Un système de questions-réponses automatiques dans le domaine légal : le cas des réglementations maritimes

Cheikh KACFAH EMANI* — **Yannis HARALAMBOUS****

* cheikh.kacfah@imt-atlantique.fr, *IMT Atlantique, CS 83818, 29238 Brest Cedex 3 et DECIDE, UMR CNRS 6285 Lab-STICC.*

** yannis.haralambous@imt-atlantique.fr, *IMT Atlantique, CS 83818, 29238 Brest Cedex 3 et DECIDE, UMR CNRS 6285 Lab-STICC.*

RÉSUMÉ. *Nous présentons les premiers travaux du projet REIZHMOR dont le but est la modélisation de textes juridiques et réglementaires autour de la navigation maritime. Après une présentation du corpus, constitué par les arrêtés préfectoraux et interpréfectoraux référencés dans un volume des* Instructions nautiques *du Shom (Service hydrographique et océanographique de la marine), nous décrivons l'élaboration d'un système de questions-réponses fondé sur des requêtes SPARQL adressées à une base de connaissances, avec une attention particulière portée aux difficultés spécifiques du langage juridique.*

ABSTRACT. *We present the first steps of the REIZHMOR project, the goal of which is to model legal and regulatory texts on sea navigation. After a short presentation of the corpus, consisting of prefectoral decrees referenced in a volume of the* Sailing Directions *of the Hydrographic and Oceanographic Service of French Navy, we describe the development of a Question-Answering system based on SPARQL queries to a knowledge base, giving particular attention to the specific difficulties of legal language.*

MOTS-CLÉS : *textes réglementaires, textes juridiques, navigation maritime, ontologie légale, système questions-réponses, SPARQL, règles.*

KEYWORDS: *Regulatory texts, legal texts, sea navigation, legal ontology, Question-Answering system, SPARQL, rules.*

1. Introduction

De nombreux textes régissent le domaine de la navigation maritime (de Cet Bertin, 2008). En ce qui concerne la navigation dans les eaux françaises et dans les eaux internationales, ces textes proviennent de divers corpus tels que les règles et recommandations internationales applicables aux transports et activités maritimes rédigées par l'Organisation maritime internationale (MARPOL, SOLAS, etc.), le code des ports maritimes, les arrêtés des préfectures maritimes, etc. En guise d'aide à la navigation, à ce corpus s'ajoutent les *Instructions nautiques* du Shom (Service hydrographique et océanographique de la marine) ainsi que les *Avis aux navigateurs*. Les *Instructions nautiques* sont des textes accompagnant les cartes marines ; elles complètent ces dernières en fournissant des informations telles que la réglementation, les données météorologiques, les canaux à très hautes fréquences ou même des informations culturelles et linguistiques (Sauvage-Vincent, 2017 ; Haralambous *et al.*, 2017). Les *Avis aux navigateurs* sont des informations de sécurité maritime qui mettent à jour ou complètent les documents nautiques, de façon permanente ou temporaire. Nous appellerons ce corpus, les *réglementations maritimes*.

Dans cet ensemble complexe de réglementations, il est primordial pour le navigateur – qui est aux commandes d'un type de navire précis, à un moment donné et dans un espace donné – de connaître l'ensemble des réglementations qui sont pertinentes pour lui. D'un autre côté, il est important pour les autorités chargées de la surveillance des espaces maritimes et des équipements afférents à ceux-ci, de relever les infractions et les situations à risque de la part des navigateurs. Enfin, les réglementations évoluant sans cesse, il est important pour les rédacteurs juridiques de s'assurer que l'ensemble des réglementations ne présente pas de contradiction et que l'ajout de nouveaux textes préserve sa cohérence.

Le Shom est en train de mettre en place une base de connaissances (Sauvage-Vincent, 2017 ; Haralambous *et al.*, 2017) à partir des informations contenues dans les *Instructions nautiques* et les autres ouvrages qu'il publie. Il est prévu que cette base de connaissances soit accessible aux navigateurs et leur propose un certain nombre de services contextuels, c'est-à-dire tenant compte de la nature et de la situation du navire, de sa position et des conditions météorologiques. Le but du projet REIZHMOR (mot-valise signifiant « loi maritime » en breton), démarré en avril 2017 et financé par le Shom, est d'ajouter un module juridique et réglementaire à cette base de connaissances, de manière à exploiter également le corpus réglementaire maritime.

Dans cet article, nous nous intéressons à la tâche qui consiste à répondre automatiquement aux questions qui peuvent être posées en langage naturel à la base de connaissances. Ce procédé est appelé *réponses automatiques aux questions* (RAQ), (*Question Answering* en anglais). Notre but est de poser les bases d'un système pour mener à bien cette tâche.

En tant que preuve de concept, nous avons sélectionné et traité un corpus (§ 2), nous nous sommes appuyés sur des travaux existants concernant une ontologie maritime en les adaptant à nos besoins (§ 3.1), et nous avons élaboré un système de

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

PRÉFECTURE MARITIME DE LA MANCHE ET DE LA MER DU NORD

Cherbourg, le 23 février 2017

PRÉFECTURE MARITIME DE LA MANCHE ET DE LA MER DU NORD
Division « action de l'État en mer »

ARRÊTÉ PRÉFECTORAL N° 03/2017

RÉGLEMENTANT LE DÉPÔT D'ENGINS SUSPECTS TROUVÉS EN MER ET FIXANT LES ZONES DE DÉPÔT TEMPORAIRE ET DE NEUTRALISATION DE CES ENGINS AUX ABORDS DES PRINCIPAUX PORTS DE LA FAÇADE MANCHE - MER DU NORD

-

Le vice-amiral d'escadre Pascal Ausseur
préfet maritime de la Manche et de la mer du Nord,

Vu la convention des Nations Unies sur le droit de la mer du 10 décembre 1982 ;
Vu le code des transports ;
Vu le code pénal ;
Vu l'ordonnance n° 2012-1218 du 2 novembre 2012 portant réforme pénale en matière maritime ;
Vu le décret n° 76-225 du 4 mars 1976 fixant les attributions respectives du ministre de l'Intérieur et du ministre de la Défense en matière de recherche, de neutralisation, d'enlèvement et de destruction des munitions et explosifs ;
Vu le décret n° 2004-112 du 6 février 2004 modifié, relatif à l'organisation de l'action de l'État en mer ;

Considérant que tout engin trouvé en mer ayant les caractéristiques ou l'aspect d'un matériel susceptible de contenir des matières explosives doit être considéré comme dangereux.

ARRÊTE

Article 1er : Définition des engins suspects.

Sont considérés comme engins suspects tous les objets trouvés en mer, immergés ou flottants, susceptibles de contenir des matières explosives dangereuses, notamment ceux ayant les caractéristiques ou l'aspect d'un matériel de guerre (bombe, obus, mine, torpille).

Préfecture maritime de la Manche et de la mer du Nord – CC 01 – 50115 Cherbourg-en-Cotentin Cedex
Tél : 02.33.92.60.61 – Fax : 02.33.92.59.26 – pc.aem@premar-manche.gouv.fr

Figure 1. *Exemple de la première page d'un arrêté*

réponses automatiques aux questions fondé sur des patrons et débouchant sur des requêtes SPARQL (§ 6).

Dans la section suivante, nous décrivons notre corpus.

2. Le corpus textuel : les arrêtés référencés dans l'*Instruction nautique* C2A

Pour rester aussi près que possible de la base de connaissances du Shom dont la source principale sont les *Instructions nautiques*, nous avons décidé de sélectionner en tant que corpus un ensemble d'arrêtés préfectoraux et interpréfectoraux référencés et en partie reproduits dans un même volume des dites *Instructions*. Nous avons choisi le volume C2A (Shom, 2010), qui traite les côtes nord et ouest de la France, entre la frontière belge et la pointe de Penmarc'h. Nous avons ainsi récupéré et traité 75 arrêtés préfectoraux et interpréfectoraux (préfectures maritimes de Brest et de Cherbourg) qui s'étalent chronologiquement entre le 10 juin 1963 et le 23 février 2017.

2.1. *Structure d'un arrêté préfectoral*

Les arrêtés de notre corpus ont tous la même structure logique et visuelle (fig. 1) qui peut être résumée comme suit : (1) un intitulé (le plus souvent une phrase utilisant un verbe au participe présent : « Arrêté interdisant la navigation à proximité de... », cf. § 2.1.1) ; (2) un ou plusieurs signataires (indiquant la fonction du signataire, et souvent aussi son identité : « Le vice-amiral d'escadre Pascal Ausseur, préfet maritime de la Manche et de la mer du Nord ») ; (3) les *visas* : il s'agit de références vers d'autres textes commençant invariablement par le mot « VU » (souvent écrit en majuscules et/ou en gras), cf. § 2.1.2 ; (4) les « SUR DEMANDE » ou « SUR PROPOSITION » : il s'agit de personnes ou d'institutions ayant formulé une demande ou une proposition qui est à l'origine de l'arrêté ; (5) les « CONSIDÉRANT » : à la suite des visas, ils indiquent les motivations de l'arrêté (par exemple : « CONSIDÉRANT que tout engin trouvé en mer [...] doit être considéré comme dangereux. ») ; (6) le mot « ARRÊTE » (ou « ARRÊTENT » dans le cas de plusieurs signataires), invariablement écrit en majuscules ; (7) les *articles* de l'arrêté, dont l'avant-dernier concerne souvent la gestion des infractions à l'arrêté et le dernier les personnes ou institutions chargées de son exécution ; (8) des éventuelles *annexes*. Dans la majorité des cas, les annexes précisent, par des listes de coordonnées géographiques ou par des cartes, la zone d'application de l'arrêté.

