationalisation de la tâche de modélisation. Sur cette base, nous avons élaboré une interface facilitant la tâche de l'ingénieur de la connaissance. Cette interface permet une modélisation en ligne de la ressource ontologique. Elle consiste, schématiquement, en un formulaire Web dans lequel peuvent être saisis les termes éléments de la ressource ontologique à modéliser, ceux que nous appelons les termes *nœuds*. Ce formulaire active un programme Web dynamyque Perl qui consulte la ressource ontologique et renvoie la liste des termes qui se trouvent y être liés au terme en question. En face de chacun de ces termes, un ensemble de boutons radio permet de qualifier le lien établi entre ce terme et le terme en cours de modélisation. Le choix est exclusif : une seule des qualifications que nous définissions dans la section précédente peut être sélectionnée. Quand chacune des relations est qualifiée, un bouton de validation actionne un autre programme Web qui collecte toutes les données dans un tableau associatif rémanent, ou *base de données*. Ainsi, les données de l'ontologie, les concepts et les termes reliés par des liens sémantiquement qualifiés, sont stockés dans une base de donnée.

Pour chacun des termes liés, en plus des qualifications telles que définies dans la section précédente, un autre choix a été aménagé. Il permet de mettre une relation *en suspens*, pour éventuellement y revenir plus tard.

Il s'avère en effet que, nous étant attelée à la tâche de modélisation, nous en avons perçu les difficultés. La principale limite réside sur le fait qu'une telle modélisation repose sur l'interprétation de connaissances approfondies du domaine. De telles interprétations peuvent par ailleurs faire l'objet de controverses théoriques. Ainsi, la modélisation n'est pas tâche aisée. Non pas que le choix de qualifications soit trop restreint, mais plutôt qu'une hésitation peut aisément émerger entre deux, voire plus, qualifications pour un même terme lié. Cette hésitation peut être résolue par la suite, lors de la modélisation d'un autre concept. Nous avons alors aménagé un autre choix en face de chaque terme lié, ce choix est appelé NON_DÉCIDÉ. Il va de soi que l'interface d'aide à la modélisation a été configurée de façon à ce qu'on puisse revenir sur un concept déjà modélisé. Ce faisant, le système rappelle les choix entre les différentes qualifications de liens qui avaient été faits et enregistre, le cas échéant, la nouvelle modélisation.

Par ailleurs, l'analyse de la tâche de modélisation a mis au jour le fait que certains termes se trouvent liés à d'autres dans la *ressource ontologique* alors même qu'ils ne sont pas valides. Ce fait est la conséquence majeure de notre choix méthodologique de ne pas procéder à une validation manuelle des éléments de la ressource ontologique. Ces termes non valides, c'est-à-dire non utiles à l'usage auquel est destinée l'ontologie, sont des termes qui ne peuvent être considérés comme des termes du domaine comme par exemple *échange des actions de cette dernière société contre les actions de la société absorbante* du fait de l'imprécision qu'il comprend, ou *obligations conventionnelles en général* du fait qu'il n'a pas de sens en tant que terme du domaine. Enfin, ces termes non valides peuvent être le résultat d'une erreur de l'outil d'extraction de termes comme *loyaux coûts de son contrat*. Pour ce genre de termes, la qualification A_IGNORER a été aménagée, qui permet au modéliseur de mettre de côté un terme. Concrètement, ce terme n'est en effet pas effacé ; il est simplement stocké dans un champ spécifique de la base contenant les données de l'ontologie. L'attribution de cette qualification A_IGNORER à un terme lié peut être assimilée à une sorte de validation manuelle des éléments de la ressource ontologique, bien que cette validation ait été exclue du processus d'élaboration de la *ressource ontologique* (Section 7.1.3).

Enfin, un espace est prévu sur la page HTML de modélisation pour saisir les commentaires jugés utiles par le modéliseur. Ces commentaires permettent, par exemple, de justifier un choix de modélisation ardu. Ils permettent de garder trace des choix ontologiques, ou *ontological commitments*.

(i) Les éléments à modéliser.

La tâche même de modélisation des données de la *ressource ontologique* repose sur un préalable : celui de la détermination de la liste des termes de la ressource ontologique qui vont faire l'objet d'une modélisation. Nous avons déjà établi que les termes de la ressource ontologique à même de faire l'objet d'une modélisation sont les termes appelés termes *nœuds*, c'est-à-dire les termes reliés à plus d'un autre. Sur l'ensemble des 103 994 termes éléments de la ressource ontologique, 17 688 se trouvent liés à plus d'un terme. Le Tableau 7.8 liste quelques uns de ces termes *nœuds*.

adoption plénière
cours de l'année
début du stage
directeur général de l'institut national de la propriété industrielle
durée légale
expéditions des décisions
investigations
objet du brevet

TAB. 7.8 – Quelques termes *nœuds* de la ressource ontologique

Nous ne pouvions envisager de lancer la procédure de modélisation de la ressource ontologique sur la base de cette liste classée par ordre alphabétique. Nous ne pouvions, en effet, au début de ce travail, déterminer si la modélisation des 17 688 termes *nœuds* de la ressource ontologique est une tâche réalisable, dans des délais raisonnables, ou non.

Nous avons donc choisi d'envisager la modélisation des termes *nœuds* de la ressource ontologique comme un processus progressif consistant à modéliser un ensemble de termes importants du domaine, ceux portant les concepts majeurs du droit, pour ensuite envisager les termes moins importants. Pour ce faire, nous sommes partis sur la base de ces termes que nous avons appelé, dans la Section 5.4.1, *termes fondamentaux du domaine*. Ces termes sont ceux, noms et syntagmes nominaux, qu'on trouve dans les intitulés des divisions des codes. Le présupposé est que, ces termes ayant été choisis pour intituler des sections de codes, ils sont importants, peut-être plus que les autres. Nous avons, de plus, choisi de circonscrire cette liste en adoptant un autre présupposé, celui selon lequel le Code civil peut être envisagé comme le code phare, le plus important de tous les codes car définissant des notions essentielles du droit. Nous prenons également le parti de ne considérer dans un premier temps que les noms présents dans les intitulés des divisions du Code civil. Le nombre de termes *nœuds* de la ressource ontologique qui correspondent à ces critères est de 393. Nous sommes donc partis de cette première liste pour procéder à la modélisation des termes *nœuds* qu'elle contient. Le Tableau 7.9 liste les premiers éléments de cette liste de termes.

absence
absents
acceptation
accession
acheteur
acquisition
acquêts
acte
action
administration

TAB. 7.9 – Extrait de la première liste de termes à modéliser

L'enjeu de l'expérience est de mesurer la quantité de travail qu'une telle modélisation nécessite, notamment et principalement en temps. Il réside également dans le retour d'expérience qu'on peut en faire, et les conclusions qu'on peut en déduire.

(*ii*) Résultats

Sur la base de l'expérience consistant à modéliser la liste des termes les plus importants, nous avons établi que le temps moyen de modélisation d'un terme, qui varie en fonction du nombre de termes qui lui sont liés, est d'environ 2 minutes. Nous pouvons dès lors faire de la prospective et considérer que la modélisation des 17 688 termes *nœuds* de la ressource ontologique nécessiterait 590 heures de travail. Il va sans dire qu'une telle perspective n'est guère envisageable. Nous atteignons là une limite de notre approche, principalement liée à la large couverture thématique de notre ontologie. Quoi qu'il en soit, le fait de considérer la tâche de modélisation comme un processus progressif partant des termes les plus importants du droit pour aller en importance décroissante permet d'envisager qu'une partie, la plus importante qualitativement, de la ressource ontologique peut être modélisée et constituer une ontologie en tant que telle.

Pour ce qui est du retour d'expérience sur le fond, nous avons saisi, lors de cette phase de modélisation de la ressource ontologique, la complexité d'une telle tâche. Cette complexité est essentiellement due à la subjectivité sur laquelle la modélisation repose nécessairement. Ainsi, qualifier des relations établies entre des termes relève foncièrement de l'interprétation. Dès lors, la modélisation faite par l'ingénieur de la connaissance reflète le point de vue de l'expert du domaine, qu'il se trouve être par ailleurs dans notre contexte. Les connaissances du domaine sont certes appliquées lors de la modélisation ; elles permettent par exemple d'établir qu'un *legs à titre universel* EST_UNE_SORTE_JURIDIQUE_DE *legs* tandis qu'un *legs mobilier* EST_UNE_SORTE_GÉNÉRALE_DE *legs*. Cependant, de nombreuses décisions de modélisation font l'objet de choix subjectifs, également connus sous le noms d'*ontological commitments*. Par exemple établir que *ressortissants* EST_UN_ÉLÉMENT_DE *communauté* alors que ce concept est défini, entre autre, par les relations suivantes :

– *communauté de ville* EST_UNE_SORTE_JURIDIQUE_DE *communauté* ;
– *communauté européenne* EST_UNE_SORTE_JURIDIQUE_DE *communauté* ;
– *communauté urbaine* EST_UNE_SORTE_JURIDIQUE_DE *communauté*.

D'aucun pourra arguer que le terme *ressortissant* et le concept qu'il porte sont propres au droit communautaire. La qualification établie est donc bien valable pour ce qui concerne le concept *communauté européenne*. Il est moins sûr que la qualification soit valable pour ce qui est des autres concepts, *communauté de ville* et *communauté urbaine*.

Le modèle du domaine résultant de la phase de modélisation de la ressource ontologique telle que nous la décrivons n'est donc qu'un modèle parmi d'autres envisageables, une interprétation parmi d'autres. Toute subjective qu'elle soit, la tâche de modélisation reste guidée par un impératif, celui de l'usage. La consigne donnée à l'ingénieur de la connaissance pour ce qui est de la modélisation est, chaque fois qu'une hésitation, un doute, surgit, de se poser la question de l'utilité de la qualification de la relation pour l'usager final du système d'aide à la reformulation de requête. Ainsi, notre phase de modélisation correspond à la phase de *conceptualisation* chez [FGPJ97], ou la phase appelée *normalisation* chez [AGBS00a], consistant à structurer les connaissances d'un domaine en fonction d'une application donnée.

Le retour sur expérience de la tâche de modélisation nous a amené à résoudre un certain nombre de points qui n'apparaissaient pas auparavant. Un de ces points est la contradiction pouvant être établie du fait qu'un même lien peut être modélisé deux fois. Une relation entre un terme 1 et un terme 2 peut en effet, et c'est la majorité des cas, être rencontrée lors de la modélisation du terme 1, qui se trouve lié au terme 2 et lors de la modélisation du

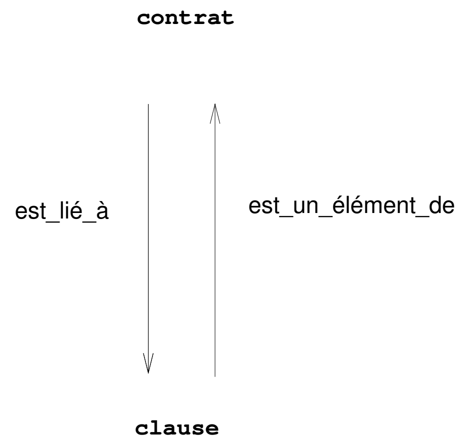

FIG. 7.6 – Double qualification du lien unissant *contrat* à *clause*

terme 2, lié au terme 1. Or aucune des qualifications des relations que nous avons définies ne connaît son opposé. Par exemple, la qualification opposée de EST_UN_ÉLÉMENT_DE serait EST_LE_TOUT_DE. Nous avons choisi, lors de la conception de l'interface d'aide à la modélisation, de ne pas proposer les relations inverses de chaque relation, de façon à ne pas alourdir, d'une part, les choix donnés au modéliseur et, d'autre part, l'interface elle-même qui fonctionne en ligne. Les qualifications inverses sont cependant supposées. Quand on établit que *clause* EST_UN_ÉLÉMENT_DE *contrat*, l'information est stockée dans la base de données contenant les éléments de l'ontologie. Dès lors, rien n'empêche de lire le lien dans l'autre sens, et de le présenter ainsi à l'utilisateur : *contrat* EST_LE_TOUT_DE *clause*.

Pour suivre le même exemple, schématisé dans la Figure 7.6, quand l'ingénieur de la connaissance se trouve à modéliser le terme *contrat*, il peut établir que le terme *clause* lui est lié par la relation EST_UN_ÉLÉMENT_DE. Lors de la modélisation du terme *clause*, la qualification de la relation devant être établie entre le terme *contrat* et *clause* est EST_LIÉ_À. La règle que nous posons est alors que les qualifications EST_UNE_SORTE_JURIDIQUE_DE, EST_UNE_SORTE_GÉNÉRALE_DE, EST_UN_ÉTAT_DE, EST_UN_ÉLÉMENT_DE, AUTRE_ACCEPTION sont *supérieures* à la qualification EST_LIÉ_À dans le cas où deux qualifications d'un même lien sont établies.

D'autres choix de modélisation ont dû être faits, par exemple, lors de la modélisation du terme *commodat*. Celui-ci se trouve lié, dans la ressource ontologique, au terme *prêt à usage*. Or il se trouve ([Cor01]) que *commodat* est l'ancienne dénomination des *prêts à usage*. Nous avons cependant choisi de qualifier le lien unissant *prêt à usage* et *commodat* de EST_UNE_SORTE_JURIDIQUE_DE.

Le résultat de la modélisation de la ressource ontologique telle que nous la décrivons ci-dessus est donc une ontologie rassemblant des concepts et des termes du domaine, des relations sémantiques entre ceux-ci et des informations sur les thématiques juridiques auxquelles ils se rattachent. Cette ontologie est partielle, puisque tous les éléments de la ressource ontologique ne font pas l'objet d'une modélisation. Cette ontologie est un modèle du domaine élaboré pour une application précise : un système d'aide à la reformulation de requête fondé sur l'ontologie.

L'ontologie que nous obtenons est une *ontologie conceptuelle*, ou ontologie *pré-computationnelle*, dans le sens où elle définit des concepts et des relations entre ces concepts dans un langage non formel et non opérationnel [Kas02a]. Les données de l'ontologie sont simplement stockées dans une base de données. L'*opérationalisation* de notre ontologie, l'étape par laquelle l'ontologie est intégrée à un système opérationnel, consiste à consulter la base de données et à en présenter les éléments appropriés à l'usager de notre système d'aide à la reformulation de requête. L'usage auquel est dédiée notre ontologie ne nécessite pas, de notre point de vue, sa spécification dans un langage formel et opérationnel de représentation des connaissances. Une telle étape nous paraît inutile du fait que la simple consultation des éléments de l'ontologie est le principal usage qui en est fait. Les services inférentiels proposés par certains langages de représentation d'ontologies ne sont pas nécessaires dans notre contexte. D'autre part, l'usage de notre ontologie se fait en ligne. La contrainte majeure est alors le temps d'accès aux éléments de l'ontologie. Le stockage de ceux-ci dans des tableaux associatifs rémanents et leur accès par le biais de programmes Web dynamiques nous permet de maximiser le temps d'accès en ligne à l'ontologie.

Troisième partie

Usages et évaluations d'une ontologie dédiée à la recherche d'informations sur le Web

L'ontologie dont nous décrivons les étapes d'élaboration est construite dans un but bien défini, celui d'être intégrée dans un système d'aide à la reformulation de requête. Le système consiste schématiquement à s'interposer entre l'usager et le moteur de recherche. En réaction à la requête de l'usager, notre système consulte l'ontologie et, sur cette base, suggère des termes plus ou moins liés à la requête permettant une reformulation de celle-ci. Dès lors, notre objet est de déterminer l'utilité que peut avoir un tel système pour l'usager, nous abordons ce point dans le Chapitre 9. Il nous a semblé que permettre la navigation dans l'ontologie était également un de ses usages envisageables. Une telle navigation dans l'ontologie, par la visualisation du réseau de concepts qu'elle constitue, lui confère des aspects pédagogiques de compréhension du domaine. Les moyens mis en œuvre pour une telle navigation et visualisation peuvent faire l'objet d'évaluations (Chapitre 8).

