Une ontologie documentaire pour la recherche d'information relationnelle

Nada Mimouni, Adeline Nazarenko et Sylvie Salotti

LIPN, CNRS (UMR 7030), Université Paris Nord
Sorbonne Paris Cité, F-93430 Villetaneuse
Nada.Mimouni, Adeline.Nazarenko,
Sylvie.Salotti@lipn.univ-paris13.fr

Résumé : Cet article présente un modèle documentaire qui prend en compte non seulement les annotations sémantiques portées par les documents, leurs structures logiques et leurs différentes versions mais aussi la structure d'une collection documentaire composée de différents types de documents reliés entre eux par des types variés de relations.

Le développement de la recherche d'information sémantique dans ses usages professionnels suppose d'exploiter tout cet ensemble de propriétés documentaires. Dans le domaine juridique, notamment, il faut pouvoir retrouver les documents d'un type particulier (par ex. des décrets émis par telle juridiction) qui portent sur une notion spécifique juridique (ex. contrat) et qui précisent un texte de loi donné. Il faut aussi pouvoir retrouver l'ensemble des textes portant sur un sujet donné (ex. le bruit) en vigueur à une date précise et la manière dont ils ont été appliqués, c'est-à-dire la jurisprudence relative à ces textes.

Au moment où les efforts de standardisation et d'ouverture donnent accès aux données publiques, il est essentiel de penser la modélisation des collections juridiques pour offrir des fonctionnalités d'interrogation avancées. L'approche proposée repose sur les standards du web sémantique. Elle a l'originalité d'intégrer les différentes propriétés documentaires dans un modèle unique qui permet de croiser les critères sémantiques, temporels et relationnels dans la recherche d'information.

Mots-clés : Collection documentaire, Modèle ontologique, RI sémantique, Intertextualité, Requêtes relationnelles.

---

0. Ce travail a été partiellement financé par le projet LEGILOCAL (FUI 2010-2013). Nous remercions nos partenaires, notamment Meritxell Fernández-Barrera, Eve Paul et Danièle Bourcier pour l'aide qu'elles nous ont apportée dans la compréhension des enjeux de la recherche d'information juridique.

1 Introduction

Avec l'émergence du web sémantique et le mouvement d'ouverture de données touchant à de plus en plus de domaines, des efforts sont faits pour rendre ces données compatibles avec les standards et normes définies dans le web sémantique (XML, RDF, SPARQL) et définir des modèles sémantiques (ontologies) pour différents domaines. Ces efforts ont pour but d'assurer l'interopérabilité des données et de faciliter leur accès et leur gestion par les utilisateurs.

Par exemple dans le domaine juridique, plusieurs standards XML juridiques ont été définis pour normaliser la structure des textes de loi, assister la production de ces textes et améliorer leur interopérabilité. En parallèle, des initiatives d'ouverture de données gouvernementales se multiplient (ex. UK Government Linked Data). Pourtant, les données mises à disposition restent souvent sous-exploitées.

Peu d'approches en effet ont été définies pour permettre la recherche d'information dans des textes juridiques normalisés et publiés. La multiplicité des sources juridiques, leur technicité, leur fort degré de structuration et d'intertextualité, les enjeux de sécurité juridique imposent des contraintes très particulières en termes de recherche d'information dans ce domaine. Il faut pouvoir retrouver tous les textes en vigueur sur un sujet donné, mais également consolider un texte de loi à une date donné en prenant en compte toutes les modifications apportées à ses différents articles, ou connaître la jurisprudence relative à tel texte juridique. Il faut pouvoir savoir quelles juridictions ont tendance à modifier ou abroger tels types de textes.

Nous défendons l'idée que le développement de fonctionnalités avancées de recherche d'information dans ce domaine repose sur l'intégration de l'ensemble des propriétés documentaires dans un modèle unique. Nous proposons une ontologie documentaire (OWL) qui permet de représenter le contenu sémantique du document (ce dont parle le document), sa structure logique, ses différentes versions et son cycle de vie, ainsi que la structure de la collection documentaire qui organise différents types de documents dans un vaste réseau de liens intertextuels de documents. Il s'agit de donner accès à la complexité des sources juridiques (Bourcier, 2011). Le but est en effet de pouvoir à terme, modéliser l'ensemble d'une collection documentaire sous la forme d'un graphe RDF puis l'interroger de manière sémantique, structurelle, temporelle et/ou relationnelle à l'aide de requêtes SPARQL.

La section 2 présente les approches et les techniques proposées pour

modéliser des documents dans le web sémantique, notamment dans le domaine juridique. La section 3 décrit les besoins auxquels la recherche d'information juridique se trouve confrontée. La section 4 présente l'ontologie documentaire que nous proposons avec les différents modules la composant et leurs dépendances. La section 5 présente des exemples d'utilisation de cette ontologie pour répondre à des requêtes relationnelles.