2.1.1. *Les intitulés*

Les intitulés des arrêtés de notre corpus sont tous, à deux exceptions près, des phrases utilisant des verbes au participe présent : « réglementant » (ou « portant règlement ») dans 45 % des cas ; « interdisant » (ou « portant interdiction ») dans 19 % des cas, l'objet de l'interdiction pouvant être la navigation, le mouillage, le dragage, le chalutage, le rejet à la mer d'objets, la plongée sous-marine, la pêche, la baignade, ou les activités aquatiques et subaquatiques ; « portant création d'une zone » dans 11 % des cas, la zone pouvant être interdite, réglementée, d'immersion de déblais de dragage ou de mouillage ; « portant définition d'une zone » dans 7 % des cas ; « instituant », « autorisant », « délimitant », « précisant » et « portant restriction » dans quelques cas isolés.

Notons que dans 7 % de cas l'intitulé décrit l'arrêté comme étant « relatif » à des sujets très spécifiques (à la circulation des navires, au compte-rendu obligatoire des navires, à l'accès aux ports, au pilotage des bateaux et à la navigation des bateaux fluviaux).

2.1.2. *Les visas*

Il s'agit de références intertextuelles, énumérées dans un ordre qui respecte *grosso modo* la hiérarchie des normes de droit français : règlements internationaux et conven-

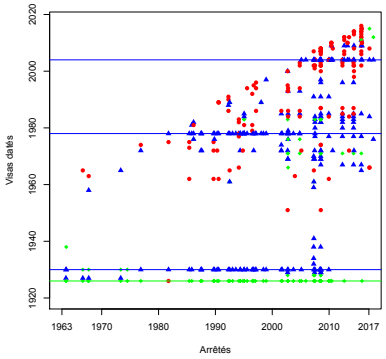


Figure 2. *Correspondance entre arrêtés et dates de visas (losanges verts : lois, triangles bleus : décrets, disques rouges : arrêtés)*

tions[1], codes, ordonnances, lois et décrets, arrêtés, demandes administratives, procès-verbaux et avis de personnes.

Dans la figure 2 nous avons représenté pour chaque arrêté (les arrêtés étant disposés chronologiquement sur l'abscisse) les dates des visas de type loi, décret et arrêté qu'il comporte. On constate que depuis les années 2000 la proportion d'arrêtés figurant dans les visas augmente, et les dates de ceux-ci sont assez proches de la date de publication de l'arrêté. La distribution des décrets reste assez uniforme, alors que les visas de lois sont plutôt épars et se concentrent dans la période après 2000 pour les arrêtés, et les années 70 pour les lois. Nous avons remarqué une présence récurrente de quatre visas pour lesquels nous avons tracé sur la figure des lignes horizontales indicatrices :

(1) la loi du 17 décembre 1926 portant code disciplinaire et pénal de la marine marchande (60 % des arrêtés du corpus) ; (2) le décret du 1er février 1930, relatif aux pouvoirs de police et à la réglementation de la pêche côtière (76 % des arrêtés du corpus) ; (3) le décret no 78-272 du 9 mars 1978 relatif à l'organisation des actions de l'État en mer (59 % des arrêtés qui lui sont postérieurs) ; (4) le décret no 2004-112 du 06 février 2004 relatif à l'organisation de l'action de l'État en mer (48 % des arrêtés qui lui sont postérieurs). À cela s'ajoute l'ordonnance royale du 14 juin 1844 concernant le service de la marine (police des rades) qui est quasiment omniprésente (83 % des arrêtés du corpus) mais que nous n'avons pas représentée sur le diagramme pour rendre la partie 1920-2017 plus lisible.

Il est intéressant de noter que même si la structure des arrêtés est restée la même, on remarque de fortes variations dans leur volume, ainsi qu'une certaine évolution au fil du temps.

1. Dans les arrêtés des années 60 nous avons observé des références indirectes vers les conventions internationales à travers de décrets les publiant. Ce n'est qu'à partir de 2002 que nous observons des références directes vers de telles conventions.

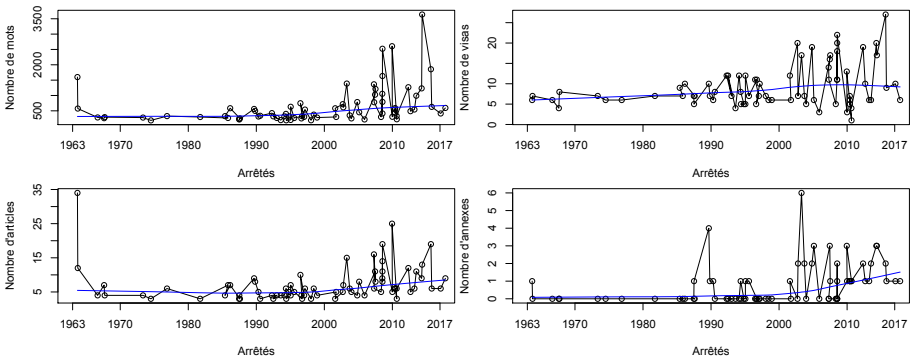


Figure 3. *Évolution du nombre de mots, de visas, d'articles et d'annexes dans le corpus d'arrêtés*

2.1.3. *Évolution de la volumétrie des arrêtés*

La figure 3 représente l'évolution du nombre de mots, de visas, d'articles et d'annexes de notre corpus d'arrêtés. Les courbes bleues représentent des courbes de régression quadratique. À l'exception de l'arrêté du 10 juin 1963 « portant règlement de police et de sécurité pour l'arsenal de Brest et pour les établissements maritimes de l'arrondissement maritime de Brest », qui est exceptionnel par sa taille (1 600 mots, 34 articles) et de l'arrêté du 9 juin 1989 « réglementant l'accès à l'île de Cézembre (Ille-et-Vilaine) » qui comporte quatre annexes, les arrêtés de la période 1960-2000 sont plutôt réduits en nombre de mots, visas, articles, annexes. À partir de l'an 2000, on constate une très grande variabilité des quatre paramètres. Globalement nous constatons une très légère augmentation pour les trois premiers paramètres, et une augmentation un peu plus significative du nombre d'annexes.

3. Modélisation ontologique du corpus

3.1. *L'ontologie du projet* e-Compliance

Un des livrables du projet européen *e-Compliance* (2014) a été une ébauche d'*ontologie maritime* (Lohrmann *et al.*, 2014). Les concepts de cette ontologie sont répartis en quatre catégories (appelées « classes ») : (1) le « légal » : les réglementations, les documents les contenant, le sens qui peut être extrait d'une réglementation. C'est dans cette catégorie que l'on trouve le concept central de l'ontologie qui est `Rule` ; (2) le « maritime » : types de navires et activités maritimes afférentes aux réglementations ; (3) l'« organisationnel » : les organisations, les juridictions et les rôles qui sont responsables des réglementations ou concernés par elles ; (4) le « territorial » : les espaces (géographique, politique, légal) dans lesquels les réglementations s'appliquent et dans lesquels une action prend place.

Dans cette ontologie, une réglementation est un ensemble de *clauses*, et à partir des clauses on peut extraire des *règles* atomiques. Une règle a trois parties : (1) une *cible*, `target`, qui est l'« objet » sur lequel la prescription ou la recommandation est faite ; (2) une *exigence*, `requirement`, qui dénote l'action requise par la cible pour être conforme à la réglementation ; (3) optionnellement, un *contexte*, `context`, précisant les conditions d'application de la règle.

Par exemple, dans la clause « l'accès de toute personne en état d'ivresse est interdit », la cible est « toute personne en état d'ivresse » et l'exigence est l'interdiction d'accès. Notons que la terminologie de cette ontologie reste limitée. Elle nécessite une extension au corpus qui nous intéresse.

3.2. *Extension de l'ontologie* e-Compliance *au corpus des arrêtés*

Après nettoyage du corpus (dans lequel nous n'avons gardé que le contenu des intitulés et des articles des arrêtés) nous avons procédé à une extraction terminologique des termes complexes à l'aide du logiciel *Acabit* (Daille, 2003). Celui-ci nous a fourni 3 650 candidats que nous avons vérifié manuellement pour aboutir à 1 164 termes complexes (du type groupe nominal) d'une longueur moyenne de 2,976 mots (le terme le plus long étant « règlement général de police, de navigation, de mouillage et de pêche »).

Parallèlement à cela nous avons extrait les arbres syntaxiques par dépendances des textes à l'aide du logiciel *MaltParser* (Nivre *et al.*, 2006). Ensuite nous avons intégré les termes complexes dans les arbres syntaxiques en écrasant les sous-arbres correspondant à des termes complexes. Finalement nous avons traduit l'information extraite dans le format OWL, en utilisant des techniques introduites par Simperl *et al.* (2008) pour l'élaboration d'une ontologie juridique. Cela nous a permis d'étendre l'ontologie e-Compliance en tenant compte des connaissances extraites de notre corpus textuel.