L'idée générale est de confronter l'ontologie que nous avons élaborée aux usages auxquels elle est destinée. A la suite de quoi, une évaluation dans l'usage de l'ontologie est envisagée. Des procédures d'évaluation d'ontologies ont déjà été définies, notamment dans [FGPJ97] et [FL99]. Celles-ci consistent essentiellement à porter un jugement technique général sur les ontologies, leur environnement logiciel et leur documentation, conformément à un cadre de référence préalablement défini. L'évaluation consiste alors en une vérification de la validité de l'ontologie et de son adéquation au système dans lequel elle est intégrée. Pour notre part, l'évaluation en usage de l'ontologie est moins stricte, dans le sens où son usage, ce à quoi elle est destinée, c'est de donner à penser à l'usager du système d'aide à la reformulation de requête. La validité de l'ontologie n'est, en conséquence, guère aisée à déterminer. Nous nous intéressons plus à l'utilité même de l'ontologie dans les usages auxquels elle est destinée, qu'à sa validité propre en tant que modèle du domaine.

Chapitre 8

Navigation et visualisation

L'ontologie dont nous décrivons l'élaboration dans la Partie II de cette thèse se présente sous la forme d'un réseau de termes reliés par des liens sémantiques qualifiés auxquels sont ajoutés des informations sur le rattachement thématique des termes. Un tel réseau constitue un modèle du domaine, une façon d'appréhender le droit français. Au-delà de l'application visée par l'ontologie et de son intégration dans un système d'aide à la reformulation de requête, celle-ci présente un certain intérêt à être visualisée en tant que telle. Nous avons donc aménagé un système permettant la navigation en ligne dans l'ontologie. Nous en présentons les principes dans la Section 8.1 pour ébaucher ensuite l'évaluation qui peut être faite d'un tel système Section 8.2.

8.1 Principe

Le système permettant la navigation en ligne dans l'ontologie permet la visualisation des informations contenues dans l'ontologie : les concepts et termes, les relations entre ceux-ci et leurs rattachements thématiques. Nous traitons dans un premier temps la visualisation du réseau que constitue les termes[1] et les relations entre ceux-ci (Section 8.1.1), pour traiter ensuite de la visualisation des thématiques rattachées à chaque terme (Section 8.1.2).

8.1.1 Le réseau de termes

L'accès à notre ontologie, sa visualisation[2], se fait par les termes de celle-ci, c'est-à-dire par les nœuds de l'arbre qu'elle constitue. Nous avons aménagé

[1]Pour faciliter notre discours dans cette Partie, nous utilisons le mot *terme* indifférem-ment pour désigner les concepts et les termes de l'ontologie.

[2]http ://ontologie.w3sites.net

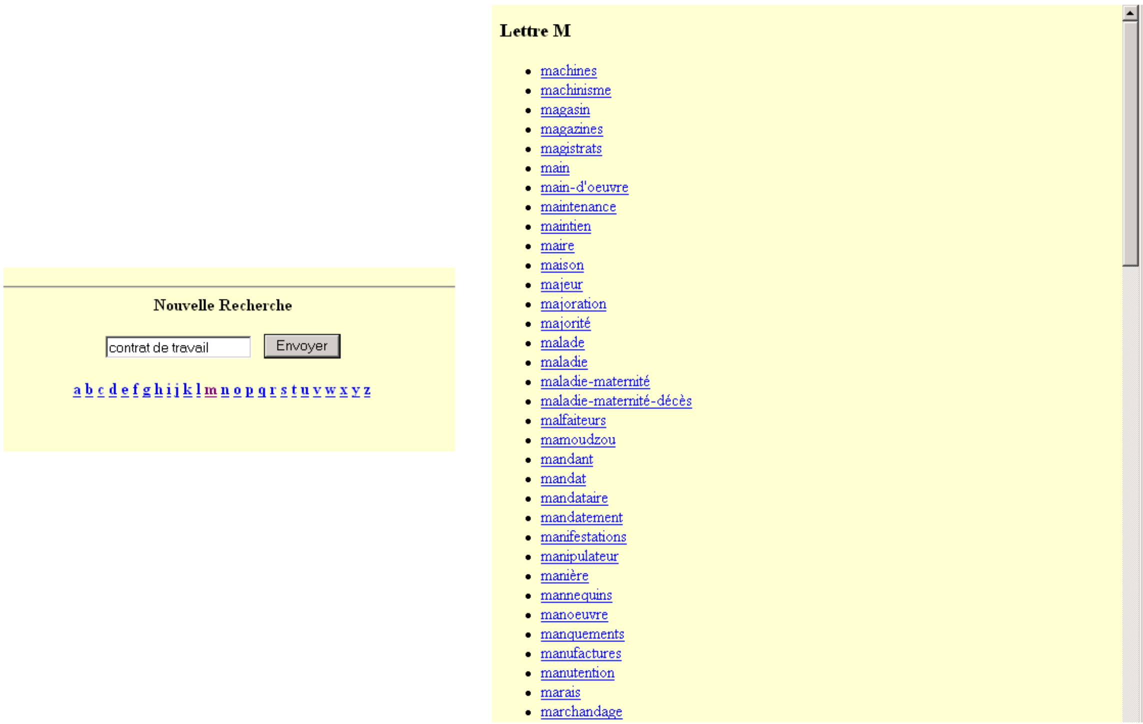

FIG. 8.1 – Le double accès aux termes de notre ontologie

un double accès à ces éléments. Le premier se fait par un formulaire Web de saisie. L'utilisateur peut y saisir les termes qu'il souhaite. A la validation de ce formulaire, une consultation Web en ligne de l'ontologie est lancée. La base de données dans laquelle les éléments de l'ontologies sont stockés est consultée. Sont alors renvoyés à l'utilisateur, sur une nouvelle page Web, le terme en question et l'ensemble des termes qui lui sont liés dans l'ontologie. Le deuxième accès se fait par le biais d'une liste de lettres. L'utilisateur peut alors sélectionner une lettre. Lui est alors renvoyée la liste des noms commençant par cette lettre correspondant au premier mot des termes de l'ontologie, ces premiers mots de termes sont aussi appelés *têtes*. De la même façon que précédemment, le terme en question et la liste des termes qui lui sont associés est renvoyée. La Figure 8.1 illustre ce double accès aux termes de notre ontologie.

La visualisation du réseau de termes se fait ainsi terme par terme. Une représentation graphique de parties du graphe, schématisant les nœuds et les arcs n'est pas envisagée. Une telle représentation nécessite en effet des applications Java ou Javascript. Nous avons souhaité demeurer dans une perspective d'accessibilité au plus grand nombre, ce qui empêche d'utiliser de telles solutions. La navigation dans notre ontologie se fait donc *à plat*.

La technique des fenêtres, ou *frames*, est utilisée, qui permet de partager une page en deux zones distinctes. Le découpage est ici vertical et la fenêtre de gauche concerne les termes sélectionnés et leurs termes liés tandis que la fenêtre de droite permet l'affichage de toute information relative à ces termes.

Pour chaque terme de l'ontologie, est renvoyée la liste des termes qui lui sont liés. Cette liste constitue une sorte de réseau sémantique autour du terme appelé dès lors concept. Nous envisageons les termes de l'ontologie comme des moyens d'accéder à des documents. Est alors constitué un système d'accès à l'information sur la base des concepts de l'ontologie. Ce procédé peut être rapproché du système expliqué dans [MBC$^+$00] procédant à l'accès à des documents par les composants des *diagrammes de connaissances*. Notre système est cependant bien plus simple, notamment par le fait que seuls les éléments concepts de l'ontologie permettent d'accéder aux documents et que ces accès se font uniquement par l'index des moteurs de recherche, aucun lien n'est établi *à la main* entre des termes et des documents. Dans notre système, pour chaque terme sélectionné et pour chacun des termes qui s'y trouvent liés, est aménagé un accès aux documents contenant le terme en question. L'indexation[3] est faite par plusieurs moteurs de recherche sur des corpus de documents différents. L'accès à cette fonctionnalité se fait par un bouton situé en face de chaque terme (terme sélectionné ou terme lié) se présentant sous la forme d'une loupe. En cliquant sur ce bouton, l'utilisateur se voit proposer un ensemble d'informations complémentaires sur le terme, notamment des informations sur les thématiques auxquelles il se rapporte (voir Section 8.1.2). Il peut alors sélectionner une fonction intitulée *Recherche sur ce terme*.

En cliquant sur ce lien, l'utilisateur se voit proposer plusieurs choix, illustrés dans la Figure 8.2. Le premier concerne le moteur de recherche qu'il souhaite utiliser. Le second concerne le corpus de documents dans lequel il souhaite lancer sa recherche. Deux moteurs peuvent en effet être utilisés.

(i) Le moteur Pertimm[4] a été installé sur les documents du site droit.org.

[3]Nous nous traitons pas ici de l'*indexation conceptuelle* au sens classique, notamment définie dans [Pri00] comme : *toute connaissance ajoutée à un document pouvant servir dans le cadre de calculs sous-tendus par l'exploitation de ces documents, pourvu que cette connaissance soit utilisable aussi bien par l'homme que par la machine.*

[4]http ://www.pertimm.com

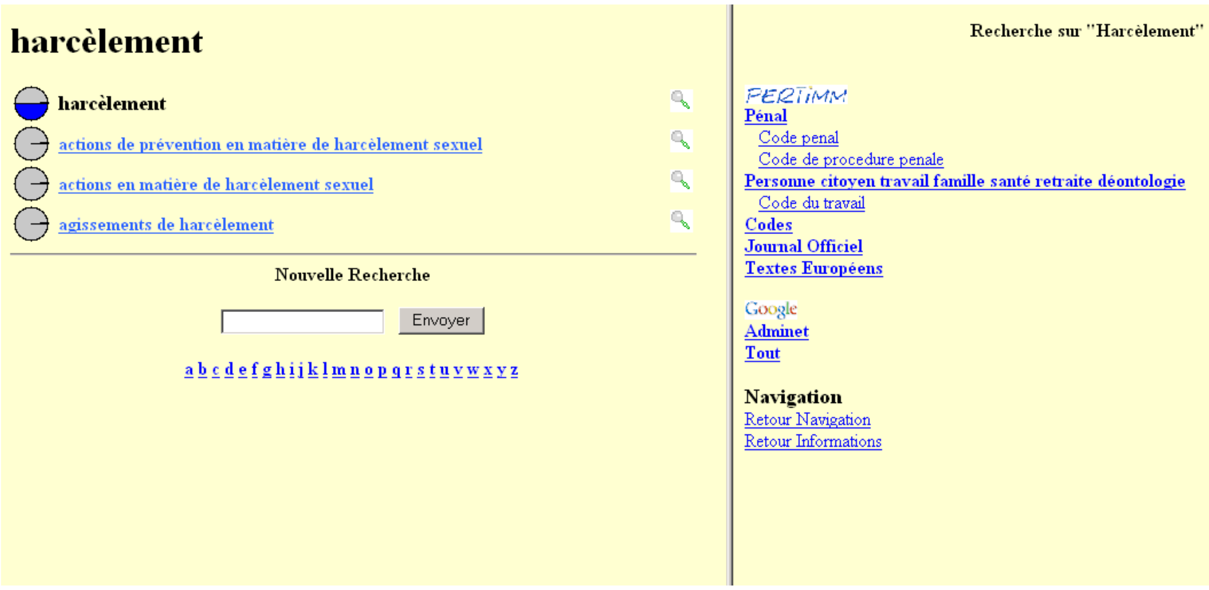

FIG. 8.2 – Termes associés et recherches possibles sur le terme *harcèlement*

Ce moteur permet de circonscrire la recherche à un ensemble donné de documents. Nous avons donc aménagé plusieurs façons d'utiliser le moteur :

– la première consiste à accéder aux articles des codes par les grands thèmes que nous définissons dans la section suivante ;
– la seconde permet d'accéder aux documents du Journal Officiel ;
– et la troisième aux documents européens (Directives et règlements).

L'accès aux articles des codes par des termes eux-mêmes extraits de ces codes peut paraître quelque peu redondant. Cela permet cependant d'accéder aux textes des articles dans une perspective de recherche d'information, de même pour la recherche dans les documents du Journal Officiel et dans les directives et règlements communautaires. La séparation de chacun de ces corpus de recherche permet de cibler la recherche qui devient par là plus précise. Un inconvénient à l'utilisation du moteur Pertimm sur les termes de notre ontologie réside dans la façon différente qu'il a de concevoir les termes, et principalement les groupes de mots, au regard de celle de notre outil d'extraction de termes Syntex. Le moteur Pertimm procède à une indexation des documents qui lui sont soumis non seulement par les mots seuls mais également par les groupes de mots qu'il identifie. Le moteur Pertimm utilise une logique différente de notre extracteur de termes complexes. Ces différences résident essentiellement dans les règles grammaticales qui ne sont pas appliquées de la même façon et par l'identification des mots *stop* qui peut également se faire de manière différente. Dès lors, des requêtes lancées sur Pertimm avec des termes générés par Syntex peuvent se révéler infructueuses alors même que le terme existe bien dans les documents. Ce problème se rencontre par exemple pour le terme *intention de rompre le contrat de travail*.

(*ii*) Le second moteur utilisé pour les recherches à partir des termes de l'ontologie est le moteur Google. Google[5] indexe une grande partie des pages du Web. A ce titre, de nombreuses pages, mais pas toutes, des sites admi.net et droit.org sont indexées par Google. De plus, le moteur rend possible la restriction de requêtes à un nom de domaine donné. Nous utilisons le moteur Google sur les termes de notre ontologie, les recherches étant limitées au site admi.net. Cette solution n'est cependant pas totalement fiable puisque toutes les pages du site ne sont pas indexées. Nous proposons également de faire une recherche par Google sur les termes de l'ontologie dans le Web général. Cette dernière possibilité donne des résultats d'autant plus satisfaisants que le terme est précis. Ainsi, une requête qui utilise le terme *acte* de l'ontologie sur la totalité du Web renvoie des documents assez divers issus de sites d'associations professionnelles ou de conception de sites Web. Par contre, une requête sur le terme *acte authentique* renvoie des documents de sites de notaires, de l'Association pour le Développement de l'Informatique Juridique[6] ou du site de la Mission de recherche Droit et Justice.

La représentation graphique que nous proposons du réseau de termes que constitue notre ontologie se fait par la spécification du terme *nœud* sélectionné et la liste des termes qui lui sont associés. Faute de temps, nous n'avons pu implémenter l'affichage dans l'interface des qualifications sémantiques des relations établies entre termes. Dans une première interface, nous nous contentons de séparer les relations issues de l'analyse des formes (désignées *spécialisation*) et les relations issues des autres méthodes (désignées *autres voies*).

Une telle spécification de la sémantique des relations entre termes a d'incontestables aspects d'apprentissage. Elle permet à l'utilisateur d'appréhender l'articulation existant entre concepts juridiques.

8.1.2 Les thématiques

A chaque concept ou terme de l'ontologie correspond des informations quantifiées sur leurs rattachements thématiques. Ces thématiques correspondent au domaine du droit traité par les codes desquels proviennent les termes en question. En effet, la méthode d'élaboration d'ontologie que nous

[5]http ://www.google.fr
[6]http ://www.adij.asso.fr

utilisons fait que les termes sont directement issus des documents de notre corpus de référence. Ce corpus de référence rassemble un ensemble des codes du droit français. Or chaque code traite d'un domaine particulier du droit, le droit civil pour le Code civil, le droit pénal pour le Code pénal etc. Disposant pour chaque terme des codes dans lesquels il apparaît et du nombre d'occurrences, nous pouvons affecter à chaque terme (ou concept) de l'ontologie, des thématiques. Le Tableau 8.1 liste ainsi un ensemble de termes et leurs thématiques.

créancier	civil commerce justice administrative organisation judiciaire pénal procédure civile procédures fiscales procédure pénale propriété intellectuelle travail
espaces naturels	environnement rural urbanisme
ouvrages publics	civil expropriation rural

TAB. 8.1 – Quelques termes de l'ontologie et leurs thématiques

L'ensemble des thématiques correspondant à autant de codes, c'est-à-dire 58, nous ne pouvions envisager de présenter chacune de ces thématiques pour chaque terme, d'autant plus que certains, ceux correspondant à des concepts génériques du droit[7], rassemblent plus de 30 thématiques différentes. Nous avons donc rassemblé ces 58 thématiques en 8 groupes de thèmes plus larges, à chacun étant attribué un titre. Ces 8 thèmes sont les suivants :

– Civil ;
– Pénal ;
– Personne Citoyen Travail Famille Santé Retraite Déontologie ;
– Commerce Economie Marchés Assurances Consommation ;

[7]Comme, par exemple, *acte*, *contrat*, ou *personne*.