2 Modélisation des documents dans le web sémantique

L'essor du web sémantique et du web de données repose sur l'évolution des technologies sémantiques qui assurent l'interopérabilité des données (section 2.1) mais aussi sur le développement des ressources pour l'annotation sémantique des documents (section 2.2). Dans ce contexte, un effort est fait pour développer des ontologies documentaires mais nous montrons que les modèles existants sous-estiment la dimension intertextuelle et ne permettent pas de modéliser l'ensemble des propriétés documentaires de manière homogène (section 2.3), ce qui constitue un frein à l'essor des méthodes de recherche d'information sémantique.

2.1 Langages et standards du web sémantique

Pour améliorer l'interopérabilité des données, le langage de balisage XML est utilisé pour représenter la structure des documents et de multiples modèles de documents ont été définis pour modéliser différents types de documents. Dans le domaine juridique, ces standards XML sont des éléments clés de la standardisation des contenus des sources de loi. Différents standards ont été développés par différents états européens, qui peuvent s'articuler avec le standard européen CEN-Metalex [1]. Au niveau international, un ensemble de DTD a été développé par la Chambre des représentants des Etats-Unis et le projet AKOMANTOSO [2] a produit des DTD pour les documents parlementaires, législatifs et judiciaires de plusieurs pays africains. Les détails de ces standards XML ont été décrits dans (Sartor, 2011).

Les données dans le web sémantique sont stockées sous forme de graphes de données, *c.a.d* sous la forme de triplets RDF (*RDF triple store*). Le format RDF (*Resource Description Framework* [3]) définit des déclarations

1. http ://www.metalex.eu/
2. http ://www.akomantoso.org/
3. http ://www.w3.org/RDF/

comprenant un sujet, un prédicat (*property*) et un objet, c'est-à-dire un triplet (sujet,prédicat,objet). Des URIs (*Uniform Resource Identifiers*) sont utilisés pour donner un identifiant unique au sujet, au prédicat et à l'objet.

Pour associer une sémantique aux modèles de données RDF, on peut définir les URIs dans des schémas (RDFS [4]) ou des ontologies (OWL [5]). RDFS et OWL sont des spécifications W3C. En OWL, on définit un concept comme une classe d'individus partageant les mêmes caractéristiques. Les individus sont reliés par des rôles ou des relations (*Object properties*) et ils peuvent comporter des attributs valués (*Datatype properties*). SKOS [6] (Simple Knowledge Organization System) est une famille de langages formels permettant aussi une représentation standard des thésaurus, classifications ou tout autre type de vocabulaire contrôlé et structuré.

Les triples stores RDF sont interrogés à l'aide du langage (protocole) SPARQL [7], de même que les bases de données relationnelles peuvent être interrogées à l'aide du langage SQL. SPARQL est un standard W3C et il est actuellement à sa version 1.1.

2.2 Vocabulaires conceptuels et annotation sémantique

L'approche classique de recherche d'information sémantique (comme par exemple dans AquaLog (Lopez *et al.*, 2007), KnOWLer (Ciorascu *et al.*, 2003) ou MELISA (Abasolo & Gomez, 2000)) dépasse les méthodes à base de mots clés en exploitant les annotations sémantiques qui sont apposées sur les documents pour en modéliser le contenu.

Les termes utilisés comme annotations sont définis dans des vocabulaires ou des ontologies qui sont eux-mêmes définis en SKOS ou OWL. Les ontologies de domaine permettent d'associer aux contenus des documents une description sémantique à la fois explicite et formelle, ce qui facilite l'exploitation sémantique des contenus par des outils automatiques et améliore l'interopérabilité des sources. Dans le domaine juridique, on s'appuie notamment sur des ontologies comme DOLCE (Gangemi *et al.*, 2005) ou LKIF core (Hoekstra *et al.*, 2009). Une initiative récente, Linked Open Vocabularies (LOV) [8] vise à rassembler et fournir un seul point d'en-

4. http ://www.w3.org/2001/sw/wiki/RDFS

5. http ://www.w3.org/2001/sw/wiki/OWL

6. http ://www.w3.org/2001/sw/wiki/SKOS

7. http ://www.w3.org/2001/sw/wiki/SPARQL

8. Publiée le 26/04/2013, http ://lov.okfn.org/dataset/lov/, par Mondeca, Inserm, DataLift project et Open Knowledge Foundation

trée pour les vocabulaires ouverts liés (ontologies RDFS ou OWL) utilisés dans *Linked Data Cloud*. Les vocabulaires sont listés et décrits individuellement par des métadonnées, organisés dans des classes de vocabulaires et inter-liés par le vocabulaire dédié VOAF (Vocabulary Of A Friend [9]).