4. Les réponses automatiques aux questions et leurs difficultés

Avec Hirschman et Gaizauskas (2001) nous disons qu'un système de RAQ est « un système qui permet à des utilisateurs de poser leurs questions en langage naturel, en utilisant leur propre terminologie, et pour lesquelles ils attendent des réponses précises, obtenues en interrogeant une base de connaissances ». La problématique de la RAQ a été abordée très tôt dans le domaine de l'intelligence artificielle dans les années 70 (Lopez *et al.*, 2011). Néanmoins, malgré l'ancienneté du domaine, plusieurs défis restent encore à relever. Dans le sous-domaine du Web sémantique (Berners-Lee *et al.*, 2001), il existe plusieurs dizaines de travaux abordant la RAQ. Le récent état de l'art proposé par Höffner *et al.* (2016) permet de se rendre compte de la diversité des travaux dans ce domaine[2] et surtout d'exhiber les défis inhérents à tous les systèmes

2. Ainsi, Höffner *et al.* (2016) ont identifié plus de 72 publications proposant 62 systèmes de RAQ. Ces publications couvrent la période de novembre 2010 à juillet 2015.

de RAQ. Ainsi, (Höffner *et al.*, 2016) regroupent les difficultés que doivent surmonter les systèmes de RAQ en sept catégories :

1) la *variété lexicale* ;

2) l'*ambiguïté* induite par la polysémie des termes ;

3) le *multilinguisme* ;

4) la *complexité de la question.* S'il est facile de répondre à une question « simple », dans le sens où elle peut être modélisée par un seul triplet RDF[3] (*e.g.* « quel est l'intitulé de l'arrêté 03/2017 ? »), les questions complexes peuvent, quant à elles, faire appel à plusieurs triplets RDF qu'il faut identifier correctement pour les combiner de manière adéquate et si nécessaire inclure des filtres, des fonctions d'agrégation ou des tris ;

5) les *bases de connaissances distribuées.* Dans certains cas, il est nécessaire de combiner les informations de plusieurs sources pour pouvoir répondre à une question de l'utilisateur. Adresser ce problème peut nécessiter l'exploitation d'alignements déjà existants entre différentes ressources ou alors l'utilisation d'un moteur d'inférence pour les obtenir à la volée ;

6) les *questions procédurales, temporelles ou spatiales.* Les bases de connaissances fondées sur les triplets RDF se prêtent difficilement aux questions temporelles (par exemple concernant l'ordonnancement des événements), aux questions spatiales (par exemple sur le degré de superposition d'entités géographiques) et aux questions procédurales (c'est-à-dire celles qui demandent une liste d'étapes, autrement dit : une procédure) afin de résoudre un problème ;

7) les *patrons de questions.* Pour les questions complexes (voir item 4 ci-dessus), plusieurs approches ont recours aux *patrons* (cf. définition ci-dessous) syntaxiques et/ou sémantiques pour extraire le sens de la question.

Dans un domaine donné, nous nous proposons de décrire l'ensemble des questions valides en tant que langage formel : toute question est alors la séquence de feuilles d'un arbre de dérivation, à condition qu'elles soient des symboles terminaux de la grammaire, et donc des membres de l'alphabet du langage. Nous appelons *patron* d'une question une séquence de feuilles d'un arbre de dérivation, *constituée de symboles non terminaux* (par exemple, pour la phrase « la fille mange la pomme », des patrons possibles sont « GN GV », « GN V GN », etc.). Intuitivement on peut dire qu'un patron est le résultat d'une suite de règles de production appliquées à l'axiome de départ sans qu'on « aille jusqu'au bout », c'est-à-dire sans qu'on n'aboutisse à des symboles terminaux.

Une approche de RAQ privilégiée par les chercheurs est la génération de patrons de requêtes SPARQL à partir de patrons de questions (le formalisme SPARQL étant choisi parce qu'il fait partie intégrante de nombreux outils du Web sémantique).

3. Un triplet RDF est un triplet de ressources « sujet, prédicat, objet » modélisant une phrase du type « sujet, verbe, complément ».

5. Particularités de la RAQ liées au domaine réglementaire

En plus des difficultés mentionnées dans la section précédente, un système de RAQ ciblant une base de connaissances réglementaires doit faire face à des spécificités induites par les particularités du langage juridique (qui peut être considéré comme étant un langage contrôlé, cf. § 5.1) et le décalage d'articulation conceptuelle existant dans les ontologies légales (cf. § 5.2).

5.1. *Le langage réglementaire juridique en tant que langage contrôlé*

Pour le lecteur non spécialiste, le langage juridique, et plus particulièrement, le langage législatif (Cornu, 2005, titre 2, chap. 1) semble rigide (au sens où, du moins dans le cadre de la réglementation, l'utilisation de la périphrase est très limitée et que tout mot ajouté ou supprimé est susceptible d'avoir un impact fort sur la sémantique de la phrase), voire quasi formel et proche du langage des mathématiques. Néanmoins, cette « formalité » apparente ne fait pas du langage juridique un véritable langage formel.

Pour l'évaluer néanmoins en tant que *langage contrôlé*, nous utilisons la classification PENS de Kuhn (2014).

5.1.1. *La précision*

Selon le cas (type de document réglementaire, auteur du document, cadre juridique), on peut dire que du point de vue de la *précision* on se situe entre le P^2 (langages imprécis : *degree of ambiguity and vagueness is considerably lower than in natural languages, and their interpretation depends much less on context*[4]) et le P^3 (langages à interprétation fiable : *syntax is heavily restricted, though not necessarily formally defined. The restrictions are strong enough to make automatic interpretation reliable*[5])[6].

4. Traduction : le degré d'ambiguïté et d'imprécision est considérablement inférieur à celui des langages naturels, et leur interprétation dépend beaucoup moins du contexte.

5. Traduction : la syntaxe est fortement réduite même si elle n'est pas nécessairement formellement définie. Les restrictions sont suffisamment fortes pour rendre l'interprétation automatique fiable.

6. Pour illustrer la variabilité (qui est faible mais néanmoins présente) du langage juridique dans notre corpus, prenons un article que l'on retrouve obligatoirement dans chaque arrêté et qui exprime le fait que les infractions à l'arrêté seront sanctionnées selon un certain nombre de textes indiqués. Le sens de cet article est invariable, pourtant sa formulation peut prendre plusieurs formes : « Les infractions au présent arrêté sont passibles des peines prévues par... », « Les infractions au présent arrêté exposent leurs auteurs aux poursuites et aux peines prévues à... », « Les infractions au présent arrêté [...] feront l'objet de poursuites conformément à... », « Ces infractions sont punies des peines prévues par les mêmes codes », « Les infractions au présent arrêté sont prévues et réprimées par... ». On constate donc que le même sujet (« les infractions ») peut être utilisé avec les verbes « être passible », « exposer ses auteurs », « être puni », « être prévu et réprimé ». La variabilité est surtout lexicale, mais on constate aussi, dans certains des cas ci-dessus, l'emploi d'une métonymie : on dit que « les infractions sont punies »

5.1.2. *L'expressivité*

Les critères qui permettent d'évaluer l'expressivité d'un langage contrôlé sont : la quantification universelle de premier ordre (sur les individus), l'arité des relations, l'existence de structures de règle (si ... alors), l'existence de la négation, la quantification universelle du deuxième ordre (sur les concepts et les relations). Clairement, l'éloquence de ses auteurs et la richesse de la langue française situent le langage juridique en E^5 (langages d'expressivité maximale).

5.1.3. *La naturalité*

Le langage juridique est un cas typique de langage N^4 (*languages with sentences that can be considered valid natural sentences. Speakers of the respective natural language recognize the statements as sentences of their language and are able to correctly understand their essence without instructions or training*[7]) à cela près que les francophones non spécialistes ne possèdent pas forcément le vocabulaire spécifique et ne peuvent donc accéder que partiellement à la sémantique du texte (il s'agit de l'« écran linguistique » dont parle Cornu (2005, p. 12)).

5.1.4. *La simplicité*

Le langage juridique est un langage S^1 (langage très complexe) puisqu'il a la complexité de la langue française *sans restriction syntaxique ou sémantique*[8].

Nous concluons donc que le langage juridique, considéré en tant que langage contrôlé, se situe dans la zone $P^{2\text{-}3}E^5N^4S^1$ de la classification de Kuhn (2014)[9].

5.2. *Le décalage d'articulation ontologique*

Le second défi que doivent relever les systèmes de RAQ et qui est prégnant dans le domaine juridique est la différence entre l'articulation des concepts et des rela-

---

alors que ce sont, en réalité, leurs auteurs qui le sont. Enfin on constate de la variabilité au niveau de l'utilisation du verbe « prévoir » : dans un cas ce sont les peines qui sont prévues dans les textes, dans un autre cas ce sont les infractions qui le sont. Cela montre les défauts de formalité d'une formule qui pourtant semble figée et est souvent répétée à l'identique, tel un *mantra*, pendant des décennies.

7. Traduction : langages avec des phrases qui peuvent être considérées comme des phrases valides de langage naturel. Les locuteurs du langage naturel correspondant reconnaissent les assertions en tant que phrases de leur langue et sont capables de comprendre leur essence correctement sans instructions ou formation préalable.

8. Citons Cornu (2005, p. 316) : « le langage juridique français ne s'oppose pas à la langue française ; il la met en œuvre ».