– Impôts ;
– Aménagement Construction Collectivités Rural Environnement ;
– Domaine et prérogatives de l'Etat ;
– Justice Juridictions.

Chacun des 8 thèmes regroupe un ensemble de codes, donc de thématiques juridiques. La liste des thèmes et les codes correspondant est disponible en Annexe.

Pour la visualisation dans l'interface de navigation des thèmes rattachés à chaque terme de l'ontologie, nous avons choisi la symbolique des couleurs dans un disque, plus connue sous l'appelation *camembert*. A chacun des 8 grands thèmes est associée une couleur. A chaque concept ou terme de l'ontologie est associé un disque dont le partitionnement se fait selon les fréquences d'apparitions dans les codes correspondant aux grands thèmes. La création des camemberts pour chaque terme se fait de façon dynamique, par un programme Web consultant les données de l'ontologie dans les tableaux associatifs, en utilisant la bibliothèque graphique Perl *GD*.

Ainsi, lors de la navigation dans l'ontologie, l'usager, qu'il ait eu accès au concept en question par la liste alphabétique ou par le formulaire de saisie libre, dispose des informations thématiques sur le concept en question et pour chacun des termes qui lui sont liés. L'écran est alors séparé en deux, la partie de gauche listant les termes et leurs camemberts respectifs, la partie de droite rappelant les 8 grands thèmes, leurs symboliques de couleurs et les codes auxquels ils correspondent. La Figure 8.3 illustre ces points.

Expliciter les thèmes se rattachant à chaque élément de l'ontologie a, nous semble-t-il, des aspects pédagogiques de compréhension du domaine. On constate ainsi que certains termes sont plus génériques que d'autres, du fait qu'ils rassemblent un grand nombre de thèmes. Par exemple, à *acquisition*, *administration* ou *gestion* correspondent des *camemberts* thématiques bariolés. D'autres termes se trouvent être beaucoup plus spécifiques comme *legs à titre universel*, *jeunes salariés*, *société d'économie mixte de construction* ou *litispendance* ; chacun correspond à un seul thème juridique, leurs *camemberts* sont donc monochromes. Ces informations sur les thématiques rattachées aux termes permettent de voir dans quels sous domaines du droit un même terme est employé, donc valable. Une personne un peu avertie du domaine pourra également en déduire que différents sens peuvent alors être attachés à un même terme, en fonction des domaines concernés. Ainsi, une *personne* ne sera pas envisagée de la même façon selon le thème en question. Dans le thème *civil*, la personne sera envisagée du point de vue des droits civils, de la nationalité, du mariage, en tant que partie à un contrat etc. Dans

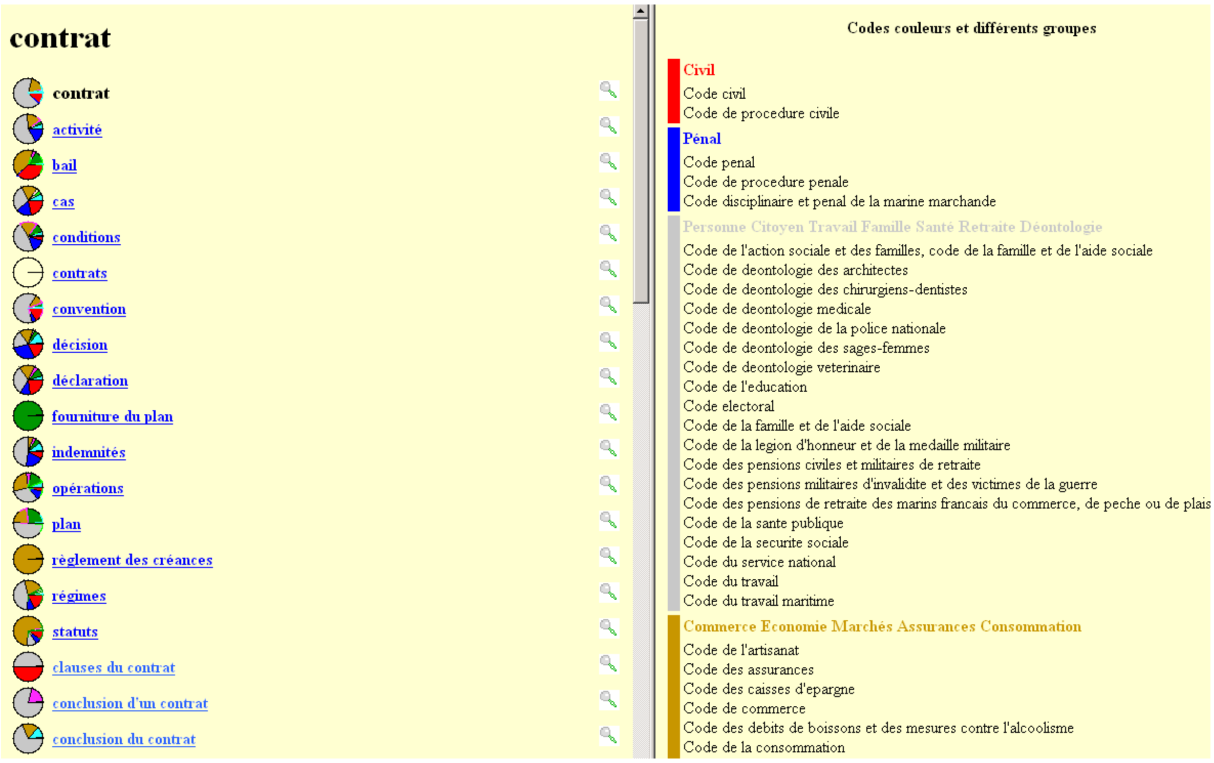

FIG. 8.3 – Termes et leurs thématiques associés au terme *contrat*

le thème *pénal*, elle sera envisagée en tant qu'auteur ou victime d'un délit etc.

8.2 Evaluation

8.2.1 Principes d'évaluation

Le premier élément d'évaluation d'une telle interface de navigation dans l'ontologie repose sur le temps d'accès aux données. Ainsi, quand un terme est choisi dans la liste ou lorsqu'il est saisi dans le formulaire prévu à cet effet, un programme Web dynamique consulte la base de données dans laquelle est stockée l'ontologie et affiche les résultats. Le temps d'exécution de ce programme est variable, il dépend du terme sélectionné. En effet, la base de données contient toutes les informations relatives à l'ontologie. Ainsi, les liens de chaque terme sont stockés, indépendamment de leur nombre. Or, comme nous l'expliquons dans la Section 6 on ne peut se permettre, pour des questions de lisibilité et donc d'accessibilité à l'information, d'afficher un nombre trop important de termes liés à un même terme. Ce nombre est donc réduit par différents moyens définis dans la Section 6. L'exécution de l'algorithme concerné se fait à la consultation de la base, de façon dynamique. De plus, la construction des *camemberts* thématiques pour chaque terme se fait également en ligne. Ces deux processus se déroulent dans un temps moyen variant globalement en fonction du nombre de termes liés au terme en question. Le maximum est atteint pour de rares termes présentant plus de 200 termes liés, le temps d'affichage de la page des termes liés et de leurs *camemberts* est de 17 secondes, moins d'une seconde est nécessaire pour afficher la plupart des termes, le tout sur une ligne à 56 Kbits.

Pour ce qui est de la technique de représentation de l'information que nous adoptons, des listes de termes, à chacun étant associées diverses fonctionnalités, le choix que nous avons fait est essentiellement guidé par l'accessibilité technique et matérielle à ces informations. Notre souci est que ces informations ne soient pas trop longues à charger pour l'ordinateur client. Pour une étude détaillée sur les différentes techniques de visualisation de l'information, notamment les techniques de mapping ou de visualisation dynamique, voir [FHRT01].

Notre interface de navigation dans l'ontologie est en effet tout à fait comparable à celles développées pour la navigation dans des thésaurus ou dans des réseaux conceptuels. Une étude globale ([RLK00]) mène une comparaison de quelques uns de ces travaux, en souligne les caractéristiques, tout en notant que peu s'attachent à une réelle évaluation de leurs interfaces. Les conclusions globales de cette étude sont notamment les suivantes :

- D'une part, la plupart des travaux menés sur l'interfaçage de réseaux de termes partent du principe qu'une représentation spatiale, sous forme de graphe par exemple, est toujours plus lisible qu'une représentation textuelle, par liste de termes. Or, les auteurs de l'étude notent que ce principe n'est pas forcément vrai, bien que (la remarque est de nous) des évaluations quantifiées aient montré que les utilisateurs adhérent à ce genre d'interfaces graphique [CHSS98]. Nous notons à cet égard que la taille du réseau est, dans le choix d'une représentation spatiale ou textuelle, un paramètre important. La représentation spatiale d'un réseau de terme important nécessite l'aménagement de mécanismes de zoom sur le graphe.
- D'autre part, le type de représentation que nous avons choisi, consistant à considérer un terme après l'autre et d'y associer ses termes fils, peut conduire au phénomène de désorientation des utilisateurs dès lors qu'ils ont saisi plus d'un terme. Il leur devient difficile de se représenter mentalement le réseau de termes puisqu'ils n'en ont qu'une vue parcellaire.
- Enfin, pour ce qui est de la représentation sur l'interface des relations établies entre les termes, certaines études avancent que la spécification de leur signification, de leur étiquette, est nécessaire, notamment du fait du rôle explicatif alors rempli. Nous notons à cet égard la nécessité de spécifier les significations des relations établies entre les termes. Elle l'est dans une perspective d'apprentissage, de pédagogie. Simplement, on pourrait dire que *trop d'information risque de tuer l'information* et qu'un équilibre doit être trouvé entre l'information fournie à l'utilisateur, celle à laquelle il accède effectivement, et celle qui lui est utile.

8.2.2 Futurs travaux

L'évaluation d'interfaces telle que celle que nous venons de décrire pour la visualisation et la navigation dans l'ontologie relève d'études d'ergonomie dont résultent des données essentiellement qualitatives. De telles études tentent de déterminer la façon dont les usagers de ces interfaces perçoivent les informations qui leurs sont fournies. Ces évaluations se font en terme d'utilisabilité, ou *usability*, elles déterminent la façon dont les usagers s'approprient le site en question pour ce qui est de la navigation, de l'utilisation des outils de recherche, de la compréhension du rubricage et de la signalétique. De telles études font l'objet de plus en plus d'intérêt dans la communauté de recherche sur le Web, en témoigne, par exemple, le thème principal de la revue *IEEE Internet Computing* d'avril 2002 ([NHR02]) : *Usability and the Web*. Ces études utilisent différents outils comme les questionnaires fournis aux utilisateurs (voir [CDN88]), les interviews, le fait de filmer les réactions des utilisateurs, de mesurer plus précisément les mouvements de l'œil sur l'écran etc. Nous n'avons pas procédé à une telle évaluation.

Chapitre 9

Aide à la reformulation de requêtes

L'usage principal auquel est dédiée notre ontologie est d'être intégrée dans un système d'aide à la reformulation de requête. Nous en décrivons les principes dans la Section 9.1. Un tel usage de notre ontologie peut faire l'objet d'une procédure d'évaluation, nous y consacrons la Section 9.2.

9.1 Principe

Nous décrivons dans un premier temps (Section 9.1.1) le principe général de la reformulation de requête telle que nous l'envisageons. Dans un deuxième temps (Section 9.1.2), le système tel que nous l'avons conçu intégrant l'ontologie est explicité.

9.1.1 La reformulation de requête

La reformulation est une méthode qui tente de résoudre les problèmes posés par les différences pouvant résider entre le langage de la requête de l'utilisateur d'un moteur et le langage des documents interrogés. Le problème est alors de faire correspondre les termes de la requête et les termes des documents. Pour faire correspondre ces deux langages, il peut être opéré à différents niveaux ([ABD+81]) :

- au niveau des documents de la base en fonction du langage de requête ;
- au niveau de la requête, connaissant les liens entre le langage de requête et le langage des documents ;
- et enfin à la fois au niveau de la requête et des documents en produisant une représentation normalisée couvrant les deux langages.

Notre démarche se place dans le deuxième cas pour, disposant d'une base de connaissances sur le langage de la base de documents et par là sur le langage du domaine (l'ontologie), agir au niveau de la requête pour la faire correspondre un maximum avec ce langage.

Globalement, un procédé de reformulation de requête consiste à ajouter ou substituer aux termes d'une requête initiale des termes qui lui sont plus ou moins liés dans le but d'améliorer la recherche d'information. Ce procédé interfère donc entre la requête initiale d'un utilisateur et le moteur de recherche en question. Deux méthodes peuvent alors se distinguer :

- soit la reformulation de la requête se fait automatiquement, sans intervention de l'utilisateur. On parle de système d'expansion de requête, ou *query expansion system* ;
- soit les éléments de reformulation sont suggérés à l'utilisateur qui procède lui-même à la reformulation. On parle alors de système interactif d'aide à la reformulation de requête, ou *interactive query expansion system*.

Nous nous plaçons dans le deuxième cas pour élaborer un système se contentant de suggérer des termes lié à sa requête à l'utilisateur du moteur de recherche. L'idée sur laquelle notre système repose est que l'utilisateur est le plus à même de déterminer quels sont les termes qui lui sont utiles pour reformuler une requête. Notre système d'aide à la reformulation de requête est ainsi un système qui *donne à penser* à l'utilisateur. Une étude ([MP01b]) a spécifiquement montré que les moteurs de recherche procédant à des modifications automatiques des requêtes des utilisateurs les perturbaient. L'étude porte essentiellement sur les moteurs de sites commerciaux qui, souvent, remplacent les espaces entre les mots de la requête par des opérateurs booléens, éliminent des mots considérés inutiles ou attribuent différents degrés de priorité aux termes de la requête. Les auteurs montrent que ce genre de procédés, quand ils sont faits à l'insu des utilisateurs, sont extrêmement perturbants pour eux. Ceci est principalement dû au fait que la représentation que les utilisateurs se faisaient du fonctionnement du moteur[1] est contestée, il en découle un sentiment de frustration face aux résultats du moteur. Nous pensons que ce raisonnement est transposable aux moteurs procédant à des reformulations automatiques de requêtes et y préférons les systèmes interactifs se contentant de suggérer des pistes de reformulation aux utilisateurs.

Les systèmes de reformulation de requête puisent les informations nécessaires à la reformulation de différentes façons. Le système peut procéder à des calculs à la volée sur les documents de la base résultats de la requête

[1] *Users' mental models of search engines*

initiale. Différents algorithmes sont alors utilisés, permettant de détecter les termes non éléments de la requête initiale qui sont néanmoins représentatifs de ces documents. Pour une description exhaustive de toutes les méthodes, voir [Eft96]. Le système peut utiliser des connaissances extérieures aux documents recherchés et stockées par ailleurs dans des structures comme des thesaurus, des terminologies, des réseaux de termes etc.

Des systèmes de reformulation de requête, comme celui installé dans SPIRIT ([ABD$^+$81]) qui repose sur l'établissement d'une représentation normalisée commune entre les requêtes et les documents, combinent diverses approches pour mener différentes types de reformulations :

- la lemmatisation consistant à représenter tous les mots des documents et requêtes sous forme canonique ;
- la détection d'erreurs de frappe ou autre notamment par comparaisons phonétiques et morphologiques entre les termes des documents et ceux des requêtes ;
- l'identification dans les documents de cadres morpho-syntaxiques particuliers comme les listes ou les définitions et leur explicitations dans les représentations internes des documents ;
- un autre niveau paradigmatique que la lemmatisation est ajouté à la représentation des requêtes, qui associe aux termes de la requête des termes liés et des synonymes. Ces éléments se basent sur des dictionnaires préalablement crées de familles de mots (par exemple *utiliser*, *utilité*, *utilisation* ...) ;
- l'utilisation de thesaurus pour le rapprochement de termes de la requêtes avec des termes des documents sémantiquement liés ;
- ou encore l'utilisation d'explications données par les utilisateurs même du moteur quand leur requête n'a pas aboutie aux résultats escomptés (les termes mêmes de ces remarques pouvant être utilisés à fin de reformulation).