Des outils d'annotation sont utilisés pour annoter sémantiquement les documents au regard d'une ontologie, c'est-à-dire pour lier certains fragments de textes (des mots, groupes de mots, phrases, etc.) à des entités de l'ontologie, le plus souvent à des instances (Amardeilh *et al.*, 2005; Uren *et al.*, 2006), mais aussi, dans certains cas, à des concepts et à des rôles (Ma *et al.*, 2013).

Le contenu d'un document ainsi que les annotations qui lui sont attachés peuvent ainsi être publiés sous forme de triplets RDF. Les annotations permettent d'identifier les entités et les concepts mentionnés dans les documents d'un domaine donné : littérature scientifique dans le domaine biomédical (Croset *et al.*, 2010) ou celui de la biodiversité (Cui *et al.*, 2010), comptes rendus hospitaliers (Minard *et al.*, 2011), etc. Dans (Mokhtari, 2010), les annotations sémantiques des documents sont stockés sous forme de triplets RDF, générés selon l'emplacement de leurs propriétés dans le texte.Dans (Croset *et al.*, 2010), la modélisation sous la forme de triplets RDF et d'URIs permet également de lier les articles scientifiques et les bases de connaissances du domaine. (Mrabet *et al.*, 2012) propose à l'inverse d'enrichir des bases de connaissances RDF/OWL en utilisant une base de documents HTML annotés par un ou plusieurs outils d'annotations.

Une fois publiées sous forme de triplets RDF, les annotations sont interrogeables par des requêtes SPARQL, même si une phase de transformation est nécessaire si la requête est formulée en langage naturel. Un système de questions réponses basé sur des patrons de requêtes (utilisés par exemple dans (Pradel *et al.*, 2012)) a été proposé comme solution intuitive et expressive au problème d'accès aux données liées publiées en RDF (Unger *et al.*, 2012).

2.3 Ontologies documentaires

Au-delà de la modélisation du contenu, des ontologies ont été produites pour modéliser les propriétés documentaires. Elles s'inspirent naturellement des langages de métadonnées définis dans la tradition des documentalistes, comme le Dublin Core. Ces ontologies sont souvent conçues pour

9. http ://lov.okfn.org/vocab/voaf/v2.2/index.html

des usages particuliers. Dans (Bouzidi *et al.*, 2011) par exemple, la modélisation doit aider la rédaction des documents réglementaires dans le domaine du bâtiment.

Ces ontologies mettent l'accent sur différents types de propriétés documentaires. L'ontologie SDO (*SALT Document Ontology* [10]) décrit la structure d'une publication scientifique, ainsi que ses propriétés identificatoires et les différentes révisions qu'elle comporte. L'ontologie d'annotation SAO (*SALT Annotation Ontology* [11]) permet de lui associer une couche d'annotation sur le contenu en lien avec des ontologies existantes, telles que FOAF, SWRC et l'ontologie bibliographique BIBO. Cette dernière (*Bibliographic Ontology* [12]) décrit en RDF des entités bibliographiques pour le web sémantique.

D'autres ontologies mettent l'accent sur le cycle de vie du document. L'ontologie PDO (*Project Documents Ontology* [13]) modélise la structure des documents de projets, en rendant compte de leurs différents statuts (rapports d'étape, rapports finaux, livrables, etc.). De la même manière, dans le domaine juridique, l'ontologie MetaLex prend en compte le statut du document (ex. document de travail) et les relations qu'ils entretiennent (`resultOf`, `generatedBy`, etc.).

L'ontologie documentaire proposée dans cet article intègre les différents types de propriétés (sémantiques, structurelles et temporelles) dans un même modèle. Elle permet aussi de rendre compte de la dimension intertextuelle qui est peu représentée dans les ontologies documentaires existantes.

3 Enjeux de la recherche d'information juridique

Notre travail se situe dans le cadre du projet Légilocal, qui vise à faciliter l'accès des citoyens aux documents juridiques des collectivités locales.

De fait, l'accès à l'information juridique est aussi problématique pour les citoyens qui essayent de comprendre la norme qui s'applique à leurs cas particulier que pour les juristes professionnels qui doivent déterminer comment la loi s'applique sur des cas de droit. Le champ du juridique pose des questions spécifiques en terme de recherche d'information.

10. http ://salt.semanticauthoring.org/ontologies/sdo
11. http ://salt.semanticauthoring.org/ontologies/sao
12. http ://uri.gbv.de/ontology/bibo/
13. http ://vocab.deri.ie/pdo-Document

En premier lieu, il est essentiel de comprendre que le tri des résultats retournés par un moteur de recherche n'est pas central dans le domaine juridique, où la recherche d'information se doit d'abord d'être exhaustive. La sécurité juridique impose en effet de prendre connaissance de tous les documents qui se rapportent à un cas particulier. Il est préférable de laisser le contrôle au juriste qui peut progressivement raffiner sa requête en fonction de ses besoins plutôt que de lui présenter un sous ensemble de documents sélectionnés en fonction d'un critère de pertinence défini *a priori*. En cela la recherche d'informations juridiques se distingue clairement des moteurs de recherche généraliste sur le web.