9. Notons qu'il existe un langage contrôlé dans le domaine législatif ayant à peu près la même classification Kuhn : $P^2E^5N^5S^1$, il s'agit du *Massachusetts Legislative Drafting Language*, qui est défini par une centaine de règles syntaxiques, sémantiques et structurelles à appliquer à la langue anglaise. Il a été introduit en 2003 par le sénat de Massachusetts (Massachusetts Senate, 2010).

tions dont on trouve des représentations lexicales dans le texte et celle des concepts et relations présents dans l'ontologie. Cette différence est un obstacle à l'alignement direct entre les concepts et relations extraits du texte et ceux prévus dans l'ontologie, nous l'appellerons *décalage d'articulation ontologique*. En guise d'illustration, prenons la règle « Tout pétrolier d'un tonnage supérieur à 150 000 tonnes doit être muni du Certificat de prévention de la pollution des eaux de mer par les hydrocarbures ». Cette exigence fait intervenir : (a) une entité représentée lexicalement par le terme `"pétrolier"`, ayant un attribut pour représenter la notion de `tonnage` et sur lequel on a imposé la condition `tonnage` > 150 ; (b) une entité pour capturer le sens du terme `"Certificat de prévention de la pollution des eaux de mer par les hydrocarbures"` ; (c) une *relation directe* entre les deux entités mentionnées ci-dessus (le premier doit être muni du deuxième).

Or, lorsque l'on considère l'ontologie maritime proposée par Lohrmann *et al.* (2014), décrite en section 3.1, le schéma conceptuel est différent : pour modéliser la règle à travers l'ontologie maritime d'e-Compliance, l'entité dénotant `"pétrolier"` est la cible et l'entité dénotant `"Certificat de prévention de la pollution des eaux de mer par les hydrocarbures"` est l'exigence ; les deux entités étant connectées à une instance de la classe `Rule`.

En utilisant la syntaxe Turtle[10], ces assertions se présentent ainsi[11] :

```
:Ship1 a :Ship;
    :shipType "pétrolier";
    :minTonnage "150".
:Certificate1 a :Certificate;
    :documentTitle "Certificat de prévention de la pollution...".
:Rule1 a :Rule,
    :hasTarget :Ship1;
    :hasRequirement :Certificate1.
```

On voit que `"pétrolier"` n'a pas le statut d'entité mais est une simple valeur d'attribut. De même, on est contraint de passer par une utilisation du concept de règle `Rule` à travers les propriétés `:hasTarget`, `:hasRequirement`, `:documentTitle`, `:shipType`, non présents lexicalement dans la règle textuelle de départ (le statut de règle étant implicite dans la phrase « Tout pétrolier... », un affirmatif que Cornu (2005, p. 274)appelle une *marque ostensible de généralité*).

Cet exemple permet de se rendre compte de fait qu'une approche de RAQ ayant affaire à une base de connaissances construite sur une ontologie légale doit être capable de résoudre le décalage d'articulation ontologique inhérent à celle-ci. Dans la section

10. `https://www.w3.org/TR/turtle/`.

11. L'objectif de notre travail n'est pas de proposer une modélisation des règles ou une alternative aux modélisations existantes. Ainsi, l'ontologie support de notre approche, à savoir l'ontologie e-Compliance, est utilisée telle que présentée par ses auteurs, à quelques ajouts de vocabulaire près.

suivante, nous posons les bases de notre approche de RAQ appliquée à une base de connaissances légales. À notre connaissance, c'est la première approche à s'intéresser à la RAQ dans le domaine des ontologies légales. En plus de s'attaquer aux spécificités du domaine réglementaire, notre approche dispose d'un mécanisme de résolution de décalages en faisant appel à des patrons de questions ce qui permet entre autres la formalisation des questions complexes.

6. Notre approche de RAQ

Considérant le corpus de questions en tant que langage formel, il est tout à fait naturel de procéder à une analyse syntaxique des questions textuelles pour obtenir l'arbre de dérivation correspondant à chaque question, et ensuite traduire cet arbre en requête SPARQL. Or, cette approche est inefficace puisqu'elle mettrait au même niveau les représentations lexicales d'objets de l'ontologie, celles décrivant les propriétés de ces objets, et *last but not least* les mots grammaticaux qui sont indépendants du domaine et ne dépendent que de la langue. Elle demanderait donc l'élaboration d'une grammaire monolithique, où la moindre omission d'un détail entraînerait l'échec de l'analyse.

Au lieu de cela nous procédons d'abord à un *chunking* (une analyse syntaxique superficielle) en identifiant en priorité dans les *chunks* les représentations lexicales d'objets de l'ontologie. Les *chunks* contenant de telles représentations sont alors associés à des sommets intermédiaires de l'arbre de dérivation, et leur ensemble peut donc, le cas échéant, correspondre à un ou plusieurs patrons de questions (§ 6.2). Ensuite on traite un autre type de *chunk*, ceux qui représentent des *descriptions d'entités* (§ 6.3). Un troisième type de *chunk* correspond aux mots grammaticaux qui régissent les questions. Ces trois types de *chunk* participent chacun à sa manière à la traduction de la question en requête SPARQL. Les étapes de notre approche sont présentées sur la figure 4.

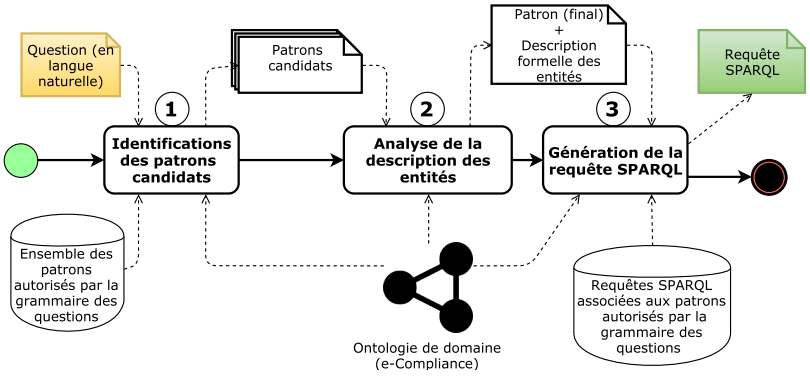


Figure 4. *Étapes de notre approche de RAQ*

6.1. *Description des différents types de questions*

En nous inspirant de travaux de RAQ s'appuyant sur les patrons des questions (paragraphe 4, point 7), nous avons étudié la structure d'un certain nombre de questions de notre domaine d'application. Ces questions, présentées en annexe B, proviennent de la section « Perspectives » du projet e-Compliance[12] et de séances de travail avec les parties prenantes du projet REIZHMOR. En étudiant le corpus de questions, nous leur avons attribué une catégorie en fonction de l'objectif. Ces catégories sont celles de questions : (1) permettant de connaître la valeur d'une propriété ; (2) servant à obtenir la liste des entités ayant une certaine (valeur de) propriété ; (3) permettant de retrouver les réglementations s'appliquant à certaines situations (décrites dans la question) ; (4) permettant de retrouver les textes ou articles qui prescrivent un comportement donné ; (5) servant à identifier les conditions ou les contextes dans lesquels on peut ou non effectuer des actions données ; (6) permettant d'exhiber les entités cibles d'une réglementation ; (7) relatives aux procédures et aux protocoles ; (8) concernant la structure ou les métadonnées des textes réglementaires ; (9) concernant l'application d'un texte ou d'un article.

Environ 63 % des questions du corpus appartiennent aux catégories (1) à (6) ; les 37 % restant se trouvent dans les catégories (7) à (9). Les patrons de questions que nous proposons correspondent aux questions des catégories (1) à (7). Leur extension aux patrons des catégories de questions restantes est un travail en cours. Comme annoncé, les patrons de questions que nous proposons s'appuient sur un langage formel dont nous donnons en annexe A la grammaire régulière dans le formalisme EBNF (*Extended Backus-Naur Form*). Notons que :

– une question est la séquence (cf. ligne 1 du listing) d'un élément de restriction optionnel `Restriction`, d'un groupe de mots codant le type de la question `QuestionType`, du corps de la question `QuestionBody`, et d'un signe d'interrogation `QuestionMark` signifiant sa fin ;

– `Restriction` est un élément qui permet de restreindre l'espace de recherche d'une question à un ensemble de textes ; cela explique le fait qu'il ait la même définition que la référence à un texte réglementaire `ReferenceToReg` (cf. l. 2 du listing). Il fait référence à des expressions telles que : « Dans l'article 2.3 », ou « Dans l'annexe 1 de la Convention internationale pour la prévention de la pollution par les navires », etc. ;

– le corps de la question `QuestionBody`, est formé par l'un des six types de questions que nous avons mentionnés auparavant : `ValueOfProp`, `EntitiesHaving-Prop`, `RulesApplyingToCases`, `RulesPrescribingRequirements`, `Condition-ForActivities` et `TargetedEntities` (cf. l. 9 et 10) ;

– `ValueOfProp` symbolise les questions relatives à la valeur d'une propriété (catégorie (1)). Par exemple « Quelle est la date de publication de la Convention internationale de 1973 ? » ;

12. `http://www.e-Compliance-project.eu/`

– `EntitesHavingProp` est le symbole des questions sur les entités ayant une propriété dont la valeur remplit certaines conditions (catégorie (2)). Par exemple « Quels documents ont été publiés avant 2011 ? » ;