Notre démarche de reformulation agit au niveau de la requête. Disposant d'une base de connaissances sur le langage de la base de documents et par là sur le langage du domaine (l'ontologie), notre système agit au niveau de la requête pour la faire correspondre un maximum avec ce langage. Nous utilisons ainsi notre ontologie comme seule et unique source de connaissances à la reformulation de requêtes. Le système est configuré pour permettre l'interrogation de plusieurs corpus de documents juridiques du site droit.org ainsi que le Web, dans sa totalité.

9.1.2 Système d'aide à la reformulation de requête basé sur notre ontologie

Le système que nous avons conçu d'aide à la reformulation de requête basé sur notre ontologie se fonde donc sur elle pour suggérer à l'utilisateur des termes plus ou moins liés à ceux de sa requête. Le système est implémenté dans un programme Web Perl. Les différentes étapes par lesquelles procède le programme sont les suivantes :

- le découpage de la requête. La requête est envisagée mot par mot. A partir de ces mots, des termes sont fabriqués, constitués des mots seuls, des mots associés par deux, par trois etc. Notre algorithme de découpage des requêtes et de constitution de termes conserve l'ordre des mots choisi par l'utilisateur. Pour reprendre l'exemple que nous explicitons dans [BL02], sur une requête *rupture du contrat de cession*, notre programme va identifier les termes suivants : *rupture*, *contrat*, *cession*, *rupture du contrat*, *contrat de cession* et *rupture du contrat de cession* mais pas les termes *rupture de cession* ni *cession de contrat*.
- l'identification des termes connus. A partir de la liste des termes constitués sur la base de la requête, le programme identifie ceux qui sont connus dans l'ontologie, ceux auxquels correspond une entrée dans la base de données. La base de données, ou tableau associatif rémanent, étant chargée en local par le programme dans un tableau de hashage, l'accès aux données de l'ontologie est optimisé en temps. Dans la liste des termes fabriqués à partir de la requête, ne sont alors conservés que les termes connus dans l'ontologie.
- le renvoi des termes connus et de leurs termes associés. Les termes liés dans l'ontologie aux termes connus sont récupérés dans la base de donnée. Chaque terme connu est spécifié à l'utilisateur ainsi que la liste des termes qui lui sont liés dans l'ontologie. Cette liste de termes ne fait pas l'objet de pondérations, ou *ranking*, elle est simplement présentée par ordre alphabétique, ce qui, nous semble-t-il, en facilite l'accès par l'utilisateur.
- la reformulation de requête. L'utilisateur peut alors choisir dans ces listes les termes qu'il estime utile pour la formulation d'une nouvelle requête. Un nouveau formulaire est aménagé, permettant la saisie d'une nouvelle requête. Deux moteurs peuvent être choisis pour lancer cette nouvelle requête. D'une part le moteur Pertimm sur l'ensemble des documents du site admi.net et d'autre part le moteur Google sur les pages qu'il a indexé du site. Le formulaire envoyant la requête reformulée, ou non, aux moteurs, permet l'utilisation des opérateurs booléen *ET* et *OU* entre les termes, ainsi que la spécification des termes eux-mêmes

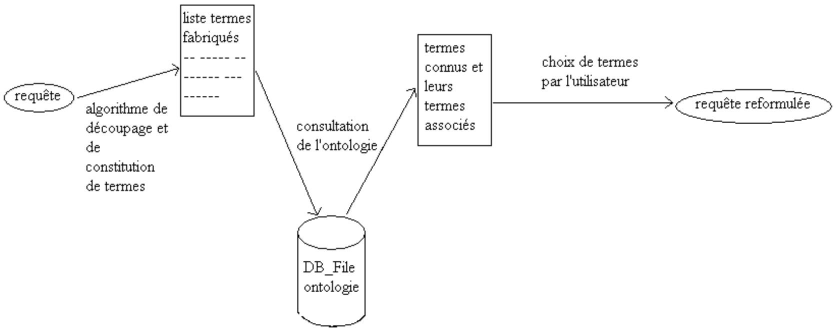

FIG. 9.1 – Le système d'aide à la reformulation de requête basé sur l'ontologie

entre guillemets. Les opérateurs booléens permettent ainsi à l'utilisateur de combiner les termes de sa requête avec les termes qui lui sont suggérés pour formuler une nouvelle requête.

La Figure 9.1 illustre les premières étapes de notre système d'aide à la reformulation de requêtes.

Pour ce qui est de la sémantique des relations établies dans l'ontologie entre les termes, nous n'avons pas pu faire les développements nécessaires pour les expliciter à l'utilisateur. Les raisons en sont des raisons de temps. Les termes liés à ceux constitués sur la base de la requête sont simplement distingués selon qu'ils contiennent le terme en question (ce sont les termes issus de l'analyse des formes) ou non (les termes issus des autres méthodes de détection de liens entre termes). Nous proposons ainsi les premiers sous l'appelation *spécialisation*, ces termes correspondent généralement aux relations EST_UNE_SORTE_JURIDIQUE_DE, EST_UNE_SORTE_GÉNÉRALE_DE, et EST_UN_ÉTAT_DE et les seconds sous l'appellation *autre voie*, ceux-ci correspondant généralement aux relations EST_UN_ÉLÉMENT_DE et EST_LIÉ_À.

9.2 Evaluation

Pour ce qui est de l'estimation du temps nécessaire à notre programme pour opérer le découpage de la requête, la constitution des termes, leur reconnaissance dans l'ontologie et la récupération des termes liés, celui-ci est fonction du nombre de mots composant la requête. Plus celle-ci en contient un nombre important, plus le temps nécessaire à la création de la page des termes suggérés sera long. Ainsi le temps d'affichage de la page pour une requête de un mot est de moins d'une seconde, celui pour une requête de sept mots est d'environ 20 secondes.

Des études ont déjà été menées, procédant à l'évaluation de systèmes d'expansion de requêtes sur la base de structures comparables à des ontologies. Voir, par exemple, [HPD00]. Ces études concernent des systèmes automatiques de reformulation de requêtes, des comparaisons en terme de rappel et précisions peuvent alors être menées entre la requête initiale et la requête reformulée. L'étude que nous citons montre en particulier que ce genre de reformulations automatiques de requêtes n'est pas satisfaisant, plus de la moitié des requêtes reformulées présentant des taux de rappel et de précision moins intéressants que les requêtes initiales, sachant que les opérateurs booléens ne sont pas utilisés.

L'évaluation d'un système d'aide à la reformulation de requête tel que nous le décrivons ne peut cependant se faire qu'en mesurant la satisfaction des usagers du système. Une telle évaluation est décrite dans [Eft00]. Elle s'attache à l'utilisateur et à la mesure de sa satisfaction de l'outil notamment par le biais de questionnaires sur les termes suggérés et sur les documents réponse. Une évaluation de notre système d'aide à la reformulation de requête a été menée, simplement, elle concernait la ressource ontologique et non pas l'ontologie en tant que telle. Cette évaluation (Section 7.2) suggère que les utilisateurs du système d'aide à l'expansion de requête estiment, pour une majorité de requêtes, que plus de la moitié des termes suggérés est utile à la reformulation d'une requête. En faisant une telle évaluation en usage de la *ressource ontologique*, nous procédons de fait à l'évaluation d'une grande partie de l'ontologie : sa structure et les termes qu'elle contient.

Nous n'avons pas procédé à l'évaluation des qualifications sémantiques des relations entre termes établies lors de la phase de modélisation. Une telle évaluation n'aurait pû être le fait que de spécialistes du domaine, elle se serait faite non pas en terme d'utilité, mais en terme de validité du modèle. Or ce type d'évaluation de notre ontologie ne nous semble pas utile au regard de l'usage auquel est destinée notre ontologie. Il nous semble dès lors que l'évaluation en usage de la ressource ontologique conduit l'évaluation en usage de l'ontologie en elle-même. Nous renvoyons donc aux résultats de l'évaluation de la ressource ontologique, Section 7.2, constituant, de fait, une évaluation de l'utilité en usage de notre ontologie.

Conclusion

Conclusion

Nous avons présenté dans cette thèse les étapes qui ont prévalu à l'élaboration d'une ontologie du droit français dédiée à la recherche d'informations sur le Web. La démarche qu'elles sous-tendent s'inscrit dans la continuité de principes méthodologiques définie dans des travaux relevant de l'ingénierie des connaissances et alliant la terminologie et l'intelligence artificielle. Ces principes sont principalement l'utilisation de textes et l'exploitation d'outils de traitement automatique de la langue ainsi que la large place donnée à l'usage et à l'application visée par l'ontologie. Notre démarche se distingue cependant de ces travaux, notamment du fait de la large couverture thématique portée par notre ontologie. En résulte des choix méthodologiques singuliers comme le choix de la non validation des résultats des outils par un expert ou le fait d'envisager les connaissances linguistiques structurées (les termes et leurs relations dans la ressource ontologique) comme l'ossature de l'ontologie.

Le principal apport de notre approche est ainsi une nouvelle dimension apportée aux travaux relevant de l'ingénierie des connaissances. L'impératif guidé par l'usage d'une ontologie présentant une large couverture thématique (notre ontologie comporte plus de 100 000 termes) ouvre de nouvelles pistes de recherche parallèles à celles conduisant à la définition d'ontologies certes précises mais de couverture restreinte. La place des tâches manuelles dans le processus d'élaboration d'ontologies peut alors être discutée et repensée.

Nos travaux ont été conduits dans une perspective majeure, celle de construire une ontologie répondant à un usage. L'idée n'est certes pas nouvelle, notamment dans ses conséquences méthodologiques, mais elle mène dans notre cas à une interrogation majeure : étant donné l'usage de recherche d'informations, est-il vraiment nécessaire de définir une ontologie ?

Cette interrogation résulte d'un constat, celui de la masse énorme de travail de modélisation nécessaire à la définition d'une ontologie sur la base de la *ressource ontologique*. A ce constat se greffe une autre interrogation corrollaire : la présentation de liens spécifiquement ontologiques à l'utilisateur est-elle réellement utile dans une tâche d'aide à la reformulation de requêtes ?

Des éléments de réponse à cette question ne peuvent être apportés que par des travaux approfondis relevant de l'ergonomie et de l'étude des interfaces homme-machine. Nous sommes simplement en mesure de fournir à cette question un point de vue empirique. Il semble que l'utilisateur n'ait pas besoin de se voir signifier des relations proprement ontologiques entre les termes qui lui sont suggérés et ceux de sa requête dans le cadre d'un système d'aide à la reformulation de requête. Ce point de vue se fonde sur l'idée que l'utilisateur est le plus à même de déterminer les termes qu'il estime utiles à une reformulation de sa requête, qu'il est en conséquence à même d'établir pourquoi ces termes lui sont utiles, ce qui le conduit à définir lui-même des liens ontologiques entre ces termes et sa requête.

Nous pouvons alors avancer que le travail de modélisation permettant d'obtenir une ontologie sur la base de la ressource ontologique n'est pas forcément nécessaire dans un contexte de recherche d'informations. Dès lors, une ressource ontologique rassemblant les termes du domaine et des liens non qualifiés sémantiquement entre ces termes peut suffire pour un système d'aide à la reformulation de requête. Nous pouvons cependant noter que c'est une des limites de notre démarche de construction d'ontologie que de prévoir un travail de modélisation peut-être trop important.

Quoi qu'il en soit, l'ontologie du droit français codifié que nous obtenons est intégrée dans un système d'aide à la reformulation de requête. Nous avons également aménagé la navigation dans l'ontologie. Le tout est accessible gratuitement en ligne à l'adresse suivante : http ://ontologie.w3sites.net. Deux utilisations de notre ontologie sont ainsi effectives : la navigation dans le modèle du domaine qu'elle constitue et son intégration à un système d'aide à la reformulation de requête.

Un autre usage de l'ontologie est envisageable, celui du classement automatique, ou *catégorisation*, de documents. Il est en effet tout à fait possible de se baser sur le réseau de termes que constitue l'ontologie, chaque terme étant associé à des informations relatives aux thématiques juridiques le concernant, pour procéder à des catégorisations automatiques de documents. Le procédé consiste alors à utiliser les mêmes outils de traitement automatique de la langue sur des documents juridiques quelconques, pourvu qu'ils soient un tant soit peu relatifs au droit français. Les termes alors identifiés ainsi que leurs proportions peuvent être comparés aux éléments de l'ontologie. En découle l'attribution de thématiques juridiques aux documents en question. Il peut ainsi tout à fait être envisagé de prendre des documents de jurisprudence et de leur faire correspondre des thématiques juridiques propres au codes. L'utilisation de notre ontologie du droit français dans un procédé de catégorisation de documents est ainsi une des perspectives de nos travaux. Nous en avons ébauché les principes généraux, auxquels s'ajoutent quelques expérimentations, dans deux articles : *Classement automatique de documents et analyse terminologique de corpus* ([Lam01b]) et *A Categorization Method for French Legal Documents on the Web* ([Lam01a]).

De ces perspectives sur le classement automatique de documents découle une autre continuation possible de nos travaux qui reflète une de leurs limites : celle de la réactualisation de l'ontologie. Cette perspective est également envisagée dans les deux articles que nous venons de citer. Le droit est en effet une discipline qui évolue, de nouvelles notions apparaissent tandis que d'autres tombent en désuétude, comme autant de témoignages de l'ancrage du droit dans l'évolution des sociétés. Ainsi, la notion d'*autorité parentale* s'est substituée à celle de *puissance paternelle*. De nouvelles notions sont créées comme *pacte civil de solidarité*. Le modèle du domaine que constitue l'ontologie, et ce, d'autant plus que l'usage auquel celle-ci est dédiée est la recherche d'information, doit pouvoir prendre en compte ces évolutions. L'ontologie que nous avons élaborée l'a été sur la base des codes du droit français dont nous disposions en avril 2001. L'ontologie reflète alors l'état du droit codifié à ce moment. Depuis lors, de nouvelles dispositions ont été intégrées dans les codes. Du fait des nombreuses étapes qu'elle comprend et de la lourdeur du processus, il ne peut être envisagé de procéder à une nouvelle inférence d'ontologie sur la base des codes modifiés. La réactualisation doit pouvoir se faire de façon assez facile et rapide, à intervalles réguliers sur la base des nouvelles dispositions de codes. Une façon d'intégrer les nouveaux concepts éventuellement créés par ces nouvelles dispositions peut alors résider dans le classement des nouveaux documents sur la base de notre ontologie. De par les thématiques juridiques qui lui seront attribuées, un document porteur d'une nouvelle disposition pourra se trouver rapproché d'un ensemble de termes éléments de l'ontologie. Ces rapprochements seront alors comme autant de pistes pour situer un nouveau terme non encore élément de l'ontologie.

Un tel processus de réactualisation peut être également envisagé pour élargir les thématiques envisagées par l'ontologie. Notre ontologie ne couvre en effet que le droit français codifié. Bien que la partie du droit qui ait fait l'objet de codification représente un *noyau* thématique, de nombreuses autres sources peuvent être considérées pour procéder à l'élargissement thématique de notre ontologie. Ainsi, les lois non codifiées, des décisions de jurisprudence, des notes de doctrines peuvent être prises en compte pour un élargissement par un processus itératif sur le fondement du *noyau* thématique que constitue notre ontologie. La prise en compte de telles sources du droit permettrait d'appréhender d'autres langages du droit, comme le langage jurisprudentiel, différent du langage législatif ainsi que le note [Cor01]. Une telle démarche permettrait de tendre un peu plus encore vers le rapprochement recherché entre le langage utilisé par les utilisateurs dans leurs requêtes et le langage des documents sur lesquels opèrent les moteurs.

Les travaux que nous avons menés conduisent donc à la définition d'une ontologie du droit français codifié dédiée à la recherche d'informations sur le Web. La démarche par laquelle nous avons procédé est décrite dans cette thèse. Cette démarche apporte une nouvelle dimension aux recherches relatives à l'ingénierie des connaissances et à l'élaboration d'ontologies, tout en comportant intrinsèquement des limites.

L'ontologie telle que nous l'avons définie est intégrée dans un système d'aide à la reformulation de requête. Ce système est librement accessible en ligne ; il constitue notre modeste contribution à la vaste problématique de l'accès de tous au droit sur Internet, illustrant ce que la recherche est en mesure d'apporter à ces questions.