La structure du document est essentielle à prendre en compte. Un texte juridique, notamment le texte d'une loi, est composé d'articles qui ont un cycle de vie autonome. Ils peuvent être modifiés ou même abrogés indépendamment de la loi considérée dans son ensemble. Il est essentiel pour un juriste de pouvoir consolider un texte de loi, c'est-à-dire retrouver toutes les modifications qui s'appliquent à ce texte, et retrouver la version en vigueur à une date donnée, parce qu'il faut pouvoir déterminer le droit qui s'applique à un moment particulier du passé. Il faut également pouvoir ajuster la granularité documentaire (texte complet *vs.* article de ce texte) aux besoins de l'utilisateur et prendre en compte la complexité du cycle de vie du document juridique qui peut être signé, publié, entré en vigueur, promulgué, modifié et abrogé à des dates différentes. Les systèmes actuels d'accès à l'information juridique, comme Normattiva [14] ou UK Legislation [15], prennent partiellement en compte ce type de propriétés quand ils proposent un accès temporel aux sources juridiques (*point in time access*).

Le plus souvent cependant, dans ces systèmes, les notions de modification ou d'abrogation qui sont en réalité des relations intertextuelles sont modélisées comme des attributs de documents. On peut savoir quel est le statut d'un document juridique mais pas quel est le texte qui lui confère ce statut. La dimension intertextuelle des collections de documents juridiques est mal prise en compte. Elle est pourtant centrale dans la compréhension du raisonnement juridique : un texte ne s'interprète pas isolément, indépendamment de la jurisprudence et des interprétations auxquelles il a donné lieu, des textes qui sont venus le modifier ou des décrets qui en précisent l'application. La dimension intertextuelle des collections juridiques est reconnue comme un facteur de complexité majeur (Bourcier, 2011) pour la compréhension du droit. Ouvrir cette complexité est aujourd'hui un défi

14. http ://www.normattiva.it/ricerca/avanzata/vigente

15. http ://www.legislation.gov.uk/search/point-in-time

majeur pour l'accès à l'information juridique [16] : cela suppose de pouvoir lancer des requêtes relationnelles sur un moteur de recherche et de retrouver non pas une liste de documents autonomes mais une liste de graphes de documents qui respectent les contraintes relationnelles formulées en entrée par l'utilisateur (*Quels sont les textes de jurisprudence relatifs au texte de loi donné avant la date d'abrogation de ce dernier ?*).

Au-delà de ces besoins particuliers au domaine juridique, il faut également fournir des outils sémantiques d'accès au contenu pour permettre aux utilisateurs de retrouver des documents à partir de leurs métadonnées d'identification (date de publication, titre, type de document, numéro d'un article, etc.) mais aussi de certaines notions clés.

4 Proposition d'une ontologie documentaire

L'ontologie que nous proposons a été conçue sur la base de cette analyse des besoins. Elle permet de représenter de manière homogène toutes les informations relatives aux documents juridiques : 1) la structure d'un document (sections, paragraphes, etc.), 2) le cadre temporel dans lequel il s'inscrit, 3) la caractérisation sémantique de son contenu à l'aide de concepts ou d'entités du domaine considéré, 4) son type (loi, décret, etc.) et 5) les relations qu'il entretient avec d'autres (modification, abrogation, jurisprudence, transposition, etc.).

Notre ontologie de documents est structurée en trois grands modules qui permettent de modéliser les propriétés ci-dessus : le module document (propriétés 1 et 2), le module sémantique (propriété 3) et le module collection (propriétés 4 et 5). Les détails des classes, propriétés et attributs dans chaque module seront décrits dans ce qui suit [17].

16. Les efforts de simplification juridique actuels portent essentiellement sur la normalisation et le contrôle du lexique, à ce jour.