– `RulesApplyingToCases` représente les questions sur les règles s'appliquant dans des cas donnés (catégorie (3)). « Quelles exigences s'appliquent aux engins nautiques ? » en est un exemple ;

– `RulesPrescribingRequirements` permet de détecter les questions sur les règles prescrivant un certain comportement (catégorie (4)). Par exemple, « Quelles exigences interdisent la pêche dans les ports du Morbihan ? » ;

– `ConditionForActivities` symbolise les questions de la catégorie (5) c'est-à-dire relatives aux conditions dans lesquelles peuvent (ou pas) s'exercer certaines activités. Comme exemple de question on a « Dans quels ports de la Manche les pétroliers peuvent-ils mouiller ? » ;

– le symbole `TargetedEntities` est celui de la catégorie des questions qui s'intéressent aux entités *ciblées* par un texte légal donné. Par exemple, nous avons : « Quels types de navires sont concernés par l'arrêté 25/67 du 5 juillet 1967 ? » ;

– dans le corps d'une question, à travers les différents symboles codant les catégories de questions que nous avons relevées, on peut retrouver la description des différentes entités présentes dans la question. Dans le cadre de l'ontologie maritime légale e-Compliance, ces entités peuvent être soit un rôle, soit une organisation, soit un navire. Les descriptions de ces trois entités sont symbolisées respectivement par `RoleDesc`, `OrganisationDesc` et `ShipDesc` ;

– la description d'une cible (*target*), par exemple `ShipDesc`, se compose d'une mention de la cible (*chunk* de premier type) et éventuellement de la description proprement dite (*chunk* de type 2). Cette description constitue un sous-langage formel décrit par l'automate de la figure 5. En guise d'illustration, on peut reconnaître des expressions telles que : « navire marchand », « navire marchand ayant un tonnage supérieur à 400 », « navire sans moteur », etc. ;

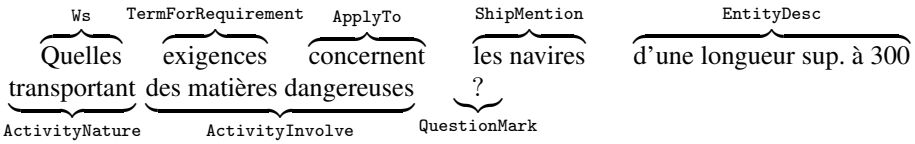
– nous allons voir comment à chaque patron de question, on peut faire correspondre un patron de requêtes SPARQL. Ainsi, le patron `Ws BePredicate TermForRequirement ApplyTo ShipType QuestionMark` a pour patron de requêtes SPARQL :

```
SELECT DISTINCT ?title ?content
WHERE {
        ?rule a :Rule; :isDerivedFromClause ?clause;
                       :hasTarget ?ship.
        ?clause a :Clause; :title ?title; rdfs:label ?content.
        ?ship a :Ship; :type "<ShipType>".}
```

Lors de l'exécution de cette requête, le *slot* `<ShipType>` sera remplacé par l'occurrence de `Shiptype` dans la question en langage naturel.

6.2. *Identification des patrons candidats*

C'est la première étape de notre approche telle que présentée dans la figure 4. Nous effectuons un *chunking* (analyse syntaxique superficielle) de la phrase où chaque *chunk* correspond à la représentation lexicale d'un certain sommet intermédiaire de l'arbre de dérivation de la phrase. Ce *chunking* détecte en priorité les représentations lexicales d'objets (classes, individus, relations) de l'ontologie de domaine (dans notre cas d'application, l'ontologie e-Compliance étendue par nos soins), ainsi que les mots grammaticaux spécifiques à la réalisation syntaxique des questions. Ainsi, la question « Quelles exigences concernent les navires d'une longueur supérieure à 300 transportant des matières dangereuses ? » est découpée de la manière suivante :



où `Ws` (l'adjectif interrogatif « quelles ») est un *chunk* de troisième type, `EntityDesc` est (comme son nom l'indique) un *chunk* de deuxième type, et tous les autres sont des *chunks* de premier type puisqu'ils correspondent à des objets de l'ontologie de domaine.

Au terme de cette étape, nous disposons d'un patron de question. Ce dernier peut contenir des symboles non terminaux représentant des descriptions d'entités qu'il est nécessaire de décoder. C'est le but de la deuxième étape. Dans l'exemple ci-dessus, on a une description d'entités (sommet `EntityDesc`) qui est « longueur supérieure à 300 ». Il faudra donc analyser cette expression.

6.3. *Analyse de la description des entités*

L'étape d'analyse de la description des entités permet de formaliser les fragments de la question qui ne correspondent ni à des concepts, individus ou relations de l'ontologie de domaine, ni à des mots grammaticaux généralement présents dans les questions relatives aux corpus réglementaires (elle correspond à l'étape 2 de la figure 4). Nous considérons, de manière heuristique, que ces fragments servent à apporter un complément d'information aux concepts, individus et relations qui les encadrent. Pour ce faire, nous nous servons d'une version augmentée de l'automate du système CANaLI de (Mazzeo et Zaniolo, 2016) (cf. figure 5). Comme cet automate nécessite la présence d'une entité, nous l'appliquons non pas à `EntityDesc` seul, mais à la paire `OpenEnt EntityDesc` où `OpenEnt` est l'entité qui précède[13] `EntityDesc` (dans notre cas, `ShipMention`, lexicalisée par « les navires »). Nous avons ajouté des transitions à l'automate permettant un spectre plus large de descriptions. Par exemple, des descriptions telles que « Navire de 5 (mètres) de long » ou « Navire long de plus de 5 (mètres) » ne sont pas reconnues par l'automate original alors que la version augmentée que nous proposons les reconnaît.

13. On peut, de manière similaire, étendre l'automate au cas où l'entité suit sa description.

En outre, ce processus doit tenir compte des variantes lexicales possibles des termes de `EntityDesc`, appelées *lexicalisations* dans le cadre du projet DBpedia (Mendes *et al.*, 2012). Ainsi, le mot « long » peut être associé à des propriétés de l'ontologie de domaine impliquant une longueur, et dans l'ontologie e-Compliance, l'adjectif « long » peut être associé à `:maxLength` et `:minLength` qui sont des propriétés de la classe `:Ship`. L'analyse de la description d'entités se décompose alors en deux sous-étapes : lexicalisation et reconnaissance par l'automate.

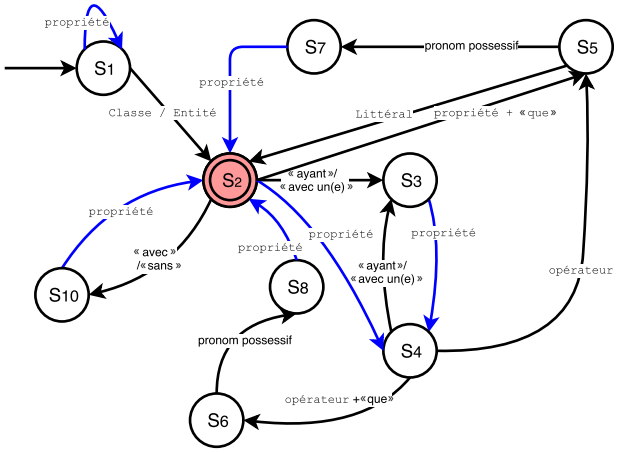


Figure 5. *Automate pour la reconnaissance d'une expression décrivant une entité (adapté de Mazzeo et Zaniolo (2016))*

6.3.1. *Lexicalisation*

Nous procédons à un *chunking* du contenu textuel de la description d'entités. Dans chacun des *chunks*, on procède à la recherche des variantes lexicales possibles. Pour cela nous nous servons les formes lexicalisées des classes[14] et des propriétés[15] de DBpedia (Unger *et al.*, 2013). Nous disposons ainsi de la version en langage naturel et des éventuelles variantes de chaque *chunk*. Nous enrichissons cet ensemble, fondé sur DBpedia, en lui ajoutant les représentations lexicales des propriétés de notre ontologie de domaine, obtenues par proximité au niveau des chaînes de caractères, par la distance de Levenshtein.

Nous illustrons cette étape avec la description « d'une longueur supérieure à 300 » qu'il nous reste à décoder après l'étape d'identification des patrons candidats (voir section 6.2). Une analyse syntaxique superficielle de ce fragment nous donne les deux

14. `https://github.com/dice-group/hawk/blob/master/resources/dbpedia_3Eng_class.ttl`
15. `https://github.com/dice-group/hawk/blob/master/resources/dbpedia_3Eng_property.ttl`

chunks « d'une longueur » et « supérieure à 300 ». Les traitements effectués à l'aide de ces deux *chunks* sont regroupés dans le tableau 1.