Futures recherches

Le travail qui vient de vous être présenté ouvre différentes voies, comme autant de questions non résolues par nous et autant de futures recherches à mener. Les principales sont les suivantes :

- poursuivre la modélisation de la ressource ontologique ou trouver une solution alternative pour la détection des significations des liens établis en termes ;
- évaluer et déterminer l'utilité de spécifier aux utilisateurs la signification des liens établis entre les termes ;
- élargir l'ontologie sur la base d'autres sources du droit. Les mêmes principes que ceux ayant conduit la construction de notre ontologie peuvent être appliqués, reste à résoudre l'intégration de nouvelles données au modèle existant ;
- évaluer l'ergonomie des interfaces de navigation et d'utilisation de l'ontologie.

Bibliographie

[AB96] H. Assadi and D. Bourigault. Acquisition et modélisation de connaissances à partir de textes : outils informatiques et éléments méthodologiques. In *Actes de la conférence RFIA'96 Reconnaissance des Formes et Intelligence Artificielle*, pages 505–514, Rennes, France, 1996.

[ABD+81] A. Andreeswky, J.-P. Binquet, F. Debili, C. Fluhr, and B. Pouderoux. Linguistic and statistical processing of texts and its application in the field of legal documentation. In Council of Europe, editor, *Proceedings of teh 6th Symposium on Legal Datda Processing in Europe*, Thessaloniki, Greece, 1981.

[AC00] N. Aussenac and A. Condamines. Entre textes et ontologies formelles : les bases de connaissances terminologiques. In Hermès, editor, *Systèmes d'information pour la gestion des connaissances : recherches en cours et approches industrielles*. Zacklad, M. and Grundstein, M., Paris, 2000.

[AGBS00a] N. Aussenac-Gilles, B. Biébow, and S. Szulman. Modélisation du domaine par une méthode fondée sur l'analyse de corpus. In *Actes de la conférence IC'2000. Journées Francophones d'Ingéniérie des connaissances*, pages 93–103, Toulouse, mai 2000. AFIA, ERSS, IRIT, GRACQ.

[AGBS00b] N. Aussenac-Gilles, B. Biébow, and S. Szulman. Revisiting ontology design : A method based on corpus analysis. In R Dieng and O. Corby, editors, *Proceedings of EKAW'2000, Knowledge Engineering and Knowledge Management. Methods, Models and Tools*, pages 172–188, Juan-les-Pins, France, october 2000. Springer.

[All97] R. Allier. *Nouveau Larousse Illustré. Dictionnaire universel encyclopédique*, chapter Ontologie. Larousse, 1897.

[Ass98] H. Assadi. *Construction d'ontologies à partir de textes techniques*. PhD thesis, Université Paris 6, 1998.

[AVGPLTP98] J. Arpierz Vega, A. Gomez-Perez, A. Lozano Tello, and H. Pinto. (onto)2agent : An ontology-based www broker to select ontologies. In *Proceedings of ECAI'1998, Workshop on Applications of Ontologies and Problem solving Methods*, Brighton, England, august 1998.

[Bac00] B. Bachimont. Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingéniérie des connaissances. In J. Charlet, M. Zacklad, G. Kassel, and D. Bourigault, editors, *Ingéniérie des connaissances. Evolutions récentes et nouveaux défis*, pages 305–323. Eyrolles, Paris, 2000.

[BB92] B. Barakat-Barbieri. *Vers une construction automatique de graphes de concepts*. PhD thesis, Ecole centrale, Paris, 1992.

[BCKN01] C. Barry, C. Cormier, G. Kassel, and J. Nobécourt. Evaluation de langages opérationnels de représentation d'ontologies. In *Actes de la conférence IC'2001, Ingéniérie des connaissances*, pages 309–327, Grenoble, France, juin 2001.

[BHR00] D. Bourcier, P. Hassett, and C. Roquilly. Droit et intelligence artificielle : une révolution de la connaissance juridique. In D. Bourcier, P. Hassett, and C. Roquilly, editors, *Droit et Intelligence artificielle*, pages 9–16. Romillat, Paris, France, 2000.

[BKL+02] A. Boukhors, A. Kaszycki, J. Laplce, S. Munerot, and L. Poublan. *XML, la synthèse. Intégrez XML dans vos architectures*. Dunod, Paris, 2002.

[BL02] D. Bourigault and G. Lame. Analyse distributionnelle et structuration de terminologie, application à la construction d'une ontologie documentaire du droit. *Traitement Automatique des Langues*, 2002.

[BM99] J. Breuker and A. Muntjewerff. Ontological modelling for designing educational systems. In *Proceedings of AIED'99, Workshop on Ontologies for Intelligent Educational Systems*, Le Mans, France, 1999.

[Bou93] D. Bourigault. Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *TAL, Traitement Automatique des Langues*, 34(2), 1993.

[Bou99] D. Bourcier. Représenter quelques règles de droit dans une machine. au-delà de faust et de frankenstein. In *Savoir innover en droit. Concepts, outils, systèmes*, pages 119–131. La documentation française, 1999.

[Bou02] D. Bourigault. Analyse distributionnelle étendue. In *Actes de TAL'2002, Traitement Automatique des Langues*, Nancy, France, juin 2002.

[Bra99] G. Braibant. La problématique de la codification. In *Savoir innover en droit. Concepts, outils, systèmes*, pages 55–65. La documentation française, 1999.

[Bru89] M.-F. Bruandet. Outline of a knowledge base model for an intelligent information retrieval system. *Information Processing and Management*, 25 :85–115, 1989.

[BS99] D. Bourigault and M. Slodzian. Pour une terminologie textuelle. *Terminologies Nouvelles*, (19) :29–32, juin 1999.

[CAG00] A. Condamines and N. Aussenac-Gilles. Entre textes et ontologies formelles, les bases de connaissances terminologiques. In *Capitalisation des connaissances.* Hermès, Paris, 2000.

[Cas02] F. Castel. Ontological computing. *Communications of the ACM*, 45(2) :61–65, february 2002.

[CDN88] J. Chin, V. Diehl, and K. Norman. Development of a tool measuring user satisfaction of the human-computer interface. In *Proceedings of SigChi'88 Human Factors in Computer Systems*, pages 213–218, 1988.

[CGP00] O. Corcho and A. Gómez-Pérez. Evaluating knowledge representation and reasoning capabilities of ontology specification languages. In *proceedings of ECAI'2000, Workshop on Applications of Ontologies and Problem-Solving Methods*, pages 31–39, Berlin, Allemagne, 2000.

[Che02] JP Chevallet. Building thesaurus from manual sources and automatic scanned texts. In *Proceedings of The 2nd International Conference Adaptive Hypermedia and Adaptive Web Based Systems*, pages 95–104, Malaga, Spain, may 2002.

[CHSS98] H. Chen, A. Houston, R. Sewell, and B. Schatz. Internet browsing and searching : User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 49(7) :582–603, 1998.

[CKZB00] J. Charlet, G. Kassel, M. Zacklad, and D. Bourigault. Ingéniérie des connaissances, recherches et perspectives. In J. Charlet and al., editors, *Ingéniérie des Connaissances ; Evolutions récentes et nouveaux défis*, pages 2–22. Eyrolles, 2000.

[Con95] P. Constant. L'analyseur linguistique sylex. In *Actes de la 5° Ecole d'été du CNET, Centre National D'Etude des Télécommunications*, 1995.

[Cor00] G. Cornu. *Linguistique juridique*. Dunod, Paris, 2000.

[Cor01] G. Cornu. *Vocabulaire juridique*. Quadrige, Presses Universitaires de France, Paris, 2001.

[Dai94] B. Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. PhD thesis, Université Paris 7, Paris, France, 1994.

[DFR89] F. Debili, C. Fluhr, and P. Radasoa. About reformulation in full-text irs. *Information Processing and Management*, 25(6) :647–657, 1989.

[DMVH+00] S. Decker, S. Melnik, F. Van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks. The semantic web : The roles of xml and rdf. *IEEE Internet Computing*, pages 63–73, september/october 2000.

[Dom98] J. Domingue. tadzebao and webonto : Discussing, browsing, and editing ontologies on the web. In *Proccedings of the Workshop on Knowledge Acquisition, Modeling and Management*, Alberta Canada, april 1998.

[DP90] S. David and P. Plante. De la nécessité d'une approche morpho-syntaxique dans l'analyse des textes. *Intelligence Artificielle et Sciences Cognitives au Québec*, 3(3) :140–154, 1990.

[Eft96] E. Efthimiadis. Query expansion. In R. Williams and E. Martha, editors, *Annual Review of Information Systems and Technology*, pages 121–187. 1996.

[Eft00] E. Efthimiadis. Interactive query expansion : a user-based evaluation in a relevance feedback environment. *Journal of* *the American Society of Information Science*, 51(11) :989–1003, 2000.

[Eng92] C. Enguehard. *Acquisition naturelle automatique d'un réseau sémantique*. PhD thesis, Université de Technologie de Compiègne, Compiègne, France, 1992.

[Fen00a] D. Fensel. Oil in a nutshell. In R. Dieng and O. Corby, editors, *Proceedings of EKAW 2000, Knowledge Engineering and Knowledge Management*, pages 1–16, Juan-les-Pins, France, october 2000. Springer-Verlag.

[Fen00b] D. Fensel. *Ontologies : A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer, 2000.

[FFR97] A. Farquhar, R. Fikes, and J. Rice. The ontolongua server : a tool for colaborative ontology construction. *International Journal Human-Computer Studies*, 47 :707–727, 1997.

[FGPJ97] M. Fernandez, A. Gomez-Perez, and N. Juristo. Methontology : From ontological art towards ontological engineering. In *Proceedings of the AAAI'97 Spring Symposium Series on Ontological Engineering, pp. 33-40, Stanford, USA*, pages 33–40, Stanford, USA, 1997.

[FHRT01] F. Féru, E. Hugues, A. Rivière, and S. Tafaro. Des interfaces pour améliorer les performances d'exploitation d'un système d'information complexe. In *Actes de NîmesTIC'2001, La relation Homme-Système : Complexe ?*, pages 115–126, Nîmes, France, décembre 2001. Ecole des mines d'Alès, Centre de Recherche LGI2P, Nîmes.

[FHvH+01] D. Fensel, I. Horrocks, F. van Harmelen, D. McGuiness, and P. Patel-Schneider. Oil : An ontology infrastructure for the semantic web. *IEEE Intelligent Systems*, 16(2) :38–45, march/april 2001.

[FL99] M. Fernandez-Lopez. Overview of methodologies for building ontologies. In V.R. Benjamins, B. Chandrasekan, A. Gomez-Perez, N. Guarino, , and M. Uschold, editors, *Proceedings of the IJCAI-99 Workshop on Ontologies and Problem Solving Methods*, august 1999.

[FMJ01] G. Falquet and C.-L. Mottaz Jiang. navigation hypertexte dans une ontologie multi-points de vue. In *Congrès scientifique NîmesTIC'2001. La relation Homme-Système : Complexe ?*, pages 77–84, Nîmes, décembre 2001. Ecole des mines d'Alès.

[GB01] N. Grabar and S. Berland. Construire un corpus web pour l'acquisition terminologique. In Unité de recherche et innovation INIST-CNRS, editor, *Actes de la conférence TIA'2001, Terminologie et intelligence artificielle*, pages 44–54, Nancy, France, mai 2001.

[GDK01] F. Gandon and R. Dieng-Kuntz. Ontologie pour un système multi-agent dédié à une mémoire d'entreprise. In *Actes de la conférence IC'2001, Ingéniérie des connaissances*, Grenoble, France, mai 2001.

[GF95] M. Gruninger and M. Fox. Methodology for the design and evaluation of ontologies. In *Proceedings of the IJCAI'95 Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal, Canada, 1995.

[GG95] N. Guarino and P. Giaretta. Ontologies and knowledge bases : Towards a terminological clarification. In N. Mars, editor, *Towards Very Large Knowledge Bases : Knowledge Building and Knowledge Sharing*, pages 25–32. IOS Press, 1995.

[gra00] *What TOVE did for DAPHNE - Using an Ontology Construction Methodology for Combining Different Sources of Knowledge*, Berlin, Germany, august 2000.

[Gre94] G. Grefenstette. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, Boston, USA, 1994.

[Gru93] T. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition Journal*, pages 199–220, 1993.

[Gua98] N. Guarino. Formal ontologies and information systems. In N. Guarino, editor, *Proceedings of FOIS'98*, pages 3–15, Trento, Italy, june 1998. IOS Press Amsterdam.

[Gua99] N. Guarino. Ontoseek : Content-based access to the web. *Intelligent Systems IEEE*, pages 70–80, may/june 1999.

[GWR99] S. Gauch, J. Wang, and S. Rachakonda. A corpus analysis approach for automatic query expansion and its extension to multiple databases. *ACM Transactions on Information Systems*, 17(3) :250–269, 1999.

[Har61] H. Hart. *The Concept of Law.* Clarendon Press, Oxford, 1961.

[Har71] Z. C. Harris. *Structures mathématiques du langage.* Dunod, 1971. Traduction C. Fuchs.

[Hea78] H.-S. Heaps. *Information retrieval, computational and theoretical aspects.* Academic Press, London, UK, 1978.

[Hea92] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING'1992, International Conference on Computational Linguistics*, pages 539–545, Nantes, France, 1992.

[Hen01] J. Hendler. Agents and the semantic web. *Intelligent Systems IEEE*, 16(2) :30–36, 2001.

[HJ02] C. Holsapple and K. Joshi. A collaborative approach to ontology design. *Communications of the ACM*, 45(2) :42–47, february 2002.

[HPD00] W. Hersh, S. Price, and L. Donohoe. Assessing thesaurus-based query expansion using the umls metathesaurus. In *Proceedings of the AMIA 2000 Annual Symposium, American Medical Informatics Association*, Los Angeles, USA, 2000.

[HV99] J. Hage and B. Verheij. The law as a dynamic interconnected system of states of affairs : a legal top ontology. *International Journal Human-Computer Studies*, 51 :1043–1077, 1999.

[JBCV98] D. Jones, T. Bench-Capon, and P. Visser. Methodologies for ontology development. In *Proceedings of IT-KNOWS Conference, XV IFIP World Computer Congress*, Budapest, Hungaria, august 1998.

[JBD+97] C. Jouis, I. Biskri, J.-P. Descles, F. Le Priol, J.-G. Meunier, W. Mustafa, and G. Nault. Vers l'intégration d'une approche sémantique linguistique et d'une approche numérique pour un outil d'aide à la construction de bases terminologiques. In *Actes JST FRANCIL 1997, Journées Scientifiques et Techniques du réseau Francophone de l'Ingérniérie de la Langue*, pages 427–432, Avignon, France, avril 1997.

[Jon95] D. Jones. Interactive thesaurus navigation : intelligence rules ok ? *Journal of the American Society for Information Science*, 46 :52–59, 1995.

[Jou93] C. Jouis. *Contributions à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes.* PhD thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris, France, 1993.

[Jou95] C. Jouis. Seek, un logiciel d'acquisition de connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe. In AFIA INRIA, editor, *Actes de JAVA'1995, Journées Acquisition, Validation*, pages 159–172, Grenoble, France, avril 1995.

[JS00] M.-P. Jacques and A.-M. Soubeille. Partage des termes, partage des connaissances ? construire une modélisation unique de plusieurs corpus. In *Actes de IC'2000, Journées Francophones d'Ingéniérie des Connaissances*, pages 313–324, Toulouse, France, mai 2000.

[Kas99] G. Kassel. Co-operative ontology construction needs to carefully articulate terms, notions and objects. In *Proceedings of the International Workshop on Ontological Engineering on the Global Information Infrastructure*, pages 57–70, Dagstuhl Castle, Germany, 1999.

[Kas02a] G. Kassel. Ontospec : une méthode de spécification semi-informelle d'ontologies. In *Actes de IC'2002, Journées Francophones d'Ingéniérie des Connaissances*, 2002.

[Kas02b] G. Kassel. Veille technologique, ingéniérie ontologique, concepts, méthodes et outils. document powerpoint présentation du domaine à des industriels, 2002.

[Kay97] D. Kayser. *La représentation des connaissances*. Hermès, 1997.

[Kim02] H. Kim. Predicting how ontologies for the semantic web will evolve. *Communications of the ACM*, 45(2) :48–54, february 2002.