17. Nous avons utilisé Protégé et OWL-DL pour créer ce modèle ontologique

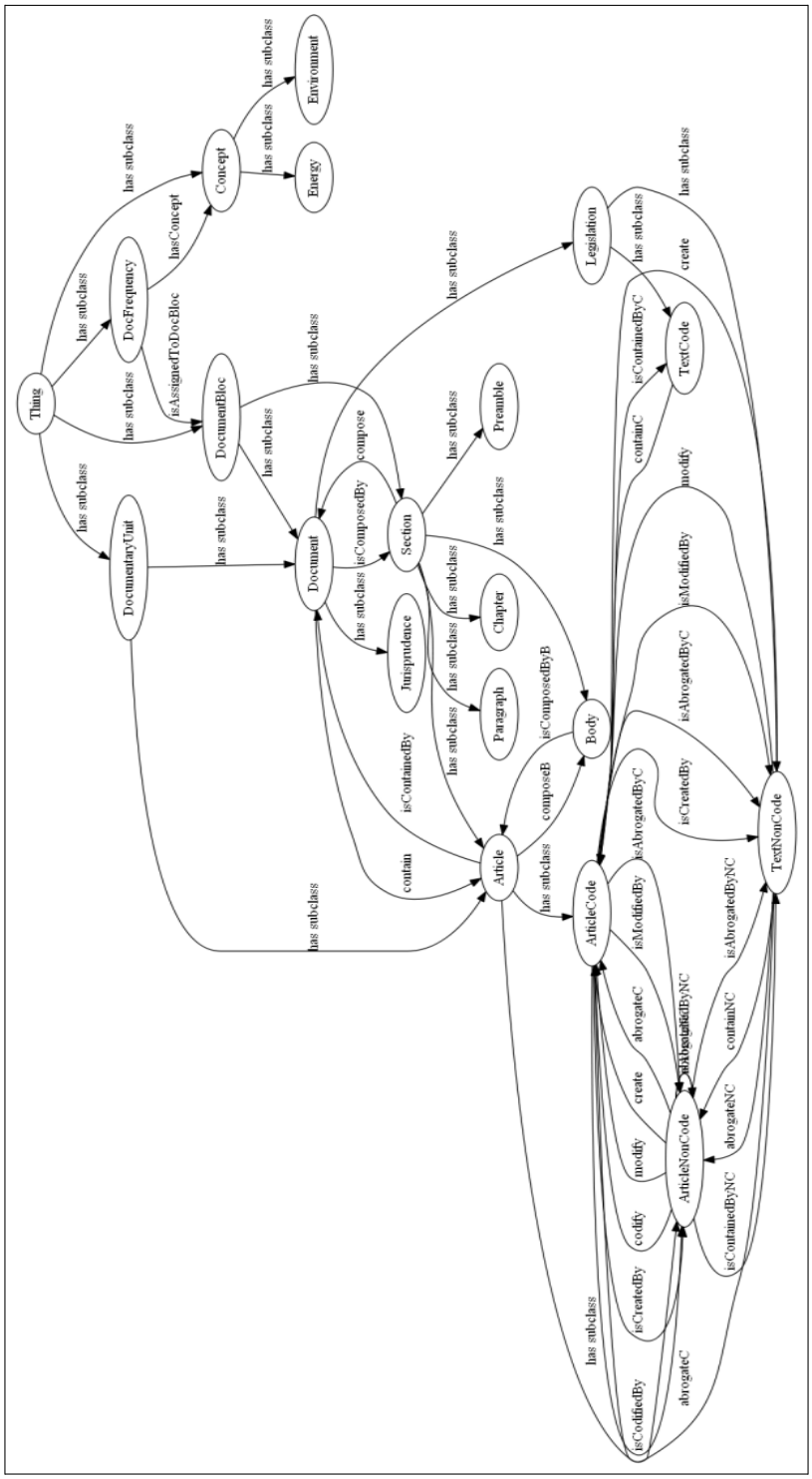

FIGURE 1 – Ontologie de collection documentaire

La figure 1 montre les différents modules de notre ontologie documentaire, sachant que la granularité de la description a été adaptée au cas d'usage Légilocal pour lequel cette ontologie a été initialement développée. Le module document est représenté par les classes `DocumentaryUnit` et `DocumentBloc`. Le module sémantique est représenté par la classe `Concept` et interagit avec le module document *via* la propriété `isAssignedToDoc`. Le module collection est représenté par l'ensemble des types de documents (`Legislation` est une sous classe de `Document`, par exemple) et un ensemble des relations intertextuelles (par ex. `modify`, `isCodifiedBy`, etc.).

Les annotations sémantiques et les relations intertextuelles peuvent porter sur n'importe quel bloc documentaire, que ce soit un document juridique complet ou un de ses composants.

4.1 Module document

Les documents juridiques possèdent une structure riche dont la sémantique est importante à prendre en compte. Les parties d'un document n'ont pas toutes la même importance : le préambule est généralement peu utile alors que les articles qui composent le texte font l'objet de requêtes particulières. L'intérêt de l'utilisateur (citoyen ou expert) porte souvent sur une partie du texte plutôt que sur le texte dans son ensemble. Cela suppose que les métadonnées d'identification et les annotations sémantiques soient attachées non pas au texte globalement mais à ses sous-parties (Hoekstra, 2011).

Le module document est représenté par les classes `DocumentaryUnit` et `DocumentBloc` et leurs sous classes `Document`, `Article` et `Section` (figure 2, figure 3). Dans le contexte de Légilocal, nous considérons en effet que l'unité documentaire de base qui peut être identifiée, qui subit des modifications et qui a un cycle de vie propre est l'article. Un article est lié au document qui le contient par les propriétés `contains` et `isContainedBy`. La structure d'un document peut être décrite plus finement par un ensemble de composants : la classe `Section` est reliée à `Document` par la propriété `compose` ; elle comporte un ensemble de sous-classes qui décrivent les différentes parties d'un document (figure 3).