	« d'une longueur »	« supérieure à 300 »
Ajout par lexicalisation	*length*	-
Ajout par alignement avec l'ontologie	`:maxLength`, `:minLength`	-
Interprétations possibles de « d'une longueur supérieure à 300 »	(1) « d'une longueur supérieure à 300 » (2) « *length* supérieure à 300 » (3) « `:maxLength` supérieure à 300 » (4) « `:minLength` supérieure à 300 »	

Tableau 1. *Obtention des variations lexicales possibles du fragment « d'une longueur supérieure à 300 »*

6.3.2. *Analyse de la description d'entités*

Au cours de cette sous-étape, nous filtrons les variantes lexicales possibles d'une description d'entités en ne gardant que celles qui sont reconnues par l'automate. Notons que nous gardons les états et les transitions des variantes lexicales qui sont reconnues par l'automate puisqu'elles nous serviront à la formalisation proprement dite de la description. Nous synthétisons les résultats de ce processus pour notre exemple, dans le tableau 2. Dans ce tableau, on a d'un côté les interprétations candidates et de l'autre les états et transitions obtenus lors des tentatives de reconnaissance de ces phrases par l'automate. Ajouter à la description proprement dite l'entité qui précède nous permet de confirmer que la description correspond effectivement à l'entité. En effet, dire qu'une description est celle d'une entité impose de devoir faire *une validation sémantique en plus d'une validation syntaxique*. Dans ce cas, la validation syntaxique repose sur les états et transitions de l'automate et la validation sémantique s'effectue en s'assurant que :

Interprétations	*États et transitions*
(1) les navires + d'une longueur supérieure à 300	$S_1 \rightarrow S_2 \rightarrow$ X
(2) les navires + *length* supérieure à 300	$S_1 \rightarrow S_2 \rightarrow$ X
(3) les navires + `:maxLength` supérieure à 300	$S_1 \rightarrow S_2 \rightarrow S_4 \rightarrow S_5 \rightarrow S_2$
(4) les navires + `:minLength` supérieure à 300	$S_1 \rightarrow S_2 \rightarrow S_4 \rightarrow S_5 \rightarrow S_2$

Tableau 2. *États et transitions pour le fragment « d'une longueur supérieure à 300 ». Le symbole* X *signifie un échec de la reconnaissance de la phrase par l'automate*

– lorsqu'on rattache une propriété à une classe alors cette dernière est incluse dans le *domaine*, au sens de la propriété `rdfs:domain`[16], de cette propriété ;

– lorsqu'un comparateur (*e.g.* « inférieur à », « plus grand que ») suit une propriété, alors cette propriété est de type comparable (*e.g.* entier, flottant, date, etc.). De même,

16. Le préfixe `rdfs` fait référence à `https://www.w3.org/2000/01/rdf-schema`

lorsqu'une (valeur de) propriété est comparée à un littéral, on vérifie que ce dernier est de type compatible à celui de la propriété. On évitera ainsi, de valider la comparaison d'un flottant et d'une date.

Ainsi pour les lignes (1) à (4) du tableau 2, on cherche à valider syntaxiquement et sémantiquement l'analyse de la description « d'une longueur supérieure à 300 » rattachée à l'entité (`OpeningEnt`) « les navires ». Rappelons que cette entité a été décodée en tant que référence à la classe `ecom:Ship` de l'ontologie de domaine. Des quatre interprétations candidates, seules la (3) et la (4) sont reconnues avec :

– entre les états S_1 et S_2 une transition avec l'entrée « les navires » qui fait référence à la classe `:Ship` ;

– entre S_2 et S_3 une transition avec l'entrée `:maxLength` (resp. `:minLength`) pour l'interprétation (3) (resp. (4)). Cette transition est aussi validée sémantiquement, car la classe `:Ship` en est le domaine des propriétés `:maxLength` et `:minLength` ;

– entre S_4 et S_5 il y a le comparateur « supérieur à » qui assure la transition. De plus il est sémantiquement compatible avec le domaine d'arrivée de `:maxLength` et `:minLength` ;

– entre S_5 et S_2 on a le littéral « 300 » qui, en outre, est en adéquation avec le domaine d'arrivée de `:maxLength` et `:minLength`.

Notons que dans cet exemple les interprétations (3) et (4) sont toutes deux reconnues par l'automate. Nous choisissons l'interprétation finale de manière aléatoire parmi ces deux interprétations.

Au terme de cette étape, nous disposons :

1) du patron de la question obtenu dès la première étape de notre approche (section 6.2) ;

2) des interprétations des descriptions d'entités du patron et des états et transitions ayant permis de valider chacune de ces interprétations.

Nous pouvons désormais générer la requête qui nous permettra d'interroger la base de connaissances réglementaires et donc de répondre à la question posée.

6.4. *Génération de la requête SPARQL*

Notre base de connaissances est composée de règles qui constituent des représentations formelles des règles textuelles provenant des arrêtés préfectoraux. Toute règle de la base de connaissances est une instance de la classe `:Rule` de l'ontologie e-Compliance. Rappelons que e-Compliance nous sert à la fois d'ontologie de domaine et d'ontologie légale. En tant qu'ontologie de domaine, et donc ontologie maritime, elle dispose de structures pour représenter le domaine maritime (les navires, les activités, les rôles et organisations, les juridictions, etc.). En tant qu'ontologie légale elle propose une modélisation ontologique des *règles* (une règle ayant un contexte, une cible et une exigence) (Lohrmann *et al.*, 2014).

Le formalisme retenu pour e-Compliance étant le langage OWL[17] (*Web Ontology Language*), le moyen adéquat pour l'interroger est de se servir du langage de requêtes standard SPARQL[18] (*SPARQL protocol and Query Language*). Ce dernier permet d'interroger des données RDF, RDFS, OWL. Il propose quatre types de requêtes :

– `SELECT` pour sélectionner tous les éléments répondant à des critères donnés ;

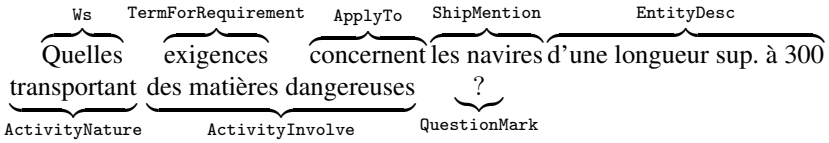
– `ASK` pour affirmer ou infirmer l'existence d'éléments répondant à des critères donnés ;

– `DESCRIBE` pour obtenir une description d'un certain nombre de ressources ;

– `CONSTRUCT` pour construire de nouvelles ressources à partir d'un ensemble d'entités répondant à des critères donnés.

Pour notre approche de RAQ, seules les requêtes SPARQL de type `SELECT` et `ASK` sont pertinentes. En effet, soit il sera question d'identifier et de retourner les éléments de réponse à une question, soit il faudra répondre si oui ou non, le cas décrit dans une question est prévu dans la base de connaissances. De même, dans la grammaire EBNF des questions (annexe A p. 69), à la ligne 3, nous prévoyons deux types de questions, symbolisés par `Select` et `Ask`.

Rappelons qu'à ce stade nous disposons du patron de la question et d'éléments de formalisation des descriptions éventuelles d'entités présentes dans le patron. Aussi rappelons qu'à chaque patron de question il y a une requête SPARQL associée. L'annexe C donne des exemples de patrons ainsi que les requêtes associées. Cet état de fait est illustré par l'étape 3 de la figure 4. Ainsi pour notre exemple, il nous faut instancier la requête SPARQL correspondant au patron de la requête. Le patron de notre question est :



La requête SPARQL associée à ce patron est :

```
1  SELECT DISTINCT ?title
2  WHERE {
3          ?rule rdf:type :Rule; :isDerivedFromClause ?clause;
4                        :hasRequirement ?req;
5                        :hasTarget ?ship. ?ship rdf:type :Ship.
6          ?clause rdf:type :Clause.
7          ?clause :isPartOf ?reg.
8          ?reg rdf:type :Regulation; :title ?title.
9          ?req :nature "transport"^^xsd:string.
10         ?req :involves ?cargo.
```

17. https ://www.w3.org/OWL/

18. https ://www.w3.org/TR/rdf-sparql-query/

```
11          ?cargo rdf:type :Cargo; :isDangerous "true"^^xsd:boolean.
12          [?ship :maxLength ?length. FILTER(?length > 300)]}
```

Cette requête a été obtenue en instanciant le patron correspondant (voir annexe C) avec :

– la classe `:Ship` comme cible (propriété `:hasTarget`) des règles pertinentes pour notre question (voir ligne 5 de la requête) ;

– la nature des activités, `"transport"` (voir ligne 9), ainsi que la nature des éléments impliqués dans ce transport à savoir une cargaison de produit dangereux (voir ligne 11) ;

– la description de la cible visée par les règles pertinentes pour la question. Cette description est donnée par la ligne 12 de la requête.

Comme nous l'avons mentionné dans la section 5.2, à cause du décalage d'articulation ontologique important entre les ontologies légales et les règles en langue naturelle, les requêtes SPARQL peuvent être relativement complexes. Le fait de disposer pour chaque patron de question du patron de la requête SPARQL correspondante permet d'appréhender ce phénomène.

7. Évaluation

Nous avons évalué la capacité de formalisation de notre approche sur un corpus constitué des questions mentionnées en annexe B. Ce corpus est constitué de trente questions en langage naturel, ciblant les textes du corpus réglementaire décrit à la section 2. Sur ces trente questions :

1) quatorze questions, soit 46,67 %, sont correctement formalisées. Les patrons correspondant à ces questions sont convenablement identifiés et instanciés ;

2) les patrons de cinq questions, soit 16,67 %, ne sont pas identifiés. Bien que ces questions appartiennent aux six catégories de questions que nous avons identifiées et pour lesquelles des patrons sont proposés, notre langage formel ne couvre pas la syntaxe des questions ;

3) onze questions, soit 36,67 %, ne sont pas identifiées, (i) soit parce que le type de la question n'est pas pris en compte dans la syntaxe (exemple : la question 27 sur le nombre d'articles de l'arrêté 96/2015), (ii) soit parce que l'ontologie maritime e-Compliance, qui sert de support à notre base de connaissances, ne formalise pas les informations nécessaires pour répondre à la question (exemple : la question 30 sur les informations à fournir en cas de découverte d'engin suspect).