[KL89] M. Kifer and G. Lausen. Flogic : A higher order language for reasoning about objects, inheritance, and schema. In *Proceedings of SIGMOD'89, International Conference on Management of Data*, pages 134–146, Portland, USA, june 1989. ACM.

[Lam00] G. Lame. Acquisition de connaissances à partir de textes, vers l'élaboration d'une ontologie du droit. In M. Ayel and J.-M. Fouet, editors, *Actes des RJCIA'2000. Cinquièmes rencontres nationales des jeunes chercheurs en intelligence artificielle*, pages 211–221, Lyon, septembre 2000. AFIA.

[Lam01a] G. Lame. A categorization method for french legal documents on the web. In H. Prakken, editor, *Proceedings of* *the 8th International Conference on Artificial Intelligence and Law*, pages 219–220, Saint-Louis MO USA, mai 2001. ACM, ACM Press.

[Lam01b] G. Lame. Classement automatique de documents et analyse terminologique de corpus. In Unité de recherche et innovation INIST CNRS, editor, *Actes de la conférence TIA'2001, quatrièmes rencontres Terminologie et Intelligence Artificielle*, pages 149–158, Nancy, France, mai 2001.

[Lam01c] G. Lame. Constructing an ir-oriented legal ontology. Second International Workshop on Legal Ontologies, JURIX'2001, Amsterdam, december 2001.

[Lef00] P. Lefevre. *La recherche d'information, du texte intégral au thesaurus*. Hermes science, Paris, 2000.

[LSR96] S. Luke, L. Spector, and D. Rager. Ontology-based knowledge discovery on the world-wide-web. In *Proceedings of AAAI'96 Workshop on Internet-based Information Systems*, 1996.

[MBC+00] L. Médini, P. Bigeon, J.-M. Charlot, M. Chaillot, O. Ferret, D. Malavieille, D. Penel, D. Scapin, and J.-M. Van craeynest. Intégration de l'indexation conceptuelle dans l'expression du besoin d'information. *Document numérique*, 4(1-2) :85–108, 2000.

[McC89] L. McCarty. A language for legal discourse, basic features. In *Proceedings of ICAIL'89, Second International Conference on Artificial Intelligence and Law*, pages 180–189, Vancouver, Canada, 1989. ACM, IAAIL.

[Mil93] G. Miller. Wordnet : a lexical database. *Communications of the ACM*, 38(11) :39–41, 1993.

[MM00] D. Moldovan and R. Mihalcea. Using wordnet and lexical operators to improve internet searches. *IEEE Internet Computing*, pages 34–43, 2000.

[MN96] J.-G. Meunier and M. Nyongwa. Extraction dynamique et connexionniste de connaissances terminologiques. In *Actes du Colloque Franco-québéquois LANCI, Extraction des connaissances et technologies de l'information*, Université du Québec à Montréal, Canada, 1996.

[Moe00] M-F Moens. *Automatic indexing and abstracting of document texts*. Kluwer, Boston, Dordrecht, London, 2000.

[Mom98] L. Mommers. On the ontological status and representation of legal concepts. In *Proceedings of Jurix'98, Conference on artificial intelligence and law*, 1998.

[Mor99] E. Morin. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques.* PhD thesis, Université de Nantes, Faculté des Sciences et Techniques, Laboratoire IRIN, 2 rue de la Houssinière, BP 92208, Nantes Cedex 3, 1999.

[MP01a] A. Moreno and C. Pérez. From text to ontology : Extraction and representation of conceptual information. In CNRS Unité de Recherche et Innovation, INIST, editor, *Actes de la conférence TIA'2001, Terminologie et Intelligence Artificielle*, pages 233–242, Nancy, France, mai 2001.

[MP01b] J. Muramatsu and W. Pratt. Transparent queries : Investigating users' mental models of search engines. In B. Croft, D. Harper, D. Kraft, and J. Zobel, editors, *Proceedings of SIGIR'01, International ACM Conference on Research and Development in Information Retrieval*, pages 217–224, New Orleans, Louisiana, USA, september 2001.

[MS00] A. Maedche and S. Staab. Mining ontologies from text. In R. Dieng and O. Corby, editors, *Proceedings of EKAW'2000, Knowledge Engineering and Knowledge Management*, pages 189–202, Juan-les-Pins, France, october 2000. Springer.

[MSO97] L. Mommers, A. Schmidt, and E. Oskamp. Controversories in the ontology and law debate. In P. Visser and R. Winkels, editors, *Proceedings of LEGONT'97, First International Workshop on Legal Ontologies*, pages 1–5, Melbourne, Victoria, Australia, july 1997.

[NHR02] C. Neuwirth and S. Harkness Regli. Usability and the web. *IEEE Internet Computing*, 6(2) :44–79, march-april 2002.

[PCCC01] J. Prager, J. Chu-Carrol, and K. Czuba. Use of wordnet hypernyms for answering what-is questions. In *Proceedings of TREC10, Text Retrieval Conference*, Gaithersberg, MD, USA, 2001.

[Pea98] J. Pearson. Terms in context. *Studies in Corpus Linguistics*, 1, 1998.

[Pra01] H. Prakken, editor. *Proceedings of the 8th International Conference on Artificial Intelligence and Law*, Saint-Louis, MI, USA, 2001. ACM, ACM Press.

[Pri00] Y. Prié. Sur la piste de l'indexation conceptuelle de documents. *Document numérique*, 4(1-2) :11–35, 2000.

[RC01] M. Ribière and P. Charlton. Ontology overview. Technical report, Motorola Labs, Paris, Commune de Saint-Aubin, 2001.

[Reb00] J. Rebeyrolle. Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. In *Actes de la conférence IC'2000. Journées Francophones d'Ingénierie des connaissances*, pages 105–114, Toulouse, mai 2000. AFIA, ERSS, IRIT, GRACQ.

[RLK00] D. Roger, J. Lavandier, and E. Kolmayer. Navigation et interfaces : cartes conceptuelles et autres outils. http ://www.enssib.fr/bibliotheque/documents/travaux/kolmayer/kolm-navig0.html, 2000.

[SBAG01] S. Szulman, B. Biebow, and N. Aussenac-Gilles. Vers un environnement intégré pour la structuration de terminologies : Terminae. In Unité de recherche et innovation INIST-CNRS, editor, *Actes de la conférence TIA'2001, Terminologie et Intelligence Artificielle*, pages 98–108, Nancy, France, mai 2001.

[SEMD00] S. Staab, M. Erdmann, A. Maedche, and S. Decker. An extensible approach for modelling ontologies in rdf(s). In *Proceedings of ECD, European Conference on Research and Advanced Technologies for Digital Libraries, Workshop on Semantic Web*, Lisbon, Portugal, septembre 2000.

[SJ71] K. Spark Jones. *Automatic keyword classification for information retrieval.* Butterworth, London, UK, 1971.

[SJ73] K. Spark Jones. Index term weighting. *Information Storage and Retrieval*, 9 :619–633, 1973.

[SL97] E. Schweighofer and F. Lachmayer. Ideas, visualisations and ontologies. In P. Visser and R. Winkels, editors, *Proceedings of LEGONT'97, Workshop on Legal Ontologies*, pages 7–13, Melbourne, Victoria, Australia, july 1997.

[SM83] G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* McGraw Hill, New-York, USA, 1983.

[Sta80] R. Stamper. Legol : Modelling legal rules by computer. In B. Niblett, editor, *Computer Science and Law*. Cmabridge University Press, Cambridge, United Kingdom, 1980.

[Sta91] R. Stamper. The role of semantics in legal expert systems and legal reasoning. *Ratio Juris*, 4(2) :219–244, 1991.

[SWJ95] G. Schreiber, B. Wielinga, and W. Jansweijer. The kactus view of the 'o' word. In *Proceedings of IJCAI'1995, Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal, Canada, august 1995.

[TND+98] Y. Toussaint, F. Namer, B. Daille, J. Royauté, and N. Hathout. Une approche linguistique et statistique pour l'analyse de l'information en corpus. In *Actes de TALN'1998, Conférence sur le Traitement Automatique des Langues Naturelles*, pages 182–191, Paris, 1998.

[Usc96] M. Uschold. Building ontologies : Towards a unified methodology. In *Proceedings of Expert System'1996, conference of the British Computer Society Specialist Group on Expert Systems*, Cambridge, England, december 1996.

[Val94] A. Valente. Making ends meet : Conceptual models and ontologies in legal problem solving. In *Proceedings of SBIA'94, Brazilian AI Symposium*, pages 1–15, Fortaleza, Brazil, 1994.

[VB94] A. Valente and J. Breuker. A functional ontology of law. In G. Bargellini and S. Binazzi, editors, *Towards a global expert system in law*. CEDAM, Padua, Italy, 1994.

[VBC98] P.R.S. Visser and T. Bench-Capon. A comparison of four ontologies for the design of legal knowledge systems. *Artificial Intelligence and Law*, 6(1) :27–57, march 1998.

[VdVSM94] P. Van de Vet, P. Speel, and N. Mars. The plinius ontology of ceramic materials. In N. Mars, editor, *Working papers of ECAI'94, Conference on Artificial Intelligence, Workshop on Implemented Ontologies*, pages 187–206, Amsterdam, The Nederlands, 1994.

[VKVBCVDH99] R. Van Kralingen, P. Visser, T. Bench-Capon, and J. Van Den Herik. A principled approach to developping legal knowledge systems. *International Journal Human-Computer Studies*, 51 :1127–1154, 1999.

[VLLM01] B. Verheij, A. Lodder, R. Loui, and A. Muntjewerff, editors. *Proceedings of JURIX'2001, International Conference on Legal Knowldge and Information Systems*, Amsterdam, the Nederlands, 2001. IOS Press.

[Win00] R. Winkels. Clime : Un projet de développement de serveurs juridiques intelligents. In D. Bourcier, P. Hassett, and C. Roquilly, editors, *Droit et Intelligence artificielle*, pages 59–72. Romillat, Paris, France, 2000.

[ZBB+96] P. Zweigenbaum, B. Bachimont, J. Bouaud, J. Charlet, and J. Boisvieux. Le rôle du lexique sémantique et de l'ontologie dans le traitement automatique de la langue médicale. In P. Le Beux and A. Burgun, editors, *Actes de CRISTAL'S 96, Colloque Contribuer à la Recherche et à l'Innovation, au Service du Traitement et de l'Analyse des Langages utilisés dans les Systèmes de Soins*, Saint-Malo, France, 1996.

[ZBHN97] P. Zweigenbaum, J. Bouaud, B. Habert, and A. Nazarenko. Coopération apprentissage en corpus et connaissances du domaine pour la construction d'ontologies. In *Actes FRANCIL'1997. Journées Scientifiques et Techniques*, pages 501–508, Avignon France, 1997. AUPELF UREF.

Table des matières

Remerciements **5**

Introduction **9**

Partie préliminaire : positionnements **15**

0.1 Contexte 15
0.2 Usage 16
0.3 Ontologie dans l'usage 17
0.4 Elaboration d'une ontologie dédiée 18
0.4.1 Choix du *corpus de référence* 19
0.4.2 Outils et techniques d'acquisition de connaissances à partir de texte 20
0.4.3 Choix de la non validation 22
0.4.4 Evaluation par l'usage de l'ontologie 23
0.4.5 Modélisation de la ressource ontologique 24
0.5 L'ingénieur de la connaissance 25

I Une ontologie du droit français dédiée à la recherche d'information **27**

1 Définition des ontologies **31**

1.1 Les ontologies en informatique, définition par l'usage 32
1.1.1 Ontologies et traitement du langage naturel 33
1.1.2 Ontologies et systèmes à base de connaissances 34
1.1.3 Ontologies et recherche d'informations 38
1.1.4 Ontologies et *Web semantique* 40
1.2 Elaboration d'ontologies 41
1.2.1 Différentes approches 41
1.2.2 Des méthodes d'élaboration d'ontologies 45

1.2.3 Des outils dédiés aux ontologies 58
1.3 Ontologies et analyse de texte 60
1.3.1 Les thesaurus, terminologies et réseaux sémantiques . . 61
1.3.2 Méthodes d'élaboration (semi)automatiques 63

2 Formalismes pour les ontologies 67
2.1 XML 68
2.2 RDF 70
2.3 OIL 71
2.4 OntoSpec 73
2.5 Autres langages 74
2.6 Conclusion sur les langages 76

3 Ontologie et droit 79
3.1 Particularités juridiques 80
3.1.1 Caractéristiques théoriques de la discipline juridique . . 80
3.1.2 Caractéristiques pratiques de la discipline juridique . . 84
3.2 Les ontologies juridiques existantes 88
3.2.1 L'ontologie fonctionnelle du droit de Valente 88
3.2.2 L'ontologie de Van Kralingen et Visser 92
3.3 Linguistique juridique 93
3.3.1 Le vocabulaire juridique 93
3.3.2 Les discours du droit 97

II Elaboration d'une ontologie à partir de textes 101

4 Un corpus de référence 105
4.1 Définition du corpus de référence 105
4.2 Choix du corpus de référence 108
4.2.1 Le Journal Officiel de la République française 108
4.2.2 Les codes du droit français 109

5 Extraction des termes 113
5.1 Outils d'extraction de termes 113
5.1.1 Méthodes d'extraction de termes 114
5.1.2 Les outils Sylex et Syntex 115
5.2 Les termes du domaine 116
5.3 Filtrage des candidats termes 117
5.3.1 Pondération des candidats termes 119
5.3.2 La liste de termes juridiques 125

5.4 Choix de la non validation 133
5.4.1 Les *termes fondamentaux du domaine* 138
5.4.2 Les termes vides 139
5.4.3 Les formes plurielles 140

6 Identification de liens entre termes 143
6.1 Analyse des relations de coordinations 144
6.1.1 La méthode 144
6.1.2 La pratique, résultats 144
6.2 Analyse distributionnelle 146
6.2.1 la méthode 146
6.2.2 En pratique, l'établissement de liens 148
6.2.3 Des seuils empiriques 151
6.2.4 La sélection du nombre de relations 153
6.3 Analyse statistique 156
6.3.1 La méthode 156
6.3.2 La pratique, résultats 158
6.4 Analyse des formes 159
6.4.1 La méthode 159
6.4.2 La pratique, résultats 160
6.5 Ensemble des relations entre termes 162

7 De la ressource ontologique à l'ontologie 165
7.1 La ressource ontologique 166
7.1.1 Définition 166
7.1.2 La ressource ontologique et les BCT 169
7.1.3 Le choix de la non validation 170
7.2 Evaluation de la ressource ontologique 173
7.2.1 Mise en place du système d'aide à la reformulation de requête 174
7.2.2 Résultats de l'évaluation 179
7.3 Phase de modélisation des connaissances 185
7.3.1 L'ensemble des qualifications possibles 185
7.3.2 La modélisation 193

III Usages et évaluations d'une ontologie dédiée à la recherche d'informations sur le Web 201

8 Navigation et visualisation 205
8.1 Principe 205

8.1.1 Le réseau de termes 205
8.1.2 Les thématiques 209
8.2 Evaluation 213
8.2.1 Principes d'évaluation 213
8.2.2 Futurs travaux 214

9 Aide à la reformulation de requêtes **217**
9.1 Principe 217
9.1.1 La reformulation de requête 217
9.1.2 Système d'aide à la reformulation de requête basé sur notre ontologie 220
9.2 Evaluation 221

Conclusion **229**

Bibliographie **229**

A Les codes et leurs abbréviations **249**

B Liste des termes du droit ... **253**

C Codes ayant servi pour l'expérimentation .. **263**

D Termes associés par l'analyse ... **265**

E Thèmes juridiques et codes s'y rapportant **269**

Annexe A

Les codes et leurs abbréviations

CACTSOC Code de l'action sociale et des familles ;
CFAMILL Code des familles ;
CARTISA Code de l'artisanat ;
CASSURA Code des assurances ;
CAVIACI Code de l'aviation civile ;
CCAISSE Code des caisses d'épargne ;
CCIVILL Code civil ;
CCOMMER Code de commerce ;
CCOMMUN Code des communes ;
CCONSOM Code de la consommation ;
CCONSTR Code de la construction et de l'habitation ;
CDABBOI Code des débits de boissons ;
CDARCHI Code de déontologie des architectes ;
CDCHIRD Code de déontologie des chirurgiens-dentistes ;
CDMEDIC Code de déontologie médicale ;
CDPOLIC Code de déontologie de la police nationale ;
CDSAGES Code de déontologie des sages-femmes ;
CDVETER Code de déontologie vétérinaire ;
CDVIMAR Code disciplinaire et pénal de la marine marchande ;
CDWOMET Code du domaine de l'Etat ;
CDXFLUV Code du domaine public fluvial et de la navigation intérieure ;
CDYANES Code des douanes ;

CEDUCAT Code de l'éducation ;

CELECTO Code électoral ;

CENVIRO Code de l'environnement ;

CEXPROP Code de l'expropriation pour cause d'utilité publique ;

CFOREST Code forestier ;

CGCTERR Code général des collectivités territoriales ;

CGIMPOT Code général des impôts ;

CGIMPX Code général des impôts, Annexes ;

CGLIVPF Livre des procédures fiscales ;

CINDCIN Code de l'industrie cinématographique ;

CJURFIN Code des juridictions financières ;

CJUSADM Code de justice administrative ;

CJUSMIL Code de justice militaire ;

CLEGHON Code de la légion d'honneur et de la médaille militaire ;

CMARPUN Code des marchés publics ;

CMINIER Code minier ;

CMONFIN Code monétaire et financier ;

CMUTUAL Code de la mutualité ;

CORGJUD Code de l'organisation judiciaire ;

CPENALL Code pénal ;

CPENSIC Code des pensions civiles et militaires de retraite ;

CPENSIM Code des pensions militaires d'invalidité et des victimes de la guerre ;

CPENSIR Code des pensions de retraite des marins français du commerce, de la pêche ou de plaisance ;

CPORMAR Code des ports maritimes ;

CPOSTES Code des postes et télécommunications ;

CPROCIV Code de procédure civile ;

CPROCPE Code de procédure pénale ;

CPROINT Code de la propriété intellectuelle ;

CROUTE Code de la route ;

CRURAL Code rural ;

CSANPUB Code de la santé publique ;

CSECSOC Code de la sécurité sociale ;
CSERVNA Code du Service national ;
CTRAVAI Code du travail ;
CTRAVMA Code du travail maritime ;
CURBANI Code de l'urbanisme ;
CVOIRIE Code de la voirie routière.