Le cycle de vie du document juridique (le texte de loi mais aussi chacun des articles qui le composent) est complexe du fait des processus de modi-

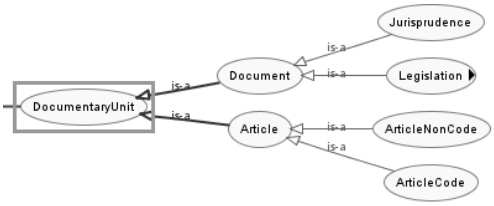

FIGURE 2 – Unité documentaire : Document ou Article

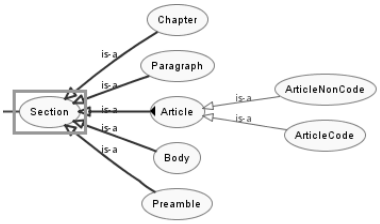

FIGURE 3 – Composants d'un document

fication des textes juridiques, de leur consolidation [18] ou de leur application par les structures gouvernementales. Plusieurs dates sont généralement associées à un document. Nous les représentons par des attributs (*Datatype properties*) dont les valeurs sont de type `date`. C'est la nature de l'attribut qui associe à la date son statut (date de création, d'entrée en vigueur, etc.).

Notons ici que nous considérons toutes les versions d'articles comme des unités documentaires différentes et que la modification d'un article est représentée par le lien existant entre l'article modifieur et l'article modifié, la date de modification étant celle de l'entrée en vigueur de l'article modifieur.

4.2 Module sémantique

Le module sémantique (voir figure 4) est classique. En pratique, on cherche généralement à réutiliser une ontologie existante. Nous définissons le prédicat `hasConcept` et `isAssignedToDoc` entre un concept du module sémantique et un bloc documentaire (document ou section). Nous avons également prévu d'associer une fonction de pondération à la

18. La consolidation consiste à intégrer dans un acte de base tous ces actes modificateurs.

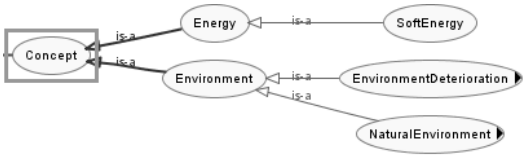

FIGURE 4 – Module sémantique : concepts du domaine

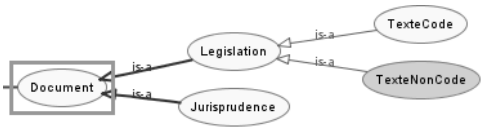

FIGURE 5 – Hiérarchie des types des documents juridiques

relation `hasConcept` pour tenir compte du nombre d'occurrences d'un concept dans un document donné. Pour modéliser cette relation ternaire nous avons créé une classe, `DocFrequency`, sur laquelle sont définis trois prédicats, `isAssignedToDoc` vers la classe bloc documentaire, `hasConcept` vers la classe concept et `hasFrequency` vers une donnée de type entier.

Comme mentionné plus haut, selon la stratégie d'annotation adoptée, les annotations sémantiques portent uniquement sur des instances du module sémantique ou renvoient aussi à des concepts et des rôles. Aucun choix n'a encore été fait à ce stade dans le projet Légilocal.

4.3 Module collection documentaire

Comme argumenté plus haut, il est essentiel de modéliser des différents types des documents juridiques. La figure 5 fait un zoom sur la hiérarchie des types de documents.

Par ailleurs, du fait de la structure hyper enchevauchée de la législation, il faut souvent consulter plusieurs textes et plusieurs versions de ces textes pour interpréter une loi. Une décision publiée comme jurisprudence doit être reliée aux textes législatifs qu'elle met en application. Il est essentiel de pouvoir modéliser cette dimension de l'intertextualité dans l'ontologie documentaire : cela doit permettre d'interroger la collection documentaire en croisant les critères sémantiques, temporels, structurels et relationnels.

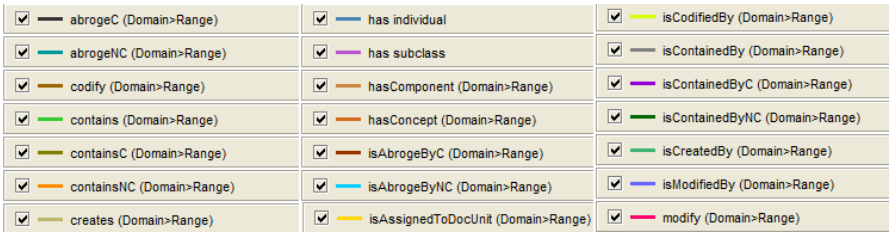

FIGURE 6 – Différents types de relations entre les entités

Dans notre ontologie, l'intertextualité est modélisée par des relations (*Object properties*) qui ont pour sujet une unité documentaire (document ou article) et pour objet une autre unité documentaire (document ou article). Par exemple le prédicat `isCreatedBy` définit la relation de création entre un article de texte non codifié [19] vers un article de texte codifié [20]. La figure 6 donne un aperçu des types de relations que nous avons codés dans ce module.