Cette première évaluation montre qu'une extension de la syntaxe ainsi que de l'ontologie – à la fois légale et maritime – e-Compliance sont à considérer dans nos travaux futurs, que nous présentons dans la section suivante. Notons toutefois que la métrique de notre évaluation doit aller au-delà du regard de la qualité de la résolution du patron d'une question. En effet, *un patron peut être bien identifié mais incorrectement instancié*. Cela arrive principalement avec la formalisation des descriptions des entités. On peut le voir dans l'exemple que nous avons déroulé pour illustrer notre approche.

Dans cet exemple, nous avons formalisé l'expression « d'une longueur supérieure à 300 » (voir tableaux 1 et 2). À l'issue de la formalisation de cette expression le terme « longueur » est formalisé par la propriété `:maxLength`. Or d'après les experts, le prédicat correct dans ce cas est `:minLength`. Disposer des validations des experts ainsi que des réponses précises aux différentes questions du corpus est un travail à faire pour améliorer notre méthode d'évaluation.

8. Travaux futurs

L'approche de RAQ que nous avons présentée fait apparaître les points d'amélioration que voici :

– *l'extension de la portée des types de questions et de leur syntaxe.* Parmi les questions qui sont à la portée de notre approche, on retrouve essentiellement celles qui sont liées au contenu des réglementations : ce qu'elles prescrivent, dans quel contexte, etc. Ce sont, par exemple, les questions relatives à la structure des textes réglementaires, ou encore liées aux relations entre les textes ou aux procédures et informations mentionnées dedans ;

– *l'extension de l'ontologie maritime e-Compliance.* Nous avons relevé le fait que l'ontologie e-Compliance, dans sa forme actuelle, n'est pas capable de représenter certaines informations contenues dans les réglementations, diminuant de ce fait la performance des approches de RAQ. Une extension de cette ontologie est nécessaire pour prendre en compte la représentation de la structure des textes réglementaires ainsi que les relations hiérarchiques entre les textes, et pour représenter plus en détail le contenu des textes (par exemple les procédures à suivre). En outre il faut augmenter la représentation des règles de cette ontologie de manière à avoir une règle au sens classique du terme, *i.e.* `antécédent` $\Longrightarrow$ `conséquent`. Cela permettrait de disposer d'un dépôt de règles formelles et d'effectuer un contrôle automatique de conformité (Yurchyshyna et Zarli, 2009 ; Kacfah Emani, 2016) ;

– *proposition d'une approche plus générale.* En nous appuyant sur l'approche de RAQ présentée, nous envisageons de proposer une méthodologie générale de RAQ vis-à-vis d'une base de connaissances légales, le cas de la réglementation maritime ne constituant qu'un cas d'application. Pour démontrer la pertinence de cette méthodologie, il faudra l'appliquer à d'autres domaines d'application ;

– *mise sur pied d'un corpus d'évaluation.* Il est nécessaire de disposer de bancs d'essai pour évaluer les approches de RAQ sur des bases de connaissances légales. De tels bancs d'essai doivent comprendre : (i) plusieurs ontologies légales multilingues – une fois peuplées, ces ontologies serviraient de bases de connaissances cibles pour les questions en langage naturel ; (ii) un nombre important de questions en langage naturel avec les réponses attendues, ainsi que les requêtes SPARQL correspondantes, validées par des experts du domaine. Ces questions doivent couvrir la diversité des types de requêtes que les utilisateurs usuels des bases de connaissances légales se posent.

9. Conclusion

Nous avons présenté les premières briques du projet REIZHMOR dont le champ d'application est celui de la réglementation maritime. Nous avons décrit le premier corpus de ce projet qui est composé d'arrêtés préfectoraux et interpréfectoraux ; au cours de l'évolution du projet, ce corpus devra être étendu à des textes de niveau national et international afin de proposer des preuves de concept robustes. Nous avons avancé une approche de réponses automatiques aux questions (RAQ) en langage naturel dans le domaine légal appliqué à la réglementation maritime. En plus d'aborder les problèmes généraux de RAQ tels que la variété lexicale, elle propose des solutions pour des problèmes spécifiques au domaine légal, telles que la résolution du décalage d'articulation ontologique à l'aide de patrons de questions. Une première évaluation de notre approche montre des résultats prometteurs avec plus de 46 % des questions correctement formalisées. Pour améliorer ces résultats, nous prévoyons d'étendre le niveau d'informations formalisées par les ontologies maritimes existantes et aussi la portée des patrons syntaxiques des questions. En outre, nous visons à proposer une méthodologie générale de RAQ pour les bases de connaissances légales.

Remerciements

Ce travail est financé par le Service hydrographique et océanographique de la marine (Shom) dans le cadre du projet REIZHMOR.

10. Bibliographie

Berners-Lee T., Hendler J., Lassila O. *et al.*, « The semantic Web », *Scientific American*, vol. 284, n° 5, p. 28-37, 2001.

Cornu G., *Linguistique juridique*, Montchrestien, Paris, 2005.

Daille B., « Conceptual structuring through term variations », *in* F. Bond, A. Korhonen, D. MacCarthy, A. Villacicencio (eds), *Proceedings ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, p. 9-16, 2003.

de Cet Bertin C., *Introduction au droit maritime*, ellipses, Paris, 2008.

Haralambous Y., Sauvage-Vincent J., Puentes J., « A Hybrid (Visual/Natural) Controlled Language », *Languages Resources and Evaluation*, vol. 51, n° 1, p. 93-129, 2017.

Hirschman L., Gaizauskas R., « Natural language question answering : the view from here », *Natural Language Engineering*, vol. 7, n° 4, p. 275-300, 2001.

Höffner K., Walter S., Marx E., Usbeck R., Lehmann J., Ngonga Ngomo A.-C., « Survey on challenges of Question Answering in the semantic Web », *Semantic Web*, vol. 9, p. 1-26, 2016.

Kacfah Emani C., Formalisation automatique et sémantique de règles métiers, thèse de doctorat, Université de Lyon, 2016.

Kuhn T., « A survey and classification of controlled natural languages », *Computational Linguistics*, vol. 40, p. 121-170, 2014.

Lohrmann P., Seizou M., Hagaseth M., Griffiths D., A European Maritime e-Compliance Cooperation Model - Ontology, Technical Report n° 2.2, Seventh Framework Program, 2014.

Lopez V., Uren V., Sabou M., Motta E., « Is question answering fit for the semantic Web ? : A survey », *Semantic Web*, vol. 2, n° 2, p. 125-155, 2011.

Massachusetts Senate, *Legislative Drafting and Legal Manual*, 2010. `https://malegislature.gov/Content/Documents/General/LegislativeDraftingManual.pdf`.

Mazzeo G. M., Zaniolo C., CANaLI : A System for Answering Controlled Natural Language Questions on RDF Knowledge Bases, Technical Report n° 160004, 2016. `http://fmdb.cs.ucla.edu/Treports/canali_tr_160004.pdf`.

Mendes P. N., Jakob M., Bizer C., « DBpedia - A Multilingual Cross-domain Knowledge Base », *Proceedings of LREC 2012, Eighth International Conference on Language Resources and Evaluation*, p. 183-1817, 2012.

Nivre J., Hall J., Nilsson J., « MaltParser : A Data-Driven Parser-Generator for Dependency Parsing », *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*, p. 2216-2219, 2006.

Sauvage-Vincent J., Un langage contrôlé pour les *Instructions nautiques* du Service Hydrographique et Océanographique de la Marine, thèse de doctorat, IMT Atlantique, 2017.

Shom, *France (côtes Nord et Ouest). De la frontière belge à la pointe de Penmarc'h*, vol. C2A of *Instructions nautiques*, 2010. Édition à jour le 21 juin 2017.

Simperl E., Tempich C., Vrandečić D., « A methodology for ontology learning », *in* P. Buitelaar, P. Cimiano (eds), *Ontology learning and population : bridging the gap between text and knowledge*, IOS Press, p. 225-249, 2008.

Unger C., McCrae J., Walter S., Winter S., Cimiano P., « A lemon lexicon for DBpedia », *Proceedings of the 2013 International Conference on NLP & DBpedia-Volume 1064*, CEUR-WS.org, p. 103-108, 2013.

Yurchyshyna A., Zarli A., « An ontology-based approach for formalisation and semantic organisation of conformance requirements in construction », *Automation in Construction*, vol. 18, n° 8, p. 1084-1098, 2009.