Annexe B

Liste des termes du droit extraits des dictionnaires juridiques en ligne

abandon, abandon de famille, abrogation, abroger, absence, abstension, abstention constructive, abstention positive, abus, acceptation, accession, accessoire, accord amiable, accord de schengen, accord européen, accord international bilatéral, accord international multilatéral, accord social, accroissement, accusé, accusatoire, acompte, acquéreur, acquêt, acquiescement, acquis communautaire, acquittement, acquitter, acte, acte administratif, acte authentique, acte conservatoire, acte d'état civil, acte d'administration, acte de disposition, acte de notoriété, acte de procédure, acte juridictionnel, acte notarié, acte sous seing privé, actif, action, action civile, action directe, action en justice, action estimatoire, action oblique, action paulienne, action personnelle, action publique, action réelle, ad hoc, ad litem, additionnelle, adhésion, adjudicataire, adjudication, administrateur, administrateur de biens , administration, administration légale, adoptabilité de l'enfant étranger, adoption, adoption internationale, adoption plénière, adoption simple, affacturage, affaire pendante, affiliation, affrètement, agence immobilière, agrégation, agrément, ags, aide à l'accès au droit, aide de préadhésion, aide juridictionnelle, aide juridique, aide sociale à l'enfance, ajournement, aléatoire, aliénation, alignement, aliments, allégation, allocation, aménagement, amende, amende civile, amiable, amiable compositeur, amicus curiae, amnistie, anatocisme, annulation, antériorité, antenne de justice, antichrèse, apatride, apostille, apparence, apparentement, appel, appel en garantie, appel nullité, appel-nullité , approfondissement, arbitrabilité, arbitrage, arbitrage international, arbitrage multipartite, arbitre, architecture européenne, argument, arrérage, arrêté, arrêté de cessibilité, arrêt de mise en accusation, arrhes, ascendant, asile, assemblée, assemblée générale de copropriété, assesseur, assignation, assises, assistance éducative, association, associations de parents adoptifs, assurance, assurance construction , assurance professionnelle , assurance rcp, assurrance trc - tous risques chantiers, astreinte, audience, audience de départage, audience foraine, audience solennelle, audiovisuel, auditeur de justice, audition, auteur, authentique, autorisation, autorité centrale, autorité de la chose jugée, autorité parentale, auxiliaire de justice, aval, avance, avancement d'hoirie, avant-dire droit, avantage acquis, avantage en nature, avenant, aveu, avocat, avocat commis d'office, avocat général, avoué, ayant cause, ayant droit, bâtonnier, bail, bail à ferme, bail commercial, bail d'habitation , bail professionnel , bail rural, bailleur, banque centrale européenne, barreau, baux, bce, bien, bien commun réservé, bien immobilier, bien propre, biens communs, biens corporels, biens immobiliers, biens incorporels, biens indivis, biens propres, bilatéral, billet à ordre, bonne foi, bornage, brevet, budget, caducité, cantonnement, capable, capacité, capacité adoptive, capacité juridique, carence, carte professionnelle, casier judiciaire, cassation, caution, cautionnement, cecos, cedh, centre d'étude et de conservation des oeufs et du sperme humains, centre d'arbitrage, centre de détention, certificat de coutume, certification, cessation des paiements, cession, cessionnaire, chambre, chambre d'accusation, chambre du conseil, chancelier, chancellerie, charge, charge de la preuve, charges récupérables, charte des droits fondamentaux, charte des services publics, charte sociale, chirographaire, chose, circulation des mineurs étrangers, circulation des personnes, citation, citation directe, citoyen, citoyenne, citoyenneté, citoyenneté de l'union, civil, classification des dépenses, clause, clause compromissoire, clause d'exemption, clause de suspension, clause pénale, clause résolutoire, clerc, cmu, cotraitant, codébiteur, code, code civil, code de commerce, code de procédure civile, code de procédure pénale, code du travail, code pénal, codicille, codification des textes législatifs, coefficient d'occupation des sols, cohérie, cohéritier, cohésion économique et sociale, coindivisaire, colégataire, collégialité, collatéral, collectivité, collectivité territoriale, collectivités territoriales, collocation, comité économique et social, comité de conciliation, comité de probation, comité des régions, comité et groupes de travail, comité politique (copo), comitologie, commandement, commandite, commerçant, commerce d'enfants, comminatoire (astreinte), commis, commis d'office, commissaire, commissaire-priseur, commission, commission de recours amiable, commission européenne, commission rogatoire, commission rogatoire internationale, commissions parlementaires, commissoire, commodat, communal, communale, communauté conjugale, communautarisation, commune, communicable, communication des causes, comourants, compétence, compétence communautaire, compétence des arbitres, compétence externe de la communauté européenne, compétence subsidiaire, compétitivité, comparution, comparution immédiate, comparution personnelle, compensation, complainte, composition de la commission, composition du parlement européen, compromis, compromis de luxembourg, compromis de vente, compte-courant, compulsoire, concentration, conception-réalisation, conciliateur, conciliation, conclusions, concordat, concubinage, concurrence, condamné, condamnation, condamnation avec sursis, condamnation définitive, condamnation par défaut, conditions suspensives, conférence, conférence européenne, conférence intergouvernementale (cig), conférence intergouvernementale bilatérale, confirmation, conflit d'intérêts, confrontation, confusion, confusion des peines, congé, congrès, conjoint, connaissement, connexité, conseil d'état, conseil de famille, conseil de famille des pupilles de l'état, conseil de l'union, conseil de prud'hommes, conseil des prud'hommes, conseil européen, conseil général, conseil supérieur de l'adoption, conseil syndical, conseiller, consentement à l'adoption, conservatoire, consignation, consolidation, consolidation des textes législatifs, consort, constat, constitution, constitutionnel, constitutionnelle, constructeur, constructibilité limitée, consultation, contentieux, contestation de paternité légitime, contrôle de l'application du droit communautaire, contrôle judiciaire, contrôle sanitaire des enfants étrangers, contrôle technique, contractant, contractante, contradiction, contradictoire, contrainte, contrainte par corps, contrat, contrat d'arbitrage, contrat d'architecte, contrat d'entreprise, contrat de location, contrat de maîtrise d'oeuvre, contrat de mariage, contravention, contre-lettre, contredit, convention, convention de croupier, convention de schengen, convention européenne des droits de l'homme (cedh), coobligé, coopération policière et judiciaire en matière pénale, coopération politique européenne (cpe), coopération renforcée, copropriété, copropriétaire, coreper, coreu (pesc), cos, coupable, cour, cour d'appel, cour d'assises, cour de cassation, cour de justice, cour des comptes, courtier, couverture maladie universelle, covendeur, créance, créancier, création d'un lien de filiation entre l'enfant et sa famille adoptive, crédit bail, crédit documentaire, crédit-bail, crédit-bailleur, crime, critère d'adhésion, critère de convergence, critère de copenhague, croupier, culpabilité, culture, curatelle, curateur, débats, débauchage, débiteur, débitrice, débours, débouter, débrayage, décès, déchéance, décision administrative, décision-cadre, décisoire, déclaration d'utilité publique, déclaration de petersberg, déclaration préalable, déclinatoire, défaut, défendeur, défense, défense collective, défenseur, déficit démocratique, délégation, délégation d'autorité parentale, délai congé, délai de carence, délai de grâce, délai de viduité, délai-congé, délais de procédure, délibéré, délinquant, délit, démembrement du droit de propriété, déni de justice, dénonciation de nouvel oeuvre, dépôt de garantie, départage, département, départemental, départementale, dépens, dépositaire, déposition, désaisissement, désaveu de paternité, déshérence, désistement, détachement, détention, détention provisoire, détenu, développement durable, développement rural, dévolution, dation en paiement, degré de juridiction, demande reconventionnelle, demandeur, denier, descendant, descente sur les lieux, destination, destination du père de famille, dette, devis, dialogue social, dilatoire, diligence, diligenter, discrétionnaire, discuter, disjonction, dispense de peine, dispositif, dispositif du jugement, distances à respecter pour les plantations, divorce, doctrine, document d'urbanisme, dol, domicile, dommage, dommage-intérêt, dommage-ouvrage, dommages aux existants, dommages de travaux publics, dommages et intérêts, dommages-intérêts, don, donataire, donateur, donation, donner acte, dotal, dotaux, droit commun, droit communautaire, droit d'alerte , droit d'initiative, droit d'usage et d'habitation, droit de délaissement, droit de pétition, droit de préemption urbain, droit de préférence, droit de propriété, droit de rétention, droit de rétrocession, droit de repentir, droit de se clore, droit de suite, droit international privé, droit privé, droit public, droits civils, droits dérivés, droits de l'homme, ducroire, dup, empêchement, emphytéose, empiétement, emplacements reservés, employeur, empreintes génétiques, enchère, endossement, enlèvements d'enfants, enquête, enquête judiciaire, enquête parcellaire, enquête préalable sur l'utilité publique, enquête préliminaire, enquête publique, enquête sociale, enrôler, enregistrement, enrichissement sans cause, ententes, entrepreneur, entreprise, environnement, envoi en possession, espace de liberté de sécurité et de justice, espaces boisés, ester en justice, estimatoire, eurocorps, eurofor, euromarfor, europe , europol, exécuteur, exécution, exécution provisoire, exception, exciper, exclusion, exequatur, exigibilité, exonération, exorbitant, expédient, expédition, expert immobilier, expert judiciaire, expertise, expertise sanguine, exploit, exposé des motifs, expropriation, expulsion, extra petita, extradition, extrajudiciaire, faculté, faire droit, fausses déclarations de naissance, faute , faute inexcusable, faux, fermage, feuille d'audience, filiale, filiation, fin de non recevoir, fisc, fiscal, fiscalité, fiscalité et adoption, flagrant délit, foncière, foncier, fonds de commerce, fonds de garantie, fonds structurels et fonds de cohésion, fongible, force exécutoire, force majeure, force publique, forclusion, forfait, formation professionnelle, former un pourvoi, fortuit, fortune de mer, foyer d'action éducative, frais d'acquisition, frais de justice, frais de notaire, frais et dépens, franchise, fraude, fruits, frustratoires, fusion, gérance, gage, garantie, garantie de parfait achèvement, garantie financière, garanties biennale et décennale, garanties dûes par les constructeurs, garde, garde à vue, garde des sceaux, gestion d'affaires, gie, globalisation de l'économie, gouvernement, grâce, gracieuse, gratification, greffe, greffier, greffier en chef, grosse, groupement d'entreprises, groupement d'intérêt économique, héritage, héritier, habilitation, haut-commissaire, hiérarchie des actes communautaires, hiérarchie des normes, hoir, hoirie, homicide, homicide involontaire, homicide volontaire, homologuer, hors de cause, huis clos, huis-clos, huissier, huissier de justice, hypothèque, hypothécaire, identité européenne de sécurité et de défense, illégal, illicite, immeuble, immigration, immobilier, immunité, impartir, impenses, implication, imputation, in solidum, inaliénabilité, inaliénable, incapable, incapacité, incarcération, incident, incompétence, incorporels, indemnité, indemnité d'immobilisation, indexation, indignité successorale, indivis, indivisaire, indivisibilité, indivisible, indivision, indu, inexécution, inexcusable, infirmer, infra petita, infraction, injonction, injonction de payer, innocence, inscription de faux, insolvabilité, insolvable, inspection, installation, instance, instance arbitrale, institution, instruction, instrumentaire, instruments juridiques communautaires, intégration de l'accord social, intéressement, intention, interdiction, interjeter, interlocutoire, intermédiaire, interprétation, intestat, intimé, intuitu personae, inventaire, investiture de la commission, irrévocabilité, irrévocabilité de l'adoption plénière, irrecevabilité, itératif, jetons de présence, jeunesse, jonction, jouissance légale, jours et vues sur les fonds voisins, juge, juge aux affaires familiales, juge départiteur, juge d'appui, juge d'instruction, juge de l'application des peines, juge de l'exécution, juge de l'expropriation, juge de la mise en état, juge des enfant, juge des enfants, juge des référés, juge des tutelles, juge du siège, jugement, jugement avant dire droit, jugement contradictoire, jugement d'adoption, jugement d'expédient, jugement par défaut, jugement sur le fond, juges non professionnels, juré, juridiction, juridiction administrative, juridiction civile, juridiction de droit commun, juridiction pénale, juridiction spécialisée, juridictionnel, juridictionnelle, juridictions, jurisprudence, jury, justice, justice et affaires intérieures, légal, légale, légalisation, légalisation des dossiers d'adoption, légataire, légataire universel, légitimation, léonin, lésion, leasing, leg, lettre, lettre de change, lettre de mission, libéralité, libération conditionnelle, libératoire, liberté surveillée, libre circulation des personnes, licenciement, licitation, lien de filiation, ligne, ligne de succession, liquidation, lisibilité des traités, litige, litisconsorts, litispendance, livre blanc, livre vert, livres verts, locataire, location, location meublée, location-accession, location-gérance, locations saisonnières, lock-out, lot, lotissements, louage, lutte contre la criminalité internationale organisée et le blanchiment de l'argent, lutte contre la drogue, lutte contre la fraude, lutte contre le racisme et la xénophobie, lutte contre le terrorisme, médiateur, médiateur européen, médiation, médiation pénale, mémoire, métayage, méthodes communautaire et intergouvernementale, maîtrise d'ouvrage, magasins généraux, magistrat, magistrat du siège, magistrats du ministère public, magistrats du parquet, magistrats du siège, mainlevée, maison centrale, maison d'arrêt, maison de justice, majeur, majeur protégé, majorité, majorité qualifiée, majorité qualifiée renforcée, malfaçon, mandant, mandat, mandat d'amener, mandat d'arrêt, mandat de comparution, mandat de dépôt, mandataire, manquement, marc le franc, marché a forfait, marchés de travaux, marchand de biens, marchand de listes, mariage, marque de fabrique, maternité de substitution, matrimonial, membres associés de l'ueo, mercuriale, mesure conservatoire, mesure d'administration, mesures conservatoires, mesures provisoires, mettre à néant, meuble, meubles corporels, meubles incorporels, milieu ouvert, mineur, mineurs demandeurs d'asile, ministère public, ministre, minorité, minute, mise à pied, mise en état, mise en accusation, mise en cause, mise en demeure, mise en examen, mission de base batiment, missions de petersberg, mitoyenneté, mobilier, mondialisation, monopole, montage de promotion immobilière, moratoire, motif surabondant, moyens et motifs, multilatéral, multipartite, mutation, mutuelle, négociations d'adhésion, nantissement, national, nationale, nationalité, naturalisé, naturalisée, naturalisation, nolissement, nomenclature, non avenu, non-lieu, norme, notaire, notification, notification d'une décision, notoire, notoriété, novation, noyau dur, nu-propriétaire, nue-propriété, nullité, nus-propriétaire, nus-propriétaires, obligation, obligation alimentaire, obligation d'information, obligation de divisibilité, obligation de réserve, obligation de reserve, observateur auprès de l'ueo, oeuvre d'adoption, office des migrations internationales, office européen de lutte antifraude, office européen de police, office public d'aménagement et de construction, officier de police judiciaire, officier ministériel, officier ministériel ou officier public, offres réelles, olaf, olographe, omission de statuer, onéreux, opac, opc - ordonnancement, pilotage, coordination, opposabilité, opposition, opposition à tiers détenteur, opting out, ordonnance, ordonnance d'expropriation, ordonnance de taxe, ordonnance pénale, ordonnance sur requête, ordre, ordre administratif, ordre judiciaire, ordre public, organisation du traité de l'atlantique nord, organisation judiciaire, organisme, organisme agréé, organisme autorisé, original, otan, otan 'rénovée', péremption, pétitoire, pac, pacs, pacte, pacte commissoire, pacte de préadhésion sur la criminalité organisée, pacte de stabilité et de croissance, pae, paiement, paraphe, parlement, parlement européen, parlements nationaux, parquet, parquet général, parrainage, partage, partage d'ascendants, partenaires associés de l'ueo, partenaires sociaux, partenariat pour l'adhésion, partie civile, passerelle communautaire, paternité, patrimoine, patrimonial, paulienne, payeur, pendante, pension, pension alimentaire, permis de construire, personnalité juridique de l'union, personne morale, perte d'une chance, pesc, pièce du dossier, pig, pignoratif, piliers de l'union européenne, placement en vue de l'adoption plénière, placement familial, plafond légal de densité, plainte, plan d'occupation des sols, pld, plumitif, police de l'audience, police judiciaire, police unique de chantier, politique économique, politique étrangère et de sécurité commune, politique agricole commune, politique commerciale commune, politique commune des transports, politique de défense commune, politique monétaire, politique sociale, pollicitation, pondération des voix au conseil, portable, pos, position commune (pesc), position dominante, possession d'état, possessoire, postulation, potestatif, potestative, pourvoi, pourvoi en cassation, pouvoir discrétionnaire, pouvoir souverain, pouvoirs publics, préavis, précarité, préciput, préciputaire, précompte, préemption, préférence, préfix, préjudice, préjudice (ou dommage), préjudice corporel, préjudice d'agrément, préjudice matériel, préjudice moral, préjudiciel, présidence de l'union, président, président de la commission européenne, présomption, présomption d'innocence, présomption de paternité, présomptions, prétention, prévenu, prêt, preneur, prescription, prestation, prestation compensatoire, prestations familiales et adoption, pretium doloris, preuve, preuve par écrit, principe d'immunité, principe de la non-discrimination, principe de subsidiarité, prise à partie, privilège, privilège de juridiction, prix forfaitaire, prix unitaire, probation, procès, procès-verbal, procès-verbaux, procédure, procédure électorale uniforme, procédure abusive, procédure civile, procédure d'ordre, procédure de codécision, procédure de coopération, procédure de divorce, procédure de l'avis conforme, procédure de l'avis simple, procédure pénale, procréation médicalement assistée avec tiers donneur, procuration, procureur, procureur de la république, procureur général, prodédure, programme d'aménagement d'ensemble, programme de l'union européenne, projet d'intérêt général, promesse de vente, promoteur, propriété, protection civile, protection des consommateurs, protocole social, provision, prud'hommes, publicité, publique, puc, pupille, pupille de l'état, purger, putatif, quérable, qualification, quasi-contrats, quasi-délit, quittance, quote-part, quotes-part, quotes-parts, quotité, règlement de copropriété, règlement national d'urbanisme, récidive, récognitif, récolement, récompense, récursoire, récusation, rédhibitoire, référé, référendaire, régimes matrimoniaux, région, régional, régionale, régions ultrapériphériques, réglement, réhabilitation, réintégrande, réintégration, réméré, répétition de l'indu, réputé contradictoire, réquisition d'emprise totale, réquisitions, réquisitoire, réserve héréditaire, résiliation, résolution, résolutoire, rétention, rétracter, rétroactif, réversion, révision, révision des traités, révocation, révocation de l'adoption simple, rôle, radiation, rapport, rapport de suivi, ratification, rcd, recel, recevabilité, recherche de paternité naturelle, recherche des origines, recherche et développement, reconduction, reconnaissance, reconnaissance de plein droit des décisions étrangères, recours, recours amiable, recours en révision, recouvrement, rectification, rectification judiciaire d'un acte d'état civil, recusation, redressement, redressement judiciaire, refonte des textes législatifs, registre du commerce, regroupement familial, relèvement, relaxe, reliquat, reliquataire, renonciation, rente, renvoi, repentir, représentation, reprise de l'acquis communautaire, requérant, requérante, requête, rescision, responsabilité civile, responsabilité civile décennale, responsabilité civile du fait des travaux, responsabilité civile professionnelle, responsabilité contractuelle, responsive, ressort, ressources propres, retenue de garantie, retrait, retranchement, revendication, revenu minimum d'insertion, risque, ristourne, rnu, rogatoire, rotation de la présidence, rupture du lien de filiation préexistant, sécurité sociale, sénat, séparation de biens, séparation de corps, séquestre, sûreté, saisie, saisie des rémunérations, saisie-attribution, saisie-vente, saisine, salaire, salarié, santé publique, satisfactoire, sauvegarde de justice, scec, scellés, schémas directeurs (sdau), schengen, sci, scpi, screening, seing privé, sem, semi-liberté, sentence, serment, serment décisoire, serment supplétoire, service central d'état civil, service public, service universel, services d'intérêt économique général, services d'intérêt général, servitude, servitude de marchepied, servitude de passage en cas d'enclave, servitude de passage le long du littoral, servitude de visibilité, servitude de vue, servitudes, servitudes d'écoulement des eaux, servitudes d'utilité publique, shob, shon, siège de l'arbitrage, sieur, signature, signification, simplification des traités, simplification législative, simulation, smic, société, société civile d'attribution, société civile de construction-vente, société civile de location ou de gestion, société civile immobilière, société d'économie mixte, solidarité, solvabilité, solvable, sommation, sommation de payer, soulte, sous-acquéreur, sous-traitance, sous-traitants, soutenir, souverain, staries, statuer, statut, stellionat, stipuler, stratégie commune (pesc), stratégie coordonnée pour l'emploi, stratégie de préadhésion, subrogé-tuteur, subrogation, subsides, subsidiaire, subsidiarité, substitut, substitut du procureur, substitut général, substitution, succession, succession et adoption, superprivilège, superstaries, supplétif, supplétoire, suppression du protocole social, surabondant, surenchère, surendettement, surseoir, sursis, sursis avec mise à l'épreuve, sursis simple, suspensif, suspicion légitime, synallagmatique, syndic, syndic de copropriété, syndicat, syndicat des copropriétaires, télécommunications, témoignages, témoignages ou présomptions, témoin, tacite, tarif de responsabilité, tarifs de chancellerie, taux du ressort, taxe, territorialité, testament, testament authentique, testament olographe, testament-partage, testamentaire, testateur, testimonial, testimoniale, ticket modérateur, tierce opposition, tierce-opposition, tiers, tiers détenteur, tontine, torts, tréfond, trésor, traduction des documents constitutifs du dossier d'adoption, trafic d'enfants et adoption internationale, traité d'amsterdam, traite, transaction, transcription du jugement d'adoption, travail d'intérêt général, travaux dispensés, travaux soumis, tribunal, tribunal administratif, tribunal correctionnel, tribunal d'instance, tribunal de commerce, tribunal de grande instance, tribunal de police, tribunal des affaires de sécurité sociale, tribunal des conflits, tribunal maritime commercial, tribunal paritaire des baux ruraux, tribunal pour enfants, troïka, trouble, troubles de voisinage, tutelle, tutelle aux prestations sociales, tutelle de l'enfant mineur, tuteur, uem, ueo, ultra petita, unanimité, unilatéral, union économique et monétaire, union de l'europe occidentale, union libre, unité de planification et d'alerte rapide, usucapion, usufruit, usufruitière, usufruitier, usure, vérification, vérité, vacant, valeur vénale, valeurs mobilières, vente, vente immobilière, verdict, viagère, viager, vice du consentement, vice-président, vices cachés, vices du consentement, victime, vider un délibéré, vie privée, violence, visa, visa d'entrée, voie d'exécution, voie de fait, voie de recours, voies d'exécution, voies de recours, voies de recours dites extraordinaires, voies de recours ordinaires, vrp, vues, warrant, zac, zad, zone d'aménagement concerté, zone d'aménagement différé, zone de conflit, ès-qualités, échéance, échange, échec de l'adoption, écriture, éducateur, éducateur de la protection judiciaire de la jeunesse, éducation, éfficacité internationale des décisions étrangères, égalité de traitement entre les hommes et les femmes, égalité des chances, élargissement, élire, émancipation, émender, émolument, équilibre institutionnel et légitimité démocratique, établissement public d'amenagement urbain, état, état civil, état civil de l'enfant adopté, état d'accueil, état d'origine, études d'esquisse, évaluation des biens, éviction, évocation.