Nous précisons que nous ne prenons pas en compte ici les visas dit "de forme", c'est-à-dire les références qui figurent dans le préambule des documents juridiques, et qui ont généralement une valeur très générale, comme les références au Code Civil, par exemple.

5 Interrogation

La modélisation d'une collection juridique revient à instancier l'ontologie [21] en produisant un ensemble de triplets RDF : sont ainsi modélisés les documents et leurs types (`Legislation`, `Jurisprudence`, etc), les articles (`ArticleCode`, `ArticleNonCode`), les concepts du domaine (`Energy`, `Environment`, etc.), les relations entre classes (`hasConcept`, `isModifiedBy`, etc). La collection est ainsi représentée comme une base de connaissances qui peut ensuite être interrogée à l'aide de requêtes SPARQL.Ces requêtes peuvent porter sur :

19. Un article non codifié est un article qui appartient à un texte réglementaire autre que les codes : articles de loi, etc.

20. Un article codifié est un article qui appartient à un code : code civil, code de l'environnement.

21. Dans le cadre du projet, le peuplement de l'ontologie se fera automatiquement avec les résultats des outils d'analyse de documents (ex. structuration, repérage de références dans le texte, annotation sémantique).

- le contenu sémantique d'un type donné de documents : la requête *Quels sont les textes qui parlent de la préservation de l'environnement ?* porte par exemple sur les classes `Concept`, et `DocumentBloc` (le concept `DocFrequency` peut également être pris en compte) ;
- l'historique d'une unité documentaire, qu'il s'agisse d'un document ou de l'un de ses articles : la requête *Comment a été abrogé l'article 22 de la loi sur l'enseignement obligatoire ?* fait appel à la classe `Article`, à la propriété `estAbrogéPar` et doit retourner tous les documents qui ont abrogé l'article 22 en question (si la version de l'article 22 considérée n'est pas précisée, tous les textes abrogatifs doivent être retournés) ;
- les types de documents et les types des liens : la requête *Donnez moi les jurisprudences qui ont appliqué l'article 4 actuellement en vigueur de la loi Sapin ?* porte par exemple sur les classes `Jurisprudence`, `ArticleNonCode` (article 4), `TexteNonCode` (loi Sapin), sur la propriété `dateVigueur` de la classe `ArticleNonCode` et sur la propriété `appliedBy` (non encore présentée dans le modèle) reliant les textes de jurisprudence aux textes de loi.

Nous pouvons aussi interroger sur la consolidation d'un texte (mise à jour d'un texte de loi : décret, loi, etc.) à une date donnée. Cela suppose un calcul un peu plus compliqué, puisqu'il faut partir de la structure du texte, identifier la liste des articles qui le composent et retrouver pour chacun la version en vigueur à la date considérée.

Nous avons collecté un ensemble de requêtes utilisateurs suite à une analyse de besoins menée auprès des juristes. En étudiant ces requêtes nous avons constaté qu'elles présentent des structures récurrentes sur lesquelles nous pouvons définir des patrons de requêtes. Cette particularité nous semble un atout en faveur d'une interface d'interrogation plus facile permettant aux utilisateurs d'exprimer leurs besoins en langage naturel puis de faire la traduction de ces requêtes sous forme de graphes (exprimés en SPARQL). Cette solution n'est pas envisageable à ce stade de notre étude. Nous comptons à court terme définir des interfaces d'interrogation en langage contrôlé ou sous forme de formulaires.

6 Discussion

Le modèle ontologique que nous avons présenté peut être affiné mais la version présentée ici correspond au degré de modélisation requis par le projet Légilocal. Ce modèle est en passe d'être adopté par l'ensemble des par-

tenaires : il permet déjà d'élargir largement le cadre de la recherche d'information sémantique à la recherche d'information fine et à la recherche relationnelle. L'ontologie proposée a été définie en s'appuyant sur une expertise du domaine. Sa validation se fera en deux étapes : étape du peuplement avec des documents du projet, étape du déploiement du moteur de recherche et de son évaluation par des communes (ou comités de public). L'ontologie que nous proposons est utilisable par toute autre application de gestion documentaire dans le domaine juridique comme par exemple l'aide à la publication ou la consolidation. Les prochaines étapes de ce travail consistent à instancier l'ontologie sur le corpus de documents de travail du projet Légilocal, à produire les triplets RDF à partir de ces données et à les stocker dans des triples stores, mais aussi, à concevoir un module de traduction des requêtes utilisateurs en requêtes SPARQL et une interface d'interrogation et de visualisation des résultats.