Annexes

A. Extrait de la syntaxe EBNF des questions adressées à une base de connaissances réglementaires (appliquée au domaine maritime)

Les numéros ① à ⑥ indiquent les symboles représentant les catégories de questions (1) à (6) présentées à la section 6.1.

```
Question = Restriction?, QuestionType, QuestionBody, QuestionMark;
Restriction = ReferenceToReg;
QuestionType = Select | Ask;
Select = Ws, BePredicate;
Ws = ("Qui"|"Quel"|"Dans quel");
BePredicate = ('est' | 'sont');
Ask = ('Est-ce que');
QuestionMark = '?';
QuestionBody = ValueOfProp | EnitiesHavingProp | RegApplyingToCases
| RulesPrescribingRequirements | ConditionForActivities | TargetedEntities;
```

```
① ValueOfProp = (Property, Of)?,  (Class | Individual);
② EntitiesHavingProp = (Class|Entity), Property, Comparator, Value;
③ RegApplyingToCases = ReferenceToReg, ApplyTo, Case;
Case = TargetCase | ActivityCase;
TargetCase = (TargetDesc, (Perform, Activity)?) | (Actity, Of, TargetDesc);
TargetDesc = TargetMention, Description?;
ActitityCase = Activity;
TargetMention = ShipMention | RoleMention | OrganisationMention;
Activity = ActivityNature, Invole?, Place?, Situation?;
Place = Jurisdiction?;
Situation = MaritimeSituation?;
④ RegPrescribingRequirements = RegHavingModality | PrescriptionOnEntity;
RegHavingModality = ReferenceToReg, BePredicate, Modality;
PrescriptionOnEntity = ReferenceToReg?, ApplyTo?, ModalityOnEntity, Requirement?;
ModaliyOnEntity = (Modality, EntityDesc)| (EntityDesc, Modality);
EntityDesc = TargetDesc | Activity;
Requirement = Activity | Doc | Equip | Role;
⑤ TargetedEntities = TargetAndType, PredicateForTarget, In, ReferenceToReg;
⑥ ConditionForActivities = (Condition, ModalityOnActivity)
| (Condition, ModalityOnEntity, TargetCase) | (Condition, Activity, ModalityonEntity);
ModalityOnActivity = Modality, Activity | Activity, Modality;
Condition = (Jurisdiction | MaritimeSituation);
ReferenceToReg = TypedRegulation | TermForRequirement;
TypedRegulation = "reglementation", (In, DocName)? | Qualifier, "regulation";
DocName = ... Name of a Legal Text ... ;
TermForRequirement = "exigence", (In, DocName)?
TargetDesc = ShipDesc | RoleDesc | OrganisationDesc;
ShipDesc = ShipMention, DescriptionOfShipItself?;
ShipMention = ShipType | SynonymOfShip;
```

B. Questions du corpus de test

① à ⑥ : catégories de questions dont les patrons sont correctement identifiés, ⑦ à ⑫ : catégories de questions dont les patrons sont correctement identifiables après extension de syntaxe, ⑬ à ㉑ : autres cas.

	Questions	**Catégories**
1	Quelle est la définition de vraquier ?	①
2	Quel est le terme officiel pour désigner un navire utilisé pour la pêche ?	⑦
3	Quelles exigences s'appliquent aux caboteurs de plus de 60 mètres de long ?	⑨
4	Quels textes se rapportent aux obligations d'un capitaine de port ?	④
5	Quelles sont les règles relatives à la maintenance des navires à grue ?	③
6	Quels arrêtés préfectoraux sont relatifs à la sécurité au port de Brest ?	③
7	Le mouillage est-il autorisé dans le goulet de Brest ?	④
8	Quels sont les textes qui réglementent l'accès au port de Port-en-Bessin ?	④
9	La navigation est-elle autorisée au large de la digue du Break pour les navires de pêche ?	④
10	Quelles autorités sont chargées de l'application de l'arrêté n^{o} 22/91 ?	㉑
11	La maintenance des navires à cargaison sèche est-elle autorisée au port de Brest ?	④

12	Peut-on pratiquer la plongée sous-marine à 200 mètres du sablier TIMAC ?	⑯
13	Quelles sont les poursuites prévues par l'arrêté n^{o} 143/92 ?	⑲
14	Dans quels ports de France un navire de charge à pont ouvert peut-il accoster ?	⑫
15	Sous quelles conditions météorologiques un transbordement en mer du Nord est-il autorisé ?	⑥
16	Quels sont les visas de l'arrêté n^{o} 61/94 ?	⑳
17	Quelles activités sont interdites dans la pointe de la Torche ?	④
18	Quelles sont les conditions pour une demande de dérogation à l'interdiction de stationnement au port de Saint-Malo ?	⑲
19	Quels articles s'appliquent aux navires transportant des hydrocarbures ou des substances dangereuses ?	⑨
20	Quels navires sont concernés par l'arrêté interpréfectoral n^{o} 2002/99 Brest 2002/58 Cherbourg ?	⑤
21	Quelles sont les zones concernées par l'arrêté n^{o} 2009/55 ?	⑤
22	L'arrêté n^{o} 22/2009 est-il encore en vigueur ?	⑳
23	En cas de naufrage, peut-on mouiller dans un centre nucléaire ?	⑥
24	Quelle est la procédure d'autorisation pour pratiquer des activités sportives au large du centre nucléaire de la Penly ?	⑲
25	Quelles sont les limites du port civil de Cherbourg ?	⑬
26	Quelle est la vitesse maximale autorisée dans la petite rade de Cherbourg ?	⑬
27	Combien d'articles compte l'arrêté 96/2015 ?	⑳
28	Est-il interdit de pratiquer des activités nautiques dans l'anse du Poulmic ?	⑥
29	Qu'est-ce qu'un engin suspect ?	①
30	Quelles informations communiquer en cas de découverte d'un engin suspect ?	⑲

Rappelons qu'à la section 6.1, nous avons classé les questions en neuf catégories. Pour les questions des catégories 1 à 6 nous avons proposé des patrons (voir les symboles numérotés ① à ⑥ en annexe A). Nous mentionnons ci-dessus que les étiquettes ⑬ à ㉑ font référence aux « autres cas ». Parmi ces cas :

– les étiquettes ⑬ à ⑱ font référence aux questions des catégories 1 à 6. Cependant, les patrons proposés ne prennent pas en compte la syntaxe de ces questions et dans le même temps, le schéma de l'ontologie de domaine support, e-Compliance, ne permet pas de représenter les éléments de réponse pour cette question. Pour illustration, considérons la question 12 « Peut-on pratiquer la plongée sous-marine à 200 mètres du sablier TIMAC ? ». Pour répondre à cette dernière, il faudrait que la juridiction de déroulement d'une activité permette d'effectuer les calculs de distance adéquats. Ceci n'est pas le cas dans la version actuelle d'e-Compliance. Comme autre exemple, on peut prendre la question 26 « Quelle est la vitesse maximale autorisée dans la petite rade de Cherbourg ? ». Pour pouvoir y apporter des éléments de réponse il faut associer à chaque juridiction un certain nombre de paramètres, dont la vitesse. De plus la valeur de ces paramètres peut être contextuelle (par exemple : vitesse en cas d'intempérie, vitesse en fonction de la saison, de l'affluence, etc.) ;

– les étiquettes ⑲ à ㉑ font référence aux questions des catégories 7 à 9. Nous n'avons pas proposé de patron pour les questions de cette catégorie, elles sont donc hors de la portée de notre approche dans sa version actuelle.

Dans les lignes suivantes, nous commentons quelques-unes de ces questions :

Question 1 : Quelle (Ws) est (Be) la définition (Property) de (Of) vraquier (Entity) ? (?)

C'est une question de catégorie 1, c'est-à-dire permettant de connaître la valeur d'une propriété d'une entité.

Question 2 : Quel (Ws) est (Be) le terme officiel (Property) pour désigner (Of) un navire utilisé pour la pêche (Entity) ? (?)

C'est aussi une question de la catégorie 1, mais elle n'est pas bien identifiée par notre approche. En effet, la syntaxe actuelle des patrons de questions ne reconnaît pas l'expression « pour désigner » comme instance du symbole `Of` servant de liaison entre la propriété et son entité. Aussi, dans cet exemple, l'entité dont on questionne la valeur de propriété n'en est pas vraiment une, mais plutôt une description d'entités.

Question 4 : Quels (Ws) textes (ReferenceToReg) se rapportent (ApplyTo) aux obligations (Modality) d'un capitaine de port (RoleMention) ? (?)

C'est une question de catégorie 4, c'est-à-dire relative à des prescriptions de comportement.

Question 10 : « Quelles autorités sont chargées de l'application de l'arrêté n^{o} 22/91 ? »

Cette question est relative à l'application d'un texte et est classée en catégorie 9. L'ontologie légale et de domaine e-Compliance ne modélise pas ce type d'information. L'extension de l'ontologie que nous prévoyons d'élaborer adressera ce type de cas.

Question 16 : « Quels sont les visas de l'arrêté n^{o} 61/94 ? »

Cette question est de catégorie 8. Elle concerne la structure des textes réglementaires. Une extension de notre ontologie support permettra de représenter et donc d'interroger ce type d'information.

Question 19 : « Quels articles s'appliquent aux navires transportant des hydrocarbures ou des substances dangereuses ? »

Cette question est de catégorie 3. Autrement dit, elle concerne les réglementations s'appliquant à certaines situations. Cependant la syntaxe des patrons de questions ne gère pas de liste de situations. En effet, cette question mentionne deux cas alternatifs : le « transport d'hydrocarbures » et le « transport de substances dangereuses ». Une extension de la syntaxe s'attaquera à ces cas.

C. Exemples de requêtes SPARQL associées aux patrons de questions

Cette ressource est disponible en ligne à l'adresse `http://perso.telecom-bretagne.eu/yannisharalambous/data/tal-58-2-annexe-C.pdf`.