Annexe C

Codes ayant servi pour l'expérimentation de l'analyse distributionnelle

Code de l'action sociale et des familles
Code civil
Code de commerce
Code de l'environnement
Code de la justice administrative
Code de l'organisation judiciaire
Code pénal
Nouveau code de procédure civile
Livre des procédures fiscales
Code de procédure pénale
Code de la propriété intellectuelle
Code du travail

Annexe D

Termes associés par l'analyse distributionnelle au terme *création*

acquisition	CGIMPOX
assemblée générale	CORGJUD
autorisation	CSANPUB, CURBANI
classement	CGCTERR
conditions	CGCTERR
construction	CURBANI
décès	CGIMPOX
développement	CGCTERR
ensemble	CURBANI
extension	CGCTERR, CSANPUB, CURBANI
extension d'une telle structure	CSANPUB
fonctionnement	CORGJUD, CGCTERR
gestion	CCOMMER
inscription	CSANPU
institution	CURBANI
membres	CORGJUD, CGCTERR, CGIMPOX
modification	CGCTERR, CSANPUB
médecin	CSANPUB
organe	CGCTERR
population	CGCTERR
produits	CSANPUB
président	CORGJUD, CGCTERR
périmètre	CGCTERR
remboursement	CGIMPOX
représentant	CGCTERR
ressort	CORGJUD
sages-femmes	CSANPUB
section	CORGJUD
services	CGCTERR
siège	CORGJUD
souscription	CGIMPOX
suspension	CGCTERR
transfert	CGCTERR
transformation	CGCTERR
vice-président	CORGJUD
établissement	CSANPUB, CSANPUB
établissements publics de santé	CSANPUB

TAB. D.1 – Termes associés et codes de provenance au terme *création* avant sélection

administration	CCOMMER
autorisation	CURBANI
conseil	CGIMPOX
extension	CURBANI
fonctionnement	CGCTERR
membres	CGCTERR
modification	CGCTERR
président	CGCTERR

TAB. D.2 – Termes associés et codes de provenance au terme *création* après sélection

Annexe E

Thèmes juridiques et codes s'y rapportant

Civil
- Code civil
- Code de procédure civile

Pénal
- Code pénal
- Code de procédure pénale
- Code disciplinaire et pénal de la marine marchande

Personne Citoyen Travail Famille Santé Retraite Déontologie
- Code de l'action sociale et des familles, code de la famille et de l'aide sociale
- Code de déontologie des architectes
- Code de déontologie des chirurgiens-dentistes
- Code de déontologie médicale
- Code de déontologie de la police nationale
- Code de déontologie des sages-femmes
- Code de déontologie vétérinaire
- Code de l'éducation
- Code électoral
- Code de la famille et de l'aide sociale
- Code de la légion d'honneur et de la médaille militaire
- Code des pensions civiles et militaires de retraite
- Code des pensions militaires d'invalidité et des victimes de la guerre
- Code des pensions de retraite des marins français du commerce, de pêche ou de plaisance

- Code de la santé publique
- Code de la sécurité sociale
- Code du service national
- Code du travail
- Code du travail maritime

Commerce Economie Marchés Assurances Consommation
- Code de l'artisanat
- Code des assurances
- Code des caisses d'épargne
- Code de commerce
- Code des débits de boissons et des mesures contre l'alcoolisme
- Code de la consommation
- Code des douanes
- Code de l'industrie cinematographique
- Code des marchés publics
- Code monétaire et financier
- Code de la mutualité
- Code des postes et télécommunications
- Code de la propriété intellectuelle

Impôts
- Code général des impôts
- Code général des impôts, annexes
- Livre des procedures fiscales

Aménagement Construction Collectivités Rural Environnement
- Code de la construction et de l'habitation
- Code de l'environnement
- Code de l'expropriation pour cause d'utilité publique
- Code forestier
- Code minier
- Code général des collectivités territoriales
- Code rural
- Code de l'urbanisme
- Code de la voirie routière
- Code des communes

Domaine et prérogatives de l'Etat
- Code de l'aviation civile
- Code du domaine de l'état

- Code du domaine public fluvial et de la navigation intérieure
- Code de la route
- Code des ports maritimes

Justice Juridictions
- Code des juridictions financieres
- Code de justice administrative
- Code de justice militaire
- Code de l'organisation judiciaire

Construction d'ontologie à partir de textes. Une ontologie du droit dédiée à la recherche d'informations sur le Web.

Résumé

L'objectif de ces travaux est la construction d'une ontologie dédiée à la recherche d'information sur le Web. La méthode que nous dégageons vise à appliquer des outils et techniques d'ingéniérie linguistique sur des textes préalablement choisis. Les termes du domaine et les relations lexicales les unissant sont alors dégagés, constituant ce que nous appellons la *ressource ontologique*. Une étape de modélisation des éléments constitutifs de cette ressource permet d'obtenir l'ontologie en tant que telle. Cette ontologie est utilisée et évaluée dans un contexte de recherche d'information sur le Web et plus spéciquement est intégrée dans un système d'aide à la reformulation de requêtes. Les principaux apports de nos travaux sont (1), par la dimension même de l'ontologie (100 000 éléments) et le large domaine qu'elle couvre (le droit français codifié), une nouvelle dimension apportée à la construction d'ontologies, (2) un renouvellement des méthodes de construction d'ontologies par la maximisation du recours à des outils et techniques automatiques se doublant d'une réflexion sur la place des tâches manuelles dans les processus d'élaboration d'ontologies, et enfin (3) une réponse concrète à la problématique juridique de l'accès au droit.

Mots clés : ontologie, ingéniérie des connaissances, ingéniérie linguistique, recherche d'informations.

Building ontology from texts. A legal ontology dedicated to information retrieval on the Web.

Abstract

The goal of this work is to build an ontology dedicated to information retrieval on the Web. To reach that goal, we have defined a method which mainly consists on performing linguistic tools and techniques on previously selected texts. Domain terms and lexical relations among them are obtained from these techniques. These elements are gathered in what we call an *ontological ressource*. Next step to obtain an ontology is to perform a conceptualization of these elements. The ontology is used and evaluated in an information retrieval context, being included in a interactive request expansion system. The main contributions of our work are (1) a new perspective given to ontology enginering with our large (100 000 elements) and wide spread ontology(covering the codified french law) , (2) a renewal of ontology enginering itself leaded by the main use of automatic tools and techniques which is doubled by a discussion on the role of manual steps in ontology enginering, and finaly, a practical answer to the accessibility of law.

Keywords : ontology, knowledge enginering, information retrieval.