Références

ABASOLO J. M. & GOMEZ M. (2000). MELISA. An ontology-based agent for information retrieval in medicine. In *Proceedings of the First International Workshop on the Semantic Web (SemWeb2000)*, p. 73–82.

AMARDEILH F., LAUBLET P. & MINEL J.-L. (2005). Document annotation and ontology population from linguistic extractions. In *Proceedings of the 3rd international conference on Knowledge capture (K-CAP '05)*, p. 161–168.

BOURCIER D. (2011). Sciences juridiques et complexité. Un nouveau modèle d'analyse. *Droit et Cultures*, **61**(1), 37–53.

BOUZIDI K. R., FARON-ZUCKER C., FIES B., CORBY O. & NHAN L.-T. (2011). Modélisation de documents réglementaires dans le domaine du bâtiment. In *Actes 12e Conférence Internationale Francophone sur l'Extraction et la Gestion de Connaissance, EGC 2011*, Bordeaux, France.

CIORASCU C., CIORASCU I. & STOFFEL K. (2003). KnOWLer - Ontological Support for Information Retrieval Systems. In *Proceedings of 26th Annual International ACM SIGIR Conference, Workshop on Semantic Web*.

CROSET S., GRABMÜLLER C., LI C., KAVALIAUSKAS S. & REBHOLZ-SCHUHMANN D. (2010). The CALBC RDF Triple Store : retrieval over large literature content. *CoRR*, **abs/1012.1650**.

CUI H., JIANG K. Y. & SANYAL P. P. (2010). From text to RDF triple store : an application for biodiversity literature. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem*, volume 47, p. 1–2 : American Society for Information Science.

GANGEMI A., SAGRI M.-T. & TISCORNIA D. (2005). A Constructive Framework for Legal Ontologies. In V. BENJAMINS, P. CASANOVAS, J. BREUKER

& A. GANGEMI, Eds., *Law and the Semantic Web*, volume 3369 of *Lecture Notes in Computer Science*, p. 97–124. Springer Berlin Heidelberg.

HOEKSTRA R. (2011). The METALEX document server : legal documents as versioned linked data. In *Proceedings of the 10th International Conference on the Semantic Web*, ISWC'11, p. 128–143, Berlin, Heidelberg : Springer-Verlag.

HOEKSTRA R., BREUKER J., BELLO M. D. & BOER A. (2009). LKIF Core : Principled Ontology Development for the Legal Domain. In *Proceedings of the 2009 conference on Law, Ontologies and the Semantic Web : Channelling the Legal Information Flood*, p. 21–52, Amsterdam : IOS Press.

LOPEZ V., UREN V., MOTTA E. & PASIN M. (2007). AquaLog : An ontology-driven question answering system for organizational semantic intranets. *Web Semant.*, **5**(2), 72–105.

MA Y., LÉVY F. & NAZARENKO A. (2013). Annotation sémantique pour des domaines spécialisés et des ontologies riches. In *Actes de la 20ème conférence du Traitement Automatique du Langage Naturel (TALN 2013)*.

MINARD A.-L., LIGOZAT A.-L. & GRAU B. (2011). Extraction de relations dans des comptes rendus hospitaliers. In *22es Journées Francophones d'Ingénierie des Connaissances*, IC 2011, p. 491–506, Chambéry, France.

MOKHTARI N. (2010). *Extraction et exploitation d'annotations sémantiques contextuelles à partir de texte*. PhD thesis, Université Sophia Antipolis.

MRABET Y., BENNACER N. & PERNELLE N. (2012). Enrichissement contrôlé de bases de connaissances à partir de documents semi-structurés annotés. In *23es Journées Francophones d'Ingénierie des Connaissances*, IC 2012, Paris.

PRADEL C., HAEMMERLÉ O. & HERNANDEZ N. (2012). Des patrons modulaires de requêtes SPARQL dans le système SWIP. In *23es Journées Francophones d'Ingénierie des Connaissances*, IC 2012, Paris, France.

SARTOR G. (2011). *Law, Governance and Technology : Legislative Xml for the Semantic Web : Principles, Models, Standards for Document Management*. Law, Governance and Technology Series, 4. Springer London, Limited.

UNGER C., BÜHMANN L., LEHMANN J., NGONGA A.-C. N., GERBER D. & CIMIANO P. (2012). Template-based question answering over RDF data. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, p. 639–648 : ACM.

UREN V., CIMIANO P., IRIA J., HANDSCHUH S., VARGAS-VERA M., MOTTA E. & CIRAVEGNA F. (2006). Semantic Annotation for Knowledge Management : Requirements and a Survey of the State of the Art. *Journal of Web Semantics*, **4